Journal of Medical Internet Research

Journal Impact Factor (JIF) (2022): 7.4
Volume 19 (2017), Issue 3 ISSN 1438-8871 Editor in Chief: Gunther Eysenbach, MD, MPH

Contents

Viewpoints

Original Papers

Viewpoint

Methodological and Ethical Challenges in a Web-Based Randomized Controlled Trial of a Domestic Violence Intervention

Laura Tarzia[1], BCA, GDipArts(Socio), PGDipArts(Socio), PhD; Jodie Valpied[1], MEd, PGDip (Psych); Jane Koziol-McLain[2], RN, PhD; Nancy Glass[3], FAAN, RN, MPH, PhD; Kelsey Hegarty[1], MBBS, FRACGP, PhD

[1]Department of General Practice, The University of Melbourne, Carlton, Australia

[2]Centre for Interdisciplinary Trauma Research, Auckland University of Technology, Auckland, New Zealand

[3]School of Nursing & Bloomberg School of Public Health, Johns Hopkins University, Baltimore, MD, United States

Corresponding Author:
Laura Tarzia, BCA, GDipArts(Socio), PGDipArts(Socio), PhD
Department of General Practice
The University of Melbourne
200 Berkeley St
Carlton, 3053
Australia
Phone: 61 390358604
Fax: 61 393476136
Email: laura.tarzia@unimelb.edu.au

Abstract

The use of Web-based methods to deliver and evaluate interventions is growing in popularity, particularly in a health care context. They have shown particular promise in responding to sensitive or stigmatized issues such as mental health and sexually transmitted infections. In the field of domestic violence (DV), however, the idea of delivering and evaluating interventions via the Web is still relatively new. Little is known about how to successfully navigate several challenges encountered by the researchers while working in this area. This paper uses the case study of I-DECIDE, a Web-based healthy relationship tool and safety decision aid for women experiencing DV, developed in Australia. The I-DECIDE website has recently been evaluated through a randomized controlled trial, and we outline some of the methodological and ethical challenges encountered during recruitment, retention, and evaluation. We suggest that with careful consideration of these issues, randomized controlled trials can be safely conducted via the Web in this sensitive area.

(J Med Internet Res 2017;19(3):e94) doi:10.2196/jmir.7039

KEYWORDS

eHealth; domestic violence; web-based trials; research design; ethics, research

Introduction

The Internet is growing in popularity as a means of both delivering health interventions [1,2] and recruiting research participants to evaluate those interventions [3]. This is perhaps unsurprising when the advantages offered by Web-based methods are considered compared with the traditional ways of engaging and retaining participants. For instance, the Internet allows a large number of participants to be accessed, even globally if necessary, at a relatively low cost. In addition, diverse or marginalized populations can be included in a study sample where transcription errors are reduced and responses can be collected more quickly [4], both being beneficial as competition for valuable research dollars increases. Practical limitations that might have been present a few years ago—such as requiring the participants to have Internet access and a degree of computer literacy—are barely relevant in today's increasingly networked society. In Australia, for example, household Internet access currently sits at 80% [5] and around 15 million people (80%) use a mobile phone [6]. An estimated 65% of Australians have a Facebook account [7]. Statistics for other developed countries such as the United States and the United Kingdom are similar [8,9].

Despite this potential, the use of online methods to recruit and engage participants in evaluating interventions is not without challenges. Due to the lack of face-to-face contact with researchers, online trials can experience poor retention rates, unreliable self-report data, and poor engagement with the intervention itself [1-3,10]. Alkhaldi and colleagues in a recent systematic review concluded that more research was needed in

this area in order to be able to understand engagement with online interventions and trials [2]. On the other hand, the Internet has demonstrated potential in facilitating research into sensitive issues [11]. For instance, Web-based methods have been used in studies evaluating digital interventions for sexual health [12,13], problem drinking [14,15], smoking cessation [16], drug use [17], and mental health problems [18]. In these contexts, the anonymity offered by the Internet may encourage participants to take part where they otherwise would not [10], potentially counteracting some of the challenges mentioned earlier. This is not to say that Web-based trials of interventions for sensitive issues are not without pitfalls.

In the context of domestic violence (DV), very little is known about the potential to deliver interventions via the Internet, or about safely recruiting research participants to a Web-based randomized controlled trial [19], although several authors have examined these issues in face-to-face DV studies [20-22]. DV is characterized by a systematic pattern of physical, emotional, sexual, or financial abuse by one intimate partner toward another with the intent to intimidate and control [23]. It is overwhelmingly perpetrated by men toward women [23] and can lead to a range of negative health outcomes, serious injury, or death. Research into DV is typically fraught with methodological and ethical issues [24]. Many abused women will not identify themselves as someone who is experiencing DV, thereby presenting a challenge to researchers seeking to recruit them [11]. There are also potential safety issues (eg, loss of confidentiality, retaliation from an abusive partner or ex-partner) involved with engaging abused women [25,26]. Once recruited into a study, women can potentially be re-traumatized by insensitive or inappropriate questioning [26,27], and researchers must be mindful of the ways in which controlling perpetrators may monitor women's movements or correspondence [24]. Due to the increased level of anonymity and privacy, as well as the ability to be accessed from anywhere, the Internet has the potential to overcome some of these issues [11], while also presenting new challenges that need to be attended to in the study design phase. This paper will use the case study of I-DECIDE, a Web-based healthy relationship tool and safety decision aid for Australian women experiencing DV [11,28], developed by the authors to illustrate how a randomized controlled trial of an intervention can be safely conducted via the Internet in this sensitive area [19]. We will focus on ethical and methodological challenges that occur specifically in Web-based DV trials, rather than on Web-based trials more broadly, as this topic has been explored in detail by others [3]. We conclude by presenting some recommendations for other researchers seeking to conduct Web-based research in this challenging and sensitive area.

The I-DECIDE Study

The I-DECIDE website is described in detail elsewhere [11], as is the study protocol for the randomized controlled trial that was used to evaluate it (ACTRN12614001306606) [28]. In brief, the intervention contained interactive modules designed to guide women through a process of self-reflection and self-management. The modules focused on healthy relationships, safety/risk and priorities, and incorporated elements of motivational interviewing [29] and nondirective problem solving [30]. The culmination of these activities was an "action plan" of strategies and resources that was individualized to women's priorities, relationship choices, and level of risk. The website was designed to be worked through once in a linear fashion (although different pathways through the website were possible depending on women's answers to certain key questions). Women were not required to complete the entire website in one sitting; however once they had completed the session, they were not able to return to change their answers. It was hypothesized that working through the intervention modules would increase women's levels of awareness about abuse occurring in their relationships, improve their self-efficacy, and enhance their sense of being supported.

The I-DECIDE website was evaluated through a pragmatic randomized controlled trial, comparing it with a website representing usual care (in this case, "usual care" was standard information on DV, a static emergency safety plan, and a list of general resources for DV). The trial was conducted almost entirely via the Internet. There were 422 women participants aged 16-50 years, who experienced fear of a current or ex-partner or any form of abuse from a current or ex-partner in the 6 months before recruitment. For both intervention and comparison groups, the study measures appeared after the informed consent and at the beginning of the website session, after which women were presented with either the intervention or comparison modules. Women who completed the baseline study measures (irrespective of whether or not they went on to complete the intervention or comparison modules) were followed up at 6 and 12 months via email prompts reminding them to log back in to the website. During follow-up visits, women completed the study measures and were given the option of going through the intervention or comparison modules a second time. Women were also able to log back in to the website at any time in between scheduled visits to access their action plan. Data collection was completed recently, with 80% of the baseline sample completing their 12-month follow-up visit. No adverse events (in this context, "adverse events" would include a woman being placed at increased risk of harm from her partner or experiencing extreme emotional distress) have been reported, and posttrial qualitative feedback from women via an open text box on the website and process evaluation interviews was overwhelmingly positive.

Ethical Issues

As with all DV research, the I-DECIDE study required a strong and carefully thought-out ethical framework to ensure the safety and well-being of its participants. Human Research Ethics Committees (HRECs) in Australia (Institutional Review Boards in the United States) have a critical role in ensuring that studies adhere to the National Health and Medical Research Council's code of conduct and respect the values of beneficence and nonmaleficence [31]. However, it has been suggested that ethics committees, which tend to operate from a predominantly biomedical framework, can find research involving so-called "vulnerable participants" challenging [26,32]. It has been suggested that they sometimes hold research teams to higher standards of ethical rigor in violence-related projects than other public health projects [32,33] and play a "gatekeeper" role toward research participants [26,34], which can make it more

difficult for researchers to gain approval for work in this area. Furthermore, Downes and colleagues in United Kingdom [32] have argued that the paternalistic attitude sometimes taken by HRECs undermines the decision-making capacity and agency of abused women to participate in research. They point out that all research has the potential to cause distress to a participant, even if the topic seems innocuous, and that women experiencing violence ought not to be prevented from participating simply because they may become upset. In fact, as Valpied et al [27] found in their qualitative exploration of abused women's experiences taking part in a randomized controlled trial of a general practice counseling intervention, the benefits of research participation—such as a sense of empowerment, catharsis and self-awareness—generally far outweigh any short-term distress.

In the case of I-DECIDE, there were not only considerations surrounding women's safety and well-being, but also ethical challenges unique to Web-based research [35,36]. For instance, determining whether and how a participant can provide informed consent via the Internet [13]; a lack of face-to-face contact between researcher and participant; determining whether data will be stored securely once collected; and how privacy can be protected [36] are the common concerns with regard to the Internet or social media research. A number of specific ethical issues arose in the I-DECIDE study where the context of DV intersected with the challenges of Web-based research. These are outlined below, along with how they were overcome.

Protecting Women on the Internet

In face-to-face studies, DV researchers can use a number of strategies in order to ensure the safety and well-being of participants [26,37]. For example, using trained interviewers to speak with women so that they can respond appropriately to any signs of distress [38], ensuring that women are contacted at a time when the perpetrator is not present and taking measures to ensure that her participation in the study remains confidential. When a study is conducted via the Internet, however, many of these safeguards cannot be monitored, and new safety strategies need to be put into place to protect participants.

For I-DECIDE, many of our strategies focused on equipping women to protect their own safety on the Internet, since the team did not connect with them via the web, by telephone, or in person unless they needed to be validated or had a technical issue. Several of these strategies also relate to how the intervention would be delivered in a "real world" setting on completion of the trial. Before signing up to the study, women were provided with browser-specific instructions on how to clear their search history and use an "incognito" or "private" session. They were also advised to use a "safe" email address that the perpetrator could not access and were directed to set up a new email account through Gmail or Yahoo if they were uncertain whether their existing account was secure. As an additional strategy, all email communication about I-DECIDE referred to a "Women's Health Study," without any reference to DV, and emails were sent from a specific women's health account to avoid any connection to DV or DV services.

The website itself was equipped with a "quick exit" bar positioned along the top of every screen. This enabled a woman to exit the website with a single click anywhere in the designated area should a perpetrator come up behind her while using I-DECIDE. On clicking the "quick exit" button, the existing browser window redirected to a generic weather website and further, a new browser tab opened up displaying the Google search engine. The data already inputted by the participant were saved, so that the next time she logged in to I-DECIDE she could continue where she left off. The landing page of I-DECIDE, which was publicly accessible, did not contain any information about DV or contain any images related to abuse. Rather, the website was identified as a "Women's Wellbeing Project" and asked the following:

> *Do you worry about whether your relationship is healthy? Do you sometimes wonder if you are safe? If you are a woman aged between 16 and 50 and you have experienced relationship issues over the last 6 months you are invited to take part in this project.*

Although a list of key DV support services were provided on the landing page, they were mixed with other women's health resources (eg, for depression and smoking) to disguise the focus of the website. Beyond the landing page, the website's content was protected through a randomly generated username and password that was sent to women's safe email address upon signing up.

Duty of Care in a Web-Based Trial

A major consideration for researchers in the DV area is ensuring that they fulfill their duty of care toward the women involved in a study [26]. In a face-to-face study, this would involve checking with participants regularly to ensure that they still consent to take part and that their safety and well-being is not being compromised [37]. If a researcher is alerted to a woman being in immediate danger, their duty of care obliges them to discuss safety options such as contacting police or DV crisis line [32,38] and compels them to alert authorities if a child is suspected to be in danger [39]. The anonymity of the Internet, however, requires new strategies to be developed to maximize women's safety and well-being [40].

In the I-DECIDE study, although we could not determine with any certainty whether women taking part in the study were at immediate risk, we used a number of approaches to maximize their safety during engagement with the Web-based tool. As part of the intervention module, for instance, women were asked to complete the Danger Assessment [41], which measures her level of risk for severe violence and homicide, and the Composite Abuse Scale [42], which identifies abusive behaviors in a relationship. Their responses to these validated tools were scored and categorized by the I-DECIDE program, and matching messages were immediately provided on the screen for the woman to review. Any woman whose responses indicated a higher level of risk were provided with feedback advising them of this and suggesting they think about contacting the police. At the end of the website, these women first received an emergency safety plan (including strategies such as collecting important documents and setting up an emergency code word with a friend or family member if she was in distress) before an action plan tailored to their individual life priorities. Women at lower levels of risk received only the tailored action plan (although they could also access the emergency safety plan if

needed). All women in the comparison group automatically received only the emergency safety plan.

To minimize the likelihood of women becoming distressed while participating in the study, the I-DECIDE website included supportive messaging and feedback at critical points. In particular, the messaging in response to the woman's levels of danger and abuse was developed very carefully in consultation with community DV services. The messaging needed to communicate the woman's level of risk without frightening or traumatizing her. The example below is the message for the highest level of risk.

> *You are experiencing some extremely dangerous things in your relationship. It's critical that you talk to someone you can trust, like a friend, family member, your GP (general practitioner/family doctor), or the police, about what is going on in your relationship. A bit later, this website will suggest some things you could try to help you increase your safety, and also to help take care of yourself during this stressful time. It will also provide you with details of some services (confidential and free) that can help you if you need. There is a national counselling hotline 1800-737-732 that you can call anonymously for advice, but for an emergency situation, call 000.*

As an additional precaution, a study phone number was set up in case women needed to contact the researchers, despite the trial being otherwise carried out entirely via the Web [19,28]. A distress protocol was developed and distributed to all project team members. The protocol included active listening and exploring with the individual whether they had someone (friend, family, or service provider) trusted that they could talk to, who would understand and be supportive. Ultimately, very few women utilized the phone number, and those who did were enquiring about study participant gift vouchers or experiencing technical difficulties.

An adapted version of the validated Consequences of Screening Tool (COST) questionnaire [43] was embedded within both arms of the I-DECIDE website to ensure that the benefits of participation outweighed the harms for women. This tool asks women about how participation affects their feelings about themselves and their relationship. The adapted COST contained 10 items, each on a five-point Likert scale that indicated harm, benefit, or neutral options. On completion of baseline data collection, the harm/benefit data were reviewed by the researchers and the independent study Data Monitoring Committee before commencement of the 6-month follow-up. Although 10.9% (37/339) of women reported that their partner or ex-partner was aware that they were answering questions about DV, at follow-up many (62%, 23/37) reported a positive outcome from this—a finding consistent with previous DV research using this tool [44]. Additionally, almost all the women (93.5%, 317/339) reported that they were glad to be a participant in the I-DECIDE project, despite 11.5% (39/339) indicating that questions in the website had made them feel that their relationship problems were their own fault and 5.3% (18/339) feeling somewhat worse about themselves as a person. Research suggests that these negative emotions are common, but usually transient [27]. For instance, Valpied and colleagues [27], who analyzed data from the COST Questionnaire in a face-to-face DV intervention trial [44], suggest that women might feel worse about themselves because they realize that an abusive relationship is unsalvageable despite how hard they have tried to make it work. They argue that this is not necessarily a negative outcome, as it could lead to taking action for safety and well-being. Feeling at fault, or otherwise bad about oneself is also not necessarily a reflection on delivering the intervention or carrying out a trial via the web. In fact, the average score out of 10 from the I-DECIDE study sample regarding how supported they felt by the website was 9.

Balancing Participant Safety With Study Visibility

Attracting over 400 women nationally to the I-DECIDE study involved a mix of recruitment strategies. In particular, the essential role of social media as a key part of our recruitment became apparent very early on in baseline data collection.

Colleagues in the United States [45], who developed the original "IRIS" safety decision aid website on which I-DECIDE is based [11] had enormous success using the free classified advertisements website, Craigslist, to attract women to their study. In New Zealand [19,46], where another version of the website underwent trial, the overwhelming majority of the sample came from ads on TradeMe, a trading post style website. In Australia, however, Craigslist is not widely used, and our ads were not successful in attracting any women to the study. Similarly, ads placed on Australian trading post website, Gumtree attracted less than 10 women to the study, despite paying for our ads to be prioritized in search listings.

As the most widely used social media platform in Australia [7] and one that has demonstrated success with other Web-based trials [47] and with hard-to-reach populations [48], Facebook was next identified as a potential source of recruitment. However, there are risks associated with the format of Facebook which needed to be overcome for a DV study. Facebook enables its users to present themselves in a user profile, accumulate "friends" who can post comments on each other's pages, and view each other's profiles. Facebook members can also join virtual groups based on common interests [49]. Facebook users can "like" or "share" content that then appears on their "newsfeed," and can "follow" updates from pages associated with organizations. Although the option of setting up a page for the I-DECIDE study clearly had potential to facilitate the wide distribution of our ads, it also meant that the ad might appear on the newsfeed of an abused woman who otherwise had no DV-related content. Instead, two approaches were used. First, a paid advertisement that appeared on the side of screen for women in the target demographic (as this appeared alongside a range of other advertisements, this was deemed less risky). Second, we sent requests to Australian DV services and a range of other women's health, fashion, and motherhood organizations to post our ad. Although this would still result in the ad appearing in women's newsfeed, the women would already have been following these pages in order to receive the update, and we therefore were not adding to their existing level of risk. For all ads, the accompanying text "Please open the link in a new browser window" and "Share only if safe to do so" was

included. Ultimately, around 60% of our sample was recruited using Facebook.

Methodological Issues

In the I-DECIDE study, women were recruited, enrolled, consented, and randomized entirely via the web. As Murray and colleagues have noted [3], Web-based trials can experience a range of methodological issues including an unrepresentative sample, poor retention rates, and an inability to validate participants leading to multiple registrations or inclusion of ineligible participants. In the DV context, these issues, as well as others unique to the field, were experienced as part of the I-DECIDE study as outlined below.

Ensuring That Participants Are Who They Say They Are

The I-DECIDE study was open only to women residing in Australia aged 16-50 years who had experienced fear of a partner or some form of abuse over the past 6 months. However, in theory, to sign up to the study a person needed only to click "Be A Part of the Project" from the publicly accessible landing page to commence the enrolment process. We were therefore presented with a challenge around how to prevent enrolment by ineligible women, "trolls," or perpetrators seeking to disrupt the study.

Koziol-McLain and colleagues in New Zealand [46] asked women to input their full name and residential address upon signing up to their study. This was then compared automatically with the New Zealand electoral roll, which enabled them to check the following (1) the participant was female and (2) the participant was a resident of New Zealand. We initially endeavored to copy this approach; however, privacy laws in Australia regarding distribution of the entire electoral roll are extremely stringent, and we were unable to gain access to this document. Instead, a research team member manually validated each individual participant through the Australian Electoral Roll website. Any participant not able to be validated through their residential address was validated via email contact, social media, or a telephone call, as a last resort. This process was time consuming, but worthwhile to ensure that our sample did not contain bogus participants.

In terms of validating the age of participants, they were asked directly on sign up, "Are you aged between 16 and 50 years." Women who answered "no" were exited from the enrolment process with a message explaining that study was limited to this demographic. To further cross-check participant age, women were asked for their date of birth once they commenced the demographic questions on the first screen of I-DECIDE. Although this approach would not have stopped the determined participants outside our target age range from lying about their age, it likely minimized the incidences of ineligible participants taking part.

It was anticipated that abused women might have concerns about providing information such as their residential address and contact telephone number. We therefore included a short explanation about why we were requesting these details and what would be done with the information:

We need to collect your full name and a valid residential address.

There are three reasons for this:

To keep you safe;

To ensure that no fraudulent participants (including men) can access the website;

To send you a gift card to thank you for your participation.

We will validate your details against the Australian Electoral Roll and will not use your details for any other purpose.

The details of the participants were recorded in a separate database from their responses to the study measures and the rest of the intervention. Although providing a name did mean that women were not anonymous, it was considered to be an acceptable trade off in order to deter bogus participants.

Engaging Women Not Ready to Identify as "DV Victims"

A major consideration for the I-DECIDE study was how to include women who may not have been ready to acknowledge that their relationship was abusive [11]. As Bender [21] noted, recruiting women through specific locations such as shelters, courtrooms, or health clinics can be problematic, as levels of violence reported from women in these settings may not be representative of the overall population. Similarly, accessing women only through DV services and the community sector would have resulted in a particular demographic of women who were already seeking help for violence in their relationships and the project was interested in reaching women who may not have been aware of or previously accessed DV services. To overcome this challenge, we contacted a wide range of organizations and individuals likely to have a large number of female followers and asked them to promote the study. As mentioned earlier, strategies included not only social media posts about the study (eg, Facebook and Twitter), but also mentions in organizational newsletters, advertisements on blogs, and media interviews with members of the research team. Another successful source of recruitment was university student portals, provided that we were able to satisfy the relevant "gatekeepers."

It is not immediately obvious how one might go about recruiting women experiencing DV without mentioning "violence" or "abuse," while at the same time not misleading women as to the nature of the study. As a compromise, we referred to "relationship safety" or "feeling afraid of a partner" in all our recruitment materials (Figure 1). This was deemed to convey the topic without being too confronting or challenging for women in an earlier stage of awareness or readiness for action [50]. Although we did not ask women directly whether they perceived their relationship as violent, the study sample includes women of diverse ages and backgrounds who had experienced different types of abuse (Table 1). More detailed participant demographics will be reported at a later date.

Table 1. Brief I-DECIDE study demographics.

Study demographics	Mean (SD) or n (%)
Age in years (n=422), mean (SD)	33.74 (SD 8.48)
Aboriginal and/or Torres Strait Islander (n=376), n (%)	40 (10.64)
Type of Abuse at 12 months (n=331), n (%)	
Emotional abuse (with or without harassment)	108 (32.7)
Physical abuse and emotional abuse (with or without harassment)	36 (10.91)
Severe combined abuse	125 (37.88)
Not positive for abuse on Composite Abuse Scale [42]	61 (18.48)

Figure 1. Study advertisement for I-DECIDE.

Creating Trust in the Absence of Face-to-Face Contact

Building rapport with research participants can be an important element of DV research that occurs in a face-to-face setting [26]. In an online setting, however, creating an impression of trustworthiness is much more difficult. In preliminary focus groups conducted during the design phase of the study, women stated that a professional design and the use of the university's logo could help to distinguish the site from unreliable information on the Internet. Similarly, the language and the tone used throughout the website were designed to sound like an "empathetic expert," such as a health professional or a counselor. The aim was to make women feel supported, empowered, and listened to, without being authoritative.

As another trust-building strategy in DV research, Logan and colleagues have emphasized the importance of engaging community partners [20]. This not only suggests to potential participants that the study has their best interests at heart, but also provides a source of referral for face-to-face support should they require it. To this end, we collaborated with several DV services and sought their feedback during the design phase of the website. Facebook posts by our partner organizations as well as by other high-profile DV organizations endorsing the study were successful at attracting women to click through to the I-DECIDE website.

Encouraging Retention Over the Long Term

Retention of participants in all trials is often challenging, but Web-based trials tend to have most issues with the follow up [3]. When conducting research with abused women, there are additional challenges to contend with. For instance, women are often forced to move house or change their email address or

telephone number if they are concerned for their safety or that of their children [51], which increases the likelihood they will be lost to follow up. To combat this, women were asked to provide the first name and either a telephone number or email address of one or two "safe contacts" on signing up to the study. It was explained to women that these contacts were needed only to stay in touch with them during the study.

In case you move house or we lose touch with you for any reason during the trial, please put in the details of two trusted people we can contact to get in touch with you. Try to think of people who are unlikely to change their details over the next 12 months (for example, a parent or relative). We will only reveal that you are participating in a 'women's health study'.

As an additional retention strategy, incentives were offered in increasing value for baseline, 6 month, and 12-month follow-up visits, with a maximum of $150 for completing all 3 visits. Researchers in the DV area consistently emphasize the importance of incentives for women as a way of recognizing their contribution [20]. Although women are often motivated to participate because of other, nonfinancial reasons (for instance, a desire to help other women [26]), it is generally considered good practice to offer them as a token of appreciation for their time and expertise. For the I-DECIDE study, incentives were offered in the form of electronic gift vouchers to a large national chain of stores.

Finally, although we had intended to use solely electronic means of communication to follow up women for 6- and 12-month visits, this was ultimately not successful. Initially, our study protocol dictated that the study tracking database would send out 3 automatic email reminders at various intervals after a woman's session fell due, followed by an SMS reminder. Response rates remained lower than expected, and we therefore added a phone call reminder by a trained, study research assistant. Although the use of a research assistant can be seen as an "intervention" [40] and not indicative of how the website is likely to be accessed in a real world setting, our experience is that without this contact, it would be difficult to engage women, given the dynamics of an abusive relationship, in follow-up visits after a 6-month period of time.

Conclusions and Implications for Future Research

This paper highlights the key ethical and methodological challenges involved with conducting a Web-based trial of a DV intervention. It highlights how the particular context of DV intersects with the constraints and opportunities of digital technologies. Our I-DECIDE case study demonstrates that trials of DV interventions can be conducted via the web, and women can be safely recruited, engaged in an intervention, and successfully retained over a 12-month period using almost exclusively Web-based strategies. Based on the lessons learned from our trial, we suggest the following recommendations for other researchers seeking to conduct trials of DV interventions using the Internet:

Publish ethical and "lessons learned" papers in conjunction with trials, so that ethics committees and other researchers can draw on this knowledge base to make informed decisions about future studies;

Ensure that the language used throughout Web-based interventions and the trial processes surrounding them is supportive and non-judgmental. Interventions should be developed in close consultation with victims/survivors and services;

Equip participants with information about Internet safety so that they can take charge of their own online footprint;

Explain why personal information is being collected;

Consider avenues of recruitment outside traditional DV services, and ensure that language is inclusive of women who may not be ready to name the abuse; and

Consider telephone contact with participants as a last resort to encourage retention in a trial over the long term.

Acknowledgments

This work is funded by an Australian Research Council Discovery Project Grant (DP 130102799). The funding body had no role in the development or review of this manuscript.

Conflicts of Interest

None declared.

References

1. Murray E. Web-based interventions for behavior change and self-management: potential, pitfalls, and progress. Med 2.0 2012 Aug;1(2):e3 [FREE Full text] [doi: 10.2196/med20.1741] [Medline: 25075231]
2. Alkhaldi G, Hamilton FL, Lau R, Webster R, Michie S, Murray E. The effectiveness of prompts to promote engagement with digital interventions: a systematic review. J Med Internet Res 2016;18(1):e6 [FREE Full text] [doi: 10.2196/jmir.4790] [Medline: 26747176]
3. Murray E, Khadjesari Z, White IR, Kalaitzaki E, Godfrey C, McCambridge J, et al. Methodological challenges in online trials. J Med Internet Res 2009;11(2):e9 [FREE Full text] [doi: 10.2196/jmir.1052] [Medline: 19403465]
4. Holmes S. Methodological and ethical considerations in designing an internet study of quality of life: a discussion paper. Int J Nurs Stud 2009 Mar;46(3):394-405. [doi: 10.1016/j.ijnurstu.2008.08.004] [Medline: 18838135]

5. Australian Bureau of Statistics. 2014. Household Use of Information Technology, Australia, 2012-2013, cat no 8146 URL: http://www.abs.gov.au/ausstats/abs@.nsf/mf/8146.0 [accessed 2016-11-23] [WebCite Cache ID 6mDypCXJz]
6. The Nielsen Company & IAB Australia. 2015. The Mobile Story: Nielsen Mobile Ratings URL: https://www.iabaustralia.com.au/uploads/uploads/2015-11/1448287200_7a87cd54d624da241efa7087c362d82e.pdf [accessed 2017-03-15] [WebCite Cache ID 6oy9ACaii]
7. Cowling D. Social Media News. 2016. Social Media Statistics Australia - October 2016 URL: http://www.socialmedianews.com.au/social-media-statistics-australia-october-2016/ [accessed 2016-11-23] [WebCite Cache ID 6mDz74yuV]
8. File T, Ryan C. United States Census Bureau. 2014. Computer and Internet Use in the United States 2013 URL: https://www.census.gov/history/pdf/2013computeruse.pdf [WebCite Cache ID 6oyB2qCSd]
9. Office for National Statistics. United Kingdom: Office for National Statistics; 2016. Internet Access - Households and Individuals URL: https://www.ons.gov.uk/peoplepopulationandcommunity/householdcharacteristics/homeinternetandsocialmediausage/bulletins/internetaccesshouseholdsandindividuals/2016 [accessed 2017-03-15] [WebCite Cache ID 6nljlbJ8E]
10. Bailey J, Murray E, Rait G, Mercer C, Morris R, Peacock R, et al. Interactive computer-based interventions for sexual health promotion. Cochrane Database Syst Rev 2010 Sep 08(9):CD006483. [doi: 10.1002/14651858.CD006483.pub2] [Medline: 20824850]
11. Tarzia L, Murray E, Humphreys C, Glass N, Taft A, Valpied J, et al. I-DECIDE: an online intervention drawing on the psychosocial readiness model for women experiencing domestic violence. Womens Health Issues 2016;26(2):208-216. [doi: 10.1016/j.whi.2015.07.011] [Medline: 26362841]
12. Bull SS, Lloyd L, Rietmeijer C, McFarlane M. Recruitment and retention of an online sample for an HIV prevention intervention targeting men who have sex with men: the Smart Sex Quest Project. AIDS Care 2004;16(8):931-943. [doi: 10.1080/09540120412331292507] [Medline: 15511725]
13. Curtis B. Social networking and online recruiting for HIV research: ethical challenges. J Empir Res Hum Res Ethics 2014;9(1):58-70. [doi: 10.1525/jer.2014.9.1.58]
14. Linke S, McCambridge J, Khadjesari Z, Wallace P, Murray E. Development of a psychologically enhanced interactive online intervention for hazardous drinking. Alcohol 2008;43(6):669-674 [FREE Full text] [doi: 10.1093/alcalc/agn066] [Medline: 18693217]
15. Linke S, Brown A, Wallace P. Down your drink: a web-based intervention for people with excessive alcohol consumption. Alcohol and Alcoholism 2004;39(1):29-32. [Medline: 14691071]
16. Shahab L, McEwen A. Online support for smoking cessation: a systematic review of the literature. Addiction 2009;104(11):1792-1804. [doi: 10.1111/j.1360-0443.2009.02710.x] [Medline: 19832783]
17. Barratt M, Lenton S. Beyond recruitment? Participatory online research with people who use drugs. IJIRE 2010;3(1):69-86.
18. Berger M, Wagner TH, Baker LC. Internet use and stigmatized illness. Soc Sci Med 2005 Oct;61(8):1821-1827. [doi: 10.1016/j.socscimed.2005.03.025] [Medline: 16029778]
19. Koziol-McLain J, McLean C, Rohan M, Sisk R, Dobbs T, Nada-Raja S, et al. Participant recruitment and engagement in automated ehealth trial registration: challenges and opportunities for recruiting women who experience violence. J Med Internet Res 2016;18(10):e281 [FREE Full text] [doi: 10.2196/jmir.6515] [Medline: 27780796]
20. Logan T, Walker R, Shannon L, Cole J. Combining ethical considerations with recruitment and follow-up strategies for partner violence victimization research. Violence Against Women 2008;14(11):1226-1251. [doi: 10.1177/1077801208323976] [Medline: 18809849]
21. Bender AK. Ethics, methods, and measures in intimate partner violence research: the current state of the field. Violence Against Women 2016 Aug 04(Online First):1-32. [doi: 10.1177/1077801216658977] [Medline: 27492926]
22. Dutton M, Holtzworth-Munroe A, Jouriles E, McDonarld R, Krishnan S, McFarlane J, et al. NCJRS.: US Department of Justice; 2003. Recruitment and retention in intimate partner violence research URL: https://www.ncjrs.gov/pdffiles1/nij/201943.pdf [accessed 2017-03-14] [WebCite Cache ID 6oyCdhRKk]
23. World Health Organization. Geneva, Switzerland; 2013. Global and regional estimates of violence against women: prevalence and health effects of intimate partner violence and non-partner sexual violence URL: http://apps.who.int/iris/bitstream/10665/85239/1/9789241564625_eng.pdf [WebCite Cache ID 6oyCnQX6i]
24. World Health Organization. Geneva, Switzerland; 2001. Putting women first: ethical and safety recommendations for research on domestic violence against women URL: http://www.who.int/gender/violence/womenfirtseng.pdf [WebCite Cache ID 6oyCwnFTX]
25. Ruiz-Perez I, Plazaola-Castano J, Vives-Cases C. Methodological issues in the study of violence against women. J Epidemiol Community Health 2007;61:26-31 [FREE Full text]
26. Burgess-Proctor A. Methodological and ethical issues in feminist research with abused women: reflections on participants' vulnerability and empowerment. Wom Stud Int Forum 2015;48:124-134.
27. Valpied J, Cini A, O'Doherty L, Taket A, Hegarty K. 'Sometimes cathartic. Sometimes quite raw': benefit and harm in an intimate partner violence trial. Aggress Violent Behav 2014;19:673-685.

28. Hegarty K, Tarzia L, Murray E, Valpied J, Humphreys C, Taft A, et al. Protocol for a randomised controlled trial of a web-based healthy relationship tool and safety decision aid for women experiencing domestic violence (I-DECIDE). BMC Pub Health 2015;15(736):1-9. [doi: 10.1186/s12889-015-2072-z]
29. Miller WR, Rose GS. Toward a theory of motivational interviewing. Am Psychol 2009 Sep;64(6):527-537 [FREE Full text] [doi: 10.1037/a0016830] [Medline: 19739882]
30. Mynors-Wallis LM, Gath DH, Day A, Baker F. Randomised controlled trial of problem solving treatment, antidepressant medication, and combined treatment for major depression in primary care. BMJ 2000 Jan 01;320(7226):26-30 [FREE Full text] [Medline: 10617523]
31. National Health and Medical Research Council, The Australia Vice-Chancellor's Committee. NHMRC. Canberra: Commonwealth of Australia; 2007. The National Statement on Ethical Conduct in Human Research URL: https://www.nhmrc.gov.au/book/national-statement-ethical-conduct-human-research [accessed 2017-03-15] [WebCite Cache ID 6oyJE6BDN]
32. Downes J, Kelly L, Westmarland N. Ethics into violence and abuse research - a positive empowerment approach. Sociol Res Online 2014;19(1):1-13.
33. Cromer LD, Newman E. Research ethics in victimization studies: widening the lens. Violence Against Women 2011 Dec;17(12):1536-1548. [doi: 10.1177/1077801211436365] [Medline: 22410769]
34. Bosworth M, Hoyle C, Dempsey M. Researching trafficked women: On institutional resistance and the limits to feminist reflexivity. Qual Inq 2011;17(9):769-779.
35. Buchanan E, Hvizdak E. Online survey tools: ethical and methodological concerns of human research ethics committees. J Empir Res Hum Res Ethics 2009 Jun;4(2):37-48. [doi: 10.1525/jer.2009.4.2.37] [Medline: 19480590]
36. Moreno MA, Goniu N, Moreno PS, Diekema D. Ethics of social media research: common concerns and practical considerations. Cyberpsychol Behav Soc Netw 2013 Sep;16(9):708-713 [FREE Full text] [doi: 10.1089/cyber.2012.0334] [Medline: 23679571]
37. Ellsberg M, Heise L. WHO. Washington DC: World Health Organization, PATH; 2005. Researching violence against women: a practical guide for researchers and activists URL: http://apps.who.int/iris/bitstream/10665/42966/1/9241546476_eng.pdf [accessed 2017-03-15] [WebCite Cache ID 6oyKOID50]
38. Ellsberg M, Heise L. Bearing witness: ethics in domestic violence research. The Lancet 2002 May;359(9317):1599-1604. [doi: 10.1016/S0140-6736(02)08521-5]
39. Allenby R, Dobbs T, Diesfeld K, Nada Raja S, Wilson D, Koziol-McLain J. Safety in online research with women experiencing intimate partner violence: what about the children? Ethics Behav 2016 Mar 30:1-17. [doi: 10.1080/10508422.2016.1140579]
40. Feather JS, Howson M, Ritchie L, Carter PD, Parry DT, Koziol-McLain J. Evaluation methods for assessing users' psychological experiences of web-based psychosocial interventions: a systematic review. J Med Internet Res 2016;18(6):e181 [FREE Full text] [doi: 10.2196/jmir.5455] [Medline: 27363519]
41. Campbell JC, Webster DW, Glass N. The danger assessment: validation of a lethality risk assessment instrument for intimate partner femicide. J Interpers Violence 2009 Apr;24(4):653-674. [doi: 10.1177/0886260508317180] [Medline: 18667689]
42. Hegarty K, Bush R, Sheehan M. The composite abuse scale: further development and assessment of reliability and validity of a multidimensional partner abuse measure in clinical settings. Violence Vict 2005 Oct;20(5):529-547. [Medline: 16248489]
43. Lock J. The development of the Consequences of Screening Tool and the psychometric assessment of three woman abuse measures (thesis). Hamilton, ON: McMaster University; 2008.
44. Hegarty K, O'Doherty L, Taft A, Chondros P, Brown S, Valpied J, et al. Screening and counselling in the primary care setting for women who have experienced intimate partner violence (WEAVE): a cluster randomised controlled trial. Lancet 2013;382(9888):249-258 [FREE Full text] [doi: 10.1016/S0140-6736(13)60052-5] [Medline: 23598181]
45. Eden KB, Perrin NA, Hanson GC, Messing JT, Bloom TL, Campbell JC, et al. Use of online safety decision aid by abused women: effect on decisional conflict in a randomized controlled trial. Am J Prev Med 2015 Apr;48(4):372-383 [FREE Full text] [doi: 10.1016/j.amepre.2014.09.027] [Medline: 25547929]
46. Koziol-McLain J, Vandal AC, Nada-Raja S, Wilson D, Glass NE, Eden KB, et al. A web-based intervention for abused women: the New Zealand isafe randomised controlled trial protocol. BMC Public Health 2015 Jan 31;15:56 [FREE Full text] [doi: 10.1186/s12889-015-1395-0] [Medline: 25637195]
47. Ramo DE, Prochaska JJ. Broad reach and targeted recruitment using Facebook for an online survey of young adult substance use. J Med Internet Res 2012 Feb 23;14(1):e28 [FREE Full text] [doi: 10.2196/jmir.1878] [Medline: 22360969]
48. King D, O'Rourke N, DeLongis A. Social media recruitment and online data collection: a beginner's guide and best practices for accessing low-prevalence and hard-to-reach populations. Can Psychol 2014;55(4):240-249.
49. Ellison N, Steinfield C, Lampe C. The benefits of facebook 'friends': social capital and college students' use of online social network sites. J Comput Mediat Commun 2007;12(4):1143-1168.
50. Cluss PA, Chang JC, Hawker L, Scholle SH, Dado D, Buranosky R, et al. The process of change for victims of intimate partner violence: support for a psychosocial readiness model. Womens Health Issues 2006;16(5):262-274. [doi: 10.1016/j.whi.2006.06.006] [Medline: 17055379]

51. Rollins C, Glass NE, Perrin NA, Billhardt KA, Clough A, Barnes J, et al. Housing instability is as strong a predictor of poor health outcomes as level of danger in an abusive relationship: findings from the SHARE Study. J Interpers Violence 2012 Mar;27(4):623-643. [doi: 10.1177/0886260511423241] [Medline: 21987519]

Abbreviations

DV: domestic violence
COST: Consequences of Screening Tool
HRECs: Human Research Ethics Committees

Edited by I Weber; submitted 01.12.16; peer-reviewed by J Spangaro, N Hackworth; comments to author 14.01.17; revised version received 08.02.17; accepted 16.02.17; published 28.03.17.

Please cite as:
Tarzia L, Valpied J, Koziol-McLain J, Glass N, Hegarty K
Methodological and Ethical Challenges in a Web-Based Randomized Controlled Trial of a Domestic Violence Intervention
J Med Internet Res 2017;19(3):e94
URL: http://www.jmir.org/2017/3/e94/
doi:10.2196/jmir.7039
PMID:28351830

Original Paper

Enlight: A Comprehensive Quality and Therapeutic Potential Evaluation Tool for Mobile and Web-Based eHealth Interventions

Amit Baumel[1,2], PhD; Keren Faber[1], MA; Nandita Mathur[1], MA; John M Kane[1,2], MD; Fred Muench[1,2], PhD

[1]Psychiatry Research, The Feinstein Institute for Medical Research, Glen Oaks, NY, United States

[2]Northwell Hofstra School of Medicine, Hempstead, NY, United States

Corresponding Author:
Amit Baumel, PhD
Psychiatry Research
The Feinstein Institute for Medical Research
75-59 263rd street
Glen Oaks, NY, 11004
United States
Phone: 1 7184708267
Fax: 1 7183431659
Email: abaumel@northwell.edu

Abstract

Background: Studies of criteria-based assessment tools have demonstrated the feasibility of objectively evaluating eHealth interventions independent of empirical testing. However, current tools have not included some quality constructs associated with intervention outcome, such as persuasive design, behavior change, or therapeutic alliance. In addition, the generalizability of such tools has not been explicitly examined.

Objective: The aim is to introduce the development and further analysis of the Enlight suite of measures, developed to incorporate the aforementioned concepts and address generalizability aspects.

Methods: As a first step, a comprehensive systematic review was performed to identify relevant quality rating criteria in line with the PRISMA statement. These criteria were then categorized to create Enlight. The second step involved testing Enlight on 42 mobile apps and 42 Web-based programs (delivery mediums) targeting modifiable behaviors related to medical illness or mental health (clinical aims).

Results: A total of 476 criteria from 99 identified sources were used to build Enlight. The rating measures were divided into two sections: quality assessments and checklists. Quality assessments included usability, visual design, user engagement, content, therapeutic persuasiveness, therapeutic alliance, and general subjective evaluation. The checklists included credibility, privacy explanation, basic security, and evidence-based program ranking. The quality constructs exhibited excellent interrater reliability (intraclass correlations=.77-.98, median .91) and internal consistency (Cronbach alphas=.83-.90, median .88), with similar results when separated into delivery mediums or clinical aims. Conditional probability analysis revealed that 100% of the programs that received a score of fair or above (≥3.0) in therapeutic persuasiveness or therapeutic alliance received the same range of scores in user engagement and content—a pattern that did not appear in the opposite direction. Preliminary concurrent validity analysis pointed to positive correlations of combined quality scores with selected variables. The combined score that did not include therapeutic persuasiveness and therapeutic alliance descriptively underperformed the other combined scores.

Conclusions: This paper provides empirical evidence supporting the importance of persuasive design and therapeutic alliance within the context of a program's evaluation. Reliability metrics and preliminary concurrent validity analysis indicate the potential of Enlight in examining eHealth programs regardless of delivery mediums and clinical aims.

(J Med Internet Res 2017;19(3):e82) doi:10.2196/jmir.7270

KEYWORDS

eHealth; mHealth; assessment; evaluation; quality; persuasive design; behavior change; therapeutic alliance

Introduction

The wide distribution of personal digital devices has changed the potential to dramatically enhance public access to health interventions; tens of thousands of health, wellness, and medical apps are now available for download from online stores [1]. From a public health perspective, the large number of available eHealth intervention programs makes it impossible to empirically evaluate them using traditional research methods. In addition, the time-consuming and costly process of product evaluation may result in the technology investigated in the trial becoming obsolete by the time the results are published [2] and may also make it difficult for health system leaders to engage with potential vendors [3,4]. Agile science tries to answer some of these challenges by focusing on an adaptable and nimble scientific process to support the collective development and evaluation of interventions [5], while taking into account the rapid pace of change in the technologies that support digital interventions [6]. However, despite developments in research methods, patients and medical professionals can browse through the Web or access mobile app stores to download and use the large number of available and unexamined programs. To complicate things further, the settings in which such programs are being utilized are different than within studies where participants are being proactively recruited, paid for filling out assessments, and have scheduled check-in appointments. As a result, some of the support that participants receive in studies does not translate to the real world, which might impact intervention outcomes. Unfortunately, existing systems of user-based ratings that collect assessments in the real world are not designed to offer a metric of medical appropriateness, safety, or effectiveness [7].

Criteria-based rating scales have been developed to address this evaluation challenge. These scales are then used by trained raters to objectively examine and score the quality of eHealth intervention programs based on core concepts, each comprised of different criteria [8-11]. The importance of using a clearly defined rating system is strengthened by the fact that without the use of such systems scoring tends not to be highly reliable [7]. In facilitating the cost-effective evaluation of available eHealth interventions, such tools can enable stakeholders to discuss programs' potential prior to empirical testing [4] and to provide information that supports user recommender systems [12-14]. Although these tools have several potential uses, we will relate to two aspects we believe to be important for the contributions of these tools to the evaluation of eHealth interventions that were not covered previously: examining quality domains that relate to programs' therapeutic potential and tool generalizability.

Quality Rating Domains Related to Programs' Therapeutic Potential

Persuasive Design and Behavior Change Principles

Persuasive design aims at understanding what influences people's behavior and decision making, and then uses this information to design compelling user interactions [15,16]. Interestingly, no previous rating scale aiming to evaluate the quality of eHealth intervention programs using different concepts has related to persuasive design or behavior change quality criteria, even though such concepts have been shown to be important in the evaluation of eHealth interventions potential. For example, Kientz et al [17] compared the performance of Nielsen and Molich's usability heuristics [18] to persuasion and demonstrated that persuasion heuristics enabled the identification of more severe and more relevant interface problems in terms of persuasive, cultural, and informational issues. Kelders et al [19] showed that elements of persuasive design uniquely explained the variance in adherence to eHealth Web-based interventions, and Webb et al [20] showed that eHealth interventions that better incorporated behavior change theories also tended to have larger effects in increasing positive health-related behaviors. Altogether, these studies suggest that the quality of program persuasive design directly impacts its therapeutic potential and therefore has to be addressed when evaluating these programs.

Therapeutic Alliance Principles

No previous quality rating scale has directly assessed the therapeutic alliance being nurtured by the eHealth intervention program. Overall, studies focusing on nontechnological interventions have shown that the therapeutic alliance is one of the most robust measures for predicting psychotherapy success (eg, [21-23]) and suggested its promise for predicting intervention quality in the medical domain (eg, [24-27]). However, evaluating the potential therapeutic alliance between users and software programs requires the reexamination and adaptation of the original concept, which applies to a therapeutic relationship between people.

Studies suggest that therapeutic alliances with eHealth intervention programs do exist and that such alliances may play a role in increasing the adherence to [28,29] and effectiveness [30] of these programs. Scholars have indicated the variance in relational factors embedded within eHealth interventions that makes some programs better at nurturing a therapeutic alliance with their users [31-33]. Although these examinations are currently of a preliminary nature, there is a need for a standardized assessment measure that can adequately capture the concept of an e-therapeutic alliance and its place in the overall picture of product quality.

Overall, there is a need to develop standardized operational definitions to assess the quality of all aspects of eHealth intervention programs [34], including those relating to therapeutic potential [4]. To best address the unique contribution of each quality domain, the complete scope of different criteria should be taken into account. Such assessment will also enable stakeholders to investigate the interactions between different quality domains and their impact on outcomes.

Tool Generalizability

Another aspect that has yet to be fully examined is whether criteria-based rating tools may enable us to reliably rate eHealth intervention programs that are developed and designed to be used in different delivery mediums (eg, mobile, personal computer, other). Such reliable rating would enable examination of programs leveraging more than one delivery medium and comparison between the qualities of programs regardless of

their delivery mediums. Recently, Gomez Quiñonez et al [35] demonstrated that a Web-based app targeting adult physical activity provided better or similar results compared to the same mHealth app. Subsequently, in a recent systematic review of digital parent training programs, only one mobile app was identified, which was used as an adjunct to treatment, whereas most computer programs were used as a standalone intervention [36]. It might be that, for certain treatment aims, different design aspects limit the potential of one delivery medium, which could be examined when using the same framework of evaluation. The nature of quality criteria is that they relate to broad principles in terms of product design, which facilitates their use in different contexts [37]. The remaining question in terms of a tool's generalizability is focused on the ability to use the same standardized measure to evaluate programs targeting different clinical aims (eg, behaviors related to medical conditions, mental health). During the development of previous eHealth-related rating scales, the reliability matrix of ratings based on different clinical aims was not demonstrated. It is important to pay attention to generalizability during the tool development phase because the development of scoring benchmarks might be influenced by the nature of reviewed programs.

Study Aims

The gaps in the literature laid the foundation for the development of Enlight, a suite of criteria-based measurements aimed at enabling scholars to objectively rate eHealth interventions based on different quality concepts regardless of their delivery medium or clinical aims. Therefore, the aims of this study were to (1) identify relevant published criteria through a comprehensive systematic review that also addresses quality criteria related to persuasive design, behavior change, and therapeutic alliance/principles (a systematic review that relates to different aspects of eHealth and mHealth interventions including all concepts mentioned was not incorporated before and would enable the establishment of the tool based on a comprehensive examination of the current know-how in this field); (2) develop Enlight, a suite of criteria-based quality measures related to separate aspects of eHealth programs; (3) establish the measures' reliability and generalizability in evaluating different delivery mediums and clinical aims; and (d) examine the intercorrelations between different quality constructs and between them and preliminary validity measures.

Methods

This study was completed in two parts, each reflecting different procedures used. The first part, "Enlight Development," included a systematic search for quality criteria, the classification of these criteria into core domains and subcategories, and the creation of the different scales. The second part involved reliability testing and further analysis of the interrelationships between the quality constructs and correlations with preliminary concurrent validity measures. As emphasized, we examined the results with respect to different delivery mediums and clinical aims.

Enlight Development

Systematic Review and Collection of Quality Criteria

The systematic review was carried out in line with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) statement guidelines [38] (see Multimedia Appendix 1 for a complete list of PsycINFO database search terms used). We conducted comprehensive computer searches of IEEE Xplore, PsycINFO, PubMed, and Science Direct databases for English articles published between January 1, 2000 and April 8, 2016, containing explicit Web- or mobile app-based quality criteria. The search time window was limited to 2000 due to rapid developments in technology [2] and to reflect technologies that largely meet the expectations of today's users [4]. For general quality criteria, we searched for papers in varying combinations for criteria (eg, "criteria" OR "principle*"), assessments (eg, "assess*" OR "measur*"), and delivery mediums (eg, "mobile*" OR "Web*"). We also searched for papers in varying combinations for criteria (eg, "criteria" OR "principle*") specifically related to persuasive design and behavior change. A manual search for additional references was conducted by examining the reference lists of identified papers and previous review articles. We also reviewed grey literature by Google searching, looking into key websites (eg, Nielsen Norman Group), and asking experts for recommendations. To identify English articles containing explicit quality rating criteria in terms of therapeutic alliance/principles, we conducted comprehensive computer searches of PsycINFO, PubMed, and Science Direct databases for articles published by April 8, 2016. We searched for papers related to quality criteria (eg, "potential" OR "criteria" OR "principle*") in the field of psychotherapy and for papers related to therapeutic alliance questionnaires.

Data Extraction and Categorization

Following the extraction of criteria from identified sources, we established a multidisciplinary advisory team to support the classification of these criteria into core domains and subcategories, and to support the development of the measures' items and categories (see Multimedia Appendix 2, Advisory Team). Because the criteria identified for therapeutic alliance ratings were not focused on eHealth interventions (but rather on a human therapist), a thematic analysis [39] was conducted to redefine these criteria in terms of eHealth interventions as a preliminary step before categorizing them and building the final scale. This step was carried out by three licensed clinical psychologists.

Enlight Testing

Identifying Relevant eHealth Intervention Programs

To establish the generalizability and transferability of the tool, we tested Enlight on programs targeting either modifiable behaviors related to chronic medical illnesses (ie, health-related behaviors) or mental health, and on programs delivered through mobile apps or websites (accessed through a personal computer). The systematic identification of relevant programs followed the PRISMA statement guidelines [38] (see Multimedia Appendix 3 for search terms used to identify free eHealth intervention programs). For health-related behaviors, we targeted behaviors considered to be among the leading preventable causes of death

due to chronic medical conditions in the United States [40]: diet, physical activity, smoking cessation, and alcohol cessation. For mental health, we focused our search on the terms depression, anxiety, mental health, and well-being. For both websites and mobile apps, we included only free-to-use programs aimed at the specific condition that were published in English.

To identify relevant mobile apps, we conducted a systematic search of the Google Play store on September 5, 2016, by using search terms relevant for each condition (eg, diet: diet or weight loss). The inclusion criteria for mobile apps were (1) English language, (2) free of charge, and (3) from Android categories "Health & Fitness" and "Medical" following a careful examination of program type in different Android categories. To identify relevant Web-based programs, a systematic search was conducted on September 5, 2016, using a Google Search query per condition (eg, depression, smoking cessation) paired with terms such as "free online" and "self-help." For each condition, we looked into the organic results found within the first two pages because studies have indicated that a negligible portion of users go beyond the second page [41,42]. If a source referring to a list of programs was found among these organic searches, we included those programs as well. The lists created through the searches for mobile apps and websites were then screened by title to remove duplicates and exclude irrelevant programs (eg, magazine). Programs with unclear titles were examined (by a person who was not one of the program quality raters) using Google Play or website home page prior to exclusion. Using a randomization website [43], 24 eHealth programs were then randomly selected for each of the four conditions—two delivery mediums (mobile/website) × two clinical aims (health-related behavior/mental health)—reaching a total of 96 programs. For example, during this process 24 mobile apps that targeted health-related behavior were randomly selected.

Raters' Training

Different raters evaluated the quality (KF) and checklist (NM) sections of the programs, with the study's leading author (AB) acting as the second independent rater for both sections. A total of 12 programs (the first three from the randomized lists for each of the four conditions) were used to pilot-test the scale. As part of this process, we also examined programs from other domains that were recommended by experts for their very high quality. This approach enabled coders to locate transcription errors and refine the coding scheme. One goal of the development was to achieve high interrater reliability at the construct level so that Enlight users could reliably present and compare construct scores between different programs. To achieve this goal, the refinement process followed the methods for developing benchmarks in thematic and projection test development within the psychology domain [44-46]. During this process, benchmarks were written, codes were refined, and a preliminary manual was developed. Once the raters had rated the programs (independently), ratings were shared simultaneously. Raters then met to discuss the ratings, and to make proper clarifications and adjustments of the scale benchmarks. To minimize potential biases, when disagreement occurred, a third person examined the blinded ratings and further discussed the ratings to enable the final refinement of codes.

Reliability Testing

In line with a previous examination in this field [11], a minimum sample size of 41 was required to establish with 87% assurance whether true interrater reliability lay within 0.15 of a sample observation of 0.80 [47,48]. Therefore, we rated 42 mobile apps (21 targeting mental health and 21 targeting health-related behaviors) and 42 website programs (with the same distribution). Accordingly, our reliability testing also included 42 programs targeting mental health and 42 programs targeting health-related behaviors, for a total of 84 eHealth intervention programs that were independently rated.

This paper presents the analysis of all Enlight categories based on raters' ratings, except for security checklist because security items are based on information retrieved from parties with access to products' servers (and not on raters' ratings). The interrater reliability for each of the quality assessment subscales was measured using an intraclass correlation coefficient (ICC) [49] utilizing the two-way mixed effects model with absolute agreement [50]. The internal consistency of the core domains were calculated using Cronbach alpha, which reflects how closely related a set of items are as a group [51]. Internal consistency was not examined for items related to objective requirements on a categorical scale (Enlight checklists) because homogeneity among the checklist items was not assumed (different checklist items might relate to different parts of the construct) [52]. Therefore, interrater reliability for each categorical item was determined. Cohen kappa, which measures the achieved agreement between two raters above and beyond the overall probability of random agreement, was applied [53].

Further Analysis

Overall, differences between delivery mediums and clinical aims in terms of quality scores and intercorrelations were examined while adjusting *P* values based on the Benjamini-Hochberg correction [54]. The correlation matrix between different quality assessments was examined using Pearson correlations. The relationship between these assessments was also examined using a conditional probability approach, which measures the probability of an event given that another event has occurred [55]. This analysis aimed to examine the percentage of products that met a certain quality standard from among a total sample of products that met another quality standard. This method enabled us to examine whether a certain range scores in one quality construct were associated with a similar range of scores in a different construct.

Preliminary concurrent validity was assessed by examining the correlations between different quality constructs and selected variables that were expected to relate either to a program's acceptability or its efficacy. Combined quality construct scores based on means of several quality constructs were also added to this analysis to examine the benefits of summarizing several concepts into a single score. Two of the selected variables, credibility checklist and programs backed by research evidence (evidence-based program), were developed as part of Enlight

and will be described in the Results section related to Enlight development.

The third variable, program popularity, aimed to examine preliminary acceptability based on the number of people choosing to use it. For websites, we recorded the Alexa traffic rank [56], which estimates a website's popularity based on a combination of mean daily visitors and page views [57]. This traffic rank was used following a preliminary step in which we compared ranks to the SimilarWeb traffic estimator [58], obtaining similar results. We excluded Web-based programs nested within larger websites from the analysis because a high percentage of users were expected to access the website for reasons other than the intervention program. The number of mobile app downloads was taken from Google Play, which presents the range of downloads (eg, 500-1000) for each app; for each program, the lower limit was documented (eg, 500).

Results

Enlight Development

The electronic and manual searches produced a total of 7903 records (see Multimedia Appendix 4 for a flow diagram). Through the first screening process, 181 papers were identified and retrieved for detailed evaluation and a total of 99 sources met all inclusion criteria (76 papers from peer-reviewed journals, 9 papers from conference proceedings, 7 manuscripts, 6 websites, and 1 book). A complete list of sources used in the criteria-gathering process is available with this paper (see Multimedia Appendix 5, sources list).

Overall, 1252 items were extracted from the sources; 143 were found to be not relevant for the evaluation of eHealth products and 633 were deemed to be duplicates, leaving a total of 476 criteria. Identified criteria were then grouped and organized in an iterative process into 10 constructs (see Table 1) and three sections to create Enlight: Classification (ie, classifying the program based on acknowledged categories), quality assessment, and checklists (Multimedia Appendix 6, Enlight).

Table 1. Frequency of explicit evaluation criteria for eHealth interventions by different constructs (N=476).

Criteria constructs	n (%)
Classification (intended users, clinical condition, program aim)	19 (4.0)
Usability (navigation, learnability, ease of use)	48 (10.1)
Visual design (aesthetics, layout, size)	35 (7.4)
User engagement (content presentation, interactive, not irritating, targeted/tailored/personalized, captivating)	45 (9.5)
Content (evidence-based content, quality of information provision, complete and concise, clarity about program's purpose)	79 (16.6)
Therapeutic persuasiveness (call for action, load reduction of activities, therapeutic rationale and pathway, rewards, real data driven/adaptive, ongoing feedback, expectations and relevance)	92 (19.3)
Therapeutic alliance (basic acceptance and support, positive therapeutic expectations, relatability)	45 (9.5)
General subjective evaluation (appropriate features to meet clinical aim, right mix of ability and motivation, likability)	36 (7.6)
Credibility[a] (owner's credibility, maintenance, strong advisory support, third-party endorsement, evidence for successful implementation, evidence-based program)	49 (10.3)
Privacy and security (terms of use, information on social platforms, security of data and transmission, documentation of data exposure, compliance, third-party endorsement)	28 (5.9)

[a] Also includes evidence-based program that is ranked and examined separately.

Quality Assessment Section

The quality assessment section was designed to capture the different qualities of eHealth interventions. It consists of 25 items divided into six core constructs related to the eHealth intervention program: usability, visual design, user engagement, content, therapeutic persuasiveness, and therapeutic alliance. Another construct, general subjective evaluation (of program's potential), asks the rater to subjectively evaluate the program as a whole following the completion of the core concept ratings. All constructs were based on heuristic evaluation to enable the examination of programs independently of empirical examination, built on a scale of 1 to 5 (1=very poor; 2=poor; 3=fair; 4=good; 5=very good), and calculated by averaging the items of which they are comprised.

Checklists Section

Overall, the checklists were based on acknowledged criteria that cover distinct domains related to product use and include credibility, evidence-based program (as a distinct part of programs' credibility), privacy explanation, and basic security. These checklists are not expected to directly impact the end user's experience of the product's efficacy; however, the criteria contained in the lists may expose the user (or provider) to acknowledged risks or benefits. These measures are calculated by aggregating the scores received in each of the respective categorical items—excluding evidence-based program because it is based on a five-point scale. The privacy explanation and basic security checklists are the only measures in which a lower score equates to better quality. The basic security checklist is the only measure that is not based on a rater's rating, but rather

on information retrieved from parties with access to the product's servers.

Enlight Testing

Reliability Testing

The electronic searches produced a total of 2227 mobile apps and 1283 Web-based programs (see Multimedia Appendix 4 for a flow diagram). Through the first screening process, 235 apps and 502 websites were excluded as duplicates and 1509 apps and 665 websites were excluded for not meeting the inclusion criteria. This left a total of 523 apps and 116 Web-based programs. From these, 96 eHealth programs were chosen through the randomization procedure, 12 of which were used for the training process and 84 for the reliability examination. The scores received by the various programs used for this paper's analyses are available in Multimedia Appendix 7 (programs' scores).

Tables 2 and 3 present the descriptive statistics, Cronbach alphas, and ICCs of the categories analyzed by Enlight. The descriptive statistics of the items constituting the quality assessment section are available in Multimedia Appendix 8 (Descriptive statistics of quality section items) and the interrater kappa reliability scores for the credibility checklist and privacy explanation checklist items (which were in the substantial to outstanding agreement range) are available in Multimedia Appendix 9 (kappa reliability scores of credibility and privacy explanation checklists items).

Table 2. Descriptive statistics, Cronbach alphas (α), and intraclass correlations (ICC) of assessment scores by different delivery mediums.

Clinical aim	Total (N=84)			Mobile (n=42)			Website (n=42)		
Quality ratings	Mean (SD)	α	ICC (95% CI)	Mean (SD)	α	ICC (95% CI)	Mean (SD)	α	ICC (95% CI)
Usability	3.31 (0.69)	.83	.91 (.86-.94)	3.46 (0.71)	.85	.82 (.68-.91)	3.17 (0.65)	.79	.96 (.92-.98)
Visual design	2.81 (0.82)	.84	.77 (.64-.85)	2.93 (0.83)	.88	.80 (.63-.89)	2.68 (0.79)	.85	.74 (.52-.86)
User engagement	2.62 (0.80)	.88	.90 (.78-.94)	2.47 (0.83)	.91	.92 (.83-.96)	2.78 (0.73)	.85	.85 (.64-.93)
Content	3.00 (0.98)	.90	.93 (.89-.96)	2.40[a] (0.87)	.91	.91 (.83-.95)	3.59[a] (0.68)	.78	.85 (.73-.92)
Therapeutic persuasiveness	2.23 (0.68)	.88	.88 (.78-.93)	2.11 (0.71)	.88	.93 (.86-.97)	2.35 (0.62)	.87	.78 (.55-.89)
Therapeutic alliance	2.20 (0.75)	.83	.89 (.72-.95)	1.99 (0.72)	.83	.87 (.72-.94)	2.40 (0.73)	.82	.87 (.54-.95)
General subjective evaluation	2.09 (0.91)	.89	.83 (.73-.89)	1.89 (0.84)	.88	.85 (.73-.92)	2.29 (0.93)	.89	.73 (.50-.86)
Credibility checklist	3.14 (1.50)	—[b]	.95 (.92-.97)	2.21[a] (1.16)	—	.95 (.90-.97)	4.07[a] (1.20)	—	.95 (.90-.97)
Evidence-based program	1.32 (0.66)	—[c]	.94 (.91-.96)	1.07[a] (0.34)	—	.92 (.86-.96)	1.57[a] (0.80)	—	.94 (.88-.97)
Privacy explanation checklist	2.76 (1.58)	—[b]	.98 (.97-.99)	3.33[a] (1.26)	—	.99 (.98-.99)	2.19[a] (1.67)	—	.97 (.95-.98)

[a] The groups (within the construct) differed significantly at Benjamini-Hochberg adjusted $P<.05$ in *t* test for two independent samples.

[b] Measure of agreement per categorical item (kappa) is presented in Multimedia Appendix 7.

[c] The score is based on one item; therefore, Cronbach alpha could not be calculated.

Table 3. Descriptive statistics, Cronbach alphas (α), and intraclass correlations (ICC) of assessment scores by different clinical aims.

Clinical aim	Health-related behaviors (n=42)			Mental health (n=42)		
Quality ratings	Mean (SD)	α	ICC (95% CI)	Mean (SD)	α	ICC (95% CI)
Usability	3.29 (0.77)	.83	.92 (.85-.96)	3.34 (0.61)	.84	.88 (.55-.96)
Visual design	2.79 (0.78)	.84	.78 (.55-.89)	2.82 (0.87)	.84	.79 (.55-.93)
User engagement	2.64 (0.79)	.90	.95 (.91-.97)	2.60 (0.81)	.84	.84 (.52-.94)
Content	2.90 (0.93)	.90	.86 (.74-.92)	3.09 (1.03)	.90	.95 (.91-.98)
Therapeutic persuasiveness	2.28 (0.66)	.86	.87 (.76-.93)	2.18 (0.70)	.90	.89 (.62-.95)
Therapeutic alliance	2.03 (0.70)	.77	.73 (.41-.87)	2.37 (0.78)	.87	.89 (.65-.95)
General subjective evaluation	2.03 (0.86)	.88	.72 (.48-.85)	2.15 (0.95)	.89	.85 (.73-.92)
Credibility checklist	2.88 (1.35)	—	.93 (.87-.97)	3.41 (1.61)	—	.96 (.92-.98)
Evidence-based program	1.21 (0.52)	—	.96 (.92-.98)	1.43 (0.77)	—	.93 (.88-.96)
Privacy explanation checklist	3.14 (1.44)	—	.97 (.94-.98)	2.38 (1.64)	—	.99 (.98-.99)

The internal consistencies of the Enlight categories were very high for the total sample (Cronbach alpha: range .83-.90, median .88) and also when separated into delivery mediums (mobile Cronbach alpha: range .83-.91, median .88; website Cronbach alpha: range .78-.89, median .85) or clinical aims (health-related behaviors Cronbach alpha: range .77-.90, median .86; mental health Cronbach alpha: range .84-.90, median .87). The interrater reliabilities of the Enlight categories were in the excellent to almost perfect agreement range for the total sample (ICC: range .77-.98, median .91) and also when separated into delivery mediums (mobile ICC: range .82-.99, median .92; website ICC: range .73-.97, median .86) or clinical aims (health-related behaviors ICC: range .72-.97, median .90; mental health ICC: range .79-.99, median .89). As can been seen in Table 2, significant differences were found between Web-based and mobile-based programs in terms of content, credibility checklist, evidence-based program, and privacy explanation checklist, all favoring the Web-based programs. No other significant differences in Enlight categories' scores were found between delivery mediums and clinical aims.

Further Analysis

The Pearson correlations between the quality assessment constructs are presented in Table 4. In the total sample, usability did not correlate with the other constructs. All other constructs exhibited significant, moderate to strong, positive correlations (*r*: range .34-.86; all $P \leq .001$). A pattern of strong, positive correlations was found between user engagement, content, therapeutic persuasiveness, and therapeutic alliance (*r*: range .68-.86; all $P<.001$). The correlations between the quality assessment constructs were similar when separated into the two clinical aims (health-related behaviors, mental health). However, some differences were found between delivery mediums. Therefore, Table 4 also presents the correlation matrix by delivery mediums. In the sample of mobile apps, usability did not correlate with the other constructs. In the sample of Web-based programs, weak-to-moderate, positive correlations were found between usability and the other constructs. Compared to the mobile app sample, the correlations between visual design and most of the other constructs in the Web-based program sample were numerically lower.

Table 4. Pearson correlations between quality assessment core concepts in the total sample and by delivery mediums and clinical aims.

Quality ratings	Usability		Visual design		Content		User engagement		Therapeutic persuasiveness	
	r	*P*	*r*	*P*	*r*	*P*	*r*	*P*	*r*	*P*
Total (N=84)										
Visual design	.36	.001								
Content	–.02	.84	.34	.001						
User engagement	.14	.21	.65	<.001	.68	<.001				
Therapeutic persuasiveness	.13	.23	.60	<.001	.69	<.001	.86	<.001		
Therapeutic alliance	.15	.16	.53	<.001	.75	<.001	.73	<.001	.72	<.001
Mobile (n=42)										
Visual design	.16		.31							
Content	–.05[a]	.76	.74	<.001						
User engagement	–.03	.87	.72	<.001	.83	<.001				
Therapeutic persuasiveness	.05[a]	.76	.73	<.001	.85	<.001	.89	<.001		
Therapeutic alliance	.002[a]	.99	.70	<.001	.81	<.001	.70	<.001	.73	<.001
Website (n=42)										
Visual design	.54	<.001								
Content	.41[a]	.008	.31	.04						
User engagement	.46[a]	.002	.67	<.001	.55	<.001				
Therapeutic persuasiveness	.34	.03	.53	<.001	.60	<.001	.81	<.001		
Therapeutic alliance	.47[a]	>.002	.51	.001	.73	<.001	.73	<.001	.70	<.001

[a] Significant differences in Pearson correlation values were found between the delivery mediums (mobile, website) using Fisher Z-transformation at Benjamini-Hoffman adjusted $P<.05$.

To further examine the relationship between usability and the other constructs in the mobile app sample, these correlations were recalculated after excluding mobile apps with very few features (n=12; see mobile apps marked with "a" in studies in Multimedia Appendix 5). These mobile apps were identified by the raters to receive high usability scores only because they were very lean and therefore easy to learn and use, and not because of specific design aspects enhancing their usability.

For the remaining sample of mobile apps (n=30), moderate positive correlations were found between usability and the other constructs (visual design: r=.55, P=.002; user engagement: r=.43, P=.02; content: r=.41, P=.03; therapeutic persuasiveness: r=.35, P=.055; therapeutic alliance, r=.57, P=.001).

To further examine the pattern of strong correlations found between user engagement, content, therapeutic persuasiveness, and therapeutic alliance independent of delivery mediums or clinical aims, a conditional probability analysis was performed by examining the percentage of programs with a score of fair or above (≥3.0) in one construct out of the sample of programs that received a score of fair or above in another construct (Figure 1). Usability and visual design were also added to Figure 1 to present the readers with an overview of all quality constructs.

As Figure 1 shows, 100% of the eHealth intervention programs that received a score of fair or above in therapeutic persuasiveness or therapeutic alliance also received this range of scores in user engagement and content. For programs receiving a score of fair or above in user engagement or content, the percentages of programs receiving the same range of scores in therapeutic persuasiveness or therapeutic alliance ranged between 33% and 64%. In effect, having a fair score in user engagement or content did not necessarily mean that the program also received a fair score in therapeutic persuasiveness or therapeutic alliance. A similar pattern appeared between user engagement and content, where having a fair or above score in user engagement meant that the program most likely had a fair or above score in content (94%), but this pattern was not apparent in the opposite direction. Finally, the figure indicates that most programs that received a score of fair or above in any construct other than usability also received the same range of scores in usability (77.6% to 88.2%).

Figure 1. Percentages of eHealth intervention programs with a fair or above score (≥3.0) in quality constructs (columns) out of the sample of programs that received a score of fair or above (≥3.0) in another construct (rows). Within this study sample, higher percentages indicate that having the examined range of scores in one construct (row) improves the chances of receiving the same range of scores in the other construct (column). Fields are colored from higher to lower percentages by the following order: red (highest), orange, yellow, and green (lowest).

Construct Condition	Usability	Visual Design	User Engagement	Content	Therapeutic Persuasiveness	Therapeutic Alliance
Usability	n/a	50.0%	45.0%	63.3%	21.7%	30.0%
Visual Design	88.2%	n/a	64.7%	70.6%	41.2%	44.1%
User Engagement	81.8%	66.7%	n/a	93.9%	48.5%	63.6%
Content	77.6%	49.0%	63.3%	n/a	32.7%	42.9%
Therapeutic Persuasiveness	81.3%	87.5%	100.0%	100.0%	n/a	81.3%
Therapeutic Alliance	85.7%	71.4%	100.0%	100.0%	61.9%	n/a

Preliminary Concurrent Validity

Table 5 presents the Pearson correlations between the quality constructs and general subjective evaluation, credibility checklist, evidence-based program (empirical research evidence), and program popularity scores. It is important to note that general subjective evaluation was placed in this table because of its applicability to a general examination of programs and from an organizational point of view; however, this variable is considered to be biased because this evaluation was done following the rater's full examination of the quality constructs. Three aggregated quality construct scores were also added to this analysis: (1) the mean of user engagement, content, therapeutic persuasiveness, and therapeutic alliance scores; (2) the mean of all quality constructs excluding therapeutic persuasiveness and therapeutic alliance (traditional total), and (3) the mean of all quality constructs (total).

Table 5. Pearson correlations between Enlight quality constructs and selected variables.

Quality ratings	General subjective evaluation (n=84)		Credibility checklist (n=84)		Evidence-based program (n=84)		Program popularity (n=70)[a]	
	r	*P*	*r*	*P*	*r*	*P*	*r*	*P*
Usability	.18	.11	–.18	.10	.01	.96	.07	.54
Visual design	.57	<.001	–.02	.88	–.06	.58	.27	.02
User engagement	.77	<.001	.22	.04	.22	.050	.40	.001
Content	.70	<.001	.60	<.001	.44	<.001	.12	.34
Therapeutic persuasiveness	.76	<.001	.22	.049	.26	.02	.41	<.001
Therapeutic alliance	.81	<.001	.31	.004	.31	.004	.21	.08
User engagement + content + therapeutic persuasiveness + therapeutic alliance	.84	<.001	.40	<.001	.35	.001	.30	.01
Traditional total[b]	.78	<.001	.26	.02	.23	.03	.30	.01
Total[c]	.83	<.001	.28	.009	.27	.01	.32	.007

[a] The analysis excluded 14 Web-based programs nested within larger websites (see programs marked with a "b" Multimedia Appendix 5) because a high percentage of users were expected to access the website for reasons other than the intervention program.

[b] Traditional total=mean of all constructs excluding therapeutic persuasiveness and therapeutic alliance.

[c] Total=mean of all constructs.

Overall, no single construct had a correlation pattern that outperformed the others; however, user engagement, content, therapeutic persuasiveness, and therapeutic alliance showed positive correlations with all variables, most of which were significant (13/16; *r*: range .22-.81; all *P*<.05). The combined scores showed significant, positive, weak-to-moderate correlations with credibility checklist, evidence-based program, and program popularity scores (*r*: range .23-.40; all *P*<.05); compared to the other combined scores, traditional total showed numerically underperformed results. All combined scores showed significant, positive correlations with all selected variables.

No significant differences in Pearson correlations were found between different delivery mediums or clinical aims using Fisher Z-transformation at Benjamini-Hoffman adjusted *P*<.05 with two exceptions: significant differences in the usability and credibility checklist correlation, and usability and evidence-based program correlation were found between delivery mediums. These correlations were negative in mobile-based programs (*r*=–.34 and *r*=–.37, respectively) and positive in Web-based programs (*r*=.27 and *r*=.31, respectively).

Discussion

Enlight is a comprehensive suite of assessments developed to evaluate the quality of eHealth intervention programs. It was developed following the first systematic review assessing different aspects of both eHealth and mHealth interventions, including persuasive design, behavior change, and therapeutic alliance principles. As part of the quality assessment section, two concepts that relate to a program's therapeutic potential and did not appear in previous scales (eg, [10,11]) were introduced: therapeutic persuasiveness and therapeutic alliance. To our knowledge, Enlight is also the first suite of heuristic-based quality measures to include separate sections, one for quality aspects that cover the user's experience and the other for those quality aspects that do not directly alter the user's experience of the program (eg, team's credibility is not part of any construct within the quality assessment section). Combined with the high interrater reliability scores at the construct level (ICC: range .77-.98, median .91), these findings suggest that Enlight differs from previous work by enabling stakeholders to objectively examine individual quality constructs; in that way, Enlight is a suite of scales rather than one quality measure.

The results indicate that it is important to examine therapeutic persuasiveness and therapeutic alliance as part of the main quality constructs of eHealth intervention programs. Most importantly, a conditional probability analysis revealed that 100% of the eHealth intervention programs that received a score of fair or above in therapeutic persuasiveness or therapeutic alliance received the same range of scores in user engagement and content. For programs with fair or above scores in user engagement and content, only 33% to 64% of them received the same range of scores in therapeutic persuasiveness or therapeutic alliance. This means that, despite the strong, positive correlations found between the four aforementioned constructs, the relationships between them are more complicated: achieving a certain standard of scores in user engagement or content does not necessarily mean that this standard will be achieved in the therapeutic constructs and therefore justifies separate ratings of these constructs. Second, when examining the correlations between combined scores and variables that were expected to relate to either the program's acceptability or its efficacy, the combined score that did not include therapeutic persuasiveness or alliance descriptively underperformed other combined scores, but the difference was small. Nevertheless, altogether these preliminary findings are congruent with findings from other studies, showing that persuasive design and behavior change

principles are important factors in understanding the potential of eHealth intervention programs [17,19,20]. These findings also correspond with previous studies suggesting that eHealth intervention programs' facilitation of a therapeutic alliance may play a role in the understanding of these programs' potential [28-33].

The analyses pointed to significant, positive correlations between the combined quality scores and credibility, evidence-based program, and program popularity scores (*r*: range .23-.41; all $P \leq .02$). At the construct level, user engagement, content, therapeutic persuasiveness, and therapeutic alliance demonstrated a pattern of positive correlations with these variables, although results were not always significant (13/16, $P<.05$). These preliminary findings relate to Enlight's concurrent validity, but they should be interpreted with caution because two of the selected variables only indirectly relate to program efficacy. Nevertheless, the program popularity score is related to a product's acceptance and the evidence-based program score is directly related to the availability of sound research evidence on the product's efficacy. These findings highlight the potential of this tool despite the need for further examination as discussed later.

Integrating the Quality Assessment Section Scores

Overall, combining quality constructs into aggregated scores by averaging them was introduced by previous scale developers [47] and is the simplest way to integrate the different constructs. As demonstrated, this integration was supported by empirical evidence showing that all combined scores had significant, positive correlations with the selected variables. However, the study results indicate that it might not be beneficial to combine usability with other scores in a straightforward manner. This is because lean programs may contain very limited content and features can be very easy to learn and use. As a result, such programs yield high usability scores, but very low scores in content, engagement, or therapeutic constructs. Our finding is congruent with previous studies suggesting that usability might need to be considered as a barrier to, rather than a facilitator of, effective interventions (eg, [59,60]). It also might be that in different cases different constructs are more important or redundant. For example, Althoff et al showed that the mobile app "Pokemon Go," which asks users to move between different physical locations to advance in the game, has contributed to an increase in users' physical activity [61]. Engagement in these kinds of games might equate to beneficial outcomes; therefore, in this kind of case, the therapeutic persuasiveness construct may become redundant. To conclude, more studies are needed to examine the relationships between the different constructs before determining how to integrate them in a way that accurately captures the potential of different interventions.

Generalizability

The reliability analysis demonstrated a similar range of interrater agreement and internal consistency in the different delivery mediums and clinical aims. Further analysis revealed similar range of correlations between quality assessment scores and scores of acceptability and efficacy in the different groups, suggesting that these quality ratings account for the same phenomena in these groups. This is the result of applying heuristic-based evaluation techniques that target general principles of quality. For example, principles of therapeutic persuasiveness, such as "therapeutic rational," or principles of content, such as "information provision," do not distinguish between delivery mediums or clinical aims. In our review process, we also did not identify important principles that relate to the quality of programs cannot be accounted for by a specific delivery medium. These results extend the work of previous criteria-based tools by pointing at the first time to the possibility of objectively rating different eHealth interventions using one tool, regardless of their delivery medium or clinical aim. Our analyses also identified some significant differences between Web-based and mobile-based programs in terms of content, credibility, evidence-based program, and privacy explanation checklist, all favoring Web-based interventions. It is important to note that Web-based interventions have been around for longer; therefore, there have been more opportunities for empirically based revisions and for scholars to become an integral part of the field.

Limitations

This study has several limitations that should be addressed. Even though correlations between Enlight quality scores and relevant variables related to concurrent validity were introduced, this does not fully demonstrate criterion validity in directly predicting a program's acceptability or efficacy. This could be examined once sound data about user analytics and outcome reports are available for a set of rated programs. A further limitation is that at this point we cannot suggest a single strategy for combining quality assessment scores or what range of scores would be good enough to create desired outcomes; rather, we suggest presenting several constructs and examining the relationships between them until there is more evidence backing a specific approach to score integration. As discussed, the high reliability demonstrated by Enlight enables to present all scores at the construct level (since they can be regarded as separate objective metrics).

Future Directions

Several future directions for research and practice were identified. First, the predictive validity of Enlight could be examined by rating different programs for which metrics of acceptance and efficacy are accessible, and investigating whether and which quality scores predict engagement and efficacy. Second, it could be beneficial to develop a model of the various relationships between different constructs once many programs have been rated. Moreover, specifically examining therapeutic alliance and therapeutic persuasiveness could be helpful when trying to assess the additional need in human support to enhance adherence [62] or to provide any other benefits in eHealth interventions. Third, it could be helpful to examine the applicability of training people to use Enlight based on a complete self-help manual that includes a training kit. Finally, it would be beneficial to examine how this kind of tool can support the decision making of health system leaders when adopting new programs. Efforts are underway to conduct such an examination at Northwell Health, New York.

Conclusions

This paper provides empirical findings that emphasize the importance of examining persuasive design and therapeutic alliance in the context of quality rating. It also demonstrates the applicability of objectively rating different eHealth interventions using one suite of measures, regardless of their delivery mediums or clinical aims, providing that raters are appropriately trained. The high reliability matrix and preliminary concurrent validity indicate the tool's potential to examine eHealth programs and the multimodal relationships between different aspects of program quality. More research is needed to establish the tool's validity for predicting the efficacy of eHealth programs.

Acknowledgments

The authors would like to thank Ofer Holtsman, Eytan Majar, Ofer Reichman, Ohad Samet, and Amit Shabtay for their support in Enlight development.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Complete PsycINFO database search terms used.

[PDF File (Adobe PDF File), 30KB - jmir_v19i3e82_app1.pdf]

Multimedia Appendix 2

Advisory team.

[PDF File (Adobe PDF File), 43KB - jmir_v19i3e82_app2.pdf]

Multimedia Appendix 3

Search terms used to identify free eHealth intervention programs.

[PDF File (Adobe PDF File), 25KB - jmir_v19i3e82_app3.pdf]

Multimedia Appendix 4

Flow diagrams.

[PDF File (Adobe PDF File), 52KB - jmir_v19i3e82_app4.pdf]

Multimedia Appendix 5

Sources list.

[PDF File (Adobe PDF File), 117KB - jmir_v19i3e82_app5.pdf]

Multimedia Appendix 6

Enlight.

[PDF File (Adobe PDF File), 172KB - jmir_v19i3e82_app6.pdf]

Multimedia Appendix 7

eHealth program scores.

[PDF File (Adobe PDF File), 72KB - jmir_v19i3e82_app7.pdf]

Multimedia Appendix 8

Descriptive statistics of quality section items.

[PDF File (Adobe PDF File), 49KB - jmir_v19i3e82_app8.pdf]

Multimedia Appendix 9

Kappa reliability scores for Credibility and Privacy Explanation Checklists items.

[PDF File (Adobe PDF File), 44KB - jmir_v19i3e82_app9.pdf]

References

1. Aitken M, Gauntlett C. Patient Apps for Improved Healthcare from Novelty to Mainstream. Parsippany, NJ: IMS Institute for Healthcare Informatics; 2013.
2. Kumar S, Nilsen WJ, Abernethy A, Atienza A, Patrick K, Pavel M, et al. Mobile health technology evaluation: the mHealth evidence workshop. Am J Prev Med 2013 Aug;45(2):228-236 [FREE Full text] [doi: 10.1016/j.amepre.2013.03.017] [Medline: 23867031]
3. Baumel A. Making the case for a feasible evaluation method of available e-mental health products. Adm Policy Ment Health 2016 Sep 12:1-4. [doi: 10.1007/s10488-016-0764-z] [Medline: 27620191]
4. Baumel A, Muench F. Heuristic evaluation of eHealth interventions: establishing standards that relate to the therapeutic process perspective. JMIR Ment Health 2016 Jan 13;3(1):e5 [FREE Full text] [doi: 10.2196/mental.4563] [Medline: 26764209]
5. Hekler EB, Klasnja P, Riley WT, Buman MP, Huberty J, Rivera DE, et al. Agile science: creating useful products for behavior change in the real world. Transl Behav Med 2016 Jun;6(2):317-328 [FREE Full text] [doi: 10.1007/s13142-016-0395-7] [Medline: 27357001]
6. Patrick K, Hekler E, Estrin D, Mohr DC, Riper H, Crane D, et al. The pace of technologic change: implications for digital health behavior intervention research. Am J Prev Med 2016 Nov;51(5):816-824. [doi: 10.1016/j.amepre.2016.05.001] [Medline: 27745681]
7. Powell AC, Torous J, Chan S, Raynor GS, Shwarts E, Shanahan M, et al. Interrater reliability of mHealth app rating measures: analysis of top depression and smoking cessation apps. JMIR Mhealth Uhealth 2016;4(1):e15 [FREE Full text] [doi: 10.2196/mhealth.5176] [Medline: 26863986]
8. Charnock D, Shepperd S, Needham G, Gann R. DISCERN: an instrument for judging the quality of written consumer health information on treatment choices. J Epidemiol Community Health 1999 Feb;53(2):105-111 [FREE Full text] [Medline: 10396471]
9. Olsina L, Rossi G. Measuring web application quality with WebQEM. IEEE Multimedia 2002;9(4):20-29.
10. Provost M, Koompalum D, Dong D, Martin BC. The initial development of the WebMedQual scale: domain assessment of the construct of quality of health web sites. Int J Med Inform 2006 Jan;75(1):42-57. [doi: 10.1016/j.ijmedinf.2005.07.034] [Medline: 16169770]
11. Stoyanov S, Hides L, Kavanagh D, Zelenko O, Tjondronegoro D, Mani M. Mobile app rating scale: a new tool for assessing the quality of health mobile apps. JMIR Mhealth Uhealth 2015 Mar 11;3(1):e27 [FREE Full text] [doi: 10.2196/mhealth.3422] [Medline: 25760773]
12. PsyberGuide. 2015. URL: http://psyberguide.org/ [accessed 2015-04-10] [WebCite Cache ID 6Xh99eQmZ]
13. Anxiety and Depression Association of America. 2016. Mental health apps URL: https://www.adaa.org/finding-help/mobile-apps [accessed 2016-11-02] [WebCite Cache ID 6lhtKNvkc]
14. JMIR Publications. 2015. Contribute to mHealth research and development of the new JMIR mHealth peer-review tool for mobile apps! URL: http://www.jmir.org/announcement/view/77 [accessed 2015-08-10] [WebCite Cache ID 6agJfjxOb]
15. Fogg B. Persuasive Technology: Using Computers to Change What We Think and Do (Interactive Technologies). San Francisco, CA: Morgan Kaufmann; 2002.
16. Persuasion in Design. 2017. URL: https://uxmag.com/articles/persuasion-in-design [accessed 2017-02-09] [WebCite Cache ID 6o9P2xMjD]
17. Kientz J, Choe E, Birch B. Heuristic evaluation of persuasive health technologies. 2010 Presented at: 1st ACM International Health Informatics Symposium; Nov 11-12, 2010; Arlington, VA.
18. Nielsen J, Molich R. Heuristic evaluation of user interfaces. 1990 Presented at: CHI '90 SIGCHI Conference on Human Factors in Computing Systems; Apr 1-5, 1990; Seattle, WA.
19. Kelders SM, Kok RN, Ossebaard HC, Van Gemert-Pijnen JE. Persuasive system design does matter: a systematic review of adherence to web-based interventions. J Med Internet Res 2012;14(6):e152 [FREE Full text] [doi: 10.2196/jmir.2104] [Medline: 23151820]
20. Webb TL, Joseph J, Yardley L, Michie S. Using the internet to promote health behavior change: a systematic review and meta-analysis of the impact of theoretical basis, use of behavior change techniques, and mode of delivery on efficacy. J Med Internet Res 2010;12(1):e4 [FREE Full text] [doi: 10.2196/jmir.1376] [Medline: 20164043]
21. Martin DJ, Garske JP, Davis MK. Relation of the therapeutic alliance with outcome and other variables: a meta-analytic review. J Consult Clin Psychol 2000 Jun;68(3):438-450. [Medline: 10883561]
22. Kazdin AE, Marciano PL, Whitley MK. The therapeutic alliance in cognitive-behavioral treatment of children referred for oppositional, aggressive, and antisocial behavior. J Consult Clin Psychol 2005 Aug;73(4):726-730. [doi: 10.1037/0022-006X.73.4.726] [Medline: 16173860]

23. Klein DN, Schwartz JE, Santiago NJ, Vivian D, Vocisano C, Castonguay LG, et al. Therapeutic alliance in depression treatment: controlling for prior change and patient characteristics. J Consult Clin Psychol 2003 Dec;71(6):997-1006. [doi: 10.1037/0022-006X.71.6.997] [Medline: 14622075]
24. Ferreira PH, Ferreira ML, Maher CG, Refshauge KM, Latimer J, Adams RD. The therapeutic alliance between clinicians and patients predicts outcome in chronic low back pain. Phys Ther 2013 Apr;93(4):470-478. [doi: 10.2522/ptj.20120137] [Medline: 23139428]
25. Kim SC, Kim S, Boren D. The quality of therapeutic alliance between patient and provider predicts general satisfaction. Mil Med 2008 Jan;173(1):85-90. [Medline: 18251337]
26. Bennett JK, Fuertes JN, Keitel M, Phillips R. The role of patient attachment and working alliance on patient adherence, satisfaction, and health-related quality of life in lupus treatment. Patient Educ Couns 2011 Oct;85(1):53-59. [doi: 10.1016/j.pec.2010.08.005] [Medline: 20869188]
27. Molassiotis A, Morris K, Trueman I. The importance of the patient-clinician relationship in adherence to antiretroviral medication. Int J Nurs Pract 2007 Dec;13(6):370-376. [doi: 10.1111/j.1440-172X.2007.00652.x] [Medline: 18021166]
28. Bickmore T, Gruber A, Picard R. Establishing the computer-patient working alliance in automated health behavior change interventions. Patient Educ Couns 2005 Oct;59(1):21-30. [doi: 10.1016/j.pec.2004.09.008] [Medline: 16198215]
29. Clarke J, Proudfoot J, Whitton A, Birch M, Boyd M, Parker G, et al. Therapeutic alliance with a fully automated mobile phone and web-based intervention: secondary analysis of a randomized controlled trial. JMIR Ment Health 2016 Feb 25;3(1):e10 [FREE Full text] [doi: 10.2196/mental.4656] [Medline: 26917096]
30. Bergman NL, Carlbring P, Linna E, Andersson G. Role of the working alliance on treatment outcome in tailored internet-based cognitive behavioural therapy for anxiety disorders: randomized controlled pilot trial. JMIR Res Protoc 2013;2(1):e4 [FREE Full text] [doi: 10.2196/resprot.2292] [Medline: 23612437]
31. Cavanagh K, Millings A. (Inter) personal computing: The role of the therapeutic relationship in e-mental health. J Contemp Psychother 2013;43(4):197-206.
32. Holter MT, Johansen A, Brendryen H. How a fully automated eHealth program simulates three therapeutic processes: a case study. J Med Internet Res 2016 Jun 28;18(6):e176 [FREE Full text] [doi: 10.2196/jmir.5415] [Medline: 27354373]
33. Barazzone N, Cavanagh K, Richards DA. Computerized cognitive behavioural therapy and the therapeutic alliance: a qualitative enquiry. Br J Clin Psychol 2012 Nov;51(4):396-417. [doi: 10.1111/j.2044-8260.2012.02035.x] [Medline: 23078210]
34. Eysenbach G, Powell J, Kuss O, Sa E. Empirical studies assessing the quality of health information for consumers on the world wide web: a systematic review. JAMA 2002;287(20):2691-2700. [Medline: 12020305]
35. Gomez QS, Walthouwer MJ, Schulz DN, de VH. mHealth or eHealth? Efficacy, use, and appreciation of a web-based computer-tailored physical activity intervention for dutch adults: a randomized controlled trial. J Med Internet Res 2016 Nov 09;18(11):e278 [FREE Full text] [doi: 10.2196/jmir.6171] [Medline: 27829576]
36. Baumel A, Pawar A, Mathur N, Kane J, Corell C. Technology assisted parent training programs for children and adolescents with disruptive behaviors: a systematic review. J Clin Psychiatry 2017:A (forthcoming).
37. Nielsen J. Enhancing the explanatory power of usability heuristics. 1994 Presented at: SIGCHI Conference on Human Factors in Computing Systems; Apr 24-28, 1994; Boston, MA.
38. Liberati A, Altman DG, Tetzlaff J, Mulrow C, Gøtzsche PC, Ioannidis JP, et al. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: explanation and elaboration. Ann Intern Med 2009 Aug 18;151(4):W65-W94. [Medline: 19622512]
39. Braun V, Clarke V. Using thematic analysis in psychology. Qual Res Psychol 2006 Jan;3(2):77-101. [doi: 10.1191/1478088706qp063oa]
40. Mokdad AH, Marks JS, Stroup DF, Gerberding JL. Actual causes of death in the United States, 2000. JAMA 2004 Mar 10;291(10):1238-1245. [doi: 10.1001/jama.291.10.1238] [Medline: 15010446]
41. Van Deursen AJ, Van Dijk JA. Using the Internet: skill related problems in users' online behavior. Interact Comput 2009;21(5):393-402.
42. iProspect. Search engine user behavior study. 2006. URL: http://district4.extension.ifas.ufl.edu/Tech/TechPubs/WhitePaper_2006_SearchEngineUserBehavior.pdf [WebCite Cache ID 6p0pvxfeq]
43. RANDOM.ORG. 2016. List randomizer URL: https://www.random.org/lists/ [accessed 2016-12-22] [WebCite Cache ID 6mwaCaO2t]
44. Westen D, Lohr N. Social Cognition and Object Relations Scale (SCORS): Manual for Coding TAT Data. Ann Arbor, MI: Department of Psychology, University of Michigan; 1990.
45. Blatt S, Chevron E, Quinlan D, Schaffer C, Wein S. The Assessment of Qualitative and Structural Dimensions of Object Representations (revised edition). New Haven, CT: Yale University; 1988.
46. Cramer P. The Development of Defense Mechanisms: Theory, Research, and Assessment. New York: Springer Science & Business Media; 2012.
47. Newcombe RG. Confidence Intervals for Proportions and Related Measures of Effect Size. Boca Raton, FL: CRC Press; 2012.

48. Zou GY. Sample size formulas for estimating intraclass correlation coefficients with precision and assurance. Stat Med 2012 Dec 20;31(29):3972-3981. [doi: 10.1002/sim.5466] [Medline: 22764084]
49. Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. Psychol Bull 1979 Mar;86(2):420-428. [Medline: 18839484]
50. Bartko JJ. The intraclass correlation coefficient as a measure of reliability. Psychol Rep 1966 Aug;19(1):3-11. [doi: 10.2466/pr0.1966.19.1.3] [Medline: 5942109]
51. Cronbach L. Coefficient alpha and the internal structure of tests. Psychometrika 1951;16(3):297-334.
52. Streiner DL. Being inconsistent about consistency: when coefficient alpha does and doesn't matter. J Pers Assess 2003 Jun;80(3):217-222. [doi: 10.1207/S15327752JPA8003_01] [Medline: 12763696]
53. Sim J, Wright CC. The kappa statistic in reliability studies: use, interpretation, and sample size requirements. Phys Ther 2005 Mar;85(3):257-268 [FREE Full text] [Medline: 15733050]
54. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Series B Stat Methodol 1995;57(1):289-300.
55. Gut A. Probability: A Graduate Course. Vol 75. New York: Springer Science & Business Media; 2012.
56. Alexa. 2017. URL: http://www.alexa.com/ [accessed 2017-01-04] [WebCite Cache ID 6nGSLql6u]
57. Olteanu A, Peshterliev S, Liu X, Aberer K. Web credibility: features exploration and credibility prediction. 2013 Presented at: 35th European conference on Advances in Information Retrieval; Mar 24-27, 2013; Moscow, Russia.
58. SimilarWeb. 2017. URL: http://www.similarweb.com/ [accessed 2017-03-10] [WebCite Cache ID 6oruwtIA6]
59. O'Brien H, Toms E. What is user engagement? A conceptual framework for defining user engagement with technology. J Assoc Inf Sci Technol 2008;59(6):938-955.
60. Atkinson M, Kydd C. Individual characteristics associated with World Wide Web use: an empirical study of playfulness and motivation. SIGMIS Database 1997 Apr 01;28(2):53-62. [doi: 10.1145/264701.264705]
61. Althoff T, White RW, Horvitz E. Influence of Pokémon Go on physical activity: study and implications. J Med Internet Res 2016 Dec 06;18(12):e315 [FREE Full text] [doi: 10.2196/jmir.6759] [Medline: 27923778]
62. Mohr DC, Cuijpers P, Lehman K. Supportive accountability: a model for providing human support to enhance adherence to eHealth interventions. J Med Internet Res 2011;13(1):e30 [FREE Full text] [doi: 10.2196/jmir.1602] [Medline: 21393123]

Abbreviations

ICC: intraclass correlation coefficient
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses

Edited by G Eysenbach; submitted 09.01.17; peer-reviewed by O Anya, J Torous; comments to author 02.02.17; revised version received 21.02.17; accepted 22.02.17; published 21.03.17.

Please cite as:
Baumel A, Faber K, Mathur N, Kane JM, Muench F
Enlight: A Comprehensive Quality and Therapeutic Potential Evaluation Tool for Mobile and Web-Based eHealth Interventions
J Med Internet Res 2017;19(3):e82
URL: http://www.jmir.org/2017/3/e82/
doi:10.2196/jmir.7270
PMID:28325712

Original Paper

Internet-Delivered Health Interventions That Work: Systematic Review of Meta-Analyses and Evaluation of Website Availability

Mary AM Rogers[1], MS, PhD; Kelsey Lemmen[1], BS; Rachel Kramer[1], BS; Jason Mann[1], MSA; Vineet Chopra[1], MSc, MD

Department of Internal Medicine, University of Michigan, Ann Arbor, MI, United States

Corresponding Author:
Mary AM Rogers, MS, PhD
Department of Internal Medicine
University of Michigan
Building 16, Room 422W, North Campus Research Complex
2800 Plymouth Road
Ann Arbor, MI, 48109-2800
United States
Phone: 1 734 647 8851
Fax: 1 734 936 8944
Email: maryroge@umich.edu

Abstract

Background: Due to easy access and low cost, Internet-delivered therapies offer an attractive alternative to improving health. Although numerous websites contain health-related information, finding evidence-based programs (as demonstrated through randomized controlled trials, RCTs) can be challenging. We sought to bridge the divide between the knowledge gained from RCTs and communication of the results by conducting a global systematic review and analyzing the availability of evidence-based Internet health programs.

Objectives: The study aimed to (1) discover the range of health-related topics that are addressed through Internet-delivered interventions, (2) generate a list of current websites used in the trials which demonstrate a health benefit, and (3) identify gaps in the research that may have hindered dissemination. Our focus was on Internet-delivered self-guided health interventions that did not require real-time clinical support.

Methods: A systematic review of meta-analyses was conducted using Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines (PROSPERO Registration Number CRD42016041258). MEDLINE via Ovid, PsycINFO, Embase, Cochrane Database of Systematic Reviews, and the Cumulative Index to Nursing and Allied Health Literature (CINAHL) were searched. Inclusion criteria included (1) meta-analyses of RCTs, (2) at least one Internet-delivered intervention that measured a health-related outcome, and (3) use of at least one self-guided intervention. We excluded group-based therapies. There were no language restrictions.

Results: Of the 363 records identified through the search, 71 meta-analyses met inclusion criteria. Within the 71 meta-analyses, there were 1733 studies that contained 268 unique RCTs which tested self-help interventions. On review of the 268 studies, 21.3% (57/268) had functional websites. These included evidence-based Web programs on substance abuse (alcohol, tobacco, cannabis), mental health (depression, anxiety, post-traumatic stress disorder [PTSD], phobias, panic disorders, obsessive compulsive disorder [OCD]), and on diet and physical activity. There were also evidence-based programs on insomnia, chronic pain, cardiovascular risk, and childhood health problems. These programs tended to be intensive, requiring weeks to months of engagement by the user, often including interaction, personalized and normative feedback, and self-monitoring. English was the most common language, although some were available in Spanish, French, Portuguese, Dutch, German, Norwegian, Finnish, Swedish, and Mandarin. There were several interventions with numbers needed to treat of <5; these included painACTION, Mental Health Online for panic disorders, Deprexis, Triple P Online (TPOL), and U Can POOP Too. Hyperlinks of the sites have been listed.

Conclusions: A wide range of evidence-based Internet programs are currently available for health-related behaviors, as well as disease prevention and treatment. However, the majority of Internet-delivered health interventions found to be efficacious in RCTs do not have websites for general use. Increased efforts to provide mechanisms to host “interventions that work” on the Web and to assist the public in locating these sites are necessary.

(J Med Internet Res 2017;19(3):e90) doi:10.2196/jmir.7111

KEYWORDS

Internet; public health; randomized controlled trial; computer-assisted therapy; global health

Introduction

Background

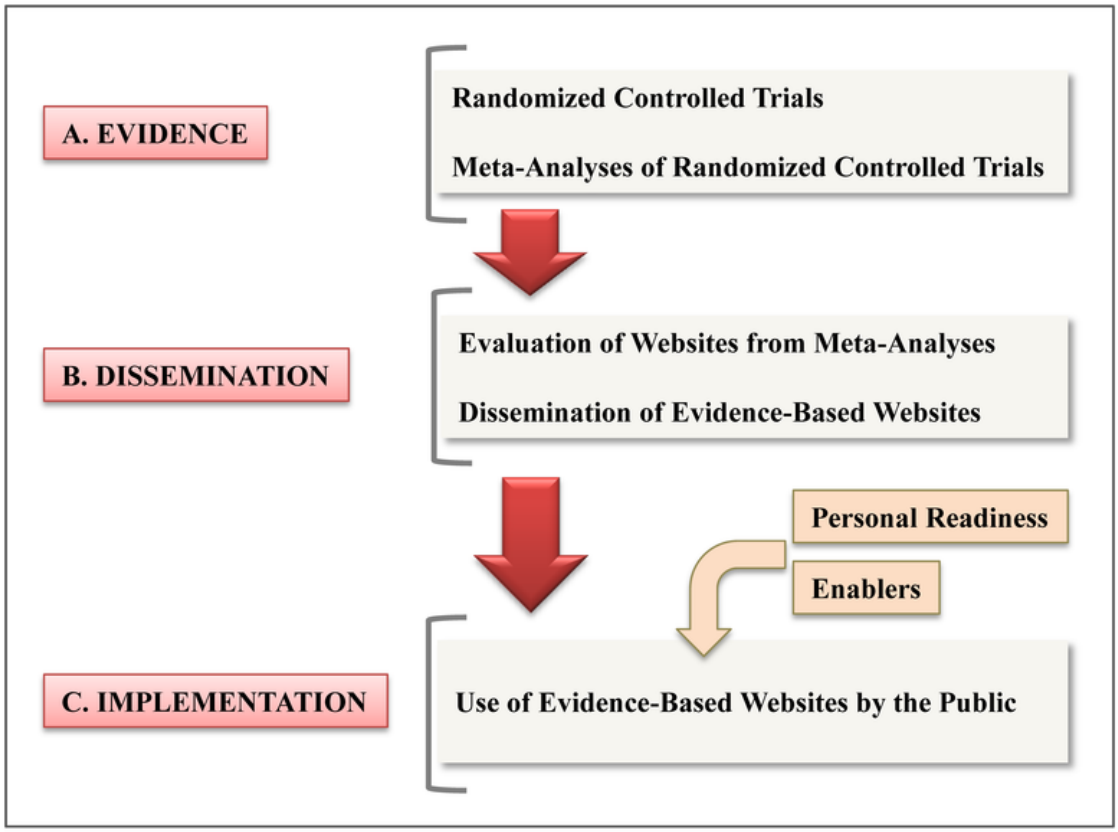
The World Health Organization recognizes that implementation of population-based strategies to improve health is critical [1,2]. Likewise, the Institute of Medicine's list of suggestions for action includes the implementation of population-based strategies to improve health [3]. The need for population approaches to solve health problems was recently reviewed by David Hunter as he stated, "As countries struggle to transform their health systems to cope with rising demand, aging populations, and largely avoidable lifestyle related illnesses within limited budgets, policy makers are desperate for the right kind of evidence" [4]. With such broad goals in mind, it is surprising that evidence-based mechanisms are not yet fully engaged so that Internet-delivered health interventions can be exploited to achieve these goals.

Although there are numerous websites that contain health-related information, the ability of the consumer—or the patient—to find scientifically robust (ie, evidence-based) health interventions is not fully known. Data from the Pew Research Center indicates that 72% of adults who use the Internet have searched for health-related information in the previous year (based on 2012 survey data) [5]. Furthermore, there is insufficient information to assist the public in deciphering which sites contain useful information that could help them stay healthy, ameliorate risky behaviors, recognize early disease, or assist with treatment of their existing disorders.

The central question is, "Which Internet-delivered health interventions actually work?" For scientists, the answer to this question can be addressed by evaluating the results from randomized controlled trials (RCTs). In fact, the efficacy of some Internet-delivered interventions has already been assessed by investigators. Yet, there is not yet a fully formed mechanism to link these results with the individuals who may wish to use this information.

To expedite this process, there are necessary preparatory steps before implementation. Our translational model is shown in Figure 1 and illustrates the steps. Many RCTs and meta-analyses of RCTs of Internet-delivered health-related interventions are already published and, therefore, some evidence is available. We now, through this report, present the results from the evaluation step of dissemination. That is, we conducted a systematic review of published meta-analyses of RCTs on Internet-delivered health-related interventions. We evaluated this evidence and generated a list of evidence-based websites currently available for use. We were especially interested in Internet-delivered therapies that do not require real-time interaction with a therapist or other health care provider. That is, the application was housed on the Internet for general use by the public.

Figure 1. Translating research into implementation for Internet-delivered health interventions.



Aims of the Study

The aims of this study were to (1) discover the range of health-related topics that were addressed through Internet-delivered interventions, (2) generate a list of current websites used in the trials which demonstrated a health benefit, and (3) identify gaps in the research that may have hindered dissemination.

Methods

Inclusion Criteria

The Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines were used [6]. To be eligible for inclusion, studies were required to be meta-analyses of RCTs with at least one intervention that was Internet-delivered and reported a health-related outcome. Within each meta-analysis, we required that there be at least one self-guided intervention without therapist or clinician support. For this review, we excluded group-based interventions (ie, trials that enrolled groups of people to experience the intervention together). There were no restrictions on the types of individuals in the trials or the type of health outcomes.

Search Strategy

A comprehensive search strategy was developed with a biomedical research librarian and was undertaken to identify articles for inclusion. The following electronic databases were searched: MEDLINE via Ovid, PsycINFO, Embase, Cochrane Database of Systematic Reviews, and the Cumulative Index to Nursing and Allied Health Literature (CINAHL). Search strategies utilized a combination of keywords and MeSH headings (Table 1). The last date of the search was June 14, 2016. There were no language restrictions on the search.

Table 1. Search strategy for the systematic review.

Database	Search terms
MEDLINE via Ovid	1. (internet or web or "web-delivered" or online).ti.
	2. ("meta-analysis" or "metaanalysis" or "meta-analytic").ab,ti.
	3. Random*.ab,ti.
	4. 1 AND 2 AND 3
PsycINFO	1. AB (web OR internet OR web-based OR online OR web-delivered)
	2. AB (meta-analysis or meta-analytic or meta-analysis)
	3. AB random*
	4. 1 AND 2 AND 3
	5. Restrict 4 to "meta-analysis"
Embase	1. (web:ti OR internet:ti OR online:ti OR 'web based':ti OR 'web delivered':ti)
	2. 'meta analysis':ab,ti
	3. Random*ab,ti
	4. 1 AND 2 AND 3
Cochrane Database of Systematic Reviews	1. web OR internet OR online OR "web-based" OR "web-delivered"(record title)
CINAHL (Cumulative Index to Nursing and Allied Health Literature)	1. TI (web OR internet OR web-based OR online OR web-delivered)
	2. TI (meta-analysis OR metaanalysis OR meta-analytic)
	3. TI (random*)
	4. 1 AND 2 AND 3

Screening of Articles

Three authors (KL, RK, and MR) independently reviewed the title and abstract of each record to determine eligibility. Any disagreements regarding inclusion or exclusion were resolved by a discussion between two authors (MR and KL). Full papers of the selected title and abstracts were reviewed independently by three authors (KL, RK, and MR) and disagreements regarding inclusion or exclusion were resolved by a discussion between two authors (KL and MR). All RCTs within each meta-analysis were screened for eligibility (self-guided Internet-based health-related intervention).

Analyses

The purpose of the analyses was to combine the results from across all the meta-analyses so that the results could be summarized and the Internet programs could be located. Individual RCTs within each meta-analysis were grouped by topic: Substance Abuse, Mental Health, Diet and Physical Activity, Disease Management, Disease Prevention, and Childhood Health Problems. A health benefit was defined as a statistically significant improvement in any health-related outcome within an RCT; all trials assessed outcomes through inferential statistics with alpha set at .05, 2-tailed. The concurrent control groups did not receive the Internet-delivered intervention (generally a wait list) unless specifically stated. Measures of efficacy were calculated when data were available within each RCT; for binary outcomes, number needed to treat (NNT) was calculated when absolute measures were reported. For outcomes measured using continuous scales, mean changes were listed (intervention relative to control).

Each RCT was reviewed for the name of the intervention and the website that housed the intervention. Functional websites of such evidence-based interventions (demonstrating a health benefit) were located on November 18, 2016. We defined functional website as those sites which housed the program which was tested in the RCTs and was available for general use.

Results

There were 363 records identified through the search (Figure 2) which yielded 304 records after removing the duplicates. The abstracts were reviewed and 162 were excluded because they did not meet the eligibility criteria. Full-text articles were reviewed for the 142 remaining articles and 36 were excluded due to noninvolvement of the Internet, 15 were excluded due to therapist or clinician support only, and 20 were excluded due to either protocol only, no health outcomes, group-based interventions only, not a meta-analysis, or did not include RCTs. There were 71 meta-analyses of Internet-based interventions that met eligibility criteria and were included in this study [7-76].

Within the 71 meta-analyses, there were 1733 studies. Of these studies, there were 268 unique RCTs that were self-help Internet-based interventions; and of the 268 studies, there were 57 trials demonstrating a health benefit with a functional website [77-138]. The topics covered are listed in Table 2.

Table 2. Number of randomized controlled trials with functional websites of self-help Internet-delivered health interventions.

Category	Topic	Number of RCTs reviewed	Number of RCTs with websites[a]	RCTs with websites[a] (%)
Substance abuse				
	Alcohol	72	8	11
	Smoking (tobacco)	30	7	23
	Cannabis	3	2	67
	Drug use (general)	1	0	0
	Total	106	17	16.0
Mental health				
	Depression	15	4	27
	Anxiety	16	5	31
	Post-traumatic stress disorder	11	2	18
	Phobias	9	3	33
	Panic disorders	5	2	40
	Obsessive compulsive disorder	1	1	100
	Mental health (general)	3	0	0
	Eating disorders	4	0	0
	Infertility distress	2	0	0
	Total	66	17	26
Diet and physical activity				
	Diet	13	7	54
	Physical activity	33	6	18
	Total	46	13	28
Disease management				
	Insomnia	13	4	31
	Chronic pain	10	2	20
	Diabetes	5	0	0
	Fatigue	1	0	0
	Tinnitus	2	0	0
	Total	31	6	19
Disease prevention				
	Cardiovascular risk	9	2	22
	Cancer prevention (skin)	1	0	0
	Sexual health	7	0	0
	Total	17	2	12
Childhood health problems				
	Childhood behavior problems	1	1	100
	Encopresis	1	1	100
	Total	2	2	100
	Grand total	268	57	21.3

[a]Websites in which there was a health benefit demonstrated in an RCT (randomized controlled trial).

Internet self-help for substance abuse was the most frequent topic in RCTs, with alcohol having the greatest number of trials. Of the 72 trials on alcohol use, there were 8 with functioning websites. Tobacco use was also a common subject for

interventions with 7 websites (from the 30 trials reviewed). Mental health interventions were available, including anxiety (5 websites on generalized anxiety disorder, 3 on phobias, 2 on panic disorders, 2 on post-traumatic distress disorder, and 1 on obsessive compulsive disorder [OCD]) and depression (4 websites). There were 46 RCTs reviewed for diet and physical activity interventions, and 13 of those yielded a functioning website. There were fewer RCTs on disease management, with insomnia and chronic pain yielding 13 and 10 trials, respectively. Within the meta-analyses, there were a few RCTs specifically on cardiovascular risk factors (blood pressure, cholesterol, and hyperlipidemia) and several on sexual health (sexually transmitted disease, sexual dysfunction, unintended pregnancy). However, there was only one RCT on self-help regarding cancer prevention (for skin cancer). Finally, there were two trials targeted to parents of children with health problems—one on behavioral problems and one on encopresis.

In most instances, the Internet-delivered interventions were offered only to the study participants in the context of the RCT; websites to deliver the intervention after the conclusion of the study were not available. For example, *Student Bodies* was an efficacious Internet-delivered program for eating disorders in girls but the program was not available for general use [139-141]. In the area of sexual health, there were several efficacious Internet-delivered programs regarding sexually transmitted diseases [142-144]. However, these sites were only available for study participants during the course of the research study. Overall, in only 21.3% (57/268) of instances, there was a functional website for the interventions after the conclusion of the trial.

We compiled a list of websites of the Internet-delivered interventions providing a health benefit and these are shown in Table 3 with the name of the program, hyperlink to the site, cost, and the languages utilized for delivery of the program.

Table 3. Evidence-based websites of Internet-delivered health-related interventions.

Target population	Name	Websites	Cost	Language
Alcohol				
Adult drinkers	Check Your Drinking	http://www.checkyourdrinking.net/CYD/CYD-ScreenerP1_0.aspx	Free	English, French, Portuguese, Spanish
Male adults	Drinktest.nl	http://www.drinktest.nl/	Free	Dutch
Adult at-risk drinkers	Balance	http://akan.no/verktoy/balance/	Free	Norwegian
Adult problem drinkers	Drinking Less	http://minderdrinken.nl/	Free	Dutch
Universities	Alcohol eCHECKUP TO GO (eCHUG)	http://www.echeckuptogo.com/programs/alcohol	Commercial	English
Universities	Alcohol Edu	https://everfi.com/health-wellness/	Commercial	English
University students	MyStudentBody	https://www.mystudentbody.com	Commercial	English
University students	Alcohol-Wise	https://web.3rdmilclassrooms.com/courses/college/alcohol-wise	Commercial	English
Tobacco				
Smokers	Smokefree	https://smokefree.gov/	Free	English, Spanish
Smokers	QuitCoach	http://www.quitcoach.org.au/	Free	English
Smokers	Stop-tobacco	http://www.stop-tobacco.ch/en/	Free	French, German, Italian, English, Spanish, Portuguese
Smokers	Dejar de Fumar (Give Up Smoking)	http://www.apsiol.uned.es/dejardefumar/	Free	Spanish
Smokers	Guía para dejar de fumar (Guide to Quitting Smoking)	https://www.aecc.es/Comunicacion/publicaciones/Documents/Guia_dejar_fumar.pdf	Free	Spanish
Smokers	Slutta (Quit)	https://helsenorge.no/rus-og-avhengighet/snus-og-roykeslutt/	Free	Norwegian
Smokers	QuitNet	https://quitnet.meyouhealth.com/#/	Some features free, some require a fee	English
Cannabis				
Cannabis users	Reduce Your Use: How to Break the Cannabis Habit	https://reduceyouruse.org.au/sign-up/	Free	English
Cannabis users	Quit the Shit	www.drugcom.de/?id=quittheshit	Free	German
Depression				
Individuals with depression	MoodGYM	https://moodgym.anu.edu.au/welcome	Free	English, Finnish, Norwegian, Dutch
Individuals with depression	BluePages	http://bluepages.anu.edu.au/	Free	English, Norwegian
Individuals with depression	Deprexis	http://www.deprexis.com/	Commercial	English, German
Individuals with depression	Kleur Je Leven (Color Your Life)	http://www.kleurjeleven.nl/	Commercial	Dutch
Generalized anxiety disorder				
Individuals with anxiety	Mental Health Online	https://www.mentalhealthonline.org.au/Default.aspx	Free	English
Individuals with anxiety	This Way Up	https://thiswayup.org.au/	Some free, some commercial	English

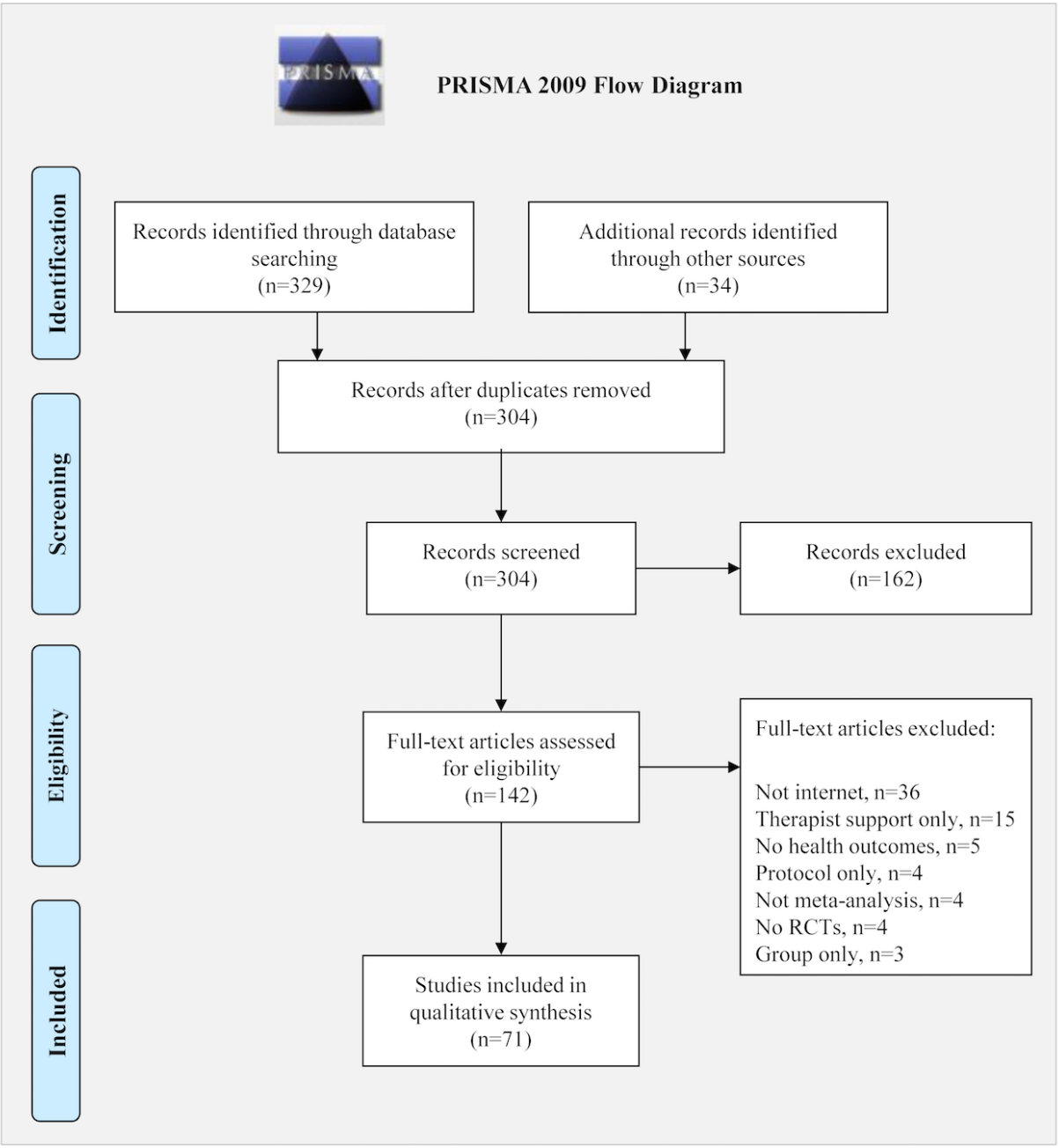
	Target population	Name	Websites	Cost	Language
	Workers	Stress and Mood Management	http://centerforworkforcehealth.com/programs/stress-and-mood-management/	Commercial	English
	Colleges, universities	My Student Body—Stress	https://www.mystudentbody.com/	Commercial	English
	Individuals with anxiety	Internet-based Mindfulness Treatment	https://www.mindfulnesscenter.se/en	Commercial	Swedish, English, Norwegian
Post-traumatic stress disorder					
	Individuals who experienced trauma	My Trauma Recovery	http://mytraumarecovery.com/	Free	English
	Individuals who experienced disasters	My Disaster Recovery	http://disaster.bluesunsupport.com/	Free	English, Spanish, Mandarin
Panic disorder and phobias					
	Individuals with panic disorders or Phobias	Mental Health Online	https://www.mentalhealthonline.org.au/Default.aspx	Free	English
	Individuals with panic disorders or phobias	This Way Up	https://thiswayup.org.au/	Some free, some commercial	English
	Adults with glossophobia (fear of public speaking)	Talk to Me	http://www.internetmeayuda.com/mhpEnglish/saludo.htm	Free	Spanish, English
Obsessive compulsive disorder					
	Individuals with OCD[a]	This Way Up	https://thiswayup.org.au/	Some free, some commercial	English
Diet and physical activity					
	Adults	Gezond Leven Check (Healthy Living Check)	http://www.gezondlevencheck.nl/	Free	Dutch
	Sedentary overweight adults	Active Living Every Day (ALED-I)	http://www.activeliving.info/demo/demo_osg_welcome.cfm	Free (fee for extra materials)	English
	Adults with diabetes	My Path to Healthy Life	http://mypathtohealthylife.com/	Free	English, Spanish
	Adults in workplace	Food Smart	http://centerforworkforcehealth.com/index.cfm/programs/food-smart/	Commercial	English
	Overweight and obese adults	The Biggest Loser Club	https://www.biggestloserclub.com/	Commercial	English
	University students	My Student Body—Nutrition	https://www.mystudentbody.com/	Commercial	English
	Managers of organizations	ExecuPrev	https://www.execuprev.com	Commercial	English
	Employees in workforce	DASH[b] for Health	http://www.dashforhealth.com/	Commercial	English
Insomnia					
	Adults with chronic insomnia	Insomnie	http://www.insomnie.nl/	Free	Dutch
	Adults with insomnia	SHUTi	http://www.myshuti.com/	Commercial	English
	Individuals with sleep problems	Sleepio	https://www.sleepio.com/	Commercial	English
	Adults with chronic insomnia	**REST**ORE	http://cobalttx.com/Products/restore.html	Commercial	English
Chronic pain					
	Adults with chronic pain	painACTION	http://www.painaction.com/#	Free	English
	Adults with chronic pain	Chronic Pain Management Program	https://pain.goalistics.com/	Commercial	English

	Target population	Name	Websites	Cost	Language
Hypertension and hyperlipidemia					
	Adults with hypertension	Blood Pressure Action Plan	http://www.heartandstroke.on.ca/site/c.pvI3IeNWJwE/b.3582093/k.8AB3/Blood_Pressure_Action_Plan.htm	Free	English
	Employees in workforce	DASH[b] for Health	http://www.dashforhealth.com/	Commercial	English
Childhood health problems					
	Parents with children who have behavior problems	Triple P Online	http://www.triplep-parenting.com	Commercial	English, Spanish, Dutch, German
	Parents of children with encopresis	U Can POOP Too	http://www.ucanpooptoo.com/	Free	English

[a]OCD: obsessive compulsive disorder.

[b]DASH: Dietary Approaches to Stop Hypertension.

Figure 2. The Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) flow diagram.



Substance Abuse

From this review, we found that there are eight currently available evidence-based websites on alcohol use. All but three were conducted using college or university students; some are specifically targeted to universities, offering a suite of programs regarding substance abuse, health, and wellness. The most common techniques utilized in these interventions (Table 4) were personalized and normative feedback, as well as goal setting. Some sites included more tailored feedback and interactive journaling. The health benefit observed in the trials was generally a reduction in alcohol consumption, although some trials showed a reduction in the consequences of heavy drinking such as impairment in control and fewer embarrassing actions. The length of the programs varied—some being rather brief screening tools and others encompassing 6 months of structured activities [77-91]. The freely available websites for alcohol emanated from various European countries. Screen shots of each home page of the websites are given in Multimedia Appendix 1.

For tobacco use, there are several free evidence-based websites for reducing or quitting smoking. Four of these sites were available in English, 4 in Spanish, and 1 in Norwegian; the *Stop-tobacco* program was available in multiple languages. Often they were supported with governmental or public health funding such as *smokefree* in the United States, *Stop-tobacco*

in Switzerland, and *Quit* in Norway. In general, the primary outcome was greater abstinence rates of smoking which were achieved through structured cognitive behavioral techniques, including motivational materials, personalized and tailored advice, goal setting, feedback mechanisms, and self-monitoring.

There were a few RCTs on curbing cannabis use which reported a reduction in the frequency or quantity of use. The mechanisms used to achieve the health benefits were similar to those used for tobacco use, relying heavily on cognitive behavioral approaches (listed in Table 4). One free Internet-delivered intervention *Reduce Your Use* comes from Australia and another was developed in Germany (*Quit the Shit*).

Table 4. Health benefits of evidence-based websites of Internet-delivered health-related interventions.

Name	Website	Intervention techniques	Health benefits
Alcohol			
Check Your Drinking	http://www.checkyourdrinking.net/CYD/CYD-ScreenerP1_0.aspx	Personalized and normative feedback	Reduction in weekly alcohol consumption
Drinktest.nl	http://www.drinktest.nl/	Personalized and normative feedback	Reduction in alcohol consumption
Balance	http://akan.no/verktoy/balance/	Personalized and normative feedback	Reduction in alcohol consumption
Drinking Less	http://minderdrinken.nl/	Goal setting, analysis of drinking behavior, maintenance, and relapse prevention	Reduction in alcohol consumption
Alcohol eCHECKUP TO GO (eCHUG)	http://www.echeckuptogo.com/programs/alcohol	Individualized feedback, recognition of harms	Reduction in alcohol consumption
Alcohol Edu	https://everfi.com/health-wellness/	Cognitive-behavioral skills, normative drinking, motivational information	Reduction in alcohol consumption, reduction in heavy alcohol use, reduction in consequences (eg, embarrassing actions, impaired control)
MyStudentBody	https://www.mystudentbody.com	Tailored feedback, normative feedback, educational tools on behavior and consequences	Reduction in binge drinking, reduction in alcohol consumption among persistent heavy drinkers
Alcohol-Wise	https://web.3rdmilclassrooms.com/courses/college/alcohol-wise	Information on social norms, interactive journaling, educational feedback	Reduction in peak number of drinks and blood alcohol concentration
Tobacco			
Smokefree	https://smokefree.gov/	Motivational materials, step-by-step quitting guide, task charts, self-monitoring tools, personal calendar	Greater abstinence rates of smoking
QuitCoach	http://www.quitcoach.org.au/	Personalized, tailored advice. Feedback with suggestions and encouragement.	Greater abstinence rates of smoking
Stop-tobacco	http://www.stop-tobacco.ch/en/	Tailoring with stages of change, coping methods, self-change strategies, feedback	Greater abstinence rates of smoking
Dejar de Fumar (Give Up Smoking)	http://www.apsiol.uned.es/dejardefumar/	Education, self-monitoring, self-control, relapse prevention, coping skills, lifestyle change	Greater abstinence rates of smoking in program completers, decrease in number of cigarettes in smokers
Guía para dejar de fumar (Guide to Quitting Smoking)	https://www.aecc.es/Comunicacion/publicaciones/Documents/Guia_dejar_fumar.pdf	Noninteractive smoking cessation guide, cigarette counter, online journal	Increase in cigarette quit rates
Slutta (Quit)	https://helsenorge.no/rus-og-avhengighet/snus-og-roykeslutt/	Personalized, adaptive messages with feedback. Coordinated with steps in behavioral change.	Greater short-term abstinence rates of smoking
QuitNet	https://quitnet.meyouhealth.com/#/	Advice, setting quit date, individually tailored feedback, problem solving skills, support	Increase in abstinence rates of smoking
Cannabis			
Reduce Your Use: How to Break the Cannabis Habit	https://reduceyouruse.org.au/sign-up/	Tracking use, information regarding attitudes, goal setting, expenditures, motivational feedback	Reduction in days of use and quantity of use
Quit the Shit	www.drugcom.de/?id=quittheshit	Personalized feedback, goal setting, information on strategies	Reduction in frequency of use and quantity of use

Name	Website	Intervention techniques	Health benefits
Depression			
MoodGYM	https://moodgym.anu.edu.au/welcome	Cognitive behavioral therapy	Reduction in symptoms of depression
BluePages	http://bluepages.anu.edu.au/	Information on depression, symptoms, prevention, and treatments, sources of help (used in conjunction with MoodGYM)	Reduction in symptoms of depression
Deprexis	http://www.deprexis.com/	Cognitive behavioral therapy	Reduction in symptoms of depression
Kleur Je Leven (Color Your Life)	http://www.kleurjeleven.nl/	Cognitive behavioral therapy	Improvement in short-term depressive symptoms in frequent users of the site
Generalized anxiety disorder			
Mental Health Online	https://www.mentalhealthonline.org.au/Default.aspx	Cognitive behavioral therapy	Reduction in general anxiety levels
This Way Up	https://thiswayup.org.au/	Cognitive behavioral therapy	Reduction in anxiety in those who completed the program
Stress and Mood Management	http://centerforworkforcehealth.com/programs/stress-and-mood-management/	Cognitive behavioral therapy	Reduction in stress
My Student Body—Stress	https://www.mystudentbody.com/	Cognitive behavioral therapy	Decreased anxiety and family problems
Internet-based Mindfulness Treatment	https://www.mindfulnesscenter.se/en	Mindfulness	Decreased anxiety, severity of insomnia, and depression. Increased quality of life.
Post-traumatic stress disorder			
My Trauma Recovery	http://mytraumarecovery.com/	Interactive modules: seeking help, relaxation, social support, coping, self-talk, triggers, and memories	Reduction in post-traumatic symptom severity
My Disaster Recovery	http://disaster.bluesunsupport.com/	Interactive modules: seeking help, relaxation, social support, coping, self-talk, triggers, and memories	Decreased worry
Panic disorder and phobias			
Mental Health Online	https://www.mentalhealthonline.org.au/Default.aspx	Cognitive behavioral therapy	Reduction in frequency of panic and anticipatory fear of panic
This Way Up	https://thiswayup.org.au/	Cognitive behavioral therapy	Reduction in social phobia in those who completed the program
Talk to Me	http://www.internetmeayuda.com/mhpEnglish/saludo.htm	Cognitive behavioral therapy	Decrease in fear and avoidance of public speaking
Obsessive compulsive disorder			
This Way Up	https://thiswayup.org.au/	Cognitive behavioral therapy	Reduction in symptoms of OCD[a], distress and depression
Diet and physical activity			
Gezond Leven Check (Healthy Living Check)	http://www.gezondlevencheck.nl/	Precaution adoption process model including tailored, personalized feedback, normative behaviors, and suggestions	Reduction in saturated fat intake and increase in physical activity
Active Living Every Day (ALED-I)	http://www.activeliving.info/demo/demo_osg_welcome.cfm	Self-paced program, interactive activities, behavior modification strategies	Increased daily steps in people with low baseline activity. Reduction in waist circumference.
My Path to Healthy Life	http://mypathtohealthylife.com/	Goal setting, tracking progress, feedback, resources, interaction, motivational tips	Increase in healthy eating habits and physical activity. Decrease in fat intake.

Name	Website	Intervention techniques	Health benefits
Food Smart	http://centerforworkforcehealth.com/index.cfm/programs/food-smart/	Information on diet, stress, fitness or physical activity. Self-tailored content. Interactive activities.	Improvement in dietary self-efficacy, dietary attitudes, and dietary stage of change
The Biggest Loser Club	https://www.biggestloserclub.com/	Social cognitive theory. Self-efficacy, goal setting, self-monitoring, outcome expectations, interaction.	Reduction in body mass index, weight, waist circumference, and waist-to-height ratio
My Student Body—Nutrition	https://www.mystudentbody.com/	Information targeted to students, rate myself assessment with feedback, diet and physical activity information, resources	Increase in fruit and vegetable intake
ExecuPrev	https://www.execuprev.com	Animated and interactive learning on health and executive leadership, self-assessment, simulations, advice from experts, coaching and webinars	Improvement in dietary self-efficacy and dietary attitudes. Decrease in waist circumference in women.
DASH[b] for Health	http://www.dashforhealth.com/	Advice on nutrition and physical activity. Feedback on 24-h food recall, weight, blood pressure. Progress reports.	Decrease in weight for obese or overweight individuals. Increased intake of fruits and vegetables. Lowered consumption of carbonated beverages.
Insomnia			
Insomnie	http://www.insomnie.nl/	Cognitive behavioral therapy	Improvement in sleep efficiency, total sleep time, sleep onset latency, wake after sleep onset and number of nocturnal awakenings. Decrease in anxiety and depression.
SHUTi	http://www.myshuti.com/	Cognitive behavioral therapy	Improvement in sleep efficiency, insomnia severity, sleep onset latency, soundness of sleep, restored feeling on awakening, and general fatigue
Sleepio	https://www.sleepio.com/	Cognitive behavioral therapy	Improvement in sleep efficiency, total sleep time, sleep onset latency, wake after sleep onset, sleep quality, and daytime functioning
RESTORE	http://cobalttx.com/Products/restore.html	Cognitive behavioral therapy	Improvement in sleep quality, insomnia severity, and daytime fatigue
Chronic pain			
painACTION	http://www.painaction.com/#	Self-management strategies and cognitive behavioral therapy	Reduction in pain intensity, depression, anxiety, stress. Increased improvement in pain, coping, use of social support.
Chronic Pain Management Program	https://pain.goalistics.com/	Self-directed interactive learning, integrating social networking, and self-management tools	Reduction in pain severity, pain-related interference, emotional burden, perceived disability, catastrophizing, and pain-induced fear. Decreased depression, anxiety, and stress.
Hypertension and hyperlipidemia			
Blood Pressure Action Plan	http://www.heartandstroke.on.ca/site/c.pvI3IeNWJwE/b.3582093/k.8AB3/Blood_Pressure_Action_Plan.htm	Stages of readiness, setting priorities, motivation, tailored advice, self-direction	Decrease in systolic blood pressure and total cholesterol in persons completing the program
DASH for Health	http://www.dashforhealth.com/	Advice on nutrition and physical activity. Feedback on 24-h food recall, weight, blood pressure. Progress reports.	Decrease in systolic blood pressure in individuals with hypertension

Name	Website	Intervention techniques	Health benefits
Childhood health problems			
Triple P Online	http://www.triplep-parenting.com	Interactive, self-directed positive parenting skills. Goal setting, evaluation, self-efficacy, and personal agency skills. Video-based modeling, experiential learning, prompting. Customizable output. Cultural sensitivity.	Reduction in the frequency and intensity of child behavioral and emotional problems. Increase in child adjustment. Decrease in dysfunctional parenting styles and parental anger. Increase in parents' confidence.
U Can POOP Too	http://www.ucanpooptoo.com/	Behavioral approach with reinforcement for spontaneous toilet use and clean pants, instructions and modeling of behaviors and actions, education	Reduced fecal soiling, increased defecation in the toilet, and increased unprompted trips to the toilet

[a]OCD: obsessive compulsive disorder.

[b]DASH: Dietary Approaches to Stop Hypertension.

Mental Health

There were 4 functional websites that were intended to help individuals with depression. *MoodGYM* and *BluePages* are generally used together, the first for the delivery of cognitive behavioral therapy and the following as an adjunct. *MoodGYM* was more effective when the entire program was completed—not brief interventions [145]. *MoodGYM* contains five sequential modules which are completed at the pace of each user. There were a few commercial sites as well— *Deprexis* (9-week program) and *Color Your Life* (8 weeks with a 9th-week booster). Both of these programs led to a reduction in the symptoms of depression. Cognitive behavioral therapy was the mechanism utilized in each of these interventions which was delivered in a modular, stepwise manner over several months. It is important to note that one of the Internet-delivered programs for treating insomnia (*Insomnie*) also led to a decrease in depression and anxiety.

There were several websites that addressed generalized anxiety, most delivering cognitive behavioral therapy and one delivering mindfulness therapy. *Mental Health Online* and *This Way Up* both emanate from Australia. *Stress and Mood Management* is from the Center for Workforce Health and is a commercial program targeting workers. On the site, a suite of programs is offered on various health-related topics. *My Student Body*, likewise, offers a suite of programs, one being on *Stress* targeted to colleges and universities. Cognitive behavioral therapy was the mechanism used in all of the anxiety interventions except for one which used mindfulness. Mindfulness treatment was available in several languages which reduced anxiety, depression, and the severity of insomnia.

Post-traumatic stress disorder (PTSD) was addressed at two partner websites—one focusing on various trauma recoveries and the other directed to individuals who experienced disasters. It was tested in an RCT for hurricane survivors [112] and is available in English, Spanish, and Mandarin. The Internet sites for PTSD used various coping strategies and behaviors which were based on social cognitive theory. The modules included social support, self-talk, relaxation, trauma triggers, unhelpful coping, and professional help.

There are functional websites for the treatment of specific phobias and panic disorders as well. Both *Mental Health Online* and *This Way Up* address these disorders through cognitive behavioral therapy. In addition, *This Way Up* offers a 10-week program for individuals with OCD which was effective in reducing the symptoms of OCD, as well as reducing distress and depression in those with OCD.

Talk to Me was developed by Spanish psychologists to treat fear of public speaking and is available for use. This 2-month program using cognitive behavioral techniques was effective in decreasing the fear and avoidance of public speaking. This same group developed *Without Fear* which is an Internet-delivered program for fear of small animals (spiders, cockroaches, mice).

Diet and Physical Activity

There were several websites on diet and physical activity interventions. In general, these programs included interactive components with goal setting and personalized feedback. Often self-monitoring and tracking of progress were included. Some of the commercial websites were found to be efficacious in terms of reducing body mass index or weight. *The Biggest Loser Club* was efficacious (decreased weight, body mass index, waist circumference) with a 12-week program [123]. Additional support (periodic reminders) did not improve the basic Internet program [123].

Dietary Approaches to Stop Hypertension (DASH) for Health is another evidence-based program with a focus on diet and physical activity. For those who completed 12 months of use, overweight or obese individuals lost weight (mean decrease of 4 pounds) [126]. Overall, people with hypertension lowered their systolic blood pressure by an average of 7 mmHg [126]. It also led to increased consumption of fruits and vegetables and lower consumption of carbonated beverages. The program included weekly education, motivation, and mechanisms for self-monitoring with progress reports [126].

The Center for Workforce Health includes a suite of programs, some of which specifically address diet and physical activity. The RCTs captured in this review indicated that the two modules entitled *Stress and Mood Management* and *Food Smart* showed

health benefits. There was a reduction in stress after the completion of *Stress and Mood Management* (3-month program), and after the *Food Smart* program, there was improvement in dietary self-efficacy, dietary attitudes, and dietary stage of change [108,122].

My Student Body also contains a suite of Internet programs. Our review indicated that the packages for *Nutrition*, *Alcohol* use, and *Stress* were efficacious in various RCTs. These are now combined and sold commercially, generally to colleges and universities. *My Student Body—Nutrition* specifically increased the intake of fruit and vegetables.

Healthy Living Check, *Active Living Every Day*, and *My Path to Healthy Life* addressed both diet and physical activity. Completion of these programs led to various health benefits including a reduction in the intake of saturated fat, reduction in waist circumference, and an increase in physical activity. *My Path to Healthy Life* was targeted to adults with diabetes mellitus [121], and completion of this program led to a decrease in fat intake and an increase in physical activity.

Some health interventions were paired with other interventions. *ExecuPrev* (*LeadWell LiveWell*) paired a leadership intervention with cardiovascular disease prevention. This program decreased waist circumference in women, although it did not affect body mass index overall [125].

Disease Management

All of the meta-analyses of RCTs on insomnia indicated that Internet-based cognitive behavioral therapy for insomnia was efficacious. In general, the therapy was delivered over several months through a series of modules and included self-monitoring through sleep diaries. The content of the therapy often included sleep information, sleep hygiene, relaxation, stimulus control, sleep restriction, and various cognitive techniques such as restructuring, paradox, mindfulness, imagery, putting day to rest, and thought stopping. Only some of the applications, however, are currently available. *Insomnie* was developed in the Netherlands and is available in Dutch. *SHUTi*, *Sleepio*, and *RESTORE* are available commercially in English. Generally, completion of these programs takes weeks to months, with specified activities required during each step of the program. The main health benefits were improvement in sleep efficiency and sleep quality, with a decrease in the severity of insomnia.

There were two evidence-based sites for chronic pain: *painACTION* and the *Chronic Pain Management Program*. The free site *painACTION* offers programs in back pain, migraines, neuropathic pain, and pain due to cancer or arthritis. *painACTION* is a 4-week course followed by 5 monthly boosters and includes self-management education in which problem solving skills were taught to reach specific goals [133,134]. The 6-week *Chronic Pain Management Program* covered four domains: cognitive (thinking better), behavioral (doing more), social (relating better), and emotional (feeling better). Both pain-related sites utilize various cognitive behavioral approaches with self-management strategies and interactive elements. These programs led to a reduction in the intensity or severity of pain, as well as a reduction in stress, anxiety, and depression.

Disease Prevention

The *Blood Pressure Action Plan* (now called the *Heart and Stroke Foundation Health e-Support* program) resulted in lower systolic blood pressure, lower pulse pressure, and lower total cholesterol in those individuals who completed the 4-month program [136]. *DASH for Health* also lowered systolic blood pressure [126]. These programs involved setting priorities and included self-monitoring, progress reports, and tailored advice.

Childhood Health Problems

There were two efficacious programs on childhood health problems. *Triple P Online* (*TPOL*) assists parents in addressing behavioral programs in children through teaching positive parenting skills. It was shown to decrease problematic child behavior, dysfunctional parenting styles, parental anger, and to improve parent's confidence. It has been studied quite extensively and is used in 25 countries throughout the world (with availability in English, Spanish, Dutch, and German). The techniques used included goal setting, evaluation, self-efficacy, personal agency skills, with video-based modeling, experiential learning, prompting, and customizable output. The other site for childhood problems is entitled *U Can POOP Too* which addresses encopresis. It has been shown to reduce fecal soiling and improve toileting skills through various cognitive behavioral approaches including reinforcement and modeling of behaviors and actions.

Measures of Efficacy

The principal measure of efficacy was NNT; these were calculated for binary outcomes and are shown in Table 5. For continuous outcomes, changes in mean differences between the intervention and control groups are shown. Overall for substance abuse (alcohol, tobacco, and cannabis), the effect was moderate with NNTs of 9-26 for avoidance or reduction in use over a short-term period (up to 6 months). For mental health problems, the degrees of effect were commonly reported using conventional scales within each field. In general, the effects were moderate with a decrease in depressive symptoms, anxiety, or stress. There were two interventions (Deprexis and Mental Health Online for panic disorders) which demonstrated particular efficacy (ie, low NNTs of 4 and 2, respectively). The efficacy of the interventions for diet and physical activity, although significant, was modest (eg, 2.1 kg mean weight reduction compared with a 0.4 kg increase in controls). There were several efficacious interventions for insomnia; the severity of symptoms, in general, decreased moderately. For example, SHUTi showed an 8-point relative reduction in severity on the Insomnia Severity Index. In addition, Sleepio also demonstrated an increase in daytime performance (2.5 points on a 5-point scale). The intervention painACTION was particularly efficacious, with an NNT of 4 for back pain and a NNT of 3 for migraine headaches. The interventions targeted to parents of small children were also very efficacious. Triple P Online had an NNT of 3 for clinical improvement in behavioral problems in children and U Can POOP Too had an NNT of 4 for prevention of fecal accidents.

Table 5. Measures of efficacy for Internet-delivered health-related interventions.

Target population	Name	Measures of efficacy
Alcohol		
Adult drinkers	Check Your Drinking	After a 3-month period, the Internet group drank 2.4 fewer drinks per week (on average) than the control (information without Internet).
Male adults	Drinktest.nl	For every 9 people who completed the Internet program, 1 person reduced their drinking levels below recommended levels after 1 month, compared with controls (information without Internet).
Adult at-risk drinkers	Balance	After use of an Internet program for 6 months, people who had an intensive Internet therapy drank approximately 3 fewer drinks per day than people in the brief self-help program.
Adult problem drinkers	Drinking Less	For every 9 people who completed the Internet program, 1 person reduced their drinking levels below recommended levels after 6 weeks, compared with controls (informational brochure).
Universities	Alcohol eCHECKUP TO GO (eCHUG)	Students who were heavy drinkers and completed the Internet intervention reduced their drinking by 8 drinks per week, were less likely to drink to intoxication, and had fewer alcohol-related problems after 3 months than students who did not use this intervention. Binge drinkers who completed the Internet intervention had 5 less drinks (on average) on any given night and were less likely to drink to intoxication after 3 months than students not using the intervention.
Universities	Alcohol Edu	First year students who participated in the Internet intervention showed a reduction in alcohol use and binge drinking within a 30-day period compared with those who did not use the intervention.
University students	MyStudentBody	Heavy drinkers who used the Internet intervention had (on average) one-half drink less than those not using the intervention. Women, in particular, lowered their alcohol intake with the Internet intervention.
University students	Alcohol-Wise	Students at a public urban university who completed the Internet intervention drank (on average) 2 fewer drinks per week than the control (without the intervention).
Tobacco		
Smokers	Smokefree	For every 9 people who used the Internet intervention (without email support), 1 person abstained from smoking for 3 months compared with people who used an abbreviated version of the website. For every 18 people who used the Internet intervention (without email support), 1 person abstained from smoking for 7 months compared with people who used an abbreviated version of the intervention.
Smokers	QuitCoach	For every 15 people who used the Internet intervention with Web-based structured planning, 1 person abstained from smoking for 6 months compared with people who did not use structured planning.
Smokers	Stop-tobacco	For every 26 people who completed the Internet intervention, 1 person abstained from smoking for 7 days compared with a modified program (with less information regarding risks and coping).
Smokers	Slutta (Quit)	For every 24 people who used the tailored Internet intervention, 1 person abstained from smoking for 3 months compared with people who used a nontailored website.
Cannabis		
Cannabis users	Reduce Your Use: How to Break the Cannabis Habit	After 3 months, people who used the Internet intervention had 3 fewer days of cannabis use (per month) compared with people not using the intervention.
Cannabis users	Quit the Shit	After 3 months, people who used the Internet intervention had 4 fewer days of cannabis use (per 30-day period) compared with people not using the intervention.
Depression		
Individuals with depression	MoodGYM with BluePages	Symptoms of depression decreased by 4 points (on the Centre for Epidemiologic Studies Depression Scale) for people using the Internet intervention, whereas symptoms of depression increased by 3 points for people who were not using the intervention. For every 6 people with depression using the Internet intervention, 1 person could be classified as not having depression after completing the intervention (compared with a control group not using the intervention).
Individuals with depression	Deprexis	For every 4 people with depression who completed the Internet intervention, 1 person recovered from their depression (compared with people who did not use the Internet intervention).

Target population	Name	Measures of efficacy
Generalized anxiety disorder		
Workers	Stress and Mood Management	People using the Internet intervention decreased their stress level to a greater extent (1-point relative decrease on the Symptoms of Distress scale) than people not using the intervention.
Individuals with anxiety	Internet-based Mindfulness Treatment	After completing the Internet intervention, people decreased their anxiety level (7-point relative decrease on the Beck Anxiety Inventory scale) compared with people who used a discussion forum (control). For every 3 people who completed the Internet intervention, 1 person recovered from their anxiety (compared with people who used a discussion forum).
Post-traumatic stress disorder		
Individuals who experienced disasters	My Disaster Recovery	People who completed the Internet intervention worried less (6-point relative decrease on the Penn State Worry Questionnaire) compared with people who did not use the intervention.
Panic disorder and phobias		
Individuals with panic disorders or phobias	Mental Health Online	After completing the Internet intervention, 67% of the people did not experience any panic attacks in the previous week; this compared with 11% for those who did not use the intervention. The number needed to treat was 2.
Adults with glossophobia (fear of public speaking)	Talk to Me	People who completed the Internet intervention reduced their fear and avoidance behaviors (3-point relative reduction for fear and 5-point relative reduction for avoidance on 10-point scales) compared with people who did not complete the intervention.
Diet and physical activity		
Overweight and obese adults	The Biggest Loser Club	People who completed the Internet intervention lost 2.1 kg weight, whereas people who did not complete the intervention added 0.4 kg. People who completed the Internet intervention reduced their waist circumference by 2.6 cm, whereas people who did not complete the intervention added 0.3 cm to their waist circumference.
Managers of organizations	ExecuPrev	Women who completed the Internet intervention reduced their waist circumference by 1.3 inches more than women who did not complete the intervention.
Insomnia		
Adults with chronic insomnia	Insomnie	In people who used the Internet intervention, sleep efficacy increased by 3% above that of people who did not use the intervention (measured at 48 weeks after the start of the intervention). Symptoms of depression also decreased by 3 points (on the Centre for Epidemiologic Studies Depression Scale) for people using the Internet intervention relative to people not using the intervention.
Adults with insomnia	SHUTi	For people who used the Internet intervention, the severity of insomnia (measured by the Insomnia Severity Index, a 28-point scale) decreased by 8 points relative to the people who did not use the intervention.
Individuals with sleep problems	Sleepio	In people who used the Internet intervention, sleep efficacy increased by 10% above that of people who did not use the intervention (measured at 8 weeks after the start of the intervention). Daytime performance also improved (by 2.5 points on a 5-point scale) in those who used the Internet intervention compared with those who did use the intervention.
Adults with chronic insomnia	RESTORE	People who used the Internet intervention improved their sleep quality by 0.5 points (on a 5-point scale) after 4 weeks, whereas the sleep quality of those who did not use the intervention decreased by 0.2 points.
Chronic pain		
Adults with chronic pain	painACTION	For every 4 people with back pain who used the Internet intervention, 1 person experienced improvement compared with a control (text-based material). For every 3 people with migraines who used the Internet intervention, 1 person experienced improvement compared with the control (usual treatment).
Childhood health problems		

Target population	Name	Measures of efficacy
Parents with children who have behavior problems	Triple P Online	For every 3 parents who completed the Internet intervention, 1 had a child who experienced clinical improvement in behavioral problems (control group was usual use of the Internet without the intervention). For every 2 parents who completed the Internet intervention, 1 experienced clinical improvement in parental confidence in dealing with behavior problems in their children.
Parents of children with encopresis	U Can POOP Too	For every 4 parents who completed the Internet intervention, 1 had a child who had no fecal accidents (comparison group was usual care).

Discussion

Principal Findings

In this systematic review, we developed a list of Internet health-related programs that demonstrated an evidence-based health benefit. The majority of programs dealt with substance abuse, mental health, or diet and physical activity. In addition, there were Internet programs dealing with disease management such as insomnia and chronic pain, as well as evidence-based Internet therapies for childhood health problems. There were some interventions with considerable efficacy (NNT<5); these included painACTION, Mental Health Online for panic disorders, Deprexis, Triple P Online, and U Can POOP Too.

There were several characteristics of successful Internet-delivered health interventions. First, most of the programs were rather intensive; they required assignments and engagement by the user over the course of weeks to months. For a number of the programs, not only were there interactive elements that prompted personalized feedback and self-monitoring, but also there were assignments that required the user to implement actions when they were not on the Internet such as tracking their sleeping habits via a diary, recording their eating habits throughout the day, or conducting physical activities throughout the week. In all of the therapies, educational materials were presented but these were often adjuncts to the main therapeutic approaches—not the principal tactic. Often the interventions followed cognitive behavioral strategies that were well-grounded in the psychological literature. Thus, most of the successful interventions were not truncated bits of information delivered in a short period of time. They were well-thought out progressive modules of engagement with multilayers of targeted approaches. Many also encouraged individuals to seek professional assistance if further help was needed.

Perhaps the most desirable aspect of having Internet evidence-based programs is the sheer magnitude of the audience. There were 3.5 billion Internet users in the world by December 2016 with a steady increase over the past decade [146]. By providing evidence-based programs, the potential to ameliorate some health problems or behaviors is enormous—even if the completion rates are rather low. The challenge is to determine whether these types of programs work equally well when translated into other languages and delivered to people with different social and cultural backgrounds. More information is needed regarding the triggers of personal readiness to use such programs and what factors appear to serve as enablers to use.

We found that 25 Internet programs were free to the public although some require registration. The availability of free health information removes a key barrier to the public, particularly individuals with lower incomes. The Pew Research Center found that 26% of Internet users who wanted health information were asked to pay, but only 2% of them actually paid for the information [5]. Request for payment resulted in lower-income individuals giving up the search, whereas wealthier individuals sought other avenues for the information [5]. This is of consequence because uninsured and poor individuals tend to have disproportionately higher rates of some health-related behaviors that such programs may help to abate [147].

Therapies that are Internet-based offer an attractive option for certain types of conditions due to easy access and low cost. In some locations, there may be insufficient numbers of clinicians who provide specific therapies, such as cognitive behavioral therapy for insomnia. These programs may also resolve other access problems, such as long wait times or lack of transportation to services. Such programs may be a choice for the first-line of engagement and, if the problem is not resolved, further in-person visits could be arranged. Many of these successful programs provide links to additional resources and some have specific information for health care professionals. The substance abuse websites are particularly strong in providing such links.

Another desirable feature of Internet programs is the ability to reach individuals who shun public places and therefore, are less likely to seek face-to-face care. Our review indicates that there are evidence-based programs for several phobias including social phobias (including shyness), panic attacks, and OCD. Moreover, such programs may reduce the likelihood of social stigma which sometimes occurs when seeking traditional avenues for assistance. For individuals with such problems, Internet programs may have the potential to provide the first step to eventual engagement with medical and neighborhood communities.

Although we anticipated that the Internet would be a valuable location for programs related to sexual health, it was surprising not to find any current evidence-based websites on the prevention of sexually transmitted disease through this review. There were several evidence-based programs for HIV prevention and unintended pregnancy that yielded a health-related benefit, but the programs that were tested did not yield a functional site to continue the program after the completion of the trials.

One of the challenges of Internet-delivered therapies relates to the constraints of the modality itself. There may be problems for individuals with vision problems or those with specific

functional disabilities. However, adaptive approaches may be possible to deliver audio programs for those who are blind and modifications may be available for those with specific motor-related disabilities. There are many case studies of computer technologies which have advanced the functional capabilities of those with various limitations; these include approaches which alter input devices, the use of assistive tools for processing, and restructuring the output [148,149].

Limitations

There are several limitations to this systematic review. There may be some evidence-based websites on health that were missed; we only included those RCTs that were part of a meta-analysis. Therefore, continuous updating will be necessary. This review is but the first step in this process; the development of mechanisms for continuous review is the next. Another limitation was that our focus was on self-help Internet programs. In this review, we did not include Internet-delivered health interventions that integrally involved clinicians, peer-to-peer therapies, or group therapies; an exhaustive review of each of these programs would be helpful for future research studies, so that the breadth of this field could be appreciated and any deficiencies identified. Moreover, this review is meant to initiate the process of dissemination of evidence-based websites and, therefore, additional steps will be necessary. We consider that this process will eventually become analogous to the procedures utilized during dissemination and implementation of conventional medical therapies. That is, RCTs are conducted and reviewed through meta-analysis. This is followed by professional guidelines and recommendations for use, based on the RCT evidence. This, then, typically yields studies which evaluate implementation in the wider population or within specific subgroups. Because these interventions are housed on the Internet, mechanisms for dissemination will involve Internet engagement but will likely require participation of public health professionals, policy makers, and providers of health care.

There are some precautions, however, when delivering Internet therapies directly to the public. Researchers understand that the demonstration of an overall benefit in an RCT relates to a group effect and that this does not necessarily indicate that every single person will receive a benefit. Therefore, part of the implementation process to the public should involve education regarding the limitations of evidence-based Web interventions. They do not guarantee a specific result; they only promise a greater likelihood of a benefit if the therapy is completed.

Conclusions

We identified several evidence-based health interventions that are currently available on the Internet. They include therapies related to substance abuse, mental health, diet and physical activity, disease management, disease prevention, and childhood health-related problems. Unfortunately, most of the Internet-delivered health interventions that were efficacious through RCTs were not available after the conclusion of the trials. The challenge is to find avenues through governments, organizations, universities, and interested corporations to host the evidence-based Internet programs and to notify the public of their locations. If this process is expanded, such therapies provide hope of a cost-effective mechanism to achieve healthier populations globally.

Acknowledgments

We thank Ms Marisa Conte for her assistance with the literature search.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Home pages of evidence-based online programs.

[PDF File (Adobe PDF File), 6MB - jmir_v19i3e90_app1.pdf]

References

1. Yehualashet YG, Horton J, Mkanda P, Vaz RG, Afolabi O, Gashu SG, et al. Intensified local resource mobilization for the polio eradication initiative: the experience of World Health Organization in Nigeria during 2008-2015. J Infect Dis 2016 May 01;213 Suppl 3:S101-S107 [FREE Full text] [doi: 10.1093/infdis/jiv535] [Medline: 26912380]
2. Gómez-Batiste X, Caja C, Espinosa J, Bullich I, Martínez-Muñoz M, Porta-Sales J, et al. The Catalonia World Health Organization demonstration project for palliative care implementation: quantitative and qualitative results at 20 years. J Pain Symptom Manage 2012 Apr;43(4):783-794. [doi: 10.1016/j.jpainsymman.2011.05.006] [Medline: 22265127]
3. National Academy of Sciences. 2012. For the public's health: investing in a healthier future URL: http://www.nationalacademies.org/hmd/Reports/2012/For-the-Publics-Health-Investing-in-a-Healthier-Future.aspx [accessed 2017-03-17] [WebCite Cache ID 6p1kDsndo]
4. Hunter DJ, Visram S. Better evidence for smarter policy making. BMJ 2016 Dec 01;355:i6399. [Medline: 27908919]
5. Fox S, Duggan M. Pew Research Center. Health online 2013 URL: http://www.pewinternet.org/files/old-media//Files/Reports/PIP_HealthOnline.pdf [accessed 2016-12-05] [WebCite Cache ID 6mX6XKUG7]

6. Moher D, Liberati A, Tetzlaff J, Altman DG. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. Ann Intern Med 2009 Aug 18;151(4):264-9, W64. [Medline: 19622511]
7. Andersson G, Cuijpers P. Internet-based and other computerized psychological treatments for adult depression: a meta-analysis. Cogn Behav Ther 2009;38(4):196-205. [doi: 10.1080/16506070903318960] [Medline: 20183695]
8. Andrews G, Cuijpers P, Craske MG, McEvoy P, Titov N. Computer therapy for the anxiety and depressive disorders is effective, acceptable and practical health care: a meta-analysis. PLoS One 2010;5(10):e13196 [FREE Full text] [doi: 10.1371/journal.pone.0013196] [Medline: 20967242]
9. Aneni EC, Roberson LL, Maziak W, Agatston AS, Feldman T, Rouseff M, et al. A systematic review of internet-based worksite wellness approaches for cardiovascular disease risk management: outcomes, challenges & opportunities. PLoS One 2014;9(1):e83594 [FREE Full text] [doi: 10.1371/journal.pone.0083594] [Medline: 24421894]
10. Angeles R, Howard M, Dolovich L. The effectiveness of web-based tools for improving blood glucose control in patients with diabetes Mellitus: A meta-analysis. Can J Diabetes 2011;35:344-352.
11. Antwi F, Fazylova N, Garcon M, Lopez L, Rubiano R, Slyer J. The effectiveness of web-based programs on the reduction of childhood obesity in school-aged children: A systematic review. JBI Libr Syst Rev 2012;10(42 Suppl):1-14. [doi: 10.11124/jbisrir-2012-248] [Medline: 27820152]
12. Arnberg FK, Linton SJ, Hultcrantz M, Heintz E, Jonsson U. Internet-delivered psychological treatments for mood and anxiety disorders: a systematic review of their efficacy, safety, and cost-effectiveness. PLoS One 2014;9(5):e98118 [FREE Full text] [doi: 10.1371/journal.pone.0098118] [Medline: 24844847]
13. Bailey JV, Murray E, Rait G, Mercer CH, Morris RW, Peacock R, et al. Interactive computer-based interventions for sexual health promotion. Cochrane Database Syst Rev 2010(9):CD006483. [doi: 10.1002/14651858.CD006483.pub2] [Medline: 20824850]
14. Bandelow B, Reitt M, Röver C, Michaelis S, Görlich Y, Wedekind D. Efficacy of treatments for anxiety disorders: a meta-analysis. Int Clin Psychopharmacol 2015 Jul;30(4):183-192. [doi: 10.1097/YIC.0000000000000078] [Medline: 25932596]
15. Beintner I, Jacobi C, Taylor CB. Effects of an Internet-based prevention programme for eating disorders in the USA and Germany--a meta-analytic review. Eur Eat Disord Rev 2012 Jan;20(1):1-8. [doi: 10.1002/erv.1130] [Medline: 21796737]
16. Black N, Mullan B, Sharpe L. Computer-delivered interventions for reducing alcohol consumption: meta-analysis and meta-regression using behaviour change techniques and theory. Health Psychol Rev 2016 Sep;10(3):341-357. [doi: 10.1080/17437199.2016.1168268] [Medline: 26999311]
17. Buhrman M, Gordh T, Andersson G. Internet interventions for chronic pain including headache: a systematic review. Internet Interv 2016(4):17-34.
18. Cadigan JM, Haeny AM, Martens MP, Weaver CC, Takamatsu SK, Arterberry BJ. Personalized drinking feedback: A meta-analysis of in-person versus computer-delivered interventions. J Consult Clin Psychol 2015 Apr;83(2):430-437 [FREE Full text] [doi: 10.1037/a0038394] [Medline: 25486373]
19. Car J, Lang B, Colledge A, Ung C, Majeed A. Interventions for enhancing consumers' online health literacy. Cochrane Database Syst Rev 2011(6):CD007092. [doi: 10.1002/14651858.CD007092.pub2] [Medline: 21678364]
20. Carey KB, Scott-Sheldon LA, Elliott JC, Garey L, Carey MP. Face-to-face versus computer-delivered alcohol interventions for college drinkers: a meta-analytic review, 1998 to 2010. Clin Psychol Rev 2012 Dec;32(8):690-703 [FREE Full text] [doi: 10.1016/j.cpr.2012.08.001] [Medline: 23022767]
21. Chen Y, Madan J, Welton N, Yahaya I, Aveyard P, Bauld L, et al. Effectiveness and cost-effectiveness of computer and other electronic aids for smoking cessation: a systematic review and network meta-analysis. Health Technol Assess 2012;16(38):1-205, iii [FREE Full text] [doi: 10.3310/hta16380] [Medline: 23046909]
22. Cheng SK, Dizon J. Computerised cognitive behavioural therapy for insomnia: a systematic review and meta-analysis. Psychother Psychosom 2012;81(4):206-216. [doi: 10.1159/000335379] [Medline: 22585048]
23. Civljak M, Stead LF, Hartmann-Boyce J, Sheikh A, Car J. Internet-based interventions for smoking cessation. Cochrane Database Syst Rev 2013;7:CD007078. [doi: 10.1002/14651858.CD007078.pub4] [Medline: 23839868]
24. Cowpertwait L, Clarke D. Effectiveness of web-based psychological interventions for depression: a meta-analysis. Int J Ment Health Addiction 2013;11:247-268.
25. Cuijpers P, Marks IM, van SA, Cavanagh K, Gega L, Andersson G. Computer-aided psychotherapy for anxiety disorders: a meta-analytic review. Cogn Behav Ther 2009;38(2):66-82. [doi: 10.1080/16506070802694776] [Medline: 20183688]
26. Davies CA, Spence JC, Vandelanotte C, Caperchione CM, Mummery WK. Meta-analysis of internet-delivered interventions to increase physical activity levels. Int J Behav Nutr Phys Act 2012;9:52 [FREE Full text] [doi: 10.1186/1479-5868-9-52] [Medline: 22546283]
27. Davies EB, Morriss R, Glazebrook C. Computer-delivered and web-based interventions to improve depression, anxiety, and psychological well-being of university students: a systematic review and meta-analysis. J Med Internet Res 2014;16(5):e130 [FREE Full text] [doi: 10.2196/jmir.3142] [Medline: 24836465]
28. DeSmet A, Shegog R, Van RD, Crombez G, De BI. A systematic review and meta-analysis of interventions for sexual health promotion involving serious digital games. Games Health J 2015 Apr;4(2):78-90. [doi: 10.1089/g4h.2014.0110] [Medline: 26181801]

29. Dèttore D, Pozza A, Andersson G. Efficacy of technology-delivered cognitive behavioural therapy for OCD versus control conditions, and in comparison with therapist-administered CBT: meta-analysis of randomized controlled trials. Cogn Behav Ther 2015;44(3):190-211. [doi: 10.1080/16506073.2015.1005660] [Medline: 25705787]
30. Ebert DD, Zarski A, Christensen H, Stikkelbroek Y, Cuijpers P, Berking M, et al. Internet and computer-based cognitive behavioral therapy for anxiety and depression in youth: a meta-analysis of randomized controlled outcome trials. PLoS One 2015;10(3):e0119895 [FREE Full text] [doi: 10.1371/journal.pone.0119895] [Medline: 25786025]
31. Eccleston C, Fisher E, Craig L, Duggan GB, Rosser BA, Keogh E. Psychological therapies (Internet-delivered) for the management of chronic pain in adults. Cochrane Database Syst Rev 2014;2:CD010152. [doi: 10.1002/14651858.CD010152.pub2] [Medline: 24574082]
32. Foster C, Richards J, Thorogood M, Hillsdon M. Remote and web 2.0 interventions for promoting physical activity. Cochrane Database Syst Rev 2013 Sep 30;9:CD010395 [FREE Full text] [doi: 10.1002/14651858.CD010395.pub2] [Medline: 24085594]
33. Gould RL, Coulson MC, Howard RJ. Cognitive behavioral therapy for depression in older people: a meta-analysis and meta-regression of randomized controlled trials. J Am Geriatr Soc 2012 Oct;60(10):1817-1830. [doi: 10.1111/j.1532-5415.2012.04166.x] [Medline: 23003115]
34. Graham AL, Carpenter KM, Cha S, Cole S, Jacobs MA, Raskob M, et al. Systematic review and meta-analysis of Internet interventions for smoking cessation among adults. Subst Abuse Rehabil 2016;7:55-69 [FREE Full text] [doi: 10.2147/SAR.S101660] [Medline: 27274333]
35. Grunenberg E, Lin J, Baumeister H. [Effectiveness of web-based psychological interventions for weight loss--a systematic review]. Rehabilitation (Stuttg) 2013 Jun;52(3):182-187. [doi: 10.1055/s-0033-1343489] [Medline: 23761207]
36. Hamel LM, Robbins LB, Wilbur J. Computer- and web-based interventions to increase preadolescent and adolescent physical activity: a systematic review. J Adv Nurs 2011 Feb;67(2):251-268. [doi: 10.1111/j.1365-2648.2010.05493.x] [Medline: 21198800]
37. Hedman E, Ljótsson B, Lindefors N. Cognitive behavior therapy via the Internet: a systematic review of applications, clinical efficacy and cost-effectiveness. Expert Rev Pharmacoecon Outcomes Res 2012 Dec;12(6):745-764. [doi: 10.1586/erp.12.67] [Medline: 23252357]
38. Hofmann SG, Wu JQ, Boettcher H. Effect of cognitive-behavioral therapy for anxiety disorders on quality of life: a meta-analysis. J Consult Clin Psychol 2014 Jun;82(3):375-391 [FREE Full text] [doi: 10.1037/a0035491] [Medline: 24447006]
39. Kodama S, Saito K, Tanaka S, Horikawa C, Fujiwara K, Hirasawa R, et al. Effect of Web-based lifestyle modification on weight control: a meta-analysis. Int J Obes (Lond) 2012 May;36(5):675-685. [doi: 10.1038/ijo.2011.121] [Medline: 21694698]
40. Kuester A, Niemeyer H, Knaevelsrud C. Internet-based interventions for posttraumatic stress: A meta-analysis of randomized controlled trials. Clin Psychol Rev 2016 Feb;43:1-16. [doi: 10.1016/j.cpr.2015.11.004] [Medline: 26655959]
41. Li L, Xiong L, Zhang S, Yu Q, Chen M. Cognitive-behavioral therapy for irritable bowel syndrome: a meta-analysis. J Psychosom Res 2014 Jul;77(1):1-12. [doi: 10.1016/j.jpsychores.2014.03.006] [Medline: 24913335]
42. Linetzky M, Pergamin-Hight L, Pine DS, Bar-Haim Y. Quantitative evaluation of the clinical efficacy of attention bias modification treatment for anxiety disorders. Depress Anxiety 2015 Jun;32(6):383-391. [doi: 10.1002/da.22344] [Medline: 25708991]
43. Liu S, Dunford SD, Leung YW, Brooks D, Thomas SG, Eysenbach G, et al. Reducing blood pressure with Internet-based interventions: a meta-analysis. Can J Cardiol 2013 May;29(5):613-621. [doi: 10.1016/j.cjca.2013.02.007] [Medline: 23618507]
44. Lustria ML, Noar SM, Cortese J, Van Stee SK, Glueckauf RL, Lee J. A meta-analysis of web-delivered tailored health behavior change interventions. J Health Commun 2013;18(9):1039-1069. [doi: 10.1080/10810730.2013.768727] [Medline: 23750972]
45. Maon S, Edirippulige S, Ware R, Batch J. The use of web-based interventions to prevent excessive weight gain. J Telemed Telecare 2012 Jan;18(1):37-41. [doi: 10.1258/jtt.2011.110306] [Medline: 22101608]
46. Muresan V, Montgomery G, David D. Eemotional outcomes and mechanisms of change in online cognitive-behavioral interventions: a quantitative meta-analysis of clinical controlled studies. J Technol Hum Serv 2012;30:1-13.
47. Murray E, Burns J, See TS, Lai R, Nazareth I. Interactive health communication applications for people with chronic disease. Cochrane Database Syst Rev 2005(4):CD004274. [doi: 10.1002/14651858.CD004274.pub4] [Medline: 16235356]
48. Myung S, McDonnell DD, Kazinets G, Seo HG, Moskowitz JM. Effects of web- and computer-based smoking cessation programs: meta-analysis of randomized controlled trials. Arch Intern Med 2009 May 25;169(10):929-937. [doi: 10.1001/archinternmed.2009.109] [Medline: 19468084]
49. Noar SM, Black HG, Pierce LB. Efficacy of computer technology-based HIV prevention interventions: a meta-analysis. AIDS 2009 Jan 2;23(1):107-115. [doi: 10.1097/QAD.0b013e32831c5500] [Medline: 19050392]
50. Nyenhuis N, Golm D, Kröner-Herwig B. A systematic review and meta-analysis on the efficacy of self-help interventions in tinnitus. Cogn Behav Ther 2013;42(2):159-169. [doi: 10.1080/16506073.2013.803496] [Medline: 23777192]

51. Pal K, Eastwood SV, Michie S, Farmer A, Barnard ML, Peacock R, et al. Computer-based interventions to improve self-management in adults with type 2 diabetes: a systematic review and meta-analysis. Diabetes Care 2014 Jun;37(6):1759-1766. [doi: 10.2337/dc13-1386] [Medline: 24855158]
52. Portnoy DB, Scott-Sheldon LA, Johnson BT, Carey MP. Computer-delivered interventions for health promotion and behavioral risk reduction: a meta-analysis of 75 randomized controlled trials, 1988-2007. Prev Med 2008 Jul;47(1):3-16 [FREE Full text] [doi: 10.1016/j.ypmed.2008.02.014] [Medline: 18403003]
53. Reger MA, Gahm GA. A meta-analysis of the effects of internet- and computer-based cognitive-behavioral treatments for anxiety. J Clin Psychol 2009 Jan;65(1):53-75. [doi: 10.1002/jclp.20536] [Medline: 19051274]
54. Richards J, Thorogood M, Hillsdon M, Foster C. Face-to-face versus remote and web 2.0 interventions for promoting physical activity. Cochrane Database Syst Rev 2013 Sep 30(9):CD010393. [doi: 10.1002/14651858.CD010393.pub2] [Medline: 24085593]
55. Richards D, Richardson T, Timulak L, McElvaney J. The efficacy of internet-delivered treatment for generalized anxiety disorder: A systematic review and meta-analysis. Internet Interv 2015;2:272-282.
56. Riper H, Spek V, Boon B, Conijn B, Kramer J, Martin-Abello K, et al. Effectiveness of E-self-help interventions for curbing adult problem drinking: a meta-analysis. J Med Internet Res 2011;13(2):e42 [FREE Full text] [doi: 10.2196/jmir.1691] [Medline: 21719411]
57. Riper H, Blankers M, Hadiwijaya H, Cunningham J, Clarke S, Wiers R, et al. Effectiveness of guided and unguided low-intensity internet interventions for adult alcohol misuse: a meta-analysis. PLoS One 2014;9(6):e99912 [FREE Full text] [doi: 10.1371/journal.pone.0099912] [Medline: 24937483]
58. Rooke S, Thorsteinsson E, Karpin A, Copeland J, Allsop D. Computer-delivered interventions for alcohol and tobacco use: a meta-analysis. Addiction 2010 Aug;105(8):1381-1390. [doi: 10.1111/j.1360-0443.2010.02975.x] [Medline: 20528806]
59. Samoocha D, Bruinvels DJ, Elbers NA, Anema JR, van der Beek AJ. Effectiveness of web-based interventions on patient empowerment: a systematic review and meta-analysis. J Med Internet Res 2010;12(2):e23 [FREE Full text] [doi: 10.2196/jmir.1286] [Medline: 20581001]
60. Scott-Sheldon LA, Lantini R, Jennings EG, Thind H, Rosen RK, Salmoirago-Blotcher E, et al. Text messaging-based interventions for smoking cessation: a systematic review and meta-analysis. JMIR Mhealth Uhealth 2016;4(2):e49 [FREE Full text] [doi: 10.2196/mhealth.5436] [Medline: 27207211]
61. Seo D, Niu J. Evaluation of Internet-Based Interventions on Waist Circumference Reduction: A Meta-Analysis. J Med Internet Res 2015 Jul 21;17(7):e181 [FREE Full text] [doi: 10.2196/jmir.3921] [Medline: 26199208]
62. Seyffert M, Lagisetty P, Landgraf J, Chopra V, Pfeiffer PN, Conte ML, Rogers Mary A M. Internet-delivered cognitive behavioral therapy to treat insomnia: a systematic review and meta-analysis. PLoS One 2016;11(2):e0149139 [FREE Full text] [doi: 10.1371/journal.pone.0149139] [Medline: 26867139]
63. Shahab L, McEwen A. Online support for smoking cessation: a systematic review of the literature. Addiction 2009 Nov;104(11):1792-1804. [doi: 10.1111/j.1360-0443.2009.02710.x] [Medline: 19832783]
64. Spek V, Cuijpers P, Nyklícek I, Riper H, Keyzer J, Pop V. Internet-based cognitive behaviour therapy for symptoms of depression and anxiety: a meta-analysis. Psychol Med 2007 Mar;37(3):319-328. [doi: 10.1017/S0033291706008944] [Medline: 17112400]
65. Spijkerman MP, Pots WT, Bohlmeijer ET. Effectiveness of online mindfulness-based interventions in improving mental health: A review and meta-analysis of randomised controlled trials. Clin Psychol Rev 2016 Apr;45:102-114 [FREE Full text] [doi: 10.1016/j.cpr.2016.03.009] [Medline: 27111302]
66. Spohr SA, Nandy R, Gandhiraj D, Vemulapalli A, Anne S, Walters ST. Efficacy of sms text message interventions for smoking cessation: a meta-analysis. J Subst Abuse Treat 2015 Sep;56:1-10. [doi: 10.1016/j.jsat.2015.01.011] [Medline: 25720333]
67. Stinson J, Wilson R, Gill N, Yamada J, Holt J. A systematic review of internet-based self-management interventions for youth with health conditions. J Pediatr Psychol 2009 Jun;34(5):495-510 [FREE Full text] [doi: 10.1093/jpepsy/jsn115] [Medline: 19029142]
68. Tait RJ, Spijkerman R, Riper H. Internet and computer based interventions for cannabis use: a meta-analysis. Drug Alcohol Depend 2013 Dec 1;133(2):295-304. [doi: 10.1016/j.drugalcdep.2013.05.012] [Medline: 23747236]
69. Takano A, Miyamoto Y, Matsumoto T. [A review about new approaches using the Internet and computer technology for people with drug use disorder]. Nihon Arukoru Yakubutsu Igakkai Zasshi 2015 Feb;50(1):19-34. [Medline: 26027412]
70. Tao D, Or CK. Effects of self-management health information technology on glycaemic control for patients with diabetes: a meta-analysis of randomized controlled trials. J Telemed Telecare 2013 Apr 5;19:133-143. [doi: 10.1177/1357633X13479701] [Medline: 23563018]
71. Tarver J, Daley D, Lockwood J, Sayal K. Are self-directed parenting interventions sufficient for externalising behaviour problems in childhood? A systematic review and meta-analysis. Eur Child Adolesc Psychiatry 2014 Dec;23(12):1123-1137. [doi: 10.1007/s00787-014-0556-5] [Medline: 24842197]
72. Wieland LS, Falzon L, Sciamanna CN, Trudeau KJ, Brodney S, Schwartz JE, et al. Interactive computer-based interventions for weight loss or weight maintenance in overweight or obese people. Cochrane Database Syst Rev 2012;8:CD007675 [FREE Full text] [doi: 10.1002/14651858.CD007675.pub2] [Medline: 22895964]

73. Wootton BM. Remote cognitive-behavior therapy for obsessive-compulsive symptoms: A meta-analysis. Clin Psychol Rev 2016 Feb;43:103-113. [doi: 10.1016/j.cpr.2015.10.001] [Medline: 26494179]
74. Ye Y, Zhang Y, Chen J, Liu J, Li X, Liu Y, et al. Internet-based cognitive behavioral therapy for insomnia (ICBT-I) improves comorbid anxiety and depression-a meta-analysis of randomized controlled trials. PLoS One 2015;10(11):e0142258 [FREE Full text] [doi: 10.1371/journal.pone.0142258] [Medline: 26581107]
75. Zachariae R, Lyby MS, Ritterband LM, O'Toole MS. Efficacy of internet-delivered cognitive-behavioral therapy for insomnia - a systematic review and meta-analysis of randomized controlled trials. Sleep Med Rev 2016 Dec;30:1-10. [doi: 10.1016/j.smrv.2015.10.004] [Medline: 26615572]
76. Zhao J, Zhai Y, Zhu W, Sun D. Effectiveness of telemedicine for controlling asthma symptoms: a systematic review and meta-analysis. Telemed J E Health 2015 Jun;21(6):484-492. [doi: 10.1089/tmj.2014.0119] [Medline: 25393915]
77. Doumas DM, Hannah E. Preventing high-risk drinking in youth in the workplace: a web-based normative feedback program. J Subst Abuse Treat 2008 Apr;34(3):263-271. [doi: 10.1016/j.jsat.2007.04.006] [Medline: 17600650]
78. Cunningham JA, Wild TC, Cordingley J, van MT, Humphreys K. A randomized controlled trial of an internet-based intervention for alcohol abusers. Addiction 2009 Dec;104(12):2023-2032 [FREE Full text] [doi: 10.1111/j.1360-0443.2009.02726.x] [Medline: 19922569]
79. Cunningham JA, Murphy M, Hendershot CS. Treatment dismantling pilot study to identify the active ingredients in personalized feedback interventions for hazardous alcohol use: randomized controlled trial. Addict Sci Clin Pract 2014 Dec 10;10:1 [FREE Full text] [doi: 10.1186/s13722-014-0022-1] [Medline: 25539597]
80. Boon B, Risselada A, Huiberts A, Riper H, Smit F. Curbing alcohol use in male adults through computer generated personalized advice: randomized controlled trial. J Med Internet Res 2011;13(2):e43 [FREE Full text] [doi: 10.2196/jmir.1695] [Medline: 21719412]
81. Brendryen H, Lund IO, Johansen AB, Riksheim M, Nesvåg S, Duckert F. Balance--a pragmatic randomized controlled trial of an online intensive self-help alcohol intervention. Addiction 2014 Feb;109(2):218-226. [doi: 10.1111/add.12383] [Medline: 24134709]
82. Riper H, Kramer J, Smit F, Conijn B, Schippers G, Cuijpers P. Web-based self-help for problem drinkers: a pragmatic randomized trial. Addiction 2008 Feb;103(2):218-227. [doi: 10.1111/j.1360-0443.2007.02063.x] [Medline: 18199300]
83. Steiner J, Woodall W, Yeagley J. Society for Prevention Research. 2005. The E-Chug: a randomized, controlled study of a web-based binge drinking intervention with college freshman URL: https://3rdmilclassrooms.com/Documents/SPR_2005.pdf [accessed 2017-03-17] [WebCite Cache ID 6p1nhHOtX]
84. Doumas D, Andersen L. Reducing alcohol use in first-year university students: evaluation of a web-based personalized feedback program. JCC 2009;12:18-32. [doi: 10.1002/j.2161-1882.2009.tb00037.x]
85. Doumas D, Kane C, Navarro T, Roman J. Decreasing heavy drinking in first-year studentsvaluation of a web-based personalized feedback program administered during orientation. JCC 2011;14(1):5-20. [doi: 10.1002/j.2161-1882.2011.tb00060.x]
86. Alfonso J, Hall TV, Dunn ME. Feedback-based alcohol interventions for mandated students: an effectiveness study of three modalities. Clin Psychol Psychother 2013;20(5):411-423. [doi: 10.1002/cpp.1786] [Medline: 22362618]
87. Paschall MJ, Antin T, Ringwalt CL, Saltz RF. Evaluation of an Internet-based alcohol misuse prevention course for college freshmen: findings of a randomized multi-campus trial. Am J Prev Med 2011 Sep;41(3):300-308 [FREE Full text] [doi: 10.1016/j.amepre.2011.03.021] [Medline: 21855745]
88. Lovecchio CP, Wyatt TM, DeJong W. Reductions in drinking and alcohol-related harms reported by first-year college students taking an online alcohol education course: a randomized trial. J Health Commun 2010 Oct;15(7):805-819. [doi: 10.1080/10810730.2010.514032] [Medline: 21104507]
89. Hustad JT, Barnett NP, Borsari B, Jackson KM. Web-based alcohol prevention for incoming college students: a randomized controlled trial. Addict Behav 2010 Mar;35(3):183-189 [FREE Full text] [doi: 10.1016/j.addbeh.2009.10.012] [Medline: 19900763]
90. Chiauzzi E, Green TC, Lord S, Thum C, Goldstein M. My student body: a high-risk drinking prevention web site for college students. J Am Coll Health 2005;53(6):263-274 [FREE Full text] [doi: 10.3200/JACH.53.6.263-274] [Medline: 15900990]
91. Croom K, Staiano-Coico L, Lesser ML, Lewis DK, Reyna VF, Marchell TC, et al. The glass is half full: evidence for efficacy of alcohol-wise at one university but not the other. J Health Commun 2015;20(6):627-638. [doi: 10.1080/10810730.2015.1012239] [Medline: 25909233]
92. Stoddard JL, Augustson EM, Moser RP. Effect of adding a virtual community (bulletin board) to smokefree.gov: randomized controlled trial. J Med Internet Res 2008;10(5):e53 [FREE Full text] [doi: 10.2196/jmir.1124] [Medline: 19097974]
93. Fraser D, Kobinsky K, Smith SS, Kramer J, Theobald WE, Baker TB. Five population-based interventions for smoking cessation: a MOST trial. Transl Behav Med 2014 Dec;4(4):382-390 [FREE Full text] [doi: 10.1007/s13142-014-0278-8] [Medline: 25584087]
94. Borland R, Balmford J, Benda P. Population-level effects of automated smoking cessation help programs: a randomized controlled trial. Addiction 2013 Mar;108(3):618-628. [doi: 10.1111/j.1360-0443.2012.04091.x] [Medline: 22994457]
95. Etter J. Comparing the efficacy of two Internet-based, computer-tailored smoking cessation programs: a randomized trial. J Med Internet Res 2005;7(1):e2 [FREE Full text] [doi: 10.2196/jmir.7.1.e2] [Medline: 15829474]

96. Mañanes G, Vallejo MA. Usage and effectiveness of a fully automated, open-access, Spanish Web-based smoking cessation program: randomized controlled trial. J Med Internet Res 2014;16(4):e111 [FREE Full text] [doi: 10.2196/jmir.3091] [Medline: 24760951]
97. Leykin Y, Aguilera A, Torres LD, Pérez-Stable EJ, Muñoz RF. Interpreting the outcomes of automated internet-based randomized trials: example of an international smoking cessation study. J Med Internet Res 2012 Feb;14(1):e5 [FREE Full text] [doi: 10.2196/jmir.1829] [Medline: 22314016]
98. Wangberg SC, Nilsen O, Antypas K, Gram IT. Effect of tailoring in an internet-based intervention for smoking cessation: randomized controlled trial. J Med Internet Res 2011 Dec 15;13(4):e121 [FREE Full text] [doi: 10.2196/jmir.1605] [Medline: 22169631]
99. Graham AL, Cobb NK, Papandonatos GD, Moreno JL, Kang H, Tinkelman DG, et al. A randomized trial of Internet and telephone treatment for smoking cessation. Arch Intern Med 2011 Jan 10;171(1):46-53 [FREE Full text] [doi: 10.1001/archinternmed.2010.451] [Medline: 21220660]
100. Rooke S, Copeland J, Norberg M, Hine D, McCambridge J. Effectiveness of a self-guided web-based cannabis treatment program: randomized controlled trial. J Med Internet Res 2013;15(2):e26 [FREE Full text] [doi: 10.2196/jmir.2256] [Medline: 23470329]
101. Tossmann H, Jonas B, Tensil M, Lang P, Strüber E. A controlled trial of an internet-based intervention program for cannabis users. Cyberpsychol Behav Soc Netw 2011 Nov;14(11):673-679. [doi: 10.1089/cyber.2010.0506] [Medline: 21651419]
102. Lintvedt OK, Griffiths KM, Sørensen K, Østvik AR, Wang CE, Eisemann M, et al. Evaluating the effectiveness and efficacy of unguided internet-based self-help intervention for the prevention of depression: a randomized controlled trial. Clin Psychol Psychother 2013;20(1):10-27. [doi: 10.1002/cpp.770] [Medline: 21887811]
103. O'Kearney R, Gibson M, Christensen H, Griffiths KM. Effects of a cognitive-behavioural internet program on depression, vulnerability to depression and stigma in adolescent males: a school-based controlled trial. Cogn Behav Ther 2006;35(1):43-54. [doi: 10.1080/16506070500303456] [Medline: 16500776]
104. Christensen H, Griffiths KM, Mackinnon AJ, Brittliffe K. Online randomized controlled trial of brief and full cognitive behaviour therapy for depression. Psychol Med 2006 Dec;36(12):1737-1746. [doi: 10.1017/S0033291706008695] [Medline: 16938144]
105. Farrer L, Christensen H, Griffiths KM, Mackinnon A. Internet-based CBT for depression with and without telephone tracking in a national helpline: randomised controlled trial. PLoS One 2011;6(11):e28099 [FREE Full text] [doi: 10.1371/journal.pone.0028099] [Medline: 22140514]
106. Meyer B, Berger T, Caspar F, Beevers CG, Andersson G, Weiss M. Effectiveness of a novel integrative online treatment for depression (Deprexis): randomized controlled trial. J Med Internet Res 2009;11(2):e15 [FREE Full text] [doi: 10.2196/jmir.1151] [Medline: 19632969]
107. de Graaf LE, Huibers MJ, Riper H, Gerhards SA, Arntz A. Use and acceptability of unsupported online computerized cognitive behavioral therapy for depression and associations with clinical outcome. J Affect Disord 2009 Aug;116(3):227-231. [doi: 10.1016/j.jad.2008.12.009] [Medline: 19167094]
108. Billings DW, Cook RF, Hendrickson A, Dove DC. A web-based approach to managing stress and mood disorders in the workforce. J Occup Environ Med 2008 Aug;50(8):960-968. [doi: 10.1097/JOM.0b013e31816c435b] [Medline: 18695455]
109. Chiauzzi E, Brevard J, Thum C, Thurn C, Decembrele S, Lord S. MyStudentBody-Stress: an online stress management intervention for college students. J Health Commun 2008 Sep;13(6):555-572. [doi: 10.1080/10810730802281668] [Medline: 18726812]
110. Boettcher J, Aström V, Påhlsson D, Schenström O, Andersson G, Carlbring P. Internet-based mindfulness treatment for anxiety disorders: a randomized controlled trial. Behav Ther 2014 Mar;45(2):241-253 [FREE Full text] [doi: 10.1016/j.beth.2013.11.003] [Medline: 24491199]
111. Wang Z, Wang J, Maercker A. Chinese My Trauma Recovery, a Web-based intervention for traumatized persons in two parallel samples: randomized controlled trial. J Med Internet Res 2013;15(9):e213 [FREE Full text] [doi: 10.2196/jmir.2690] [Medline: 24080137]
112. Steinmetz SE, Benight CC, Bishop SL, James LE. My Disaster Recovery: a pilot randomized controlled trial of an Internet intervention. Anxiety Stress Coping 2012;25(5):593-600 [FREE Full text] [doi: 10.1080/10615806.2011.604869] [Medline: 21834731]
113. Richards J, Klein B, Austin D. Internet cognitive behavioural therapy for panic disorder: Does the inclusion of stress management information improve end-state functioning? Clinical Psychologist. Clinical Psychologist 2006;10(1):2-15.
114. Klein B, Richards J. A brief internet-based treatment for panic disorder. Behav Cogn Psychother 2001;29:113-117.
115. Titov N, Andrews G, Schwencke G, Robinson E, Peters L, Spence J. Randomized controlled trial of Internet cognitive behavioural treatment for social phobia with and without motivational enhancement strategies. Aust N Z J Psychiatry 2010 Oct;44(10):938-945. [doi: 10.3109/00048674.2010.493859] [Medline: 20932208]
116. Titov N, Andrews G, Choi I, Schwencke G, Johnston L. Randomized controlled trial of web-based treatment of social phobia without clinician guidance. Aust N Z J Psychiatry 2009;43:913-919.

117. Titov N, Andrews G, Choi I, Schwencke G, Mahoney A. Shyness 3: randomized controlled trial of guided versus unguided Internet-based CBT for social phobia. Aust N Z J Psychiatry 2008 Dec;42(12):1030-1040. [doi: 10.1080/00048670802512107] [Medline: 19016091]
118. Botella C, Gallego MJ, Garcia-Palacios A, Guillen V, Baños RM, Quero S, et al. An Internet-based self-help treatment for fear of public speaking: a controlled trial. Cyberpsychol Behav Soc Netw 2010 Aug;13(4):407-421. [doi: 10.1089/cyber.2009.0224] [Medline: 20712499]
119. Mahoney AE, Mackenzie A, Williams AD, Smith J, Andrews G. Internet cognitive behavioural treatment for obsessive compulsive disorder: a randomised controlled trial. Behav Res Ther 2014 Dec;63:99-106. [doi: 10.1016/j.brat.2014.09.012] [Medline: 25461784]
120. Oenema A, Brug J, Dijkstra A, de Weerdt I, de Vries H. Efficacy and use of an internet-delivered computer-tailored lifestyle intervention, targeting saturated fat intake, physical activity and smoking cessation: a randomized controlled trial. Ann Behav Med 2008 Apr;35(2):125-135. [doi: 10.1007/s12160-008-9023-1] [Medline: 18363076]
121. Glasgow RE, Kurz D, King D, Dickman JM, Faber AJ, Halterman E, et al. Outcomes of minimal and moderate support versions of an internet-based diabetes self-management support program. J Gen Intern Med 2010 Dec;25(12):1315-1322 [FREE Full text] [doi: 10.1007/s11606-010-1480-0] [Medline: 20714820]
122. Cook RF, Billings DW, Hersch RK, Back AS, Hendrickson A. A field test of a web-based workplace health promotion program to improve dietary practices, reduce stress, and increase physical activity: randomized controlled trial. J Med Internet Res 2007;9(2):e17 [FREE Full text] [doi: 10.2196/jmir.9.2.e17] [Medline: 17581811]
123. Collins CE, Morgan PJ, Jones P, Fletcher K, Martin J, Aguiar EJ, et al. A 12-week commercial web-based weight-loss program for overweight and obese adults: randomized controlled trial comparing basic versus enhanced features. J Med Internet Res 2012;14(2):e57 [FREE Full text] [doi: 10.2196/jmir.1980] [Medline: 22555246]
124. Franko DL, Cousineau TM, Trant M, Green TC, Rancourt D, Thompson D, et al. Motivation, self-efficacy, physical activity and nutrition in college students: randomized controlled trial of an internet-based education program. Prev Med 2008 Oct;47(4):369-377 [FREE Full text] [doi: 10.1016/j.ypmed.2008.06.013] [Medline: 18639581]
125. Bennett JB, Broome KM, Schwab-Pilley A, Gilmore P. A web-based approach to address cardiovascular risks in managers: results of a randomized trial. J Occup Environ Med 2011 Aug;53(8):911-918 [FREE Full text] [doi: 10.1097/JOM.0b013e3182258bd8] [Medline: 21785368]
126. Moore TJ, Alsabeeh N, Apovian CM, Murphy MC, Coffman GA, Cullum-Dugan D, et al. Weight, blood pressure, and dietary benefits after 12 months of a web-based nutrition education program (DASH for health): longitudinal observational study. J Med Internet Res 2008;10(4):e52 [FREE Full text] [doi: 10.2196/jmir.1114] [Medline: 19073541]
127. Sacks FM, Svetkey LP, Vollmer WM, Appel LJ, Bray GA, Harsha D, DASH-Sodium Collaborative Research Group. Effects on blood pressure of reduced dietary sodium and the Dietary Approaches to Stop Hypertension (DASH) diet. DASH-Sodium Collaborative Research Group. N Engl J Med 2001 Jan 4;344(1):3-10. [doi: 10.1056/NEJM200101043440101] [Medline: 11136953]
128. Lancee J, van den Bout J, van Straten A, Spoormaker VI. Internet-delivered or mailed self-help treatment for insomnia?: a randomized waiting-list controlled trial. Behav Res Ther 2012 Jan;50(1):22-29. [doi: 10.1016/j.brat.2011.09.012] [Medline: 22055281]
129. Ritterband LM, Bailey ET, Thorndike FP, Lord HR, Farrell-Carnahan L, Baum LD. Initial evaluation of an Internet intervention to improve the sleep of cancer survivors with insomnia. Psychooncology 2012 Jul;21(7):695-705 [FREE Full text] [doi: 10.1002/pon.1969] [Medline: 21538678]
130. Ritterband LM, Thorndike FP, Gonder-Frederick LA, Magee JC, Bailey ET, Saylor DK, et al. Efficacy of an internet-based behavioral intervention for adults with insomnia. Arch Gen Psychiatry 2009 Jul;66(7):692-698 [FREE Full text] [doi: 10.1001/archgenpsychiatry.2009.66] [Medline: 19581560]
131. Espie CA, Kyle SD, Williams C, Ong JC, Douglas NJ, Hames P, et al. A randomized, placebo-controlled trial of online cognitive behavioral therapy for chronic insomnia disorder delivered via an automated media-rich web application. Sleep 2012;35(6):769-781. [doi: 10.5665/sleep.1872] [Medline: 22654196]
132. Vincent N, Lewycky S. Logging on for better sleep: RCT of the effectiveness of online treatment for insomnia. Sleep 2009 Jun;32(6):807-815 [FREE Full text] [Medline: 19544758]
133. Chiauzzi E, Pujol LA, Wood M, Bond K, Black R, Yiu E, et al. painACTION-back pain: a self-management website for people with chronic back pain. Pain Med 2010 Jul;11(7):1044-1058. [doi: 10.1111/j.1526-4637.2010.00879.x] [Medline: 20545873]
134. Bromberg J, Wood ME, Black RA, Surette DA, Zacharoff KL, Chiauzzi EJ. A randomized trial of a web-based intervention to improve migraine self-management and coping. Headache 2012 Feb;52(2):244-261 [FREE Full text] [Medline: 22413151]
135. Ruehlman LS, Karoly P, Enders C. A randomized controlled evaluation of an online chronic pain self management program. Pain 2012 Feb;153(2):319-330 [FREE Full text] [doi: 10.1016/j.pain.2011.10.025] [Medline: 22133450]
136. Nolan RP, Liu S, Shoemaker JK, Hachinski V, Lynn H, Mikulis DJ, et al. Therapeutic benefit of internet-based lifestyle counselling for hypertension. Can J Cardiol 2012 May;28(3):390-396. [doi: 10.1016/j.cjca.2012.02.012] [Medline: 22498181]

137. Sanders MR, Baker S, Turner KM. A randomized controlled trial evaluating the efficacy of Triple P Online with parents of children with early-onset conduct problems. Behav Res Ther 2012 Nov;50(11):675-684. [doi: 10.1016/j.brat.2012.07.004] [Medline: 22982082]
138. Ritterband LM, Cox DJ, Walker LS, Kovatchev B, McKnight L, Patel K, et al. An Internet intervention as adjunctive therapy for pediatric encopresis. J Consult Clin Psychol 2003 Oct;71(5):910-917. [doi: 10.1037/0022-006X.71.5.910] [Medline: 14516239]
139. Winzelberg AJ, Eppstein D, Eldredge KL, Wilfley D, Dasmahapatra R, Dev P, et al. Effectiveness of an Internet-based program for reducing risk factors for eating disorders. J Consult Clin Psychol 2000 Apr;68(2):346-350. [Medline: 10780136]
140. Jacobi C, Morris L, Beckers C, Bronisch-Holtze J, Winter J, Winzelberg AJ, et al. Maintenance of internet-based prevention: a randomized controlled trial. Int J Eat Disord 2007 Mar;40(2):114-119. [doi: 10.1002/eat.20344] [Medline: 17080447]
141. Jacobi C, Völker U, Trockel MT, Taylor CB. Effects of an Internet-based intervention for subthreshold eating disorders: a randomized controlled trial. Behav Res Ther 2012 Feb;50(2):93-99. [doi: 10.1016/j.brat.2011.09.013] [Medline: 22137366]
142. Roberto AJ, Zimmerman RS, Carlyle KE, Abner EL. A computer-based approach to preventing pregnancy, STD, and HIV in rural adolescents. J Health Commun 2007;12(1):53-76. [doi: 10.1080/10810730601096622] [Medline: 17365349]
143. Bowen AM, Horvath K, Williams ML. A randomized control trial of internet-delivered HIV prevention targeting rural MSM. Health Educ Res 2007 Feb;22(1):120-127 [FREE Full text] [doi: 10.1093/her/cyl057] [Medline: 16849391]
144. Swartz LH, Sherman CA, Harvey SM, Blanchard J, Vawter F, Gau J. Midlife women online: evaluation of an internet-based program to prevent unintended pregnancy & STIs. J Women Aging 2011;23(4):342-359. [doi: 10.1080/08952841.2011.613255] [Medline: 22014222]
145. Christensen H, Griffiths KM, Mackinnon AJ, Brittliffe K. Online randomized controlled trial of brief and full cognitive behaviour therapy for depression. Psychol Med 2006 Dec;36(12):1737-1746. [doi: 10.1017/S0033291706008695] [Medline: 16938144]
146. Internet Live Stats. Internet users in the world URL: http://www.internetlivestats.com/internet-users/ [accessed 2016-12-05] [WebCite Cache ID 6mXGZMx2A]
147. Stringhini S, Sabia S, Shipley M, Brunner E, Nabi H, Kivimaki M, et al. Association of socioeconomic position with health behaviors and mortality. JAMA 2010 Mar 24;303(12):1159-1166 [FREE Full text] [doi: 10.1001/jama.2010.297] [Medline: 20332401]
148. Wolpaw JR, Birbaumer N, Heetderks WJ, McFarland DJ, Peckham PH, Schalk G, et al. Brain-computer interface technology: a review of the first international meeting. IEEE Trans Rehabil Eng 2000 Jun;8(2):164-173. [Medline: 10896178]
149. Brodwin M, Star T, Cardoso E. Computer assistive technology for people who have disabilities: Computer adaptations and modifications. J Rehab 2004;70(3):28-33.

Abbreviations

CINAHL: Cumulative Index to Nursing and Allied Health Literature
DASH: Dietary Approaches to Stop Hypertension
NNT: number needed to treat
OCD: obsessive compulsive disorder
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PTSD: post-traumatic stress disorder
RCT: randomized controlled trial
TPOL: Triple P Online

Edited by A Keepanasseril; submitted 06.12.16; peer-reviewed by K Pal, E Murray; comments to author 14.01.17; revised version received 23.01.17; accepted 25.02.17; published 24.03.17.

Please cite as:
Rogers MAM, Lemmen K, Kramer R, Mann J, Chopra V
Internet-Delivered Health Interventions That Work: Systematic Review of Meta-Analyses and Evaluation of Website Availability
J Med Internet Res 2017;19(3):e90
URL: http://www.jmir.org/2017/3/e90/
doi:10.2196/jmir.7111
PMID:28341617

Original Paper

The Effect of Technology-Mediated Diabetes Prevention Interventions on Weight: A Meta-Analysis

Rachel R Bian[1*], BS; Gretchen A Piatt[2*], MPH, PhD; Ananda Sen[3,4*], PhD; Melissa A Plegue[3*], MA; Mariana L De Michele[5*], MD; Dina Hafez[5*], MD; Christina M Czuhajewski[3], MSI; Lorraine R Buis[3,6], PhD, MSI; Neal Kaufman[7*], MD, MPH; Caroline R Richardson[3*], MD

[1]University of Michigan Medical School, Ann Arbor, MI, United States

[2]Department of Learning Health Sciences, University of Michigan, Ann Arbor, MI, United States

[3]University of Michigan, Department of Family Medicine, Ann Arbor, MI, United States

[4]University of Michigan, Department of Biostatistics, School of Public Health, Ann Arbor, MI, United States

[5]University of Michigan, Department of Internal Medicine, Ann Arbor, MI, United States

[6]University of Michigan, School of Information, Ann Arbor, MI, United States

[7]Geffen School of Medicine and Fielding School of Public Health, Departments of Pediatrics and Public Health, University of California, Los Angeles, CA, United States

[*]these authors contributed equally

Corresponding Author:
Caroline R Richardson, MD
University of Michigan
Department of Family Medicine
1018 Fuller St.
Ann Arbor, MI, 48104
United States
Phone: 1 734 998 7120
Fax: 1 734 998 7335
Email: caroli@umich.edu

Abstract

Background: Lifestyle interventions targeting weight loss, such as those delivered through the Diabetes Prevention Program, reduce the risk of developing type 2 diabetes. Technology-mediated interventions may be an option to help overcome barriers to program delivery, and to disseminate diabetes prevention programs on a larger scale.

Objective: We conducted a meta-analysis to evaluate the effect of such technology-mediated interventions on weight loss.

Methods: In this meta-analysis, six databases were searched to identify studies reporting weight change that used technology to mediate diet and exercise interventions, and targeted individuals at high risk for developing type 2 diabetes. Studies published between January 1, 2002 and August 4, 2016 were included.

Results: The search identified 1196 citations. Of those, 15 studies met the inclusion criteria and evaluated 18 technology-mediated intervention arms delivered to a total of 2774 participants. Study duration ranged from 12 weeks to 2 years. A random-effects meta-analysis showed a pooled weight loss effect of 3.76 kilograms (95% CI 2.8-4.7; $P<.001$) for the interventions. Several studies also reported improved glycemic control following the intervention. The small sample sizes and heterogeneity of the trials precluded an evaluation of which technology-mediated intervention method was most efficacious.

Conclusions: Technology-mediated diabetes prevention programs can result in clinically significant amounts of weight loss, as well as improvements in glycaemia in patients with prediabetes. Due to their potential for large-scale implementation, these interventions will play an important role in the dissemination of diabetes prevention programs.

(J Med Internet Res 2017;19(3):e76) doi:10.2196/jmir.4709

KEYWORDS

diabetes mellitus, type 2; weight reduction programs; technology; meta-analysis; prediabetic state

Introduction

Over 29 million Americans (approximately 9% of the US population) have diabetes, and an additional 86 million Americans have prediabetes, an asymptomatic condition associated with an increased risk of developing type 2 diabetes [1]. In 2002, the landmark Diabetes Prevention Program (DPP) demonstrated that improved diet, regular exercise, and weight loss lead to a 58% reduction in the 3-year incidence of type 2 diabetes, compared to a placebo control arm, in patients at risk for the disease [2]. Importantly, this reduction was sustained over the long-term; specifically, 10-year follow-up analysis demonstrated the incidence of type 2 diabetes in the lifestyle intervention group was reduced by 34% compared with placebo [3].

The remarkable success of the DPP generated nationwide efforts to translate the results of the original trial intervention into practice [4,5]. The National Diabetes Prevention Program (NDPP) was established in 2010 as a congressionally-authorized initiative to support the dissemination of diabetes prevention programs across the United States [6]. NDPP-recognized diabetes prevention programs consist of 16 weekly sessions (core phase) followed by 6 monthly sessions (postcore phase) delivered by a trained lifestyle coach following the curriculum of the original DPP.

Systematic review of in-person group-based DPP programs demonstrates that they can effectively promote weight loss [7]. Although community-based DPP translations are more accessible, scalable, and financially sustainable than one-on-one interventions used during the original DPP study [4,6], there are barriers to participation. These barriers include transportation, distance, work schedules, aversion to group settings, and child care needs [8-11]. To address these issues, several pilot studies have used technology-mediated interventions to promote weight loss in participants at risk for type 2 diabetes. To date, a systematic review of these studies has not been performed.

The primary purpose of this meta-analysis is to assess the effect of technology-mediated lifestyle interventions on weight loss in those at risk for developing type 2 diabetes. We also discuss reported glycemic changes associated with these interventions.

Methods

Study Selection

We followed the Preferred Reporting Items for Systematic Reviews and Meta-Analyses statement guidelines for conducting and reporting this meta-analysis (Multimedia Appendix 1) [12]. We examined studies evaluating interventions that used technology to disseminate diet and exercise lifestyle programs, with the aim to achieve weight loss and improve glycemic control in adult patients with prediabetes. A systematic review was performed on the literature published between January 1, 2002 and August 4, 2016. Only studies published after the 2002 DPP study were included. We searched 6 databases to identify relevant studies, including PubMed, EMBASE, SportDiscus, CINAHL, PyschINFO, and Web of Science. Search terms to assess lifestyle intervention and use of technology were used, including the combination of MeSH and Emtree headings and subheadings, free-text keywords, and study design filters (eg, prediabetic state, weight loss, weight reduction programs, prediabet*, diabetes prevent*, telemedicine, telephone, web, technolog*, randomized controlled trial, and controlled clinical trial). We manually searched reference lists of review articles, and experts in the field were contacted to include all possible studies. Studies targeting individuals younger than 18 years of age, pregnant patients, or patients with a diagnosis of type 2 diabetes mellitus were excluded (see Multimedia Appendix 2 for search strategy).

Article titles and abstracts were screened to determine relevance and possible inclusion in the study. Full texts of the resulting articles were then read to determine eligibility based on the inclusion criteria. Authors were contacted directly to request missing weight change data, or clarify intervention methods and participant criteria when necessary.

Studies considered for inclusion had to satisfy three criteria. First, the primary objective of the study was to deliver diet and exercise lifestyle interventions using technology (digital versatile disc [DVD], computer-based program, phone, or text messaging) with the aim to achieve weight loss. Second, the study had to target patients with a diagnosis of prediabetes, or a body mass index (BMI) >24 kilograms per meter squared (kg/m^2; or >22 kg/m^2 if Asian) and at least one additional risk factor for diabetes (prior gestational diabetes, central adiposity, or metabolic syndrome); these criteria were based on those employed in the NDPP criteria [13]. Third, the study had to be either a randomized controlled trial (RCT) or prospective cohort study—with or without comparison groups—published in an academic journal and reported in English.

Outcomes

The primary outcome was absolute weight change following the intervention. Glycemic changes (as measured by changes in oral glucose tolerance test results, fasting blood glucose levels, hemoglobin A1c levels, prediabetes prevalence, or incidence of diabetes over the intervention period) were also reported, if available.

Studies that included a core intervention phase, as well as a postcore maintenance phase, had data extracted and used for this analysis directly after the core phase was completed. This approach served to reduce heterogeneity between studies that did and did not include a maintenance phase. We assessed absolute weight change effect of technology-mediated interventions, and compared results between those interventions modeled on the DPP with those using a different curriculum. We also examined the influence of intervention duration on weight change. Average percent weight change was also reported, if available. Program attrition was also assessed by comparing the number of enrolled participants with the number of program completers.

Data Analyses

Weight change outcomes in the core phase of each intervention were assessed using a meta-analysis. These outcomes were

either directly reported in the study results, calculated by determining the within-person difference between reported weights before and after the intervention, or obtained from the authors. The focus of our study was to observe the effect of technology-mediated interventions on weight change, and since the control groups across the different papers were significantly disparate, we excluded the data reported for control arms and only extracted data from the groups receiving an intervention delivered by technology. Each treatment group (or cohort) was analyzed as one pre/postintervention study. All weights were converted to kg units. Using the standard deviation (SD) of within-person weight change outcomes was necessary for calculating the relative study-influences for the meta-analysis. Studies that did not directly report this value had SDs calculated using either the *P*-value or the CI associated with average weight change.

Heterogeneity between studies was assessed by Cochran's Q statistic, which is the weighted sum of squared deviations of the study-specific estimates from the overall one, and is distributed approximately as a chi-squared random variable with k -1 degrees of freedom (k being the number of studies in the analysis) [14]. We further measured the I^2 statistic, defined as I^2=100%*(Q-k+1)/Q, that quantifies the proportion of heterogeneity in the trial results beyond chance. Higher I^2 is indicative of greater heterogeneity. Based on the observation of significant heterogeneity across studies, study estimates were pooled using a random-effects model that allowed for some random variability between studies, as well as sampling error. Publication bias between studies was assessed visually using a funnel plot. The visual effect is supplemented by the more formal Egger's test [15], which is essentially a significance test of intercept in a weighted least-squares fit of study-specific standardized effect on precision (reciprocal of standard error). A significant intercept is indicative of publication bias. Rosenthal's fail-safe N, based on the aggregated standardized effect size, was used to estimate the number of additional (potentially unpublished) studies required to convert a significant result into a nonsignificant one. A fail-safe N>5k+10 is indicative of lack of potential publication bias [16]. An influence analysis was carried out to assess the influence of each study by recalculating the pooled estimate after deleting the study.

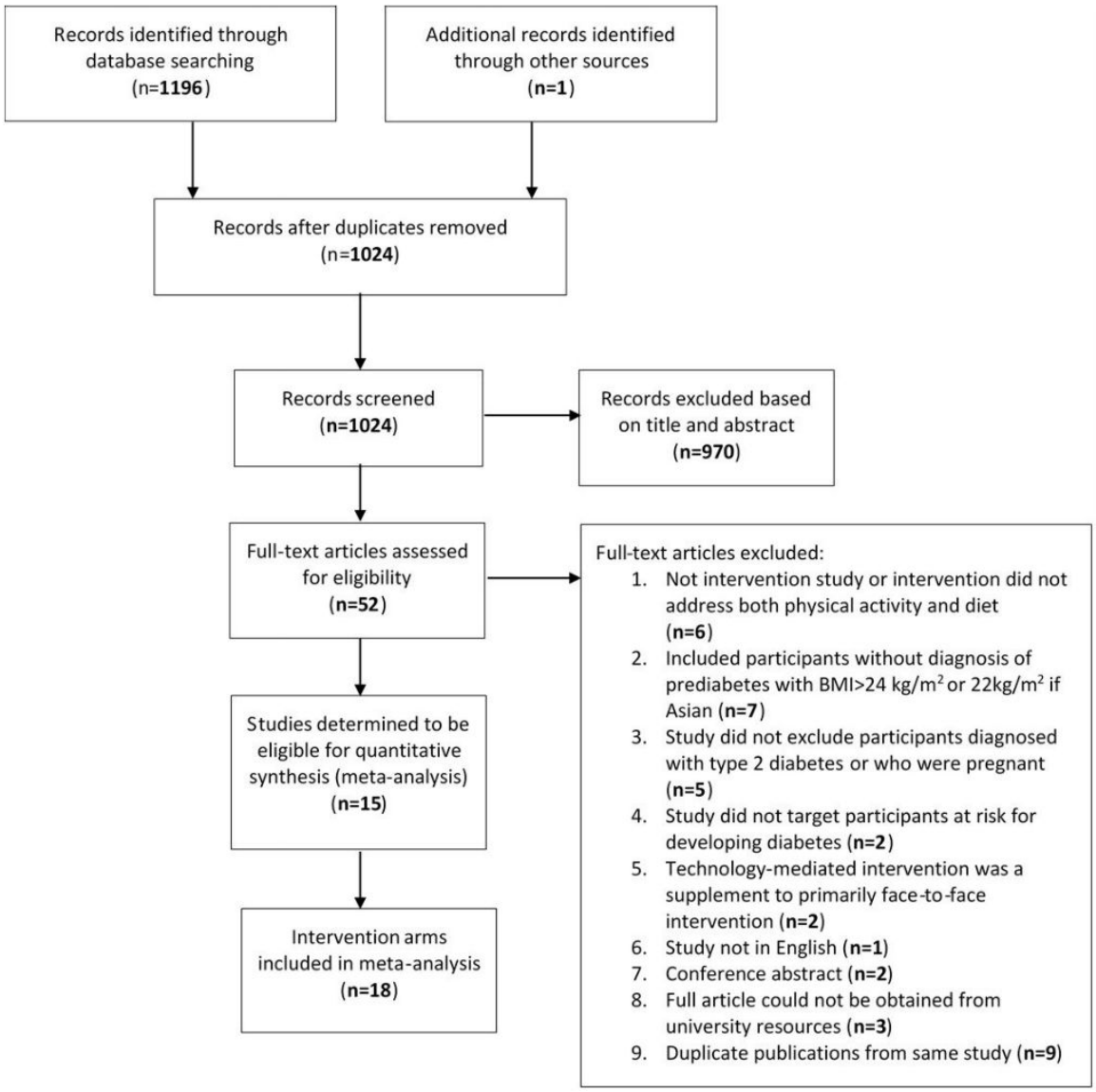
The included studies were of varied types—ranging from observational to blinded and unblinded randomized trials (with or without control)—making a fixed protocol for quality scoring impractical. Instead we performed an indirect quality-adjusted analysis via a meta-regression, adjusting for study duration, examining whether the analysis used intention-to-treat methods, and whether the intervention incorporated DPP material. Statistical analyses were carried out in STATA 13 (StataCorp LP, College Station, Texas) and R 3.3.0 (R Foundation for Statistical Computing, Vienna, Austria).

Results

Results of Systematic Literature Search

After eliminating duplicates, a total of 1024 publications were identified through the databases; one additional publication was identified from searching reference lists and through consultation with experts. Each publication was screened by title and abstract. The resulting 52 publications were reviewed in full, 37 of which were excluded because they did not meet inclusion criteria. The final systematic review included 15 publications that reported outcomes for a total of 18 intervention arms (Figure 1).

Figure 1. Flowchart showing results of systematic search strategy and selection process. BMI: body mass index.



Study Characteristics

Of the 15 studies included: 6 were RCTs in which the technology-mediated treatment arm was compared to standard care or a face-to-face intervention; 2 included control arms but assigned treatment conditions by community and without randomization; 3 randomly assigned participants to different technology-mediated intervention arms; 1 included parallel interventions arms assigned by community and without randomization; 2 were prospective cohort studies; and 1 allowed participants to self-select into the treatment arm. Multimedia Appendix 3 describes each cohort, and Table 1 shows key outcomes.

Table 1. Weight change outcomes. DVD: digital versatile disk; kg: kilogram; NA: not available or not applicable; SD: standard deviation.

Study Cohort (Year): Intervention	Attrition	Level of Utilization	Mean weight change, kg (SD)	% weight change (SD)
Aguiar et al (2016): DVD [17]	19%	NA	-4.98 (4.2)[a]	-4.85 (4.1)
Block et al (2015): Interactive voice response, email, text message, mobile app [18]	14%	After 6 months, intervention participants interacted with online program in a median of 17 of 24 weeks.	-3. 1 (3.6)[a]	-3.60 (NA)
Sakane et al (2015): Telephone [19]	18%	Mean responses to telephone calls during 1-year period: 2.8 (SD 0.6; Group A), 5.2 (SD 1.9; Group B), 8.2 (SD 3.5; Group C)	-1.1 (3.4)[a]	NA
Cha et al (2014): Internet and telephone [20]	13%	NA	-2.9 (4.3)	NA
Nicklas et al (2014): Internet [21]	11%	Median 9 videos watched	-2.6 (5.5)[a], at 6 months	NA
Sepah et al (2014): Internet [22]	15%	85% (187/220) completed at least 4 of the 16 core lessons	-5.0 (3.6)	-5.0 (NA)
Betzlbacher et al (2013): Telephone [23]	0%	All calls completed	-3.3 (4.3)	-3.7 (NA)
Ma et al (2013): DVD and email [24]	10%	NA	-4.5 (7.2)[a]	-4.9 (7.2)
Piatt et al (2013): DVD [25]	43%	Average attendance for DVD debriefing sessions 2.9/4	-5.5 (4.0)[a]	-5.7 (4.0)
Piatt et al (2013): Internet and e-counseling [25]	57%	Average 6.8 of 12 videos viewed	-6.2 (5.1)[a]	-6.3 (4.5)
Ramachandran et al (2013): Text message [26]	4%	Average number of text messages dropped from 18 to 12 messages a month	-0.1 (2.7)	NA
Weinstock et al (2013): Individual telephone [27]	43%	Average 9 of 16 sessions attended across 2 intervention cohorts	-4.6 (17.6)	-4.2 (16.9)
Weinstock et al (2013): Group telephone [27]	38%	Average 9 of 16 sessions attended across 2 intervention cohorts	-4.9 (17.7)	-4.5 (20.3)
Kramer et al (2010): DVD [28]	14%	Average 10.2 of 12 calls completed	-5.4 (5.2)[a]	-5.6 (NA)
Vadheim et al (2010): Video conference [29]	12%	Average 14.2 of 16 weeks of participation	-6.7 (3.7)	NA
Estabrooks and Smith-Ray (2008): Interactive voice response [30]	28%	10% did not complete any calls	-2.2 (2.7)	-2.6 (3.1)
Tate et al (2003): Internet only [31]	15%	NA	-2.0 (5.7)[a]	-2.2 (NA)
Tate et al (2003): Internet and e-counseling [31]	17%	NA	-4.4 (6.2)[a]	-4.8 (NA)

[a]results reported for intention-to-treat analysis

Participant Characteristics

A total of 2774 participants were enrolled in the technology-mediated interventions, of whom 2247 had follow-up data included in the final meta-analysis. Averaged across studies, enrolled participants were 49 years old and had a starting BMI of 29 (excluding 1 study that did not report baseline BMI [30]). Thirty-nine percent of the participants were female (excluding 1 study that did not report the gender composition of the intervention cohort [28]), and 34% were white (excluding 4 studies that did not report the number of white participants in the intervention cohorts [28,29]). Four studies were conducted outside of the United States [17,19,26,32] and 2 were undertaken in rural communities [25,29].

Treatment Characteristics

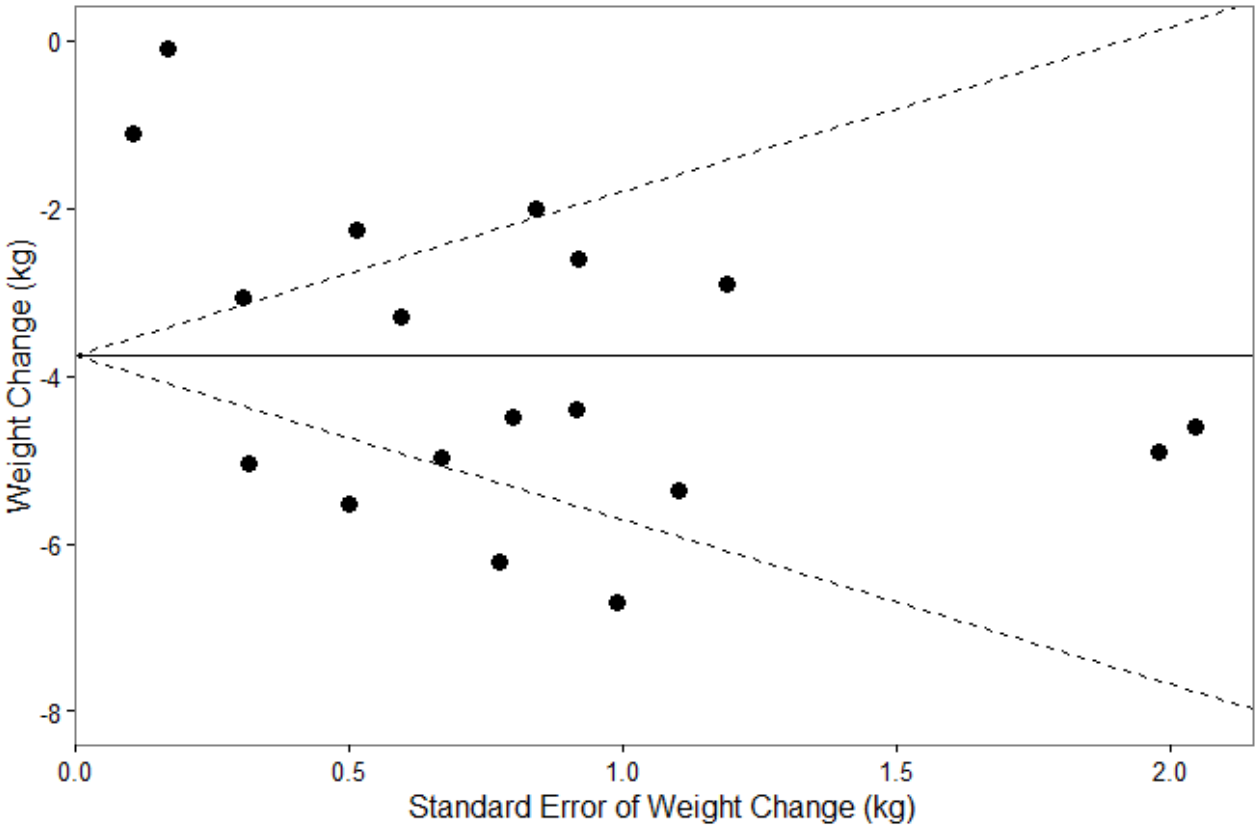
The duration of the interventions ranged from 12 weeks to 2 years. Half of the studies were modeled on the DPP [18,21,22,24,25,27-29]. The technologies employed by the 18 intervention arms included DVDs and e-videos [17,21,24,25,28], Web-based resources [18,20-22,31], videoconferencing [29], telephone (individual and conference calls) [19,20,25,27,28],

interactive voice response [18,30], text messages [26], e-counseling [21,24,25,31], email [18,24,25,31], and online group forums [22]. Supplementary print materials, such as diet and physical activity log books [17,19,23,25,28,29] and in-person group DPP [25,29], were also utilized. The lessons and messages delivered via the technology-enabled interventions centered on educating participants on how to achieve a healthy diet and exercise to reduce the risk of type 2 diabetes, and enabling behavioral changes through goal setting, self-monitoring, and logging of diet and physical activity. Video, text message, or Web-based lessons often introduced diet and physical activity concepts, while the personalized or automated phone, text message, and email messages would reinforce concepts, goals, and self-monitoring behavior.

Publication Bias

The funnel plot in Figure 2 was found to be statistically significant (P=.002) using Egger's test, indicating a potential for publication bias. However, Rosenthal's fail-safe N was above 4000 (more than 40 times the 5k+10 threshold) suggesting the potential threat from such bias to be quite small. For this reason, no corrective action was undertaken.

Figure 2. Funnel plot for publication bias analysis.



Change in Weight

Given significant heterogeneity across studies (I^2=96.1%, P<.001), a random-effects meta-analysis of the change in weight was performed utilizing 18 cohorts from the 15 studies. Figure 3 displays the change in weight and CIs for each intervention, stratified by whether the intervention was modeled from the DPP. Figure 3 also presents a pooled estimate of mean weight change using a random-effects model combining data from all 18 cohorts. These results demonstrate that technology-based interventions are effective at decreasing weight by an average of 3.76 kg (95% CI 2.8-4.7; P<.001). The DPP-based interventions resulted in marginally higher (P=.074) average weight loss (mean 4.81 kg, 95% CI 3.9-5.7) than non-DPP interventions (mean 2.44 kg, 95% CI 1.5-3.4). The contribution of each study to the overall effect ranged from 3.2-6.6%. Furthermore, the influence plot in Figure 4 indicates that no single point had exceptional influence, and the exclusion of any single cohort did not appreciably change the results of the overall estimate.

Figure 3. Forest plot of weight change from random-effects meta-analysis outcomes. DPP: Diabetes Prevention Program; DVD: digital versatile disk; IVR: interactive voice response.

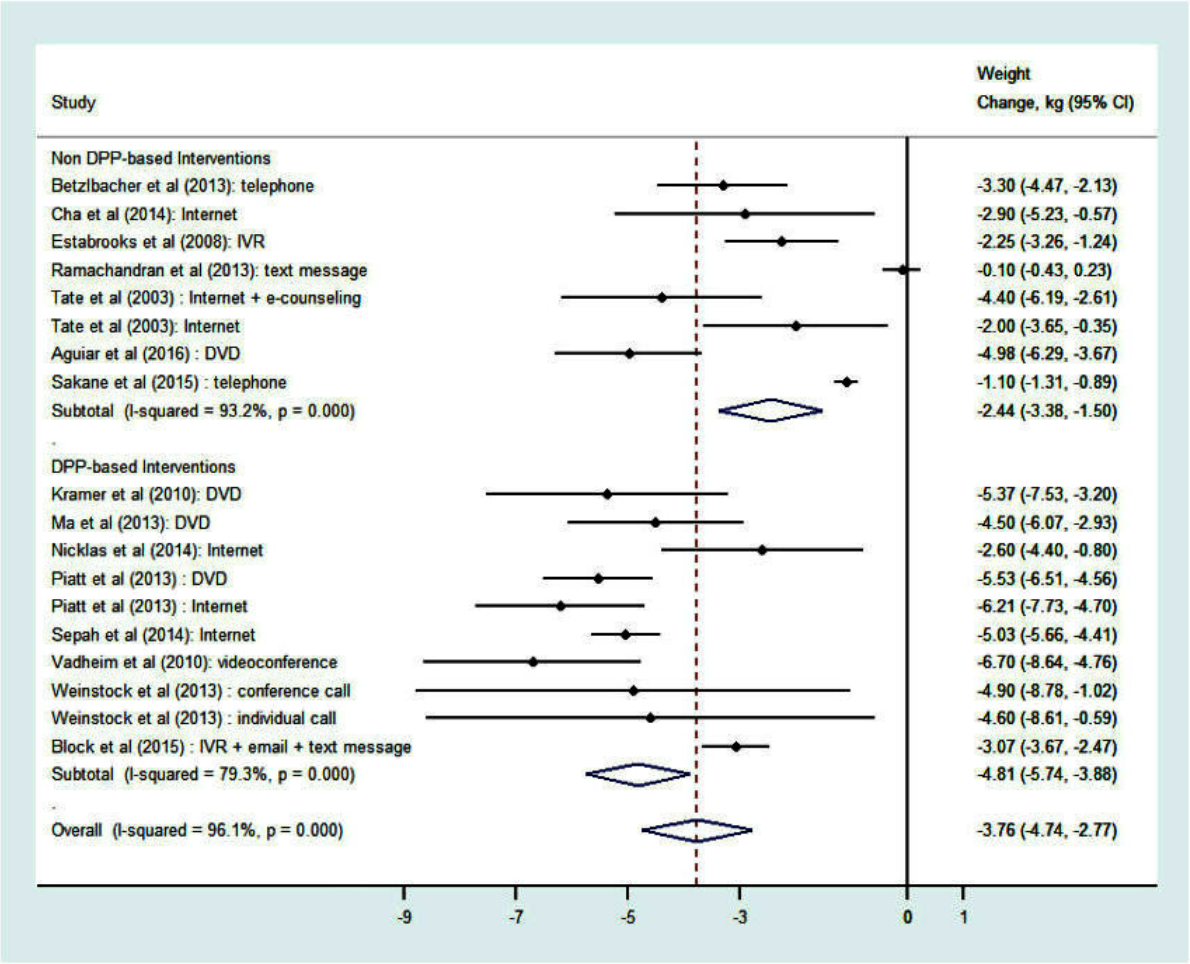
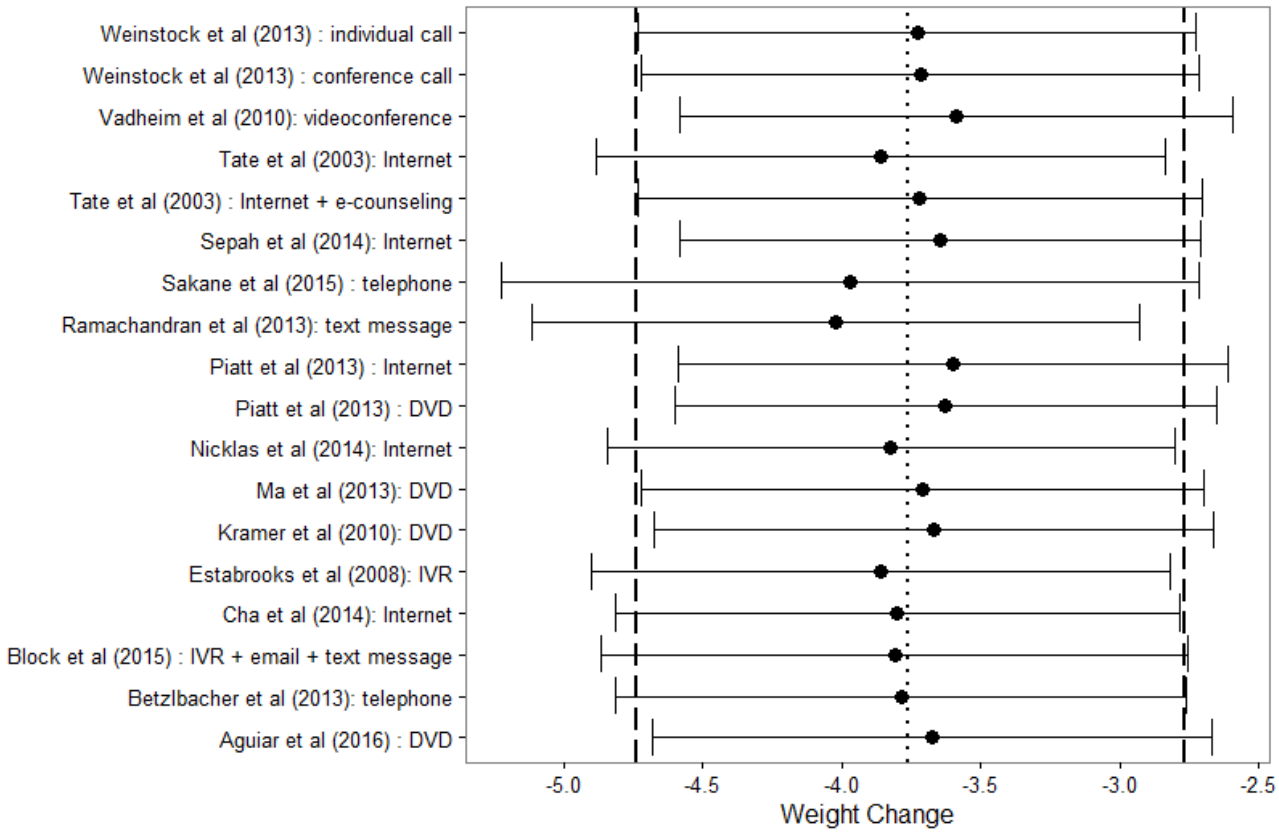


Figure 4. Influence plot of random-effects meta-analysis. DVD: digital versatile disk; IVR: interactive voice response.



Change in Glycaemia and Follow-Up Weight Outcomes

Eight studies reported a change in glycaemia over the course of the intervention [17,18,20,22-25,28]. Multimedia Appendix 4 shows the particular measure(s) used—oral glucose tolerance test, fasting blood glucose, and hemoglobin A1c—as well as baseline and postintervention mean values. Multimedia Appendix 4 also shows the reduction in prediabetes prevalence among participants. All 5 of the studies that reported baseline and postintervention prediabetes prevalence measurements reported decreases in prevalence [17,18,20,23,25]. Of the 4 cohorts reporting conversion from prediabetes to diabetes during the intervention period, rates ranged from 0-18% [21,23,24,26]. The 2 cohorts with the longest intervention durations reported the largest conversion rates [23,26].

Multimedia Appendix 5 presents change in weight measured in follow-up or maintenance periods, which were documented 8 to 36 months after the initial postcore intervention measurements. Follow-up or maintenance period durations varied by study, ranging from 12 to 48 months after baseline measurements. Some cohorts followed their core interventions with less intensive and/or optional maintenance interventions [21,22,24,27,33], while others included follow-up measurements without postcore maintenance interventions [23,25,34,35]. The relationship between follow-up time and weight change was mixed, with some shorter time frames observing a larger change, and some longer time frames observing a smaller change.

Discussion

This meta-analysis shows that technology-mediated interventions lead to clinically significant weight loss (mean 3.76 kg, 95% CI 2.8-4.7 kg; P<.001) in individuals at risk for diabetes. Additionally, as evidenced by our finding that 8 of 18 intervention arms [19,22,24,25,27,32-35] reported sustained weight loss outcomes at least one-year postintervention, the weight loss achieved through technology-mediated interventions may be sustainable. Moreover, the finding that several of the studies [17,18,20,23-25,28] reported improved glycaemia further supports the argument that these interventions are effective methods to prevent the development of diabetes.

Among American adults, there is widespread adoption of cell phones (92%) and smartphones (67%) [36]. These usage levels indicate that placing technology-mediated interventions in the hands of patients is becoming easier than ever. Moreover, from a clinical standpoint, we may soon live in a world where provider referrals to technology-mediated interventions to promote lifestyle and behavior change are commonplace. However, the marketplace is currently filled with a patchwork of technology-mediated solutions that vary widely in terms of quality, particularly in terms of health-related apps. This market is filled with products whose development often lacks professional content-expertise, theoretical underpinnings, and an appropriate evidence base to support use [37,38]. Given that we have demonstrated that quality technology-mediated interventions are effective at promoting and sustaining weight loss, more should be done to promote their use within clinical practice, so that consumers may succeed with high quality

applications, as opposed to fail with poorly conceptualized, designed, and implemented tools.

Two design components that were of particular interest in this meta-analysis were: (1) studies based on the DPP curriculum, and (2) study duration. While not quite achieving statistical significance, there was a strong trend toward more weight loss in the intervention arms that were based on the DPP curriculum. Studies investigating the efficacy of technology-based interventions often rely on short-term follow-up, which is commonly seen as a limitation of the existing body of literature. Our analysis shows no effect of intervention duration on weight loss outcomes, which is not consistent with previously reported literature [7]. It is possible that this finding is due to the heterogeneity of the technologies utilized, or the fact that some interventions included maintenance phase components while others did not.

Limitations

One limitation of our meta-analysis was the inability to compare intervention and control weight change results within the studies. This issue arose because only four studies were RCTs with a nontechnology control arm. To demonstrate whether technology-mediated interventions have comparable weight loss outcomes to in-person interventions, future studies are needed that directly compare these two cohorts. There was wide heterogeneity in intervention type, duration, population, and study attrition. Furthermore, none of the included studies had large sample sizes. Thus, conclusions could not be drawn about which method would be most efficacious. The variety of reported glycaemia changes limited our ability to perform a meta-analysis of these results. There was a fair amount of heterogeneity across the studies when examining study populations, as well as interventions used. This heterogeneity raises concerns for the statistical pooling of study results, but it strengthens the generalizability of the conclusions drawn from this systematic review, and provides important implications for the implementation of diabetes prevention programs.

Quality assessment was difficult, given the variety of study designs. Even in those studies that were RCTs, several standard quality criteria were not applicable. For example, complete blinding cannot be achieved within behavioral intervention studies. Instead, we chose three quality criteria that were relevant across studies (intention-to-treat analysis, intervention duration, and whether the study was modeled after the DPP) and included them as independent predictors in the meta-regression.

Future Directions

Those technology-mediated interventions that were modeled after the DPP tended to result in greater weight loss compared to the non-DPP modeled interventions, although this difference was not statistically significant at the .05 level. Further research is needed to test the hypothesis that weight loss outcomes can be optimized by incorporating the DPP curriculum.

Several studies investigated the effect of technology-mediated interventions on specific populations, such as younger populations or those in rural settings. Further investigation is needed using large methodologically sound comparative-effect research trials to determine which interventions are most efficacious in facilitating weight loss and glycemic improvement in specific demographic categories of participants at risk for developing diabetes. Such categories include specific age ranges, BMIs, and genders, as well as social, economic, and ethnic backgrounds. It has been shown that the success of a lifestyle intervention is largely affected by the participant's ability to choose the intervention modality [25], but factors that make an intervention more efficacious among target populations need to be explored, especially as technology allows intervention delivery to be individualized. Future studies should also evaluate whether technology should be coupled with some degree of in-person contact.

Studies are needed to examine how accessible these interventions are in low-income urban populations. For technology-mediated diabetes prevention interventions to expand the reach of diabetes prevention to a greater number of individuals at high risk for developing diabetes, research is needed with respect to intervention cost and payment models.

The inconsistency between using fasting blood glucose and hemoglobin A1c to measure metabolic control made it difficult to determine the extent to which technology-mediated interventions affected glycemic improvement. Weight loss was instead used as a direct health outcome measure. Some interventions may not result in weight loss but may improve glycemic control. Ideally, studies would include both weight loss and glycemic control (which would be measured using both fasting glucose and hemoglobin A1c) as a standardized evaluation methodology.

While a few of the studies in this analysis compared different forms of technology, further analysis is required to understand the advantages that each technology contributes to intervention outcome. In addition, the efficacy of maintenance phase interventions needs to be adequately assessed by comparing the weight loss outcomes and glycaemia of those who choose to stay in maintenance phase and those who decide to drop out. Understanding of such factors could guide the establishment of technology-mediated interventions as a potential correlate to the NDPP's current program.

Conclusion

Our meta-analysis showed that technology-mediated diabetes prevention interventions resulted in weight loss and lead to significant improvements in glycaemia. These results suggest that technology-mediated interventions could be an alternative to in-person diabetes prevention programs. The option of using technology-mediated delivery can potentially overcome barriers of access and allow expanded dissemination of such interventions.

Acknowledgments

Funding for this research was provided through the Michigan Diabetes Research Center (grant P30DK020572) as well as the Training Program in Endocrinology and Metabolism (grant T32DK007245) through the National Institutes of Health/National Institute of Diabetes, Digestive, and Kidney Disease. The funders had no role in study design, data collection, data analysis, decision to publish, or preparation of the manuscript.

Conflicts of Interest

NK is Chief Medical Officer of Canary Health (formerly DPS Health). None of the other authors declare any conflicts of interest. CRR has collaborated on research with Omada Health but has not received any financial compensation from this organization.

Multimedia Appendix 1

Preferred Reporting Items for Systematic Reviews and Meta-Analyses checklist.

[PDF File (Adobe PDF File), 31KB - jmir_v19i3e76_app1.pdf]

Multimedia Appendix 2

Search strategy.

[PDF File (Adobe PDF File), 29KB - jmir_v19i3e76_app2.pdf]

Multimedia Appendix 3

Published studies of technology-mediated diabetes prevention interventions.

[PDF File (Adobe PDF File), 69KB - jmir_v19i3e76_app3.pdf]

Multimedia Appendix 4

Change in glycaemia.

[PDF File (Adobe PDF File), 31KB - jmir_v19i3e76_app4.pdf]

Multimedia Appendix 5

Follow-up weight outcomes.

[PDF File (Adobe PDF File), 28KB - jmir_v19i3e76_app5.pdf]

References

1. Centers for Disease Control and Prevention. National Diabetes Statistics Report. Atlanta, GA: U.S. Department of Health and Human Services; 2014. Estimates of Diabetes and Its Burden in the United States URL: http://www.cdc.gov/diabetes/data/statistics/2014StatisticsReport.html [accessed 2015-05-04] [WebCite Cache ID 6YHHdxoRn]
2. Knowler WC, Barrett-Connor E, Fowler SE, Hamman RF, Lachin JM, Walker EA, Diabetes Prevention Program Research Group. Reduction in the incidence of type 2 diabetes with lifestyle intervention or metformin. N Engl J Med 2002 Feb 07;346(6):393-403 [FREE Full text] [doi: 10.1056/NEJMoa012512] [Medline: 11832527]
3. Diabetes Prevention Program Research Group, Knowler WC, Fowler SE, Hamman RF, Christophi CA, Hoffman HJ, et al. 10-year follow-up of diabetes incidence and weight loss in the Diabetes Prevention Program Outcomes Study. Lancet 2009 Nov 14;374(9702):1677-1686 [FREE Full text] [doi: 10.1016/S0140-6736(09)61457-4] [Medline: 19878986]
4. Vojta D, Koehler TB, Longjohn M, Lever JA, Caputo NF. A coordinated national model for diabetes prevention: linking health systems to an evidence-based community program. Am J Prev Med 2013 Apr;44(4 Suppl 4):S301-S306. [doi: 10.1016/j.amepre.2012.12.018] [Medline: 23498291]
5. Yoon U, Kwok LL, Magkidis A. Efficacy of lifestyle interventions in reducing diabetes incidence in patients with impaired glucose tolerance: a systematic review of randomized controlled trials. Metabolism 2013 Feb;62(2):303-314. [doi: 10.1016/j.metabol.2012.07.009] [Medline: 22959500]
6. Albright AL, Gregg EW. Preventing type 2 diabetes in communities across the U.S.: the National Diabetes Prevention Program. Am J Prev Med 2013 Apr;44(4 Suppl 4):S346-S351. [doi: 10.1016/j.amepre.2012.12.009] [Medline: 23498297]
7. Ali MK, Echouffo-Tcheugui J, Williamson DF. How effective were lifestyle interventions in real-world settings that were modeled on the Diabetes Prevention Program? Health Aff (Millwood) 2012 Jan;31(1):67-75 [FREE Full text] [doi: 10.1377/hlthaff.2011.1009] [Medline: 22232096]

8. Venditti EM, Wylie-Rosett J, Delahanty LM, Mele L, Hoskin MA, Edelstein SL, Diabetes Prevention Program Research Group. Short and long-term lifestyle coaching approaches used to address diverse participant barriers to weight loss and physical activity adherence. Int J Behav Nutr Phys Act 2014 Feb;11:16 [FREE Full text] [doi: 10.1186/1479-5868-11-16] [Medline: 24521153]
9. Laws RA, Fanaian M, Jayasinghe UW, McKenzie S, Passey M, Davies GP, et al. Factors influencing participation in a vascular disease prevention lifestyle program among participants in a cluster randomized trial. BMC Health Serv Res 2013 May;13:201 [FREE Full text] [doi: 10.1186/1472-6963-13-201] [Medline: 23725521]
10. Ackermann RT. Working with the YMCA to implement the Diabetes Prevention Program. Am J Prev Med 2013 Apr;44(4 Suppl 4):S352-S356. [doi: 10.1016/j.amepre.2012.12.010] [Medline: 23498298]
11. McTigue KM, Conroy MB, Hess R, Bryce CL, Fiorillo AB, Fischer GS, et al. Using the internet to translate an evidence-based lifestyle intervention into practice. Telemed J E Health 2009 Nov;15(9):851-858. [doi: 10.1089/tmj.2009.0036] [Medline: 19919191]
12. Moher D, Shamseer L, Clarke M, Ghersi D, Liberati A, Petticrew M, PRISMA-P Group. Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015 statement. Syst Rev 2015 Jan;4:1 [FREE Full text] [doi: 10.1186/2046-4053-4-1] [Medline: 25554246]
13. Centers for Disease Control and Prevention. Diabetes Prevention Recognition Program: Standards and Operating Procedures. 2015. URL: http://www.cdc.gov/diabetes/prevention/recognition [accessed 2015-05-04] [WebCite Cache ID 6YHIKBCjS]
14. Cochran WG. The combination of estimates from different experiments. Biometrics 1954 Mar;10(1):101. [doi: 10.2307/3001666]
15. Egger M, Davey Smith G, Schneider M, Minder C. Bias in meta-analysis detected by a simple, graphical test. BMJ 1997 Sep 13;315(7109):629-634 [FREE Full text] [Medline: 9310563]
16. Rosenthal R. Combining results of independent studies. Psychol Bull 1978;85(1):185-193.
17. Aguiar EJ, Morgan PJ, Collins CE, Plotnikoff RC, Young MD, Callister R. Efficacy of the Type 2 Diabetes Prevention Using LifeStyle Education Program RCT. Am J Prev Med 2016 Mar;50(3):353-364. [doi: 10.1016/j.amepre.2015.08.020] [Medline: 26526160]
18. Block G, Azar KM, Romanelli RJ, Block TJ, Hopkins D, Carpenter HA, et al. Diabetes prevention and weight loss with a fully automated behavioral intervention by email, web, and mobile phone: a randomized controlled trial among persons with prediabetes. J Med Internet Res 2015 Oct 23;17(10):e240 [FREE Full text] [doi: 10.2196/jmir.4897] [Medline: 26499966]
19. Sakane N, Kotani K, Takahashi K, Sano Y, Tsuzaki K, Okazaki K, et al. Effects of telephone-delivered lifestyle support on the development of diabetes in participants at high risk of type 2 diabetes: J-DOIT1, a pragmatic cluster randomised trial. BMJ Open 2015 Aug 19;5(8):e007316 [FREE Full text] [doi: 10.1136/bmjopen-2014-007316] [Medline: 26289448]
20. Cha E, Kim KH, Umpierrez G, Dawkins CR, Bello MK, Lerner HM, et al. A feasibility study to develop a diabetes prevention program for young adults with prediabetes by using digital platforms and a handheld device. Diabetes Educ 2014 Jun;40(5):626-637. [doi: 10.1177/0145721714539736] [Medline: 24950683]
21. Nicklas JM, Zera CA, England LJ, Rosner BA, Horton E, Levkoff SE, et al. A web-based lifestyle intervention for women with recent gestational diabetes mellitus: a randomized controlled trial. Obstet Gynecol 2014 Sep;124(3):563-570. [doi: 10.1097/AOG.0000000000000420] [Medline: 25162257]
22. Sepah SC, Jiang L, Peters AL. Translating the Diabetes Prevention Program into an online social network: validation against CDC standards. Diabetes Educ 2014 Jul;40(4):435-443. [doi: 10.1177/0145721714531339] [Medline: 24723130]
23. Betzlbacher A, Grady K, Savas L, Cotterill S, Boaden R, Summers L, et al. Behaviour change among people with impaired glucose tolerance: comparison of telephone-based and face-to-face advice. J Health Serv Res Policy 2013 Apr;18(1 Suppl):2-6. [doi: 10.1177/1355819612473582] [Medline: 27552773]
24. Ma J, Yank V, Xiao L, Lavori PW, Wilson SR, Rosas LG, et al. Translating the Diabetes Prevention Program lifestyle intervention for weight loss into primary care: a randomized trial. JAMA Intern Med 2013 Jan 28;173(2):113-121 [FREE Full text] [doi: 10.1001/2013.jamainternmed.987] [Medline: 23229846]
25. Piatt GA, Seidel MC, Powell RO, Zgibor JC. Comparative effectiveness of lifestyle intervention efforts in the community: results of the Rethinking Eating and ACTivity (REACT) study. Diabetes Care 2013 Feb;36(2):202-209 [FREE Full text] [doi: 10.2337/dc12-0824] [Medline: 22966092]
26. Ramachandran A, Snehalatha C, Ram J, Selvam S, Simon M, Nanditha A, et al. Effectiveness of mobile phone messaging in prevention of type 2 diabetes by lifestyle modification in men in India: a prospective, parallel-group, randomised controlled trial. Lancet Diabetes Endocrinol 2013 Nov;1(3):191-198. [doi: 10.1016/S2213-8587(13)70067-6] [Medline: 24622367]
27. Weinstock RS, Trief PM, Cibula D, Morin PC, Delahanty LM. Weight loss success in metabolic syndrome by telephone interventions: results from the SHINE Study. J Gen Intern Med 2013 Dec;28(12):1620-1628 [FREE Full text] [doi: 10.1007/s11606-013-2529-7] [Medline: 23843020]
28. Kramer MK, Kriska AM, Venditti EM, Semler LN, Miller RG, McDonald T, et al. A novel approach to diabetes prevention: evaluation of the Group Lifestyle Balance program delivered via DVD. Diabetes Res Clin Pract 2010 Dec;90(3):e60-e63. [doi: 10.1016/j.diabres.2010.08.013] [Medline: 20863586]

29. Vadheim LM, McPherson C, Kassner DR, Vanderwood KK, Hall TO, Butcher MK, et al. Adapted diabetes prevention program lifestyle intervention can be effectively delivered through telehealth. Diabetes Educ 2010 Aug;36(4):651-656. [doi: 10.1177/0145721710372811] [Medline: 20534873]
30. Estabrooks PA, Smith-Ray RL. Piloting a behavioral intervention delivered through interactive voice response telephone messages to promote weight loss in a pre-diabetic population. Patient Educ Couns 2008 Jul;72(1):34-41. [doi: 10.1016/j.pec.2008.01.007] [Medline: 18282679]
31. Tate DF, Jackvony EH, Wing RR. Effects of Internet behavioral counseling on weight loss in adults at risk for type 2 diabetes: a randomized trial. JAMA 2003 Apr 9;289(14):1833-1836. [doi: 10.1001/jama.289.14.1833] [Medline: 12684363]
32. Begg CB, Berlin JA. Publication bias and dissemination of clinical research. J Natl Cancer Inst 1989 Jan 18;81(2):107-115. [Medline: 2642556]
33. Trief PM, Weinstock RS, Cibula D, Delahanty LM. Sustained weight loss one year after group telephone intervention: 3-year results from the SHINE study. Diabetes Res Clin Pract 2014 Dec;106(3):e74-e78. [doi: 10.1016/j.diabres.2014.09.032] [Medline: 25451907]
34. Savas LA, Grady K, Cotterill S, Summers L, Boaden R, Gibson JM. Prioritising prevention: implementation of IGT Care Call, a telephone based service for people at risk of developing type 2 diabetes. Prim Care Diabetes 2015 Feb;9(1):3-8. [doi: 10.1016/j.pcd.2014.07.003] [Medline: 25175562]
35. Piatt GA, Seidel MC, Powell RO, Zgibor JC. Influence of patient-centered decision making on sustained weight loss and risk reduction following lifestyle intervention efforts in rural Pennsylvania. Diabetes Educ 2016 Jun;42(3):281-290. [doi: 10.1177/0145721716636962] [Medline: 26957534]
36. Rainie L, Zickuhr K. Pew Research Center. 2015. Americans' views on mobile etiquette URL: http://www.pewinternet.org/2015/08/26/americans-views-on-mobile-etiquette/ [accessed 2016-11-21] [WebCite Cache ID 6mBqmZyNF]
37. Subhi Y, Bube SH, Rolskov BS, Skou Thomsen AS, Konge L. Expert involvement and adherence to medical evidence in medical mobile phone apps: a systematic review. JMIR Mhealth Uhealth 2015 Jul 27;3(3):e79 [FREE Full text] [doi: 10.2196/mhealth.4169] [Medline: 26215371]
38. Nikolaou CK, Lean ME. Mobile applications for obesity and weight management: current market characteristics. Int J Obes (Lond) 2017 Jan;41(1):200-202. [doi: 10.1038/ijo.2016.186] [Medline: 27780974]

Abbreviations

BMI: body mass index
DPP: Diabetes Prevention Program
DVD: digital versatile disc
k: number of studies in the analysis
kg: kilogram
NDPP: National Diabetes Prevention Program
RCT: randomized controlled trial
SD: standard deviation

Edited by G Eysenbach; submitted 19.05.15; peer-reviewed by J Jones, A Venkatakrishnan; comments to author 19.08.15; revised version received 02.12.16; accepted 30.01.17; published 27.03.17.

Please cite as:
Bian RR, Piatt GA, Sen A, Plegue MA, De Michele ML, Hafez D, Czuhajewski CM, Buis LR, Kaufman N, Richardson CR
The Effect of Technology-Mediated Diabetes Prevention Interventions on Weight: A Meta-Analysis
J Med Internet Res 2017;19(3):e76
URL: http://www.jmir.org/2017/3/e76/
doi:10.2196/jmir.4709
PMID:28347972

Original Paper

Effectiveness of a Web-Based Health Education Program to Promote Oral Hygiene Care Among Stroke Survivors: Randomized Controlled Trial

Normaliza Ab Malik[1,2], BDS, MClinDent; Sa'ari Mohamad Yatim[3], MD, MRehab Med; Otto Lok Tao Lam[4], BSc, BDS, MEd, PhD; Lijian Jin[1], DDS, MMedSc, PhD, OdontDr, FDSRCS (Edin) ad hominem; Colman Patrick Joseph McGrath[1], BA, BDentSc (Hons), MEd, FDSRCS, DDPHRCS, MSc, FFDRCS, PhD

[1]Periodontology and Dental Public Health, The University of Hong Kong, Hong Kong SAR, China

[2]Faculty of Dentistry, Universiti Sains Islam Malaysia (USIM), Kuala Lumpur, Malaysia

[3]Hospital Serdang, Department of Rehabilitation Medicine, Selangor, Malaysia

[4]Department of Oral Rehabilitation, The University of Hong Kong, Hong Kong SAR, China

Corresponding Author:
Colman Patrick Joseph McGrath, BA, BDentSc (Hons), MEd, FDSRCS, DDPHRCS, MSc, FFDRCS, PhD
Periodontology and Dental Public Health
The University of Hong Kong
34 Hospital Road
Hong Kong SAR,
China
Phone: 86 28590513
Fax: 86 28587874
Email: mcgrathc@hku.hk

Abstract

Background: Oral hygiene care is of key importance among stroke patients to prevent complications that may compromise rehabilitation or potentially give rise to life-threatening infections such as aspiration pneumonia.

Objective: The aim of this study was to evaluate the effectiveness of a Web-based continuing professional development (CPD) program on "general intention" of the health carers to perform daily mouth cleaning for stroke patients using the theory of planned behavior (TPB).

Methods: A double-blind cluster randomized controlled trial was conducted among 547 stroke care providers across 10 hospitals in Malaysia. The centers were block randomized to receive either (1) test intervention (a Web-based CPD program on providing oral hygiene care to stroke patients using TPB) or (2) control intervention (a Web-based CPD program not specific to oral hygiene). Domains of TPB: "attitude," "subjective norm" (SN), "perceived behavior control" (PBC), "general intention" (GI), and "knowledge" related to providing oral hygiene care were assessed preintervention and at 1 month and 6 months postintervention.

Results: The overall response rate was 68.2% (373/547). At 1 month, between the test and control groups, there was a significant difference in changes in scores of attitude (P=.004) and subjective norm (P=.01), but not in other TPB domains (GI, P=.11; PBC, P=.51; or knowledge, P=.08). At 6 months, there were significant differences in changes in scores of GI (P=.003), attitude (P=.009), SN (P<.001) and knowledge (P=.001) between the test and control groups. Regression analyses identified that the key factors associated with a change in GI at 6 months were changes in SN (beta=.36, P<.001) and changes in PBC (beta=.23, P<.001).

Conclusions: The Web-based CPD program based on TPB increased general intention, attitudes, subjective norms, and knowledge to provide oral hygiene care among stroke carers for their patients. Changing subjective norms and perceived behavioral control are key factors associated with changes in general intention to provide oral hygiene care.

Trial Registration: National Medical Research Register, Malaysia NMRR-13-1540-18833 (IIR); https://www.nmrr.gov.my/fwbLoginPage.jsp

(J Med Internet Res 2017;19(3):e87) doi:10.2196/jmir.7024

KEYWORDS

oral hygiene; computer-aided learning; cerebrovascular accident; theory of planned behavior; health care providers, Internet

Introduction

Provision of oral hygiene care is often underemphasized and underpracticed in the acute hospital setting [1]. This is despite growing acceptance of the importance of oral hygiene to general health, because of its potential link with bacteremia and aspiration pneumonia [2,3]. For stroke patients in the acute hospital setting, it is recognized that oral hygiene care is of key importance to prevent complications that may compromise rehabilitation or potentially give rise to a recurrent stroke [4]. There is a growing interest in how to effectively increase the practice of providing oral hygiene care in the hospital setting through clinical interventions [5,6] and through education and training of caregivers [7,8]. Unfortunately, however, all this has met with limited success and a persistence of poor knowledge and attitudes toward providing oral hygiene care in the acute hospital setting typically prevails [9].

Despite the acknowledged importance of dental education and oral health promotion activities, the effectiveness to change practices with respect to oral hygiene care has met with limited success among carers and patients [10,11]. To this end, the need to plan and implement oral health promotion programs based on psychological models has advocated the need to "translate theory into practice" [12]. One of the most widely used theories (model) is the theory of planned behavior (TPB), which emphasizes the importance of changing the "general intention"' (GI) to perform a health behavior and its relationship to attitude (positive or negative views of a behavior), subjective norm (SN, perceived of social pressure to perform a behavior) and perceived behavior control (PBC, one's control to perform a behavior) [13,14].

Providing dental health education and oral health promotion through continuing professional development (CPD) programs is an important and practical way to promote oral hygiene care practices in hospital settings [15]. The use of Web-based and computer-aided learning (CAL) has been widely used in CPD programs for health carers, owing to its ability to implement programs across wide geographical areas at relatively low costs and because of the reported effectiveness of such programs in changing health care practices [16-20] and health behavior [21]. Increasingly, Web-based and CAL programs are being used to enhance oral hygiene care in hospital and other institutionalized setting, and there are several reports of their ability to positively bring about change in knowledge [22], attitudes [18], and oral health behavior [23].

We aimed to evaluate the effectiveness of a Web-based (CPD) program to change the GI of health carers to perform daily mouth cleaning for stroke patients using TPB in a large randomized controlled trial across Malaysia. In addition, the study aimed to identify key factors associated with changes in GI among health care workers to provide oral hygiene care to stroke patients.

Methods

Study Design and Sample

This study was a double-blind, cluster-randomized, controlled trial with 1 month and 6 months follow-ups. The study involved 10 public hospitals in Malaysia, which have participated in a survey of oral hygiene practice for stroke patients. These hospitals were selected because they provide rehabilitation services that are led by rehabilitation medicine specialists. Hospitals were first stratified by size into either large, medium, or small in terms of number of health care providers. From each stratified group, hospitals were block-randomized in groups of 4 ("ABBA") by a computer-generated randomization method. In total, 5 hospitals were assigned to the test group (277 registered nurses) and 5 hospitals were assigned to the control group (270 registered nurses). The allocation sequence was concealed from the investigator coordinating the trial (who had contact with the centers). Through concealment, the assessor was "blind" as to what group participants had been assigned and participants were also blind as to what groups they were assigned to, as both received a form of Web-based CPD.

The study population was registered nurses caring for stroke patients at the hospitals, mainly from the rehabilitation and general medical wards. All the registered nurses from these identified wards were invited to take part in this trial. Information sheet related to the study and written informed consent were given to all the nurses before commencing the study. The forms were distributed to the nurses by the ward managers or chief nurses. Nurses who provided their written consent were those who participated in the trial from the study population. Participation was voluntary and no contact was made with the nurses to ensure confidentiality and reduce the potential for "social bias."

Ethics Approval

This clinical trial was registered with the National Institutes of Health, Ministry of Health, Malaysia; NMRR-13-1540-18833(IIR). Before the commencement of the study, ethical approval was obtained from the Institute for Health Behavioral Research and Medical Research and Ethics Committee of the National Institutes of Health, Ministry of Malaysia. Permissions to conduct the study were also obtained from the directors of the respective hospitals. Recruitment and baseline assessments were from September 2014 to November 2014 at 10 hospitals across the country (both the Peninsular Malaysia and island of Borneo Malaysia). This study followed CONSORT guidelines.

Data Collection

The nurses self-completed a questionnaire on the practice of providing oral hygiene care to stroke patients, which contained 12 items specific to attitudes, SN, PBC, and GI to providing oral hygiene care related to TPB. These items were derived from the manual of "Constructing Questionnaires Based on the Theory of Planned Behavior" developed by the Centre of Health

Services Research, University of New Castle, UK (2004) [24]. Items related to direct measure of the domains were chosen and modified to the oral health context. For example, in the GI domain, "I expect to measure the blood pressure of my patients with diabetes in each consultation" was modified to "I expect to perform oral care (including denture) for patients in every session." Each domain had 3 items that were rated on a 5-point Likert scale (strongly disagree, disagree, not disagree or agree, agree, and strongly agree). Domain scores can range from 3 to 15, with higher scores reflecting more positive attitude, stronger subjective norms, greater perceived behavior control, and greater general intention to provide oral hygiene care. Sociodemographic and environmental characteristic (eg, attended oral care training, availability of oral health guidelines and oral hygiene kits, and having dental professional support in the ward) information were also obtained from the participants.

In addition, knowledge of oral health care was assessed using 5 items related to dental plaque, gum bleeding, consequences of dental plaque, how to prevent gingivitis, and how oral health affects general health [25]. Knowledge scores can range from 0 to 5, with higher scores indicative of greater oral health knowledge. Assessments were carried out preintervention and at 1 month and 6 months postintervention.

Intervention

A Web-based CPD program was developed for the test and control groups. The test group program was specific to provision of oral hygiene care to stroke patients and covered details of oral health knowledge, attitudes, subjective norms, means of behavioral control, and intention (ie, based on TPB). The test group contents include, for example, information on good oral condition and the importance of having good oral health, the consequences of poor oral hygiene, and the importance of nurse's roles and care of stroke patients. The development of the contents was guided by the definition of the TPB domains and scope of the study. The control group received an analogous Web-based CPD program related to "bundles of care" for stroke patients that included some details on oral hygiene care but not specific to TPB [26]. The CPD programs were developed by stroke physicians (rehabilitation medicine) and dentists and followed good practices of CAL for oral health [27]. Following the assignment to the groups, the participants were provided with details of the Web-based programs through a secure internet portal. Participants were reminded and encouraged to complete the Web-based CPD program every 6 weeks.

Sample Size

With the assumption that this practice is at 50% and that it will not change without education intervention, whereas there will be a 25% improvement in practices following CAL intervention (ie, 63% of nurses will practice oral care in rehabilitation). Then a proposed sample size of 247 in each group is required with sample power at 80%. Allowing for nonparticipation and a dropout rate of ~20%, thus it was prudent to attempt to recruit over 600 nurses (300 per group) in total to test the hypothesis.

Data Analysis

The changes in knowledge, attitudes, SN, PBC, and GI were determined overtime and compared between the test and control groups using Friedman two-way analysis of variance (ANOVA) and Mann-Whitney *U* test analysis, respectively. Multiple linear regression analyses were performed to determine key factors associated with changes in GI to provide oral hygiene care at 1 month and 6 months.

Results

The response rate of the trial was 68.2% (373/547); mostly loss to follow-up was because nurses were transferred to other wards or hospitals (Figure 1). The response rate among the test group was 70.4% (195/277) and among the control group was 65.0% (178/270); there was no significant difference between the response rate among those in the test and control groups ($P>.05$). The majority of nurses were female (95.7%, 357/373), had a certificate or diploma in nursing (81.5%, 304/373), worked in general medical wards (78.6%, 293/373), and reported to have worked for less than 5 years (59.0%, 220/373; Table 1).

Table 1. Health care provider and environmental characteristics (n=373)

Characteristics		n (%)
Provider characteristics		
Gender		
	Male	16 (4.3)
	Female	357 (95.7)
Years worked		
	Less than 5 years	220 (59.0)
	More than 5 years	153 (41.0)
Qualification		
	Certificate or diploma	304 (81.5)
	Post basic or degree	69 (18.5)
Working wards		
	Rehabilitation ward	80 (21.4)
	Medical ward	293 (78.6)
Environmental		
Oral care training		
	Yes	108 (29.0)
	No	265 (71.0)
Oral health care guidelines		
	Yes	296 (79.4)
	No	77 (20.6)
Oral hygiene kit		
	Yes	253 (67.8)
	No	120 (32.2)
Dental professional support		
	Yes	52 (13.9)
	No	321 (86.1)

Among all participants, there was a significant difference in knowledge scores over time (P<.05; Table 2). There was a significant improvement in knowledge scores between baseline and 1 month (P<.01), but no significant change between baseline and 6 months (P>.05), and between 1 month and 6 months (P>.05).

Table 2. Changes in knowledge and theory of planned behavior domains scores over time.

Time	General intention Mean (SD)	Attitudes Mean (SD)	Subjective norm Mean (SD)	Perceived behavior control Mean (SD)	Knowledge Mean (SD)
Baseline	10.8 (2.1)	13.1 (1.7)	9.7 (1.7)	10.5 (1.7)	3.0 (1.0)
1 month	11.0 (1.9)	12.9 (1.6)	9.9 (1.7)	10.3 (1.9)	3.2 (1.1)
6 months	10.7 (2.2)	13.0 (1.6)	9.7 (1.8)	10.3 (1.9)	3.2 (1.1)
P value	.36	.11	.05	.28	.005[a]

[a]Baseline<1 month, P=.02.

In the test group, there was a significant difference in knowledge scores over time (P<.001; Table 3). There was a significant improvement in knowledge scores between baseline and 1 month (P<.01) and between baseline and 6 months (P<.01), but no significant change between 1 month and 6 months (P>.05). There was a significant change in GI scores between baseline and 1 month (P<.05). A significant improvement in SN scores was observed between baseline and 1 month (P<.05) and between baseline and 6 months (P<.01), but no significant change between 1 month and 6 months (P>.05). However, no

significant changes over time were observed in attitude scores (P>.05) and PBC scores (P>.05).

Table 3. Test and control group changes in knowledge and theory of planned behavior (TPB) domain scores overtime.

Group		General intention Mean (SD)	Attitude Mean (SD)	Subjective norm Mean (SD)	Perceived behavior control Mean (SD)	Knowledge Mean (SD)
Test group						
	Baseline	10.6 (2.0)	12.9 (1.6)	9.5 (1.7)	10.5 (1.7)	2.9 (1.0)
	1 month	11.0 (2.0)	12.9 (1.8)	10.0 (1.8)	10.2 (2.1)	3.3 (1.1)
	6 months	10.9 (1.9)	13.2 (1.6)	10.1 (1.5)	10.2 (1.9)	3.3 (1.1)
	P value	.045[a]	.43	.001[b]	.42	<.001[c]
Control group						
	Baseline	11.0 (2.1)	13.3 (1.8)	9.8 (1.8)	10.6(1.7)	3.0 (1.0)
	1 month	11.0 (1.9)	12.8 (1.5)	9.7 (1.6)	10.5 (1.6)	3.1 (1.0)
	6 months	10.6 (2.5)	12.9(1.7)	9.3 (2.0)	10.4 (1.8)	3.0 (1.1)
	P value	.03	.002[d]	.22	.57	.16

[a]Baseline<1 month, P=.03.

[b]Baseline<1 month, P=.02; Baseline<6 months, P=.004.

[c]Baseline<1 month, P=.009; Baseline<6 months, P=.003.

[d]Baseline>1 month, P=.01.

Among the control group (Table 3), there was a significant change in GI scores (P<.05) and attitude scores (P<.01), no significant changes between baseline, 1 month, and 6 months, and significant changes between baseline and 1 month with lower score at 1 month (P<.05), respectively. No significant changes over time were observed in SN scores (P>.05), PBC scores (P>.05), and knowledge scores (P>.05) among the control group.

At 1 month, between the test and control groups, there were significant differences in the change (Δ) of attitude scores (P<.01) and SN scores (P<.05), but not in other TPB domains (P>.05) nor in knowledge scores (P>.05; Table 4). At 6 months, between the test and control group, there were significant differences in the change (Δ) of GI scores (P<.01), attitude scores (P<.01), and SN scores (P<.001), but not in PBC (P>.05). In addition, there were significant differences in the change of knowledge scores (P<.01).

Table 4. Changes in knowledge and theory of planned behavior domain between test and control groups from baseline to 1 month and baseline to 6 months.

Time		ΔGI Mean (SD)	ΔAttitude Mean (SD)	ΔSN Mean (SD)	ΔPBC Mean (SD)	ΔKnowledge Mean (SD)
1 month						
	Test	.42 (2.57)	.01 (2.23)	.45 (2.18)	−.33 (2.68)	.39 (1.37)
	Control	.02 (2.61)	−.47 (2.21)	−.13 (2.45)	−.07 (2.11)	.08 (1.46)
	P value	.11[a]	.004[a]	.01[a]	.51[a]	.08[a]
6 months						
	Test	.29 (2.45)	.24 (2.01)	.54 (2.16)	-.33 (2.51)	.41 (1.36)
	Control	−.40 (2.94)	−.35 (2.42)	−.51 (2.71)	−.12 (2.44)	−.08 (1.46)
	P value	.003[a]	.009[a]	<.001[a]	.51[a]	.001[a]

[a]*P* values derived from the Mann–Whitney *U* test.

Findings of the regression analyses are presented in Table 5. At 1 month, predictors of GI were attitude at 1 month (beta= .40, P<.001) and SN at 1 month (beta= .40, P<.001). At 6 months, predictors of GI were SN at 6 months (beta= .43, P<.001) and PBC (beta= .29, P<.001). Predictors of Δ in GI scores between baseline and 1 month were Δ (between baseline and 1 month) in attitude (beta= .31, P<.001) and Δ (between baseline and 1 month) in SN (beta= .34, P<.001). Predictors of Δ in GI scores between baseline and 6 months were Δ (between baseline and 6 months) in SN (beta= .36, P<.001) and Δ (between baseline and 6 months) in PBC (beta= .23, P<.001; Table 6).

Table 5. Multiple linear regression analyses to predict general intention to provide oral care at 1 month and 6 months.

	Model 1 (1 month)			Model 2 (6 months)		
Items	B[a]	SE[b]	*P* value	B[a]	SE[b]	*P* value
Group	−.16	.18	.38	−.01	.21	.98
Knowledge at 1 month	.01	.08	.96	.06	.09	.49
TPB[c]						
Attitude at 1 month	.40	.05	<.001	.10	.06	.12
SN[d] at 1 month	.40	.05	<.001	.43	.06	<.001
PBC[e] at 1 month	.05	.05	.34	.29	.05	<.001
	R^2-adjusted model =.23			R^2-adjusted model 1=.22		

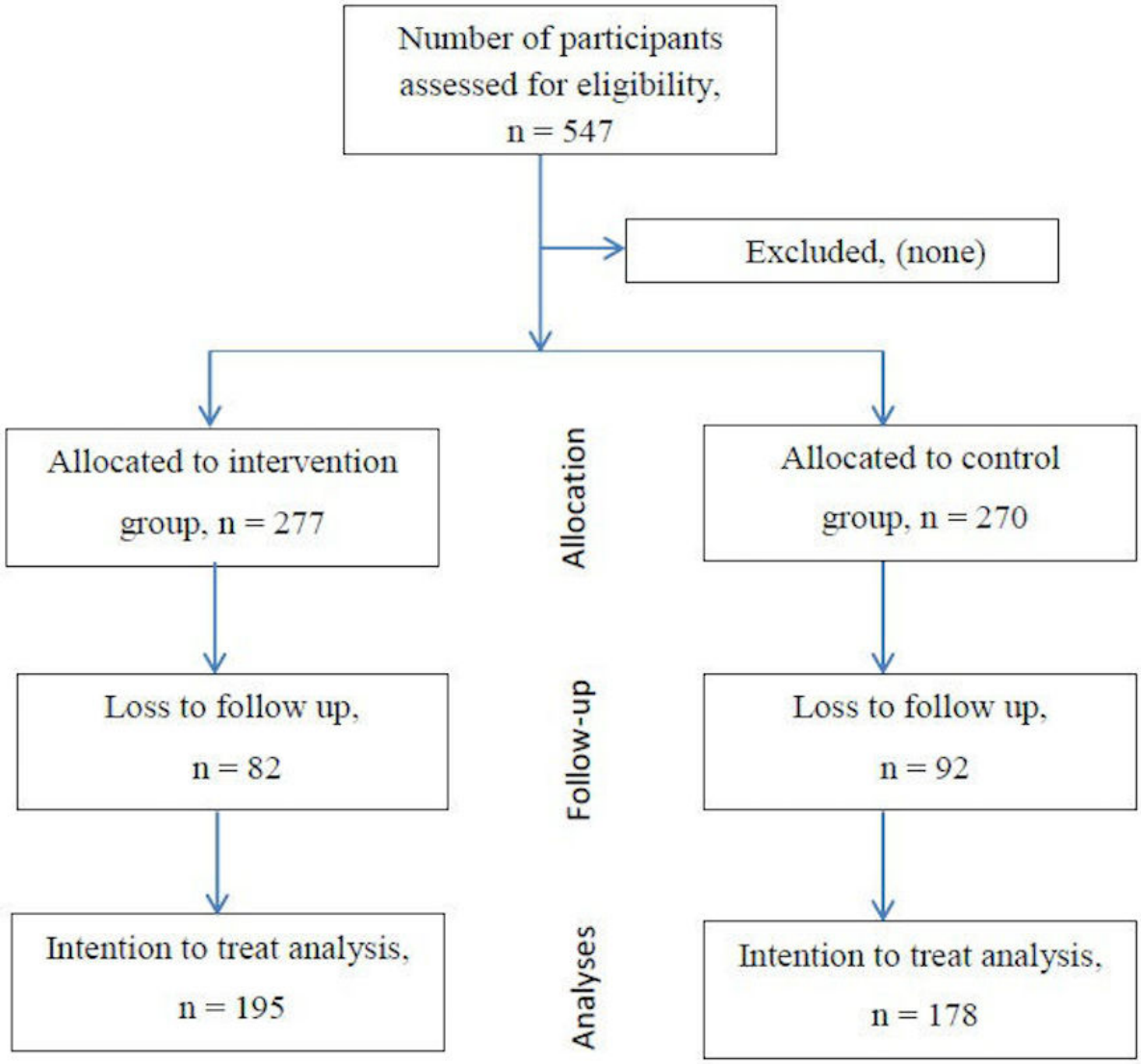
[a]B: parameter estimate.
[b]SE: standard error.
[c]TPB: theory of planned behavior.
[d]SN: subjective norm.
[e]PBC: perceived behavior control.

Table 6. Multiple linear regression analyses to predict changes in general intention between baseline and 1 month, and baseline and 6 months to provide oral care

	Model 3 (Baseline and 1 month)			Model 4 (Baseline and 6 months)		
Items	B[a]	SE[b]	*P* value	B[a]	SE[b]	*P* value
Group	.07	.25	.79	.30	.26	.25
Δ[f]Knowledge	.01	.09	.96	.01	.09	.96
TPB[c]						
ΔAttitude	.31	.06	<.001	.10	.06	.09
ΔSN[d]	.34	.05	<.001	.36	.05	<.001
ΔPBC[e]	.07	.05	.21	.23	.05	<.001
	R^2-adjusted model =.20			R^2-adjusted model 1=.21		

[a]B: parameter estimate.
[b]SE: standard error.
[c]TPB: theory of planned behavior.
[d]SN: subjective norm.
[e]PBC: perceived behavior control.
[f]Δ denotes change.

Figure 1. Flow diagram phases of the two-group randomized controlled trial.



Discussion

Principal Findings

As mentioned before, there is a need to improve the practice of oral hygiene care in the acute hospital setting, and some would argue particularly for stroke patients [28-31]. CAL has been used widely to deliver information and educate patients and caregivers in the medical field but few were reported to be related to oral health [32,33]. Thus, this study aimed to evaluate the effectiveness of a Web-based (CPD) program to change the GI of health carers to perform daily mouth cleaning for stroke patients using TPB and to identify the key factors associated with changes in GI among health care workers to provide oral hygiene care to stroke patients. Increasingly, there are calls to consider theory-based health promotion strategies in implementing programs and to provide evidence of whether theory does, in fact, dictates practice [34,35].

This study provided support for the application of TPB among stroke carers to perform oral care daily. Over time, in the test group, there was evidence of significant improvements in general intentions to perform oral hygiene care, SN, and knowledge. Even by 1 month, there was a significant improvement in GI, SN, and knowledge, and between 1 month and 6 months this persisted. Whereas in the control group there was a significant change in GI and attitudes, but it was transient; a significant difference was between 1 month and baseline but not between 6 months and baseline. At 1 month, there were significant differences in the change of attitude scores and in the change of SN scores between the test and control groups. By 6 months, significant differences in the change in GI, attitudes, and SN to provide oral hygiene care were evident in

all TPB domains except in PBC. In addition, a significant difference in the change of knowledge scores over the 6-month period was also evident between the test and control groups. This provides evidence of the effectiveness of the Web-based intervention and suggests that its effectiveness increases over time. Nonetheless, PBC did not significantly change. It is plausible that this may take longer to change or needs to be supported by environment changes [36].

Regression analyses identified that the key factors associated with GI (to provide oral hygiene care) at 1 month were attitudes (at 1 month) and SN (at 1 month); and that general intention at 6 months was associated with SN (at 6 months) and PBC (at 6 months). Furthermore, the change in GI (between baseline and 1 month) was associated with the changes that occurred in attitudes and SN; and the change in GI (between baseline and 6 months) was associated with the changes that occurred in SN and PBC. The fact that several of TPB domain scores were associated with GI at the respective time periods and that the changes in several of the TPB scores were associated with changes in GI at the respective periods provide evidence of the interrelationship between domains of TPB and GI as the theory hypothesizes. The findings of this trial support other findings with respect to predictors of GI at specific time periods [37] and changes in GI [13]. Thus in line with TPB, carers who had positive attitudes, perceived positive social pressure, and were in control of their action would have high intention of performing oral care daily to patients with stroke [13]. Of note, some domains did not change nor were some domains associated with GI or change in GI, thus the influence of the domains varies across the study [37]. This is not an unusual feature in studies as not all domain changes are associated with GI over time [38,39]. This perhaps can be attributed to the context-specific settings of the study, the type of intention, and indeed the general intention practice [40,41].

Studies have shown that a significant improvement in GI, SN, and knowledge among the participants increased their confidence in promoting oral health and educating patients [19]. A noteworthy finding of this study was that both groups were given the CAL program, but only the intervention group had detailed information on oral health related to stroke patients. Thus this contributed to a significant increase in specific knowledge and GI among the intervention group compared with the control group [18].

The study benefits from its relatively large sample size, the diverse geographical areas within a country (major centers for stroke rehabilitation), and being a double-blind randomized controlled clinical trial. The response rate was close to 70% and loss to follow-up was largely attributed to nurses having changed working environments because of being assigned to other wards or hospitals. There was no significant difference in response rate between the test and control groups.

In this trial, participants were block-randomized by hospital (stratified by size) rather than by participants to avoid the potential of bias of participants discussing the Web-based CPD program within centers. The use of Web-based CPD in both the test and control group interventions allowed for a double-blind trial. In the test group, the program was built on TPB and included specific information to promote increases in knowledge, attitudes, SN, PBC, and GI to provide oral hygiene care. Although the control group program did contain some information on oral hygiene care, it was limited and not based on the TPB.

Going forward, it would be worth investing if changes in the practice of provision of oral hygiene care do occur and that the practice is maintained over time. Furthermore, studies can be done to examine the pathways that influence changes in behavior by considering specific beliefs of TPB attributes and the impact of Web-based CPD. The impact of CPD depends on various factors such as individual carer characteristics (eg, personal compliance or interest to update themselves with new information) [42,43], and environmental factors (eg, lack of Internet access or software compatibility) [22,44]. It would also be useful to monitor engagement in the Web-based learning (which at the time keeping anonymity) to determine how this effects outcomes or indeed other methods of learning.

Conclusions

A Web-based CPD program based on TPB was effective in increasing GI to perform oral hygiene care for stroke survivors in the acute hospital setting. In addition, the program was effective in changing attitudes and SN. Furthermore, the study found that changing SNs and PBC are key factors associated with changes in GI to provide oral hygiene care. These findings support and have implications for the use of theory-based health education CPD programs and oral health promotion programs.

Acknowledgments

We would like to thank all the personnel involved in this study for their assistance and contribution. We would also like to thank the Director General of Health Malaysia for his permission to have this paper published. This study was funded by The University of Hong Kong.

Conflicts of Interest

None declared.

References

1. Friese CR, Kalisch BJ, Lee KH. Patterns and correlates of missed nursing care in inpatient oncology units. Cancer Nurs 2013;36(6):E51-E57 [FREE Full text] [doi: 10.1097/NCC.0b013e318275f552] [Medline: 23151503]

2. El-Solh AA, Pietrantoni C, Bhat A, Okada M, Zambon J, Aquilina A, et al. Colonization of dental plaques: a reservoir of respiratory pathogens for hospital-acquired pneumonia in institutionalized elders. Chest 2004 Nov;126(5):1575-1582. [doi: 10.1378/chest.126.5.1575] [Medline: 15539730]
3. Azarpazhooh A, Leake JL. Systematic review of the association between respiratory diseases and oral health. J Periodontol 2006 Sep;77(9):1465-1482. [doi: 10.1902/jop.2006.060010] [Medline: 16945022]
4. Brady M, Furlanetto D, Hunter RV, Lewis S, Milne V. Staff-led interventions for improving oral hygiene in patients following stroke. Cochrane Database Syst Rev 2006 Oct 18(4):CD003864. [doi: 10.1002/14651858.CD003864.pub2] [Medline: 17054189]
5. Lam OL, McMillan AS, Samaranayake LP, Li LS, McGrath C. Effect of oral hygiene interventions on opportunistic pathogens in patients after stroke. Am J Infect Control 2013 Feb;41(2):149-154. [doi: 10.1016/j.ajic.2012.02.020] [Medline: 22818804]
6. Lam OL, McMillan AS, Samaranayake LP, Li LS, McGrath C. Randomized clinical trial of oral health promotion interventions among patients following stroke. Arch Phys Med Rehabil 2013 Mar;94(3):435-443. [doi: 10.1016/j.apmr.2012.10.024] [Medline: 23127306]
7. Kuo Y, Yen M, Fetzer S, Chiang L, Shyu Y, Lee T, et al. A home-based training programme improves family caregivers' oral care practices with stroke survivors: a randomized controlled trial. Int J Dent Hyg 2016 May;14(2):82-91. [doi: 10.1111/idh.12138] [Medline: 25759177]
8. Smith CJ, Horne M, McCracken G, Young D, Clements I, Hulme S, et al. Development and feasibility testing of an oral hygiene intervention for stroke unit care. Gerodontology 2016 May 16;34(1):110-120. [doi: 10.1111/ger.12232] [Medline: 27198495]
9. Horne M, McCracken G, Walls A, Tyrrell PJ, Smith CJ. Organisation, practice and experiences of mouth hygiene in stroke unit care: a mixed-methods study. J Clin Nurs 2015 Mar;24(5-6):728-738. [doi: 10.1111/jocn.12665] [Medline: 25134638]
10. Kay E, Locker D. A systematic review of the effectiveness of health promotion aimed at improving oral health. Community Dent Health 1998 Sep;15(3):132-144. [Medline: 10645682]
11. McGrath C, Zhang W, Lo EC. A review of the effectiveness of oral health promotion activities among elderly people. Gerodontology 2009 Jun;26(2):85-96. [doi: 10.1111/j.1741-2358.2008.00232.x] [Medline: 19490131]
12. Kay E, Vascott D, Hocking A, Nield H, Dorr C, Barrett H. A review of approaches for dental practice teams for promoting oral health. Community Dent Oral Epidemiol 2016 Aug;44(4):313-330. [doi: 10.1111/cdoe.12220] [Medline: 26892435]
13. Ajzen I. The Theory of Planned Behaviour. Organ Behav Hum Decis Process 1991;50(2):179-211. [doi: 10.1016/0749-5978(91)90020-T]
14. Ajzen I. The theory of planned behaviour: reactions and reflections. Psychol Health 2011 Sep;26(9):1113-1127. [doi: 10.1080/08870446.2011.613995] [Medline: 21929476]
15. Shah N, Mathur VP, Kathuria V, Gupta T. Effectiveness of an educational video in improving oral health knowledge in a hospital setting. Indian J Dent 2016;7(2):70-75 [FREE Full text] [doi: 10.4103/0975-962X.184646] [Medline: 27433049]
16. Bachman MW, Lua MJ, Clay DJ, Rudney JD. Comparing traditional lecture vs. computer-based instruction for oral anatomy. J Dent Educ 1998 Aug;62(8):587-591. [Medline: 9745649]
17. Messecar DC, Van Son C, O'Meara K. Reading statistics in nursing research: a self-study CD-ROM module. J Nurs Educ 2003 May;42(5):220-226. [Medline: 12769427]
18. Albert D, Barracks SZ, Bruzelius E, Ward A. Impact of a Web-based intervention on maternal caries transmission and prevention knowledge, and oral health attitudes. Matern Child Health J 2014 Sep;18(7):1765-1771. [doi: 10.1007/s10995-013-1421-8] [Medline: 24357082]
19. George A, Duff M, Ajwani S, Johnson M, Dahlen H, Blinkhorn A, et al. Development of an online education program for midwives in australia to improve perinatal oral health. J Perinat Educ 2012;21(2):112-122 [FREE Full text] [doi: 10.1891/1058-1243.21.2.112] [Medline: 23449750]
20. Yardley L, Miller S, Schlotz W, Little P. Evaluation of a Web-based intervention to promote hand hygiene: exploratory randomized controlled trial. J Med Internet Res 2011;13(4):e107 [FREE Full text] [doi: 10.2196/jmir.1963] [Medline: 22155673]
21. Sassen B, Kok G, Mesters I, Crutzen R, Cremers A, Vanhees L. A web-based intervention for health professionals and patients to decrease cardiovascular risk attributable to physical inactivity: development process. JMIR Res Protoc 2012;1(2):e21 [FREE Full text] [doi: 10.2196/resprot.1804] [Medline: 23612470]
22. George A, Duff M, Johnson M, Dahlen H, Blinkhorn A, Ellis S, et al. Piloting of an oral health education programme and knowledge test for midwives. Contemp Nurse 2014;46(2):180-186. [doi: 10.5172/conu.2014.46.2.180] [Medline: 24787251]
23. Jerreat M, Youssouf N, Barker C, Jagger DC. Denture care of in-patients: the views of nursing staff and the development of an educational programme on denture care. J Res Nurs 2007 Mar 01;12(2):193-199. [doi: 10.1177/1744987106074504]
24. Francis JJ, Eccles MP, Johnston M, Walker A, Grimshaw J, Foy R, et al. Web.fmk. 2004. Constructing questionnaires based on the theory of planned behaviour: A manual for health services researchers URL: http://web.fmk.edu.rs/files/blogs/2010-11/Psihologija/Socijalna/TPB.pdf [accessed 2017-03-04] [WebCite Cache ID 6ohk2hZSE]
25. Al-Omiri MK, Al-Wahadni AM, Saeed KN. Oral health attitudes, knowledge, and behavior among school children in North Jordan. J Dent Educ 2006 Feb;70(2):179-187 [FREE Full text] [Medline: 16478932]

26. Resar R, Pronovost P, Haraden C, Simmonds T, Rainey T, Nolan T. Using a bundle approach to improve ventilator care processes and reduce ventilator-associated pneumonia. Jt Comm J Qual Patient Saf 2005 May;31(5):243-248. [Medline: 15960014]
27. Ab MN, Zhang J, Lam OL, Jin L, McGrath C. Effectiveness of computer-aided learning in oral health among patients and caregivers: a systematic review. J Am Med Inform Assoc 2017 Jan;24(1):209-217. [doi: 10.1093/jamia/ocw045] [Medline: 27274013]
28. Salamone K, Yacoub E, Mahoney A, Edward K. Oral care of hospitalised older patients in the acute medical setting. Nurs Res Pract 2013;2013:827670 [FREE Full text] [doi: 10.1155/2013/827670] [Medline: 23819046]
29. Purvis T, Hill K, Kilkenny M, Andrew N, Cadilhac D. Improved in-hospital outcomes and care for patients in stroke research: an observational study. Neurology 2016 Jul 12;87(2):206-213. [doi: 10.1212/WNL.0000000000002834] [Medline: 27306625]
30. Ajwani S, Jayanti S, Burkolter N, Anderson C, Bhole S, Itaoui R, et al. Integrated oral health care for stroke patients - a scoping review. J Clin Nurs 2016 Aug 18:- Epub ahead of print. [doi: 10.1111/jocn.13520] [Medline: 27538382]
31. Edwards M. Staff training improved oral hygiene in patients following stroke. Evid Based Dent 2008;9(3):73. [doi: 10.1038/sj.ebd.6400593] [Medline: 18927563]
32. Lewis D. Computers in patient education. Comput Inform Nurs 2003;21(2):88-96. [Medline: 12802949]
33. Schittek M, Mattheos N, Lyon HC, Attström R. Computer assisted learning. A review. Eur J Dent Educ 2001 Aug;5(3):93-100. [Medline: 11520331]
34. Moore SE, Holaday B, Meehan N, Watt PJ. Exploring mHealth as a new route to bridging the nursing theory-practice gap. Res Theory Nurs Pract 2015;29(1):38-52. [Medline: 25799695]
35. Newton JT. Interdisciplinary health promotion: a call for theory-based interventions drawing on the skills of multiple disciplines. Community Dent Oral Epidemiol 2012 Oct;40(Suppl 2):49-54. [doi: 10.1111/j.1600-0528.2012.00720.x] [Medline: 22998305]
36. Ajzen I. From Intentions to Actions: A Theory of Planned Behavior. In: Action Control. Berlin Heidelberg: Springer Berlin Heidelberg; 1985:11-39.
37. Armitage CJ, Conner M. Efficacy of the theory of planned behaviour: a meta-analytic review. Br J Soc Psychol 2001 Dec;40(Pt 4):471-499. [Medline: 11795063]
38. Plotnikoff RC, Lubans DR, Costigan SA, McCargar L. A test of the theory of planned behavior to predict physical activity in an overweight/obese population sample of adolescents from Alberta, Canada. Health Educ Behav 2013 Aug;40(4):415-425. [doi: 10.1177/1090198112455642] [Medline: 22984208]
39. Andrykowski MA, Beacham AO, Schmidt JE, Harper FW. Application of the theory of planned behavior to understand intentions to engage in physical and psychosocial health behaviors after cancer diagnosis. Psychooncology 2006 Sep;15(9):759-771. [doi: 10.1002/pon.1007] [Medline: 16308888]
40. McDermott MS, Oliver M, Svenson A, Simnadis T, Beck EJ, Coltman T, et al. The theory of planned behaviour and discrete food choices: a systematic review and meta-analysis. Int J Behav Nutr Phys Act 2015 Dec 30;12:162 [FREE Full text] [doi: 10.1186/s12966-015-0324-z] [Medline: 26715190]
41. McDermott MS, Oliver M, Simnadis T, Beck EJ, Coltman T, Iverson D, et al. The theory of planned behaviour and dietary patterns: a systematic review and meta-analysis. Prev Med 2015 Dec;81:150-156. [doi: 10.1016/j.ypmed.2015.08.020] [Medline: 26348455]
42. Peels DA, Bolman C, Golsteijn RH, De Vries H, Mudde AN, van Stralen MM, et al. Differences in reach and attrition between Web-based and print-delivered tailored interventions among adults over 50 years of age: clustered randomized trial. J Med Internet Res 2012;14(6):e179 [FREE Full text] [doi: 10.2196/jmir.2229] [Medline: 23246790]
43. Norman CD, Skinner HA. eHealth Literacy: essential skills for consumer health in a networked world. J Med Internet Res 2006 Jun;8(2):e9 [FREE Full text] [doi: 10.2196/jmir.8.2.e9] [Medline: 16867972]
44. Naganandini S, Rao R, Kulkarni SB. Survey on the use of the Internet as a source of oral health information among dental patients in Bangalore City, India. Oral Health Prev Dent 2014;12(2):141-147. [doi: 10.3290/j.ohpd.a31218] [Medline: 24624382]

Abbreviations

CAL: computer-aided learning
CPD: continuing professional development
GI: general intention
PBC: perceived behavior control
TPB: theory of planned behavior
SN: standard norm

Edited by G Eysenbach; submitted 20.11.16; peer-reviewed by M Horne, I Watkins; comments to author 05.01.17; revised version received 23.01.17; accepted 08.02.17; published 31.03.17.

Please cite as:
Ab Malik N, Mohamad Yatim S, Lam OLT, Jin L, McGrath CPJ
Effectiveness of a Web-Based Health Education Program to Promote Oral Hygiene Care Among Stroke Survivors: Randomized Controlled Trial
J Med Internet Res 2017;19(3):e87
URL: http://www.jmir.org/2017/3/e87/
doi:10.2196/jmir.7024
PMID:28363880

Original Paper

Trajectories of Depressive Symptoms Among Web-Based Health Risk Assessment Participants

Richard Bedrosian[1], PhD; Matt Hawrilenko[2], MA; Heather Cole-Lewis[1], MPH, PhD

[1]Johnson & Johnson Health & Wellness Solutions, Johnson & Johnson, New Brunswick, NJ, United States

[2]Clark University, Department of Psychology, Worcester, MA, United States

Corresponding Author:
Heather Cole-Lewis, MPH, PhD
Johnson & Johnson Health & Wellness Solutions
1 Johnson and Johnson Plaza
New Brunswick, NJ,
United States
Phone: 1 732 524 6462
Email: hcolelew@its.jnj.com

Abstract

Background: Health risk assessments (HRAs), which often screen for depressive symptoms, are administered to millions of employees and health plan members each year. HRA data provide an opportunity to examine longitudinal trends in depressive symptomatology, as researchers have done previously with other populations.

Objective: The primary research questions were: (1) Can we observe longitudinal trajectories in HRA populations like those observed in other study samples? (2) Do HRA variables, which primarily reflect modifiable health risks, help us to identify predictors associated with these trajectories? (3) Can we make meaningful recommendations for population health management, applicable to HRA participants, based on predictors we identify?

Methods: This study used growth mixture modeling (GMM) to examine longitudinal trends in depressive symptomatology among 22,963 participants in a Web-based HRA used by US employers and health plans. The HRA assessed modifiable health risks and variables such as stress, sleep, and quality of life.

Results: Five classes were identified: A "minimal depression" class (63.91%, 14,676/22,963) whose scores were consistently low across time, a "low risk" class (19.89%, 4568/22,963) whose condition remained subthreshold, a "deteriorating" class (3.15%, 705/22,963) who began at subthreshold but approached severe depression by the end of the study, a "chronic" class (4.71%, 1081/22,963) who remained highly depressed over time, and a "remitting" class (8.42%, 1933/22,963) who had moderate depression to start, but crossed into minimal depression by the end. Among those with subthreshold symptoms, individuals who were male (P<.001) and older (P=.01) were less likely to show symptom deterioration, whereas current depression treatment (P<.001) and surprisingly, higher sleep quality (P<.001) were associated with increased probability of membership in the "deteriorating" class as compared with "low risk." Among participants with greater symptomatology to start, those in the "severe" class tended to be younger than the "remitting" class (P<.001). Lower baseline sleep quality (P<.001), quality of life (P<.001), stress level (P<.001), and current treatment involvement (P<.001) were all predictive of membership in the "severe" class.

Conclusions: The trajectories identified were consistent with trends in previous research. The results identified some key predictors: we discuss those that mirror prior studies and offer some hypotheses as to why others did not. The finding that 1 in 5 HRA participants with subthreshold symptoms deteriorated to the point of clinical distress during succeeding years underscores the need to learn more about such individuals. We offer additional recommendations for follow-up research, which should be designed to reflect changes in health plan demographics and HRA delivery platforms. In addition to utilizing additional variables such as cognitive style to refine predictive models, future research could also begin to test the impact of more aggressive outreach strategies aimed at participants who are likely to deteriorate or remain significantly depressed over time.

(J Med Internet Res 2017;19(3):e96) doi:10.2196/jmir.6480

KEYWORDS

depression; prodromal symptoms; health risk appraisal; health surveys; prevention

Introduction

Depressive Symptoms

Depressive symptoms play a significant role in determining the potential success of population health management efforts across the care continuum [1,2]. Comorbid depression increases risk at all levels of health, being associated with poor treatment adherence, impaired self-management skills, more frequent complications from illness, and poorer medical outcomes [1-3].

The negative impact of depressive symptoms is not confined to those who suffer from formal psychiatric disorders such as major depression. Research indicates that those with subthreshold depressive symptoms, who might not meet formal diagnostic criteria for depression at a given time, may still experience significant behavioral impairment and reduced quality of life, and can be at high risk for experiencing clinical levels of distress in the future [4,5].

Although all levels of depressive symptomatology have been found to increase risk of further difficulties, the use of statistical techniques such as growth mixture modeling (GMM) [6] has enabled researchers to map qualitatively distinct trajectories of depressive symptomatology over time and identify risk factors associated with them [7,8]. The output of GMM mapping is illustrated by a study of depression among low-income women recruited during pregnancy. Mora et al [7] found five distinct patterns: (1) always or chronic depressive symptomatology, (2) antepartum only, (3) postpartum, resolving after the first year postpartum, (4) late, present at 25 months postpartum, and (5) never having depressive symptomatology. Membership in the trajectory classes was influenced by several variables, including education, race, health behavior, and psychosocial characteristics such as ambivalence about pregnancy [7].

In theory, results of such longitudinal studies could be used to identify high-risk individuals and tailor preventative services for them. However, various types of longitudinal analyses conducted with different populations and using different measures report different trajectories associated with varying constellations of risk factors, suggesting that sample demographics, recruitment strategies, treatment context, and choice of measures affect the results of such investigations [7-9]. This variability in results suggests that risk profiles may vary across different populations. Consequently, to be effective, screening algorithms and preventative interventions may need to be tailored for specific populations (eg, pregnant women and elderly patients in primary care), based on data drawn from samples with similar characteristics.

Health Risk Assessments

One unique group that may provide important insights and opportunities for intervention consists of adults who participate in mass screenings via Web-based health risk assessments (HRAs). HRAs are being administered to large segments of the population, particularly by employers. A recent report by the Kaiser Family Foundation found that 53% of large US employers (n≥200) and 23% of small employers (n<200) used HRAs with their employee populations, typically providing various incentives for participation [10]. Moreover, health plans may encourage or require members to take HRAs directly, or may create incentives for providers to administer them in clinical setting (ie, as part of the annual Medicare wellness visit). HRAs typically assess modifiable health risks, often including items covering depression, stress, and other behavioral health topics.

Because they are typically disseminated by employers rather than in clinical settings, HRAs constitute a unique channel for health screening, representing an opportunity to reach a different population, at different points in the disease continuum, as compared with patients being assessed in traditional health care settings. Employees and spouses are often incentivized on a yearly basis to take HRAs, thus providing longitudinal data on participants, both those with subthreshold symptoms and those with more significant distress.

This study used GMM [6] to examine longitudinal trends in depressive symptomatology among participants in an HRA used by employers and health plans in the United States. The primary research questions were the following: (1) Can we observe longitudinal trajectories in the HRA population like those that have been observed in other study samples? (2) Do HRA variables, which primarily reflect modifiable health risks, help us to identify predictors associated with these trajectories? (3) Can we make meaningful recommendations for population health management, applicable to HRA participants, based on the predictors we identify?

Methods

Participants

The sample initially consisted of deidentified data from 91,852 unique adult participants who completed the HealthMedia SUCCEED HRA [11]. We analyzed data only from those who took the HRA two or more times, as at least two assessment points are necessary to contribute information to growth trajectories, resulting in 22,963 participants in the final sample. Individuals participated in the HRA (at no cost to them) through employers or health care plans deploying the HRA as part of their population health offerings or health benefit structure. Table 1 presents the demographic characteristics of the sample.

Table 1. Sample characteristics (N=22,963).

Variable	Present sample[a], n (%)
Gender (n=22,907)	
Female	15,337 (66.95)
Male	7570 (33.05)
Age (n=22,912)	
18-24	576 (2.51)
25-34	4333 (18.91)
35-44	5130 (22.39)
45-54	6543 (28.56)
55-64	5690 (24.83)
65 and older	640 (2.79)
Ethnicity (n=22,812)	
Asian	469 (2.06)
Black	3628 (15.90)
Hawaiian	9 (0.04)
Hispanic	1308 (5.73)
Multiracial	209 (0.92)
Native American	74 (0.32)
Pacific Islander	39 (0.17)
White	16,857 (73.90)
Other	219 (0.96)
Marital status (n=22,963)	
Single	2917 (12.70)
Dating	1261 (5.49)
Married	15,056 (65.57)
Divorced	3284 (14.30)
Widowed	445 (1.94)
Education (n=22,798)	
Some high school	343 (1.50)
High school graduate	3623 (15.89)
Some college	6809 (29.87)
College graduate	12,023 (52.74)

[a]Sample characteristics are based on raw, not multiply imputed data.

Measures

All variables for this study were extracted from participant HRA responses. The HRA was voluntary: participants were informed that their responses would be aggregated for data analyses and customer reporting, but would not be individually shared with their health plan or employer. The HRA was administered in a Web format in nearly all instances, but was available in paper format as well.

Risk-related questions were typically derived from various validated scales, including the Cohen Perceived Stress Scale (PSS) [12], the Center for Epidemiological Studies Depression Scale (CES-D) [13-15] along with validated single-item measures of quality of life and health [16,17]. A modified version of the Work Productivity Activity Impairment (WPAI) questionnaire was used to measure worksite productivity impairment [18], a supplemental distal outcome measure for the study. To minimize the impact of transient acute sickness on productivity, WPAI questions referred to the past 4 weeks (rather than the past 7 days). The WPAI yields an estimate of total productivity impairment due to health, based on the combination of absenteeism and presenteeism. All predictor variables were measured at baseline. Data from this HRA have been extensively analyzed previously for other purposes [11,19].

Dependent Variable

The 10-item true-false version of the CES-D [13-15] was used to assess depressive symptoms. The score was computed as a sum of 10 items. CES-D scores ranged from 0 to 10 with higher scores indicating more depressive symptoms. The reliability in this sample was acceptable (Cronbach alpha=.76) and mean scores were like other nonclinical populations using a score of 4 as the cut-point for clinical significance [18,19].

Analytic Strategy

The goal of this analysis was to identify subgroups of HRA participants with distinct depression trajectories and identify predictors that make useful discriminations between these subgroups. We used GMM [6] to accomplish this objective. The following elements are described below: (1) three key analytic decisions made before fitting the growth mixture model, (2) model-building and class enumeration strategy, (3) the procedure for including predictors and distal outcomes of trajectory class membership, and (4) handling of missing data.

Analytic Decisions Before Growth Modeling

We narrowed our time horizon to 15 months, the tail end of the time period at which participants completed their “one year” assessments. Longer-term follow-up data were too sparse to provide generalizable findings.

To account for varying lengths of time between observations, we discretized time, segmenting it into 5 waves based on the patterning of responses: baseline, 0.5-3 months, 3-6 months, 6-9 months, and 9-15 months. The number and temporal width of buckets were chosen after examining the patterning of responses, with the aim of balancing granularity with adequate covariance coverage. That is, we wanted buckets of time that were narrow enough to pool participant data into the same time point, but large enough to capture enough participants to have adequate overlap between time points for the model to be empirically identified. We chose the bucketing strategy because it was the most amenable option for generating nonlinear trajectories, which we anticipated would reflect the episodic nature of depression [20]. Other strategies (eg, multilevel) of handling nonequidistant assessments were not desirable because they are limited to modeling smooth, polynomial forms of time and would likely require more assessments per individual than offered by this dataset.

Because these data were skewed with a strong floor effect, we modeled depression as an ordered categorical variable. Growth mixture models are highly sensitive to distributional assumptions, and violating these assumptions can yield inaccurate results [21]. Discretizing the measure mitigated this problem by removing distributional and linearity assumptions. Previous research identified a cutoff score of 4 and above to indicate the presence of depression [14,15]. To increase granularity, we included two additional cutoffs to make a total of 4 categories. Scores of 0 and 1 were considered indicative of minimal depression, 2-3 subthreshold, 4-6 moderate, and 7-10 severe depression. We used a latent response variable specification, which models the observed outcome as a discretized form of an underlying continuous latent response variable [22]. This specification generates thresholds where the continuous latent variable maps onto the observed ordinal outcome. For our 4 categories of depression, latent response variable values below threshold 1 map to minimal depression, values between thresholds 1 and 2 map to subthreshold depression, values between thresholds 2 and 3 map to moderate depression, and values above threshold 3 map to severe depression.

Mixture Models and Class Enumeration

To model nonlinearity, we used freeloading growth curves. Factor loadings for baseline and endpoint were fixed at 0 and 1, respectively, and loadings for the remaining time points were freely estimated. We followed the class enumeration process and reporting guidelines outlined by Masyn [23], using a variety of fit statistics to guide model selection. Because the Bayesian information criterion (BIC) and consistent Akaike information criterion (CAIC) sometimes asymptote rather than peak, we used elbow plots to assess relative benefits of selecting a model with a large number of classes.

We were interested in identifying predictors that statistically significantly differentiated between latent classes where distinctions were of substantive interest, and distal consequences of latent class membership whose means varied across class. To develop a parsimonious prediction model, we initially screened candidate predictors by including single predictors as auxiliary variables to preview their relationship with the latent class variable, using an alpha criterion of $P<.01$. Once variables were screened into the prediction model, no trimming was performed. For the final model, variables were added stepwise to understand incremental changes. For both predictors and the distal outcome, we accounted for measurement error in the latent class structure using the manual 3-step approach described by Asparouhov and Muthén [24]. Continuous predictors, excluding age and sleep hours, were divided by their standard deviations to enhance interpretability.

Missing Data Strategy

This study had the benefit of a large N, but the drawback of sparse reporting between baseline and 1 year. Although there were a smaller mean number of assessments per individual than typically seen in growth curve models, the large N enabled us to generate subgroup trajectories using the information between baseline and 1 year, because there was adequate covariance coverage between all time buckets to identify the model. Coverage between adjacent time points was low, ranging from 1.2% to 16.6%, with N postbaseline ranging from 16.45% (3778/22,968; 6-9 months) to 46.00% (10,565/22,968; 9-15 months). Given the low covariance coverage, generalizability of findings is contingent on the validity of our assumptions about missing data. Therefore, understanding any patterning present in missing data is critical for assessing generalizability.

We identified three types of mechanisms that drove missingness in this analysis.

The first was our bucketing strategy. We chose the number of buckets post hoc, so the more thinly we sliced the buckets, the larger number of buckets and consequently, the more “missing” observations we would have. Using 5 buckets yielded substantial proportions of wave-by-wave missingness, given that

participants averaged 2.10 responses over the course of follow-up, with only 9% of participants completing assessments at 3 or more waves. Missingness driven purely by design decisions and not participant characteristics meets criteria for missing at random. We used maximum likelihood estimation with robust standard errors, which produces unbiased estimates under conditions of missing at random, meaning that estimates are unbiased if missingness is either random or related only to variables that are included in the model.

The second mechanism was driven by the number of observations from each participant. The possibility that participants who provided more observations might have different characteristics than those who provided fewer data points could potentially bias estimates. To assess this mechanism, we dummy coded the number of assessments completed by each participant into 2 groups, participants with 2 versus participants with 3+ assessments, and compared them using logistic regression. We included all demographic and predictor variables in the study as covariates. We performed the analysis on the overall number of time points, and because it was possible that risk factors influenced the timing of completing an HRA, we also performed it wave-by-wave. By explicitly modeling this missingness mechanism, we could then include any significant predictors in the final model that determined class structure. By incorporating this information related to missingness in the model, maximum likelihood estimation would ensure estimates were unbiased by the mechanisms we identified.

The third issue was missing individual predictor variables. The proportion of this missingness was generally low, with the three largest proportions being hours of sleep (12.89%, 2961/22,968), physical activity (10.14%, 2328/22,968), and alcohol use (5.29%, 1214/22,968). Because these variables were exogenous, the maximum likelihood estimator did not contribute to their estimation. We generated 10 multiply imputed datasets (imputing only the X variables) using information from all study variables.

All analyses were conducted in Mplus 7.3 (Muthen & Muthen) [25].

Results

Baseline Comparisons

We compared individuals included and excluded from the study on baseline variables. Due to the large sample size, many statistically significant differences emerged, but all effect sizes were below the cutoff traditionally deemed a “very small effect” (*d*=.10). The largest differences were the following: Individuals in the excluded group had higher depression scores at baseline (Cohen *d*=.08), lower alcohol use (*d*=.08), and higher health quality (*d*=.04). Individuals included in the study were somewhat more likely to be married (65.57% vs 61.89%). Broadly, these suggest that individuals completing more HRAs had slightly poorer health and more health risk factors. Because all variables were included in the final model, these between-group differences did not bias model estimates.

Class Enumeration

We found a 5-class model to provide the best blend of fit, parsimony, and interpretability. Fit statistics are presented in Table 2. We explored freeing within-class intercept and slope variances, but freeing either resulted in empirical under-identification and/or unstable solutions. Even across many random starts that used information from previous models to aid in convergence, likelihood values did not replicate, suggesting that the freeloading factor loadings consumed much of the variability within these data. Therefore, we only included models in the class enumeration process where within-class variances were fixed at 0. Fit improved at each model tested through 6 classes, and the 7-class model did not generate a stable solution. All fit criteria clearly favored the 5-class model over the 4-class model. Although the 6-class model had incrementally better fit than the 5-class model—and likelihood ratio tests also favored the 6-class model—the decreases in the BIC, CAIC, and approximate weight of evidence (lower values imply better fit) were markedly lower than between the 4- and 5-class models, suggesting the additive explanatory power of the sixth class was low. This 6-class model also showed evidence of class splitting, meaning that one class from the 5-class model was split into two qualitatively similar classes in the 6-class model. Additionally, only a very small proportion of likelihood values replicated, decreasing our confidence in the validity of the 6-class model. For this reason, we did not include it as a finalist in our candidate models or calculate any of the Bayesian statistics for comparative fit. We used the approximate correct model probability (cmP $\hat{}_A$), which is an approximation that a given model is correct out of a set of observed models, to compare the 4 and 5-class models; along with the other fit statistics, it strongly favored the 5-class model. Entropy for the 5-class model was .95, meaning the posterior classification of individuals into latent classes was fairly precise with individuals relatively cleanly separated between classes.

Table 2. Fit statistics for growth mixture models.

Model	# parameters	Loglikelihood	BIC[a]	CAIC[b]	AWE[c]	Bayes factor (*k* vs *k* +1)	cmP_A[d]
1-class[e]	8	−36,256	72,593	72,601	72,697	<.01	–
2-class	14	−37,384	74,909	74,923	75,092	<.01	–
3-class	20	−35,123	70,448	70,468	70,709	<.01	–
4-class	26	−34,386	69,033	69,059	69,372	<.01	0
5-class	32	−33,424	67,170	67,202	67,587	–	1
6-class	38	−32,975	66,332	66,370	66,828	–	–
7-class	Empirically underidentified						

[a]BIC: Bayesian information criterion.

[b]CAIC: consistent Akaike information criterion.

[c]AWE: approximate weight of evidence.

[d]cmP_A: Probability that model is true among set of all models being compared.

[e]The intercept was not constrained to 0 in the 1 class model, allowing it to serve as a better benchmark. Values for the Lo-Mendell-Rubin and Bootstrapped Lo-Mendell-Rubin likelihood ratio tests are not included because they were all significant at *P*<.001 and not useful in distinguishing between models.

Table 3. Parameters for 5-class growth mixture model. The class proportions are based on model-estimated data, and differ slightly than those based on modal class assignment.

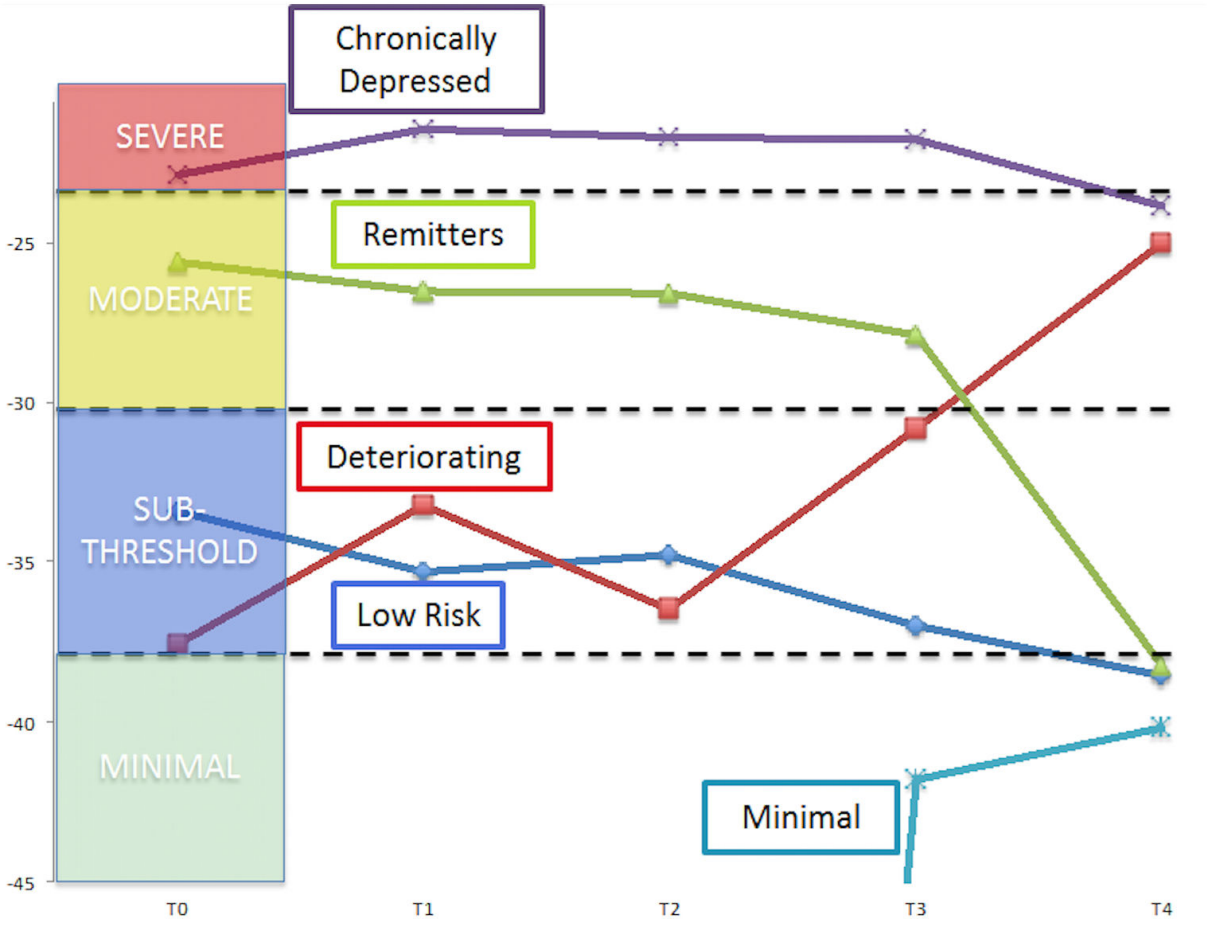
Parameters	Minimal (63.66%)	Low risk (18.65%)	Deteriorating (4.69%)	Remitting (8.05%)	Chronic (4.96%)
Intercept	−51.05	−38.14	−42.31	−30.35	−27.64
Slope	6.12	−5.14	12.56	−12.63	−0.98
Factor loadings					
Baseline	0.00[a]	0.00[a]	0.00[a]	0.00[a]	0.00[a]
0.5-3 months	−14.61	0.36	0.35	0.07	−1.51
3-6 months	−16.52	0.26	0.09[b]	0.08	−1.24
6-9 months	0.73	0.70	0.54	0.18	−1.17
9-15 months	1.00[a]	1.00[a]	1.00[a]	1.00[a]	1.00[a]
Thresholds					
Minimal	−43.01	All thresholds constrained constant across classes			
Subthreshold	−35.02				
Moderate	−27.90				
Severe	–				

[a]Denotes parameters constrained constant across latent classes.

[b]This parameter is significant at *P*=.02. All other parameters are significant at *P*<.001.

The 5-class solution consisted of a “minimal depression” class (63.66%) whose scores were low and consistent across all time waves, a “low risk” class (18.65%) whose condition remained subthreshold across time, a “deteriorating” class (4.69%) who began at subthreshold but approached severe depression by the end of the study, a “chronic” class (4.96%) who remained highly depressed across the whole study, and a “remitting” class (8.05%) who had moderate depression to start, but crossed the threshold into minimal depression by the end of the follow-up period. Growth parameters are presented in Table 3 and latent trajectory classes are graphically depicted in Figure 1. Note that an ordinal logistic model is used, so the table and figure contain thresholds where the continuous latent variable is cut into each category of depression; the thresholds themselves are not otherwise interpreted.

Figure 1. Latent class growth trajectories. The broken line for the minimal class indicates where the threshold drops below the bounds of the figure.



Predictors and Consequences of Latent Class Membership

All continuous candidate predictor variables are listed in Table 4, with means and standard deviations broken out by modal class assignment, which refers to the most likely class assignment for each individual. Table 5 lists all categorical candidate predictor variables and relative percentages within each class assignment. Because Tables 4 and 5 are broken out by modal class assignment, the sample proportions of class membership differ slightly from the model based estimates (Table 3,Figure 1).

Table 4. Descriptive statistics for continuous predictor variables by modal class assignment.

Continuous variables	Full sample N=22,963	Minimal n[a]=14,676	Low-risk n=4568	Deteriorating n=705	Remitting n=1933	Chronic n=1081
	Mean (SD[b])	Mean	Mean	Mean	Mean	Mean
Age	45.84 (11.72)	47.09	44.22	41.43	43.72	42.44
Life quality	2.87 (0.79)	3.04	2.69	2.72	2.4	2.09
Health quality	2.63 (0.83)	2.81	2.45	2.47	2.17	1.92
Sleep quality	6.8 (2.14)	7.44	5.98	6.29	5.24	4.61
Sleep hours	7.07 (1.21)	7.24	6.84	6.94	6.67	6.51
Stress	4.04 (1.06)	3.92	4.12	4.14	4.4	4.61
Alcohol use	9.5 (1.34)	9.60	9.39	9.35	9.21	9.16
Physical activity	0.88 (0.32)	0.91	0.86	0.89	0.81	0.80

[a]Class *n* s are based on modal class assignment and differ slightly from the model-estimated class proportions presented in Table 3.

[b]SD: standard deviation.

Table 5. Descriptive statistics for categorical predictor variables by modal class assignment.

Categorical variables	Full sample N=22,963	Minimal n=14,676	Low risk n=4568	Deteriorating n=705	Remitting n=1933	Chronic n=1081
	% (n/N)	% (n/N)	% (n/N)	% (n/N)	% (n/N)	% (n/N)
Depression						
Minimal	65.28 (14,959/22,914)	100 (14,638/14,638)	0.02 (1/4560)	45.3 (319/705)	0.05 (1/1932)	0 (0/1079)
Subthreshold	21.02 (4817/22,914)	0 (0/14,638)	96.97 (4422/4560)	54.8 (386/705)	0.21 (4/1932)	0.46 (5/1079)
Moderate	10.22 (2341/22,914)	0 (0/14,638)	2.83 (129/4560)	0 (0/705)	91.67 (1771/1932)	40.87 (441/1079)
Severe	3.48 (797/22,914)	0 (0/14,638)	0.18 (8/4560)	0 (0/705)	8.07 (156/1932)	58.67 (633/1079)
Gender						
Female	66.95 (15,337/22,907)	63.39 (9276/14,634)	70.12 (3196/4558)	74.9 (528/705)	76.85 (1484/1931)	79.05 (853/1079)
Male	33.05 (7570/22,907)	36.61 (5358/14,634)	29.88 (1362/4558)	25.1 (177/705)	23.15 (447/1931)	20.95 (226/1079)
Relationship status						
Single	12.54 (2917/23,268)	10.76 (1598/14,857)	14.9 (692/4645)	14.6 (104/712)	17.19 (337/1961)	17.02 (186/1093)
Dating	5.42 (1261/23,268)	4.78 (710/14,857)	6.29 (292/4645)	8.6 (61/712)	6.43 (126/1961)	6.59 (72/1093)
Married	64.71 (15,056/23,268)	68.45 (10,170/14,857)	59.74 (2775/4645)	59.4 (423/712)	55.74 (1093/1961)	54.44 (595/1093)
Divorced	14.11 (3284/23,268)	13.02 (1934/14,857)	15.24 (708/4645)	15.3 (109/712)	16.57 (325/1961)	19.03 (208/1093)
Widowed	1.91 (445/23,268)	1.78 (264/14,857)	2.17 (101/4645)	1.1 (8/712)	2.65 (52/1961)	1.83 (20/1093)
Pregnant	1.33 (305/22,914)	1.24 (181/14,638)	1.69 (77/4560)	1.0 (7/705)	1.45 (28/1932)	1.11 (12/1079)
Lives with others						
Lives alone	12.47 (2863/22,963)	11.6 (1702/14,676)	13.22 (604/4568)	13.6 (96/705)	15.26 (295/1933)	15.36 (166/1081)
Child 0-2	6.65 (1527/22,963)	6.22 (913/14,676)	7.6 (347/4568)	9.7 (68/705)	6.88 (133/1933)	6.11 (66/1081)
Child 2-12	26.6 (6109/22,963)	24.75 (3633/14,676)	28.26 (1291/4568)	36.0 (254/705)	29.9 (578/1933)	32.65 (353/1081)
Child 12-18	20.32 (4666/22,963)	19.34 (2838/14,676)	22.2 (1014/4568)	24.5 (173/705)	20.23 (391/1933)	23.13 (250/1081)
Adult	74.59 (17,128/22,963)	76.49 (11,225/14,676)	72.66 (3319/4568)	70.5 (497/705)	69.99 (1353/1933)	67.9 (734/1081)
Weight						
Healthy or underweight[a]	33.33 (7654/22,963)	35.45 (5203/14,676)	31.61 (1444/4568)	30.2 (213/705)	27.06 (523/1933)	25.07 (271/1081)
Overweight	32.27 (7410/22,963)	33.78 (4958/14,676)	30.58 (1397/4568)	28.7 (202/705)	29.85 (577/1933)	25.53 (276/1081)
Obese	27.84 (6393/22,963)	25.78 (3783/14,676)	29.93 (1367/4568)	31.2 (220/705)	32.07 (620/1933)	37.28 (403/1081)
Extremely obese	6.56 (1506/22,963)	4.99 (732/14,676)	7.88 (360/4568)	9.9 (70/705)	11.02 (213/1933)	12.12 (131/1081)
In treatment for depression	9.01 (2069/22,963)	4.57 (671/14,676)	11.01 (503/4568)	16.3 (115/705)	19.76(382/1933)	36.82 (398/1081)

[a]A minimal proportion of individuals reported underweight body mass index (BMI) in this sample.

After examining the class structure, we were interested in predictors that differentiated the low risk from the deteriorating class, or the remitting from the chronic class. Table 5 presents log odds and odds ratios from the final model, using the minimal depression class as the reference group. It also includes contrasts of interest between the low risk and deteriorating classes, as well as the chronic and remitting classes. The following variables were screened out of the final prediction model because they did not meet our criteria for distinguishing between either set of classes: living with others, weight, relationship status, being pregnant, alcohol use, and physical activity.

Table 6. Predictors and consequences of class membership with selected between-class contrasts.

Predictor variables by class	Referent class=minimal			Referent class=Low-risk			Referent class=remitting		
	Log odds	*P*	Odds ratio	Log odds	*P*	Odds ratio	Log odds	*P*	Odds ratio
Low-risk class									
Age	−0.03	<.001	0.97						
Male	−0.23	<.001	0.79						
Sleep quality[a]	−0.67	<.001	0.51						
Sleep hours	−0.10	<.001	0.91						
Stress[a]	0.20	<.001	1.22						
Life quality[a]	−0.31	<.001	0.73						
Health quality[a]	−0.13	<.001	0.88						
Treatment	0.72	<.001	2.06						
Distal outcome: WPAI[b]	Mdiff[c]=1.98, *P*<.001								
Deteriorating class									
Age	−0.05	<.001	0.95	−0.02	<.001	0.98			
Male	−0.49	<.001	0.61	−0.26	.01	0.77			
Sleep quality[a]	−0.51	<.001	0.6	0.16	.001	1.17			
Sleep hours	−0.09	.03	0.92	0.01	.77	1.01			
Stress[a]	0.22	<.001	1.25	0.03	.60	1.03			
Life quality[a]	−0.30	<.001	0.74	0.01	.91	1.01			
Health quality[a]	−0.15	.01	0.86	−0.02	.72	0.98			
Treatment	1.29	<.001	3.64	0.57	<.001	1.77			
Distal outcome: WPAI	Mdiff=57.25, *P*<.001			Mdiff=55.26, *P*<.001					
Remitting class									
Age	−0.04	<.001	0.96						
Male	−0.55	<.001	0.58						
Sleep quality[a]	−0.90	<.001	0.41						
Sleep hours	−0.12	<.001	0.88						
Stress[a]	0.43	<.001	1.53						
Life quality[a]	−0.57	<.001	0.57						
Health quality[a]	−0.28	<.001	0.76						
Treatment	1.32	<.001	3.73						
Distal outcome: WPAI	Mdiff=3.59, *P*<.001								
Chronic class									
Age	−0.06	<.001	0.94				−0.02	<.001	0.98
Male	−0.58	<.001	0.56				−0.03	.74	0.97
Sleep quality[a]	−1.12	<.001	0.33				−0.22	<.001	0.81
Sleep hours	−0.14	<.001	0.87				−0.02	.62	0.98
Stress[a]	0.55	<.001	1.73				0.12	.001	1.13
Life quality[a]	−0.94	<.001	0.39				−0.67	<.001	0.69
Health quality[a]	−0.30	<.001	0.74				−0.03	.67	0.97

Predictor variables by class	Referent class=minimal			Referent class=Low-risk			Referent class=remitting		
Treatment	2.14	<.001	8.49				0.82	<.001	2.28
Distal outcome: WPAI	Mdiff=12.70, *P*<.001						Mdiff=9.11, *P*<.001		

[a]To aid interpretation, sleep quality, stress, life quality, and health quality were divided by their baseline standard deviation, meaning the respective odds ratio are commensurate with a 1 standard deviation increase in those predictors.

[b]WPAI: workplace productivity impairment.

[c]Mdiff: mean differences between classes at final time point.

Low Risk Versus Deteriorating Class

Individuals who were older or male had lower odds of being in the deteriorating class as compared with the low risk class. Counterintuitively, better sleep quality was associated with greater odds of being in the deteriorating class. Individuals actively receiving treatment for depression or bipolar disorder were more likely to be in the deteriorating class. At the end of follow-up, those in the deteriorating class had much higher productivity impairment (Cohen *d*=3.37) than those in the low risk class.

Severe Versus Remitting Class

Age appeared to be a protective factor, with older individuals more likely to be in the remitting class than their younger counterparts. Higher stress levels and lower baseline sleep quality and life quality were all associated with a greater chance of being in the chronic class as compared with the remitting class. Hours of sleep was a significant predictor when examined on its own, but became nonsignificant when including sleep quality in the model. By the end of follow-up, individuals in the remitting class had moderately lower productivity impairment than those in the chronic class (*d*=.55).

All Classes Versus Minimal Depression Class

All variables were statistically significant in differentiating the other four classes from the minimal depression class. Individuals engaged in treatment and individuals with higher stress levels were less likely to be in the minimal class than all other classes, whereas increases in age, sleep quality, sleep hours, life quality, and health quality, as well as being male, were associated with an increased likelihood of being in the minimal class as compared with all other classes. Those in the minimal class had lower impairment scores at the end of follow-up than all other classes.

Most notably, however, individuals in the deteriorating group had much greater impairments in functioning as compared with those in the chronic group (*d*=2.72, *P*<.001), despite finishing the study at the same average level of depression.

Sensitivity and Missing Data Analyses

Missing Data Analyses

A binary logistic regression indicated that the following study variables were related to number of assessments over and above the effect of depression: age, health quality, gender, weight, and physical activity. We performed a sensitivity analysis on our latent class structure by regressing latent classes directly on these predictors, in order for the predictors to contribute to maximum likelihood estimates of the class structure and growth trajectories. By including them in the model, maximum likelihood estimation guarantees estimates unbiased by missingness related to these variables. We found the class structure to be substantively the same and class sizes to be nearly identical, with estimated class proportions identical to the second or third decimal places for all classes, indicating the main analyses were not meaningfully biased by missingness mechanisms related to these variables.

Sensitivity Analyses

We were surprised by the finding that higher sleep quality was associated with greater odds of being in the deteriorating versus low risk class and speculated that it might be driven by intercept differences, because the low risk class had a marginally larger proportion of individuals in the subthreshold category of depression than did the deteriorating class. We ran a model that controlled for this by regressing intercepts on sleep quality while also including sleep quality as a predictor of class membership. The findings were substantively the same, suggesting that sleep quality differentiated meaningfully between the two classes on both their intercept and slope components.

We conducted several post hoc analyses to examine whether the effect of age on class membership was curvilinear; it was not.

Discussion

Trajectories

The first objective of this study was to determine whether longitudinal HRA data would reveal trajectories of depressive symptomatology comparable to those found in research conducted with other populations. The trajectories identified were consistent with existing trends in the literature, but the percentages of participants in each trajectory were somewhat different than those obtained in other studies, a pattern in the previous literature described in the introduction.

As in other studies, most participants did not experience significant depressive symptomatology (“minimal depression”) [7].

Among those manifesting clinical levels of distress, there was a group whose symptomatology continued throughout the study period (“severe”), whereas another segment showed markedly few symptoms over time (“remitters”). These two trajectories replicated chronic and episodic patterns of depressive symptoms observed in studies conducted in clinical settings [5,7,9].

This study revealed that there was a group of HRA participants with subthreshold symptomatology who progressed to greater levels of distress over time (“deteriorating”), mirroring other

studies that found a similar pattern of deterioration among individuals with subsyndromal symptoms [4,7,9]. It was noteworthy that although the individuals in the deteriorating group finished the study at the same average level of depression as the chronic group, they had much greater productivity impairment, suggesting that their symptoms had greater impact on daily functioning.

Predictors

The second objective of the study was to determine whether typical HRA variables would yield predictors associated with the trajectories we observed. The results did identify some key predictors: Some seemed to reflect existing trends in the literature, whereas others did not.

Among those with subthreshold symptoms, individuals who were male and older were less likely to show symptom deterioration, whereas current depression treatment and surprisingly, sleep quality were associated with increased probability of membership in the "deteriorating" class as compared with the "low risk" class. Prior longitudinal research found a U-shaped pattern of depressive symptoms among age cohorts, starting with higher rates among young adults, decreasing during middle age, then increasing in old age [26]. Women reported greater distress in young adulthood, but the gender gap narrowed in old age [26]. Neither of these patterns was observed in our sample.

The fact that participants with subthreshold symptoms who were in treatment were more likely to be members of the "deteriorating" class may reflect several different scenarios, including the possibility that these individuals were experiencing progressively greater difficulty in daily functioning than their peers with comparable levels of symptomatology or had experienced prior depressive episodes, and therefore were more likely to seek treatment. A previous study of participants in the HealthMedia SUCCEED HRA used in this research found that those who were currently receiving depression treatment showed greater functional impairment than those with high levels of depressive symptomatology who were not receiving services [27].

The finding that higher sleep quality was associated with deterioration among participants with subthreshold symptomatology seems counterintuitive, as sleep problems play such a significant role in the clinical course of depression [28,29]. However, various studies indicate that sleeping is a commonly used coping strategy for stress reactions and symptoms of depression and anxiety [30,31] (Multimedia Appendix 1). Those findings make us wonder whether members of the "deteriorating" class had begun to use sleep to cope with prodromal symptoms when they first took the HRA. If this pattern were to appear in similar studies in the future, it may suggest that the use of sleep as a coping strategy by individuals with subthreshold distress might be a leading indicator of more serious depressive symptoms in the future.

Among participants with more significant symptomatology at the outset, those in the "severe" class tended to be younger than their counterparts in the "remitting" class. Lower baseline sleep quality and quality of life, as well as stress level were all predictive of membership in the "severe" class, results that were consistent with prior findings on the connections between those variables and both the severity and clinical course of depression [28,29,32,33].

Recommendations

The third objective of the study was to determine whether we could make meaningful recommendations for population health management, applicable to HRA participants, based on the predictors we identified. Below we offer several recommendations intended to build on and extend the results of this study:

Nearly a quarter of the population in this study experienced subthreshold depressive symptoms at baseline, and 1 in 5 of them deteriorated during succeeding years, to the point where their level of symptomatology was comparable to the chronic group, experiencing a much greater level of functional impairment at follow-up than those with chronic symptoms. Although this study did find variables that differentiated the "deteriorating" participants from those whose symptoms remained at subthreshold levels, additional research is needed before translating those results into meaningful intervention strategies for these individuals.

There is a need for further data on the convergent validity of depression scales as used in HRAs. Although high level of productivity impairment in the "deteriorating" class supports the belief that these individuals need preventative services, it would be helpful to know more about the clinical accuracy of the depression screening scores. How well do high-risk scores correlate with diagnostic judgments and needs assessments rendered by trained clinicians? Are there additional predictor variables that distinguish individuals in the "deteriorating" and "chronic" classes truly suffering from clinical levels of depressive symptoms from those who are not?

This study used data from an HRA that was administered almost exclusively via desktop computers, mostly before the implementation of the Affordable Care Act (ACA). However, HRAs are increasingly moving to mobile platforms [34] (Multimedia Appendix 2), which may well affect the psychometric properties of existing screening tools, or require shorter instruments such as the Patient Health Questionnaire-2 (PHQ-2). Use of a mobile platform may attract different demographic groups, resulting in younger and more diverse HRA populations. Changes in commercial health plan demographics have already occurred because of ACA implementation [35]: How those changes affect the profile of typical HRA participants will need to be determined.

Results of this study support the notion that mass HRA screenings conducted by health plans and employers can flag depression risk among segments of the population (eg, those who are younger or healthier) who may not be routinely screened in traditional health care settings. Health plans and employers could conduct further research on depressive symptoms among HRA participants:

They can test and refine models for predicting membership in the "chronic" and "deteriorating" classes. For example, the results of this study suggest that asking participants with

subthreshold depressive symptoms if they are using sleep as a coping strategy may increase our ability to identify those who symptoms are likely to worsen over time. Below we discuss other variables that might increase predictive power if they were included in future HRAs.

It concerned us that relatively small numbers of participants were receiving treatment for depression, even among the most highly symptomatic individuals. However, as the predictive models are refined, they should provide an opportunity for health plans and employers to test more aggressive outreach strategies (eg, outbound telephone contact) for HRA participants who are likely to belong to the "severe" and "deteriorating" classes (eg, younger participants, and those who manifest poor sleep quality, low quality of life, and/or high stress levels), to increase participation in whatever treatment services are available.

Those who design HRAs may want to include new variables that might provide more insight into the factors influencing class membership, especially those associated with chronic depressive symptoms and exacerbations in subthreshold symptomatology. HRA questions have traditionally focused on modifiable health risks [27], such as physical activity, nutrition, but typically have not assessed psychological variables known to increase vulnerability to depression, such as cognitive style or history of emotional maltreatment [36]. Moreover, HRA items have not usually evaluated protective factors that might reduce risk such as community involvement. However, as HRAs evolve, they are increasingly covering new areas such as mindfulness, vitality, and sense of mission or purpose, that might enable us to gain greater insight into the factors that decrease susceptibility to chronicity or deterioration [34] (Multimedia Appendix 2).

Limitations

The HRA participants were typically incentivized by their employers to take the assessment. Since participation rates varied widely across employers and health plans using the HRA, self-selection bias undoubtedly plays some role in this sample. The participants tend to be younger and better educated than typical US health plan members [35], and they may well be more interested in their health. As noted earlier, some results were probably affected by the dearth of elderly participants in our sample. However, because millions of people now take HRAs every year, we believe that this continues to be a population worthy of study. As noted above, shifts in technology platforms and health plan demographics will require additional research in order for predictive studies to have any value in the future.

This study utilized only self-report measures, and the results are subject to all the limitations associated with them, especially regarding the depression measure. As discussed earlier, without diagnostic interviews, we have no way of determining the relative probabilities of false positives and false negatives yielded by using this version of the CES-D with this population.

Lastly, as noted above, since the HRA focused extensively on modifiable health risks, we perhaps missed opportunities to gain greater insight into protective factors or psychological variables that increase vulnerability to depressive symptomatology, and help us better understand determinants of class membership.

Conflicts of Interest

Richard C Bedrosian was employed as Director of Behavioral Health at Johnson & Johnson Health and Wellness Solutions at the time this research was conducted. Matt Hawrilenko was contracted as a statistician at Johnson & Johnson Health and Wellness Solutions and was compensated for his work on this paper. Heather Cole-Lewis is employed as Manager, Behavioral Science at Johnson & Johnson Health and Wellness Solutions. Richard Bedrosian and Heather Cole-Lewis are shareholders of Johnson & Johnson. The HealthMedia SUCCEED health risk assessment (HRA) is property of Johnson & Johnson Health and Wellness Solutions, Inc.

Multimedia Appendix 1

Highlights, workplace stress and anxiety survey.

[HTML File, 67KB - jmir_v19i3e96_app1.html]

Multimedia Appendix 2

Whitepaper: reimagining the health assessment.

[PDF File (Adobe PDF File), 709KB - jmir_v19i3e96_app2.pdf]

References

1. Katon WJ, Unützer J. Health reform and the affordable care act: the importance of mental health treatment to achieving the triple aim. J Psychosom Res 2013 Jun;74(6):533-537 [FREE Full text] [doi: 10.1016/j.jpsychores.2013.04.005] [Medline: 23731753]
2. Katon W, Russo J, Lin EH, Schmittdiel J, Ciechanowski P, Ludman E, et al. Cost-effectiveness of a multicondition collaborative care intervention: a randomized controlled trial. Arch Gen Psychiatry 2012 May;69(5):506-514 [FREE Full text] [doi: 10.1001/archgenpsychiatry.2011.1548] [Medline: 22566583]

3. Bogner HR, Morales KH, de Vries HF, Cappola AR. Integrated management of type 2 diabetes mellitus and depression treatment to improve medication adherence: a randomized controlled trial. Ann Fam Med 2012;10(1):15-22 [FREE Full text] [doi: 10.1370/afm.1344] [Medline: 22230826]
4. Cohen CI, Goh KH, Gustave M. A prospective study of outcome and predictors of subclinical and clinical depression in an older biracial sample of psychiatric outpatients. J Affect Disord 2010 Mar;121(3):204-211. [doi: 10.1016/j.jad.2009.05.021] [Medline: 19656576]
5. Sherbourne CD, Wells KB, Hays RD, Rogers W, Burnam MA, Judd LL. Subthreshold depression and depressive disorder: clinical characteristics of general medical and mental health specialty outpatients. Am J Psychiatry 1994 Dec;151(12):1777-1784. [doi: 10.1176/ajp.151.12.1777] [Medline: 7977885]
6. Muthén B, Muthén LK. Integrating person-centered and variable-centered analyses: growth mixture modeling with latent trajectory classes. Alcohol Clin Exp Res 2000 Jun;24(6):882-891. [Medline: 10888079]
7. Mora PA, Bennett IM, Elo IT, Mathew L, Coyne JC, Culhane JF. Distinct trajectories of perinatal depressive symptomatology: evidence from growth mixture modeling. Am J Epidemiol 2009 Jan 1;169(1):24-32 [FREE Full text] [doi: 10.1093/aje/kwn283] [Medline: 19001135]
8. Choi CJ, Stone RA, Kim KH, Ren D, Schulz R, Given CW, et al. Group-based trajectory modeling of caregiver psychological distress over time. Ann Behav Med 2012 Aug;44(1):73-84 [FREE Full text] [doi: 10.1007/s12160-012-9371-8] [Medline: 22585179]
9. Nandi A, Beard JR, Galea S. Epidemiologic heterogeneity of common mood and anxiety disorders over the lifecourse in the general population: a systematic review. BMC Psychiatry 2009;9:31 [FREE Full text] [doi: 10.1186/1471-244X-9-31] [Medline: 19486530]
10. Claxton G, Rae M, Panchal N, Damico A, Whitmore H, Bostick N, et al. Health benefits in 2013: moderate premium increases in employer-sponsored plans. Health Aff (Millwood) 2013 Sep;32(9):1667-1676. [doi: 10.1377/hlthaff.2013.0644] [Medline: 23962411]
11. Lenneman J, Schwartz S, Giuseffi DL, Wang C. Productivity and health: an application of three perspectives to measuring productivity. J Occup Environ Med 2011 Jan;53(1):55-61. [doi: 10.1097/JOM.0b013e3182029110] [Medline: 21187787]
12. Cohen S, Kamarck T, Mermelstein R. A global measure of perceived stress. J Health Soc Behav 1983 Dec;24(4):385-396. [Medline: 6668417]
13. Cohen S, Kamarck T, Mermelstein R. A global measure of perceived stress. Journal of Health and Social Behavior 1983 Dec;24(4):385. [doi: 10.2307/2136404]
14. Andresen EM, Malmgren JA, Carter WB, Patrick DL. Screening for depression in well older adults: evaluation of a short form of the CES-D (Center for Epidemiologic Studies Depression Scale). Am J Prev Med 1994 Mar;10(2):77-84. [Medline: 8037935]
15. Irwin M, Artin KH, Oxman MN. Screening for depression in the older adult: criterion validity of the 10-item center for epidemiological studies depression scale (CES-D). Arch Intern Med 1999;159(15):1701-1704. [Medline: 10448771]
16. Cunny KA, Perri M. Single-item vs multiple-item measures of health-related quality of life. Psychol Rep 1991 Aug;69(1):127-130. [doi: 10.2466/pr0.1991.69.1.127] [Medline: 1961779]
17. Locke DE, Decker PA, Sloan JA, Brown PD, Malec JF, Clark MM, et al. Validation of single-item linear analog scale assessment of quality of life in neuro-oncology patients. J Pain Symptom Manage 2007 Dec;34(6):628-638 [FREE Full text] [doi: 10.1016/j.jpainsymman.2007.01.016] [Medline: 17703910]
18. Reilly MC, Zbrozek AS, Dukes EM. The validity and reproducibility of a work productivity and activity impairment instrument. Pharmacoeconomics 1993 Nov;4(5):353-365. [Medline: 10146874]
19. Striegel RH, Bedrosian R, Wang C, Schwartz S. Why men should be included in research on binge eating: results from a comparison of psychosocial impairment in men and women. Int J Eat Disord 2012 Mar;45(2):233-240. [doi: 10.1002/eat.20962] [Medline: 22031213]
20. Hosenfeld B, Bos EH, Wardenaar KJ, Conradi HJ, van der Maas HL, Visser I, et al. Major depressive disorder as a nonlinear dynamic system: bimodality in the frequency distribution of depressive symptoms over time. BMC Psychiatry 2015;15:222 [FREE Full text] [doi: 10.1186/s12888-015-0596-5] [Medline: 26385384]
21. Bauer DJ, Curran PJ. Distributional assumptions of growth mixture models: implications for overextraction of latent trajectory classes. Psychol Methods 2003 Sep;8(3):338-363. [doi: 10.1037/1082-989X.8.3.338] [Medline: 14596495]
22. Masyn K, Petras H, Liu W. Growth curve models with categorical outcomes. In: Bruinsma G, Weisburd D, editors. Encyclopedia of criminology and criminal justice. New York: Springer; 2014:2013-2025.
23. Masyn KE. Latent class analysisfinite mixture modeling. In: Little TD, editor. The Oxford handbook of quantitative methods in psychology. Oxford, UK: Oxford University Press; 2013:551-611.
24. Asparouhov T, Muthén B. Auxiliary variables in mixture modeling: three-step approaches using M. Struct Equ Modeling 2014 Jun 09;21(3):329-341. [doi: 10.1080/10705511.2014.915181]
25. Muthén L, Muthén B. Mplus user's guide, 7th edition. Los Angeles, CA: Muthén & Muthén; 2012.
26. Sutin AR, Terracciano A, Milaneschi Y, An Y, Ferrucci L, Zonderman AB. The trajectory of depressive symptoms across the adult life span. JAMA Psychiatry 2013 Aug;70(8):803-811 [FREE Full text] [doi: 10.1001/jamapsychiatry.2013.193] [Medline: 23760442]

27. Bedrosian RC, Striegel RH, Wang C, Schwartz S. Association of binge eating with work productivity impairment, adjusted for other health risk factors. J Occup Environ Med 2012 Apr;54(4):385-393. [doi: 10.1097/JOM.0b013e3182479f40] [Medline: 22453810]
28. Staner L. Comorbidity of insomnia and depression. Sleep Med Rev 2010 Feb;14(1):35-46. [doi: 10.1016/j.smrv.2009.09.003] [Medline: 19939713]
29. Skapinakis P, Rai D, Anagnostopoulos F, Harrison S, Araya R, Lewis G. Sleep disturbances and depressive symptoms: an investigation of their longitudinal association in a representative sample of the UK general population. Psychol Med 2013 Feb;43(2):329-339. [doi: 10.1017/S0033291712001055] [Medline: 22640482]
30. Morris ME, Aguilera A. Mobile, social, and wearable computing and the evolution of psychological practice. Prof Psychol Res Pr 2012 Dec;43(6):622-626 [FREE Full text] [doi: 10.1037/a0029041] [Medline: 25587207]
31. Anxiety Disorders Association of America (ADAA). Highlights: workplace stress & anxiety disorders study URL: https://www.adaa.org/workplace-stress-anxiety-disorders-survey [accessed 2017-03-09] [WebCite Cache ID 6opvke3Tt]
32. Wikman A, Wardle J, Steptoe A. Quality of life and affective well-being in middle-aged and older people with chronic medical illnesses: a cross-sectional population based study. PLoS One 2011;6(4):e18952 [FREE Full text] [doi: 10.1371/journal.pone.0018952] [Medline: 21559485]
33. Lee C, Yeh C, Lee M, Lin H, Chen VC, Hsieh M, et al. Leisure activity, mobility limitation and stress as modifiable risk factors for depressive symptoms in the elderly: results of a national longitudinal study. Arch Gerontol Geriatr 2012;54(2):e221-e229. [doi: 10.1016/j.archger.2011.06.014] [Medline: 21821296]
34. Bedrosian R, Bucher A, Steben T, Rigby S, Coluni B, Bennett S. Johnson & Johnson Health & Wellness Solutions. 2015. Reimagining the health assessment URL: https://healthandwellnesssolutions.my.salesforce.com/sfc/p/ [accessed 2017-03-09] [WebCite Cache ID 6opvoCQBb]
35. Smith J, Medalia C. Census. 2014 Sep. Health insurance coverage in the United States: current population reports URL: https://www.census.gov/content/dam/Census/library/publications/2015/demo/p60-253.pdf [accessed 2017-03-17] [WebCite Cache ID 6p1hKiq5M]
36. Liu RT, Alloy LB, Abramson LY, Iacoviello BM, Whitehouse WG. Emotional maltreatment and depression: prospective prediction of depressive episodes. Depress Anxiety 2009;26(2):174-181 [FREE Full text] [doi: 10.1002/da.20545] [Medline: 19152341]

Abbreviations

ACA: Affordable Care Act
BIC: Bayesian information criterion
BMI: body mass index
CAIC: consistent Akaike information criterion
CES-D: Center for Epidemiological Studies Depression Scale
GMM: growth mixture modeling
HRA: health risk assessment
PHQ-2: Patient Health Questionnaire-2
PSS: Perceived Stress Scale
SD: standard deviation
WPAI: work productivity activity impairment

Edited by G Eysenbach; submitted 12.08.16; peer-reviewed by K Wardenaar, S Schwartz; comments to author 29.09.16; revised version received 01.12.16; accepted 19.12.16; published 31.03.17.

Please cite as:
Bedrosian R, Hawrilenko M, Cole-Lewis H
Trajectories of Depressive Symptoms Among Web-Based Health Risk Assessment Participants
J Med Internet Res 2017;19(3):e96
URL: http://www.jmir.org/2017/3/e96/
doi:10.2196/jmir.6480
PMID:28363881

Original Paper

Evaluating the Consistency of Current Mainstream Wearable Devices in Health Monitoring: A Comparison Under Free-Living Conditions

Dong Wen[1,2], B Med; Xingting Zhang[1,2], B Med; Xingyu Liu[3], B Med; Jianbo Lei[2,4], MD, PhD

[1]Peking University Third Hospital, Beijing, China

[2]Center for Medical Informatics, Peking University, Beijing, China

[3]Department of Prosthodontics, Affiliated Hospital of Stomatology, Southwest Medical University, Luzhou, China

[4]School of Medical Informatics and Engineering, Southwest Medical University, Luzhou, China

Corresponding Author:
Jianbo Lei, MD, PhD
Center for Medical Informatics
Peking University
38 Xueyuan Rd.
Haidian District
Beijing, 100191
China
Phone: 86 (10) 8280 5901
Fax: 86 (10) 8280 5900
Email: jblei@hsc.pku.edu.cn

Abstract

Background: Wearable devices are gaining increasing market attention; however, the monitoring accuracy and consistency of the devices remains unknown.

Objective: The purpose of this study was to assess the consistency of the monitoring measurements of the latest wearable devices in the state of normal activities to provide advice to the industry and support to consumers in making purchasing choices.

Methods: Ten pieces of representative wearable devices (2 smart watches, 4 smart bracelets of Chinese brands or foreign brands, and 4 mobile phone apps) were selected, and 5 subjects were employed to simultaneously use all the devices and the apps. From these devices, intact health monitoring data were acquired for 5 consecutive days and analyzed on the degree of differences and the relationships of the monitoring measurements by the different devices.

Results: The daily measurements by the different devices fluctuated greatly, and the coefficient of variation (CV) fluctuated in the range of 2-38% for the number of steps, 5-30% for distance, 19-112% for activity duration, .1-17% for total energy expenditure (EE), 22-100% for activity EE, 2-44% for sleep duration, and 35-117% for deep sleep duration. After integrating the measurement data of 25 days among the devices, the measurements of the number of steps (intraclass correlation coefficient, ICC=.89) and distance (ICC=.84) displayed excellent consistencies, followed by those of activity duration (ICC=.59) and the total EE (ICC=.59) and activity EE (ICC=.57). However, the measurements for sleep duration (ICC=.30) and deep sleep duration (ICC=.27) were poor. For most devices, there was a strong correlation between the number of steps and distance measurements (R^2>.95), and for some devices, there was a strong correlation between activity duration measurements and EE measurements (R^2>.7). A strong correlation was observed in the measurements of steps, distance and EE from smart watches and mobile phones of the same brand, Apple or Samsung (r>.88).

Conclusions: Although wearable devices are developing rapidly, the current mainstream devices are only reliable in measuring the number of steps and distance, which can be used as health assessment indicators. However, the measurement consistencies of activity duration, EE, sleep quality, and so on, are still inadequate, which require further investigation and improved algorithms.

(J Med Internet Res 2017;19(3):e68) doi:10.2196/jmir.6874

KEYWORDS

fitness trackers; monitoring, physiologic; motor activity; activities of daily living; health status

Introduction

Importance of Monitoring Physical Activities

The amount and patterns of physical activity are closely related to health status and the rehabilitation of chronic conditions. Pedometers have a significant effect on promoting physical exercise. Being sedentary is a significant risk factor for chronic conditions; independent of physical activity and other potential confounders, it is significantly associated with the development of diabetes and other chronic conditions [1]. The longer the daily sedentary time is, the higher the risk of all-cause mortality [2]. The level of activity can be determined by the number of steps per day—5000 steps per day suggest a sedentary state, whereas 10,000 steps per day suggest an active state [3]. For healthy people, increasing the number of steps reduces the risk of death [4]. For patients with chronic conditions, increased activity shows a significant rehabilitation value; for example, middle-aged diabetic patients significantly improved their insulin sensitivity by increasing the number of steps per day during a 5-year follow-up period [5]. However, studies found that elderly populations with a high incidence of chronic conditions often overestimated their physical activity compared with the actual measurement [6]; in this regard, a pedometer can accurately estimate the amount of activity to prompt the user to engage in more physical activities and significantly lower their body mass index (BMI) and blood pressure [7]. The physical activities promoted by a pedometer have been mainly based on the social cognitive theory, self-efficacy theory, and cross-theoretical model, and it was revealed that the outcome of increased physical activity has been achieved by adopting strategies such as setting goals and participating in group activities, and so on [8].

Importance of Wearable Devices

Wearable devices have richer functions than traditional pedometers and can monitor more health indicators than pedometers do and are gradually replacing traditional pedometers. Supported by a variety of sensors and the increasing computing power, they also have other health monitoring functions [9,10]. Wearable devices for everyday health monitoring are also called “fitness trackers”. Currently, sensors in fitness trackers include three-axis accelerometers, three-axis gyroscopes, three-axis magnetic sensors, optical heart rate sensors, altimeters, ambient light sensors, temperature sensors, bioelectrical impedance sensors, and capacitive sensors. The basic mechanism of the step-counting function is that the acceleration values on 3 orthogonal axes are acquired, from which the secondary wave peaks are monitored after the waves are filtered, and the number of peaks is the number of steps [11]. Compared with traditional physical measurement measures, wearable devices reduce wear discomfort, provide appropriate interactions to facilitate habit development, and capture users’ continuous movement and health data so that the fluctuating trends of users’ health characteristics are better portrayed and potential health risks are spotted in a timely manner [12,13].

Performance of Current Wearable Devices

Wearable devices are gaining increasing market attention, but the accuracy and reliability of the monitoring of the devices remains inadequate. On one hand, the number of steps, distance, and energy expenditure (EE) have been accurately calculated in a laboratory setting. For example, Takacs et al [14] evaluated the accuracies of the number of steps and speed measurements of 30 subjects wearing multiple Fitbit Ones (Fitbit Inc) under different intensities of physical activity and found that the measurement of the number of steps was reliable and valid (ICC>.95), with an error rate of less than 1.3%; the measurement of distance was reliable but sometimes invalid, particularly at low speeds (r>.90), with a rather high error rate of up to 39.6%. LEE et al [15] investigated the accuracy of EE measurements on 60 subjects wearing 8 types of wearable devices performing activities at different intensities and showed that the mean absolute percent error of different devices on EE measurements varied between 9.3% and 23.5%; moreover, the error rate of the Fitbit product was approximately 10% and had a high correlation with the standard value (r=.81); the measurements were mostly accurate. On the other hand, in the state of normal life activities, Ferguson et al [16] assessed the reliabilities of the number of steps, activity duration, sleep duration, and EE in 21 subjects continuously wearing 7 types of devices for 48 h and found that for the number of steps and sleep duration, the measurements by the consumer products, and the professional equipment exhibited strong correlations (r>.8), but only fair correlations were obtained regarding activity duration and EE (r=.52-.91). Moreover, the measurement errors of the number of steps and sleep duration were rather low (<10%), and the measurement error of EE was fair (<30%), whereas that of activity duration was rather high (26-298%). Rosenberger et al [17] investigated the error rates of the measurements of 40 subjects wearing 9 types of devices on sedentary behavior, varying degrees of physical activity, the number of steps, and total sleep duration. They showed that all the devices exhibited a rather high error on each indicator, and no device was able to accurately acquire the activity data for 24 consecutive hours; moreover, the measurement for sleep displayed the smallest error at 8.1%, and that for moderate to vigorous physical activity (MVPA) displayed the largest error at 92%. These results suggest that various monitoring measurements presented by various wearable devices should be considered with caution.

Significance of This Study

With the rapid development of new products and new functions, the consistencies of health monitoring measurements of different wearable devices in the market must continuously be verified to reach a more uniform assessment. The aforementioned studies show that the reliability of wearable devices and the accuracy and reliability of the measurements of each indicator are still problematic, and such studies are still in the early stages; moreover, emerging products and the validation of measurements in various physical activity states must be considered. In this study, newly launched and representative smart watches, internationally renowned smart bracelets, and popular smart bracelets in China were included, and for the first time, the measurements by mobile phone apps were compared with those obtained from professional equipment. Each subject continuously wore all the devices as they participated in normal life activities, and the 5-day monitoring data were acquired. The aim of the study was to evaluate the degrees of consistency of

health monitoring measurements of the mainstream wearable devices in the market and analyzing the discrete degrees and correlations of multiple monitoring measurements of the daily monitoring data under the state of normal life activities, to provide implications to the industry in overcoming difficulties in product development, and to support consumers in choosing the right product.

Methods

Selection of Devices and Indicators

Overall, 6 mainstream wearable devices and 4 major apps were chosen for this study. Representative health tracking devices were selected from 3 product categories: smart watches, smart bracelets, and mobile phone apps. First, when choosing smart watches, given that the functions of smart watches and mobile phones were close and that mobile phones from Samsung and Apple were leading products in the market, Samsung Gear S (Samsung Inc) and Apple Watch (Apple Inc) were chosen to represent the smart watches. Second, when choosing the internationally renowned smart bracelets, according to market research data by NPD and Canalys [18-20], Fitbit had the largest market share, and Jawbone also exhibited a good market performance; therefore, Fitbit Surge (Fitbit Inc) and Jawbone Up3 (Jawbone Inc) were chosen to represent the foreign-made smart bracelets. Third, when choosing the Chinese brand smart bracelets, the shipment of the Mi Band in the health tracking devices market has been second to that of Fitbit; furthermore, according to the ranking of smart bracelet products on Zhongguancun online, one of the China's IT professional websites, Huawei was on top. Thus, the Mi Band (Mi, China) and TalkBand B2 (HUAWEI, China) were chosen to represent the Chinese brand bracelets. Regarding mobile phone apps, the ranking of health and fitness apps was used as the reference, and Pacer and Ledongli were chosen as representatives. Fourth, the mobile phones used in this study were the Samsung Note 3 and the iPhone 6 Plus. To compare the branded smart watches and the preinstalled mobile phone apps, Samsung S Health apps and Apple Health apps were also included.

With respect to indicators, the measurement indicators shared by all the devices, that is, number of steps, distance, activity duration, EE, sleep duration, and deep sleep duration were chosen as the assessment criteria in this study after listing and comparing the available monitoring indicators of all the devices.

Experimental Procedures

To ensure compliance of the study, 5 subjects from the close community of the research team were recruited via convenience sampling, and the inclusion criteria were as follows: older than 18 years of age and without major critical illnesses, not allergic to rubber straps, able to continuously wear the devices, and willing to participate in the investigation. The subjects were required to wear all devices simultaneously while maintaining normal living conditions for 5 consecutive days including weekends. We hope that the type of activities were able to represent the normal living conditions of the majority of working people, so the main types of activities were supposed to include walking, climbing stairs, sedentary, and so on, as well as light to moderate exercise. In addition, the study duration required inclusion of weekends, to ensure that the forms, amounts and intensity of activities varied from weekdays and can basically represent a complete cycle of the normal activities of the normal people in a real-life environment. As wearing 6 devices simultaneously had a certain challenge, wearing precautions and error prevention measures were instructed to subjects in details. Bracelets and watches were worn on the wrist; the Mi Band bracelet and apple watch were needed to be set correctly for left-right hand mode to ensure the position and settings to maintain consistency. Samsung watch and Fitbit Surge were needed to be manually set to enter or exit the sleep mode; the phone was placed in the pocket of clothes. In addition to the Mi Band bracelet which was said to have battery power for 20 days, other devices are unified charged every night. Data were synchronized between wearable devices and mobile app once a day and the intact data were acquired from the subjects' devices for each of the 5 consecutive days. This study was approved by the Biomedical Ethics Committee of Peking University, and the subjects were aware of the purpose and process of the study.

Data Management and Analysis

During the data collection period, the subject manually opened the corresponding app to sync the data. The monitoring data recorded by each of the apps were transcribed to an Excel (Microsoft) spreadsheet at the end point of each subject's data collection and double-checked. First, a box plot was generated for each measurement indicator to observe the data distribution, and the quartiles of the upper and lower margins, upper and lower quartiles, and median were generated for the measurements of each wearable device on each indicator based on the quartile, which made it possible to visually observe the entire picture for multiple sets of data and to compare the distribution pattern of the measurements. To determine the consistency of the different wearable devices, the intraclass correlation coefficient (ICC) was calculated for each wearable device for each indicator and was used to evaluate interobserver reliability, which ranged from 0 to 1. Values lower than .4 represented a poor reliability, and values higher than .75 represented a good reliability. To observe the discrete degree of the daily measurements by different devices, the range and coefficient of variation (CV, the ratio of the standard deviation to the mean) of the measurements of each device on each indicator were calculated to eliminate the influence of the measurement scale. Second, scatter plots were generated to investigate the relationship between different measurement indicators. In addition, for the measurements of the same brand of wearable devices and mobile phone apps, correlation coefficient analysis was applied.

Results

Comparison of the Consistencies of the Measurements of Each Indicator

To visually observe the distribution differences of the measurements of various indicators by the different wearable devices, box plots were generated from the measurements for the 6 indicators (the number of steps, distance, activity duration, EE, sleep duration, and deep sleep duration) according to the

category of the devices (Figures 1-6). One important note to explain the outliers in the box plots is that there are 3 key points in the box plots representing the third quartile Q3, median, and the first quartile Q1, respectively. The upper limit is equal to Q3+1.5×(Q3−Q1) and lower limit is equal to Q1−1.5×(Q3−Q1). Outliers are those points that are beyond the upper or lower limits. In this case, the box-plot charts were designed to observe the overall characteristics or the consistency of the multiday measurements from different wearable devices, the outliers might be resulted from the measurement of someday when the subject was observed doing unusual activity, and therefore has no special significance.

In addition, as Apple Health and Samsung S Health apps cannot be used on the same phone at the same time, there are only 3 subjects in this study using Apple's Health app, 2 subjects using Samsung S Health app. Therefore, taking into account different amount of data, Apple Health and Samsung S Health were not compared with other devices. The comparison of measurements by different devices from the same brand was performed later.

The differences in the measurements of the number of steps by different devices were rather small (Figure 1); the distance measurements by the app Ledongli were significantly lower, whereas those by the other devices only differed slightly (Figure 2). The activity duration measurements by different devices differed significantly (Figure 3). The EE measurements could be divided into 2 levels: the measurements by some of the devices (Apple Watch, Jawbone Up3, and Fitbit Surge) included resting EE and activity EE, whereas measurements by the other devices (Samsung Gear S, Huawei TalkBand B2, Mi Band, Ledongli, and Pacer) were specifically activity EE (Figure 4). The sleep durations and deep sleep durations were significantly different according to the different devices (Figures 5 and 6).

For each indicator, the ICC of the measurements by different wearable devices was calculated and is shown in Table 1. The ICCs of the measurements for the number of steps and distance were higher than .8, indicating excellent consistencies by different wearable devices; the ICCs of the measurements for activity duration, total EE and activity EE indicated only fair consistencies by different wearable devices; and the ICCs of the measurements for sleep duration and deep sleep duration were lower than .4, indicating poor consistencies by different wearable devices.

Table 1. Intraclass correlation coefficient (ICC) of the measurements by different wearable devices

Items	Intraclass correlation coefficient	95% CI
Number of steps	.89	0.83-0.94
Distance	.84	0.75-0.91
Activity duration	.59	0.44-0.75
Total EE[a]	.59	0.37-0.77
Activity EE	.57	0.41-0.74
Sleep duration	.30	0.13-0.52
Deep sleep duration	.27	0.08-0.50

[a]EE: energy expenditure.

For each indicator, the CV of the daily measurements by different devices was calculated and is shown in Figures 7-13. According to the above analysis, the consistency of the number of steps was excellent; however, the CVs of the measurements of different days fluctuated greatly (2-38%), with a range of 297-8047 steps. The consistency of the distance measurements was excellent; however, the CVs of the measurements of different days also fluctuated greatly (5-30%), with the range of .4-8.7 km. For activity duration, which only had a fair consistency of measurements, the CVs of the measurements of different days varied from 19% to 112%, with a range of 22-170 min. EE also had a fair measurement consistency; the CVs of the total EE measurement of different days varied from 1% to 17%, with a range of 5-662 kcal; the CVs of activity EE measurement of different days varied from 22% to 100%, with a range of 51-706 kcal. For sleep duration and deep sleep duration, which had poor measurement consistencies, the CVs of the measurements of different days fluctuated from 2-44% and 35-117%, respectively, with a range of 29 min to 8 h 44 min for sleep duration and a range of 2 h 20 min to 7 h 40 min for deep sleep duration.

Further analysis by comparing this study to previous studies is shown in Table 2.

Table 2. Comparison with previous studies.

Indicators	Steps	Distance	Activity duration	EE[a]	Sleep duration	Deep sleep duration
This study	Excellent consistency with CV[b] (2-38%)	Excellent consistency with CV (5-30%)	Fair consistency with CV (19-112%)	Fair consistency with CV for total EE (.1-17%) and activity EE CV (22-100%)	Poor consistency with CV (2-44%)	Poor consistency with CV (35-117%)
Previous studies	Consumer-grade wearable devices provided consistently similar step counts with research-grade devices for average daily activity (P>.05) [21]	The inter-device reliability of wearable devices in measuring distance was excellent for all treadmill speeds (ICC[c]≥.90) [14]	Consumer-level wearable devices showed moderate validity for measurement of moderate to vigorous physical activity in free-living conditions (r=.52-.91) [16]	Consumer-level wearable devices showed moderate validity for measurement of total daily EE in free-living conditions (r=.74-.81) [16]	Consumer-level wearable devices showed strong validity for measurement of sleep duration in free-living conditions (r>.8) [16]	Consumer-grade wearable devices showed good agreements with PSG[d] for sleep efficiency, and they overestimated PSG sleep efficiency slightly [22]
	The inter-device reliability of wearable devices in measuring steps in free-living conditions is good (ICC≥.90) [23]	Distance errors in wearable devices were within 5% in level walking, and they overestimated distance for stair walking by at least 45% [24]	Consumer-grade wearable devices can't accurately capture activity data across the entire 24-h day, error rates ranged from 51.8% to 92% for moderate to vigorous physical activity [17]	Consumer-grade wearable devices reasonably and reliably estimate EE during walking and running (ICC≥.95) [25]	Consumer-grade wearable devices performed consistently compared with each other (reliability=96.5-99.1%), and they overestimated sleep time by an average of 67.1 min compared with PSG [26]	Consumer-grade wearable devices performed consistently compared with each other (reliability=96.5%-99.1%), and they overestimated sleep efficiency by an average of 14.5% compared with PSG [26]

[a]EE: energy expenditure.

[b]CV: coefficient of variation.

[c]ICC: intraclass correlation coefficient.

[d]PSG: Polysomnography.

Figure 1. Box plot of the measurement distribution of the number of steps by different devices.

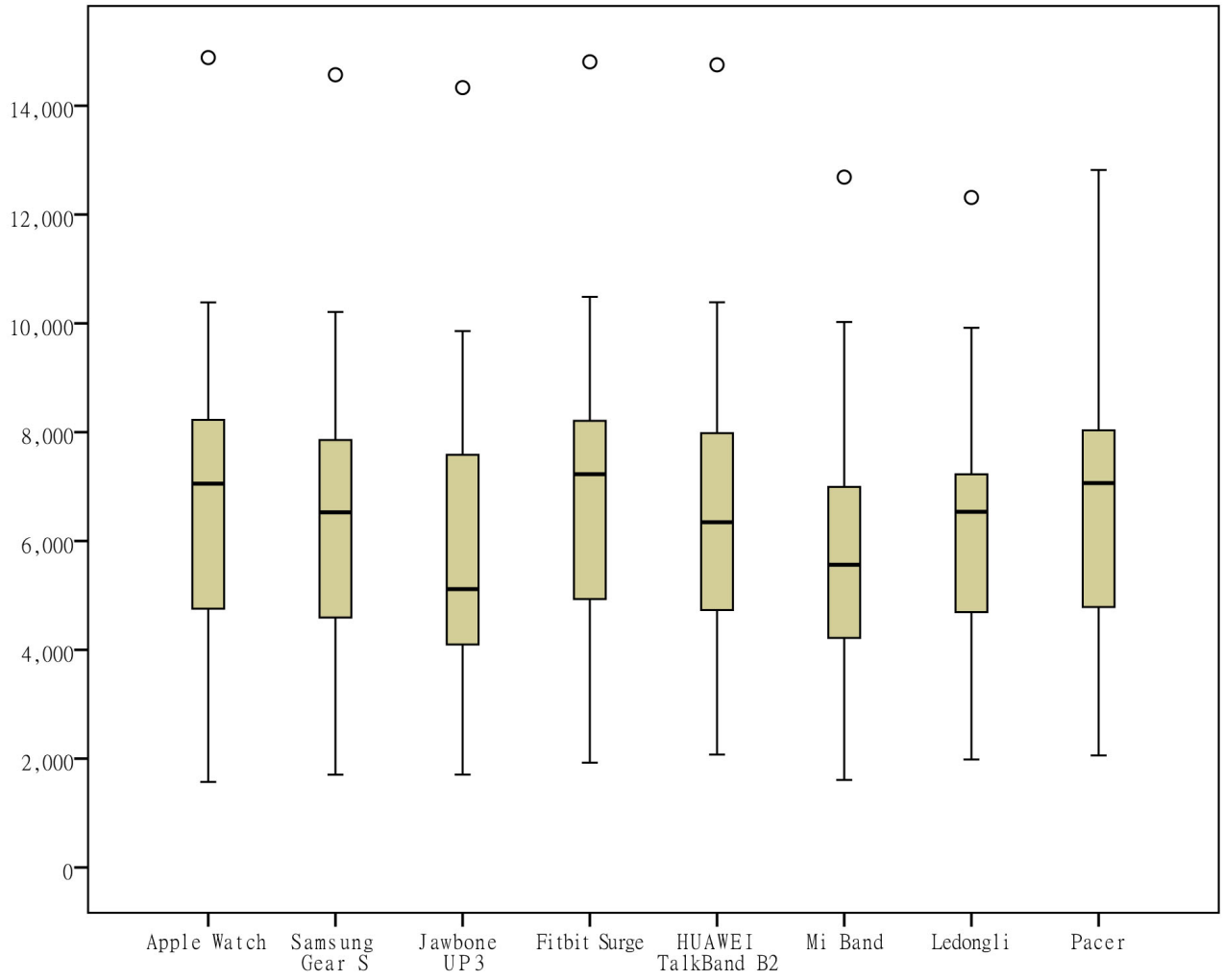


Figure 2. Box plot of the measurement distribution of distance by different devices.

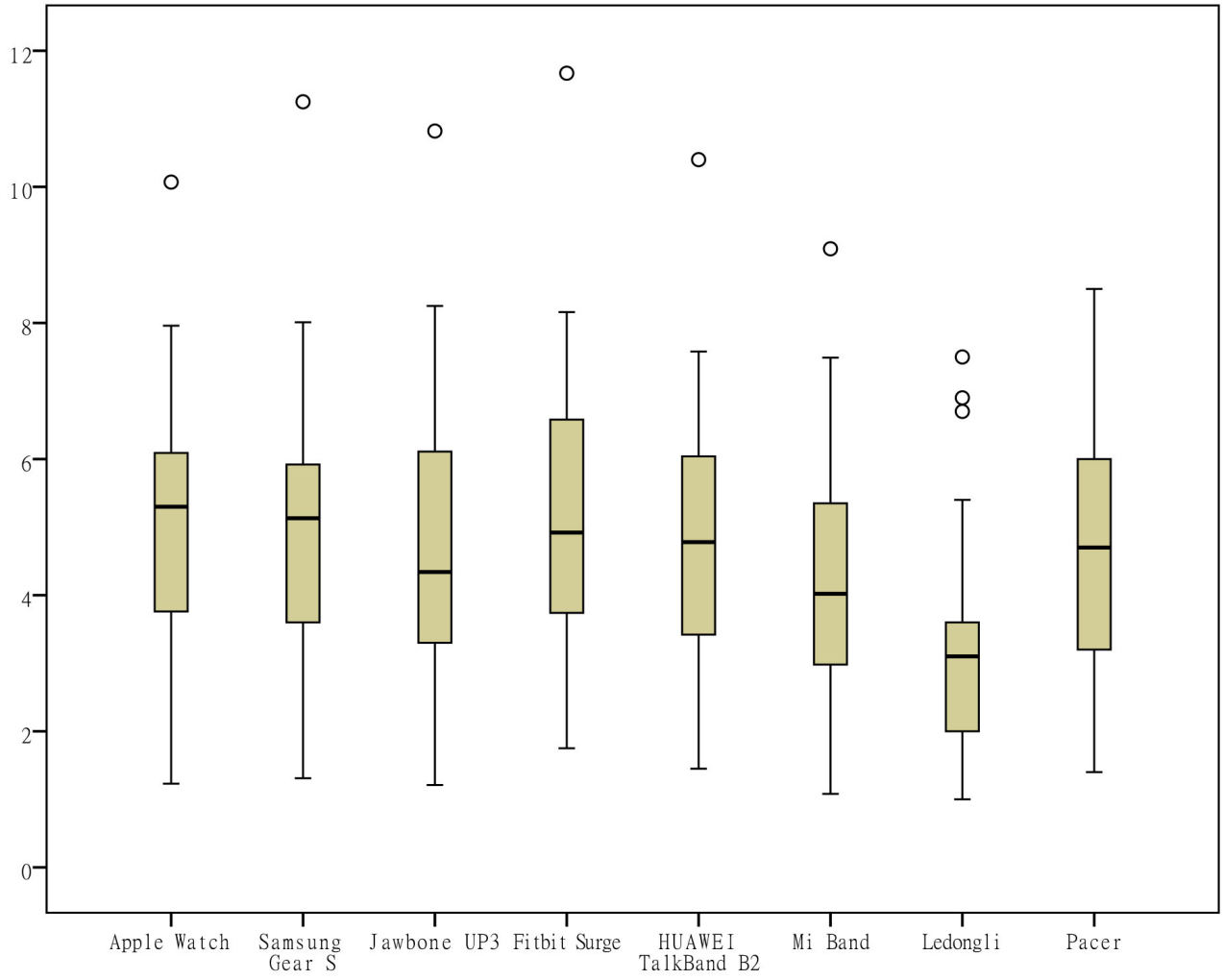


Figure 3. Box plot of the measurement distribution of activity duration by different devices.

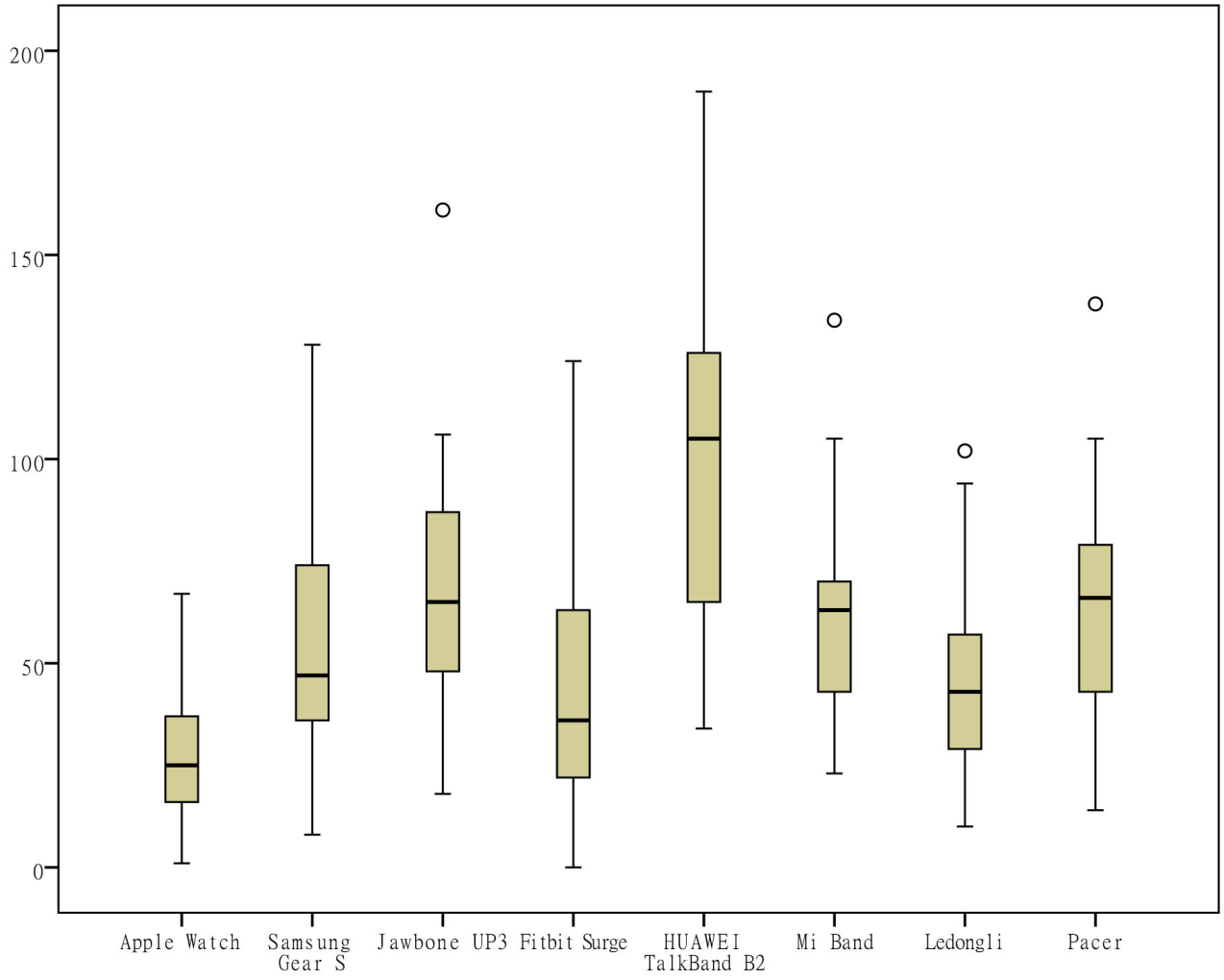


Figure 4. Box plot of the measurement distribution of energy expenditure by different devices.

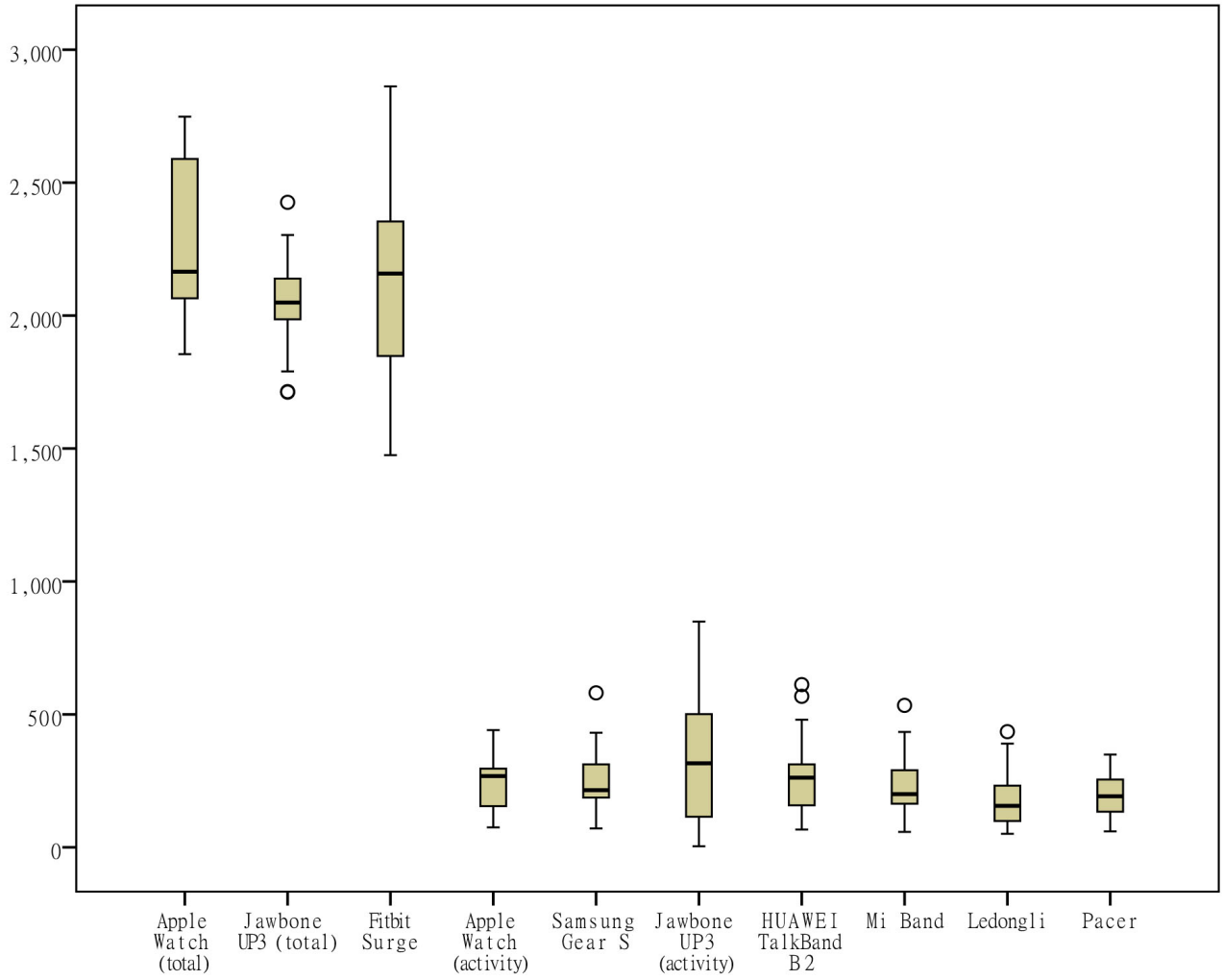


Figure 5. Box plot of the measurement distribution of sleep duration by different devices.

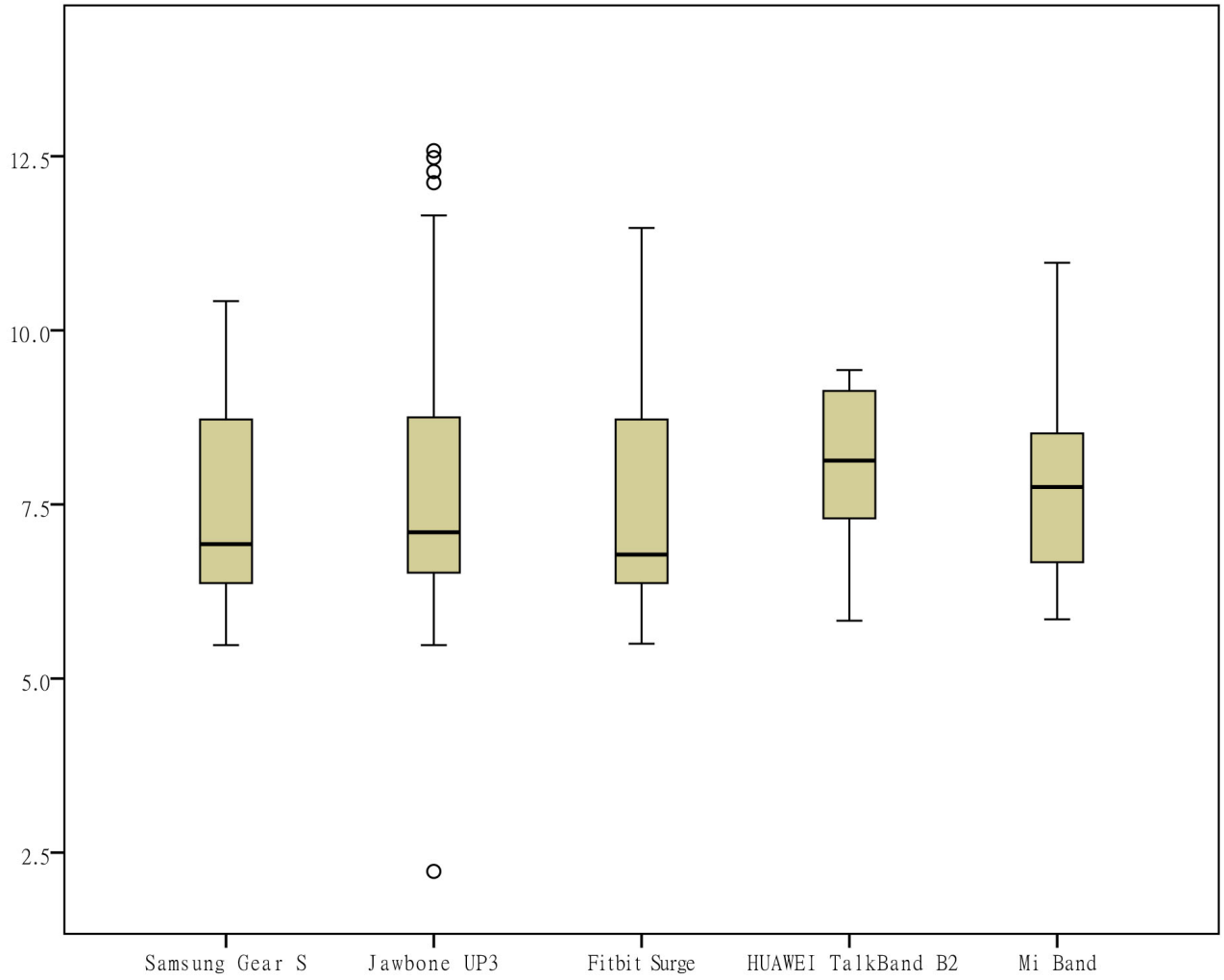


Figure 6. Box plot of the measurement distribution of deep sleep duration by different devices.

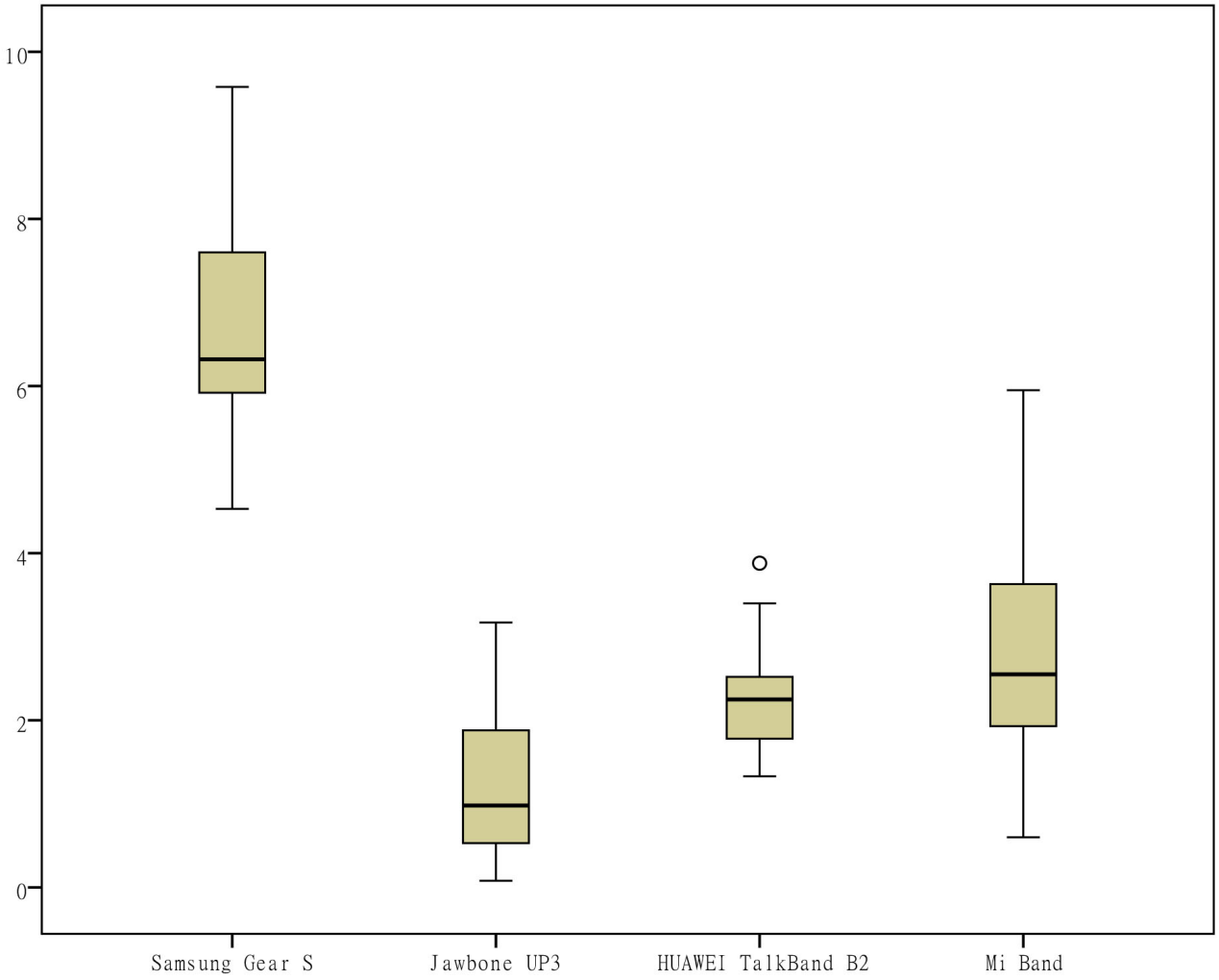


Figure 7. Coefficient of variation of daily measurements of the number of steps by different devices.

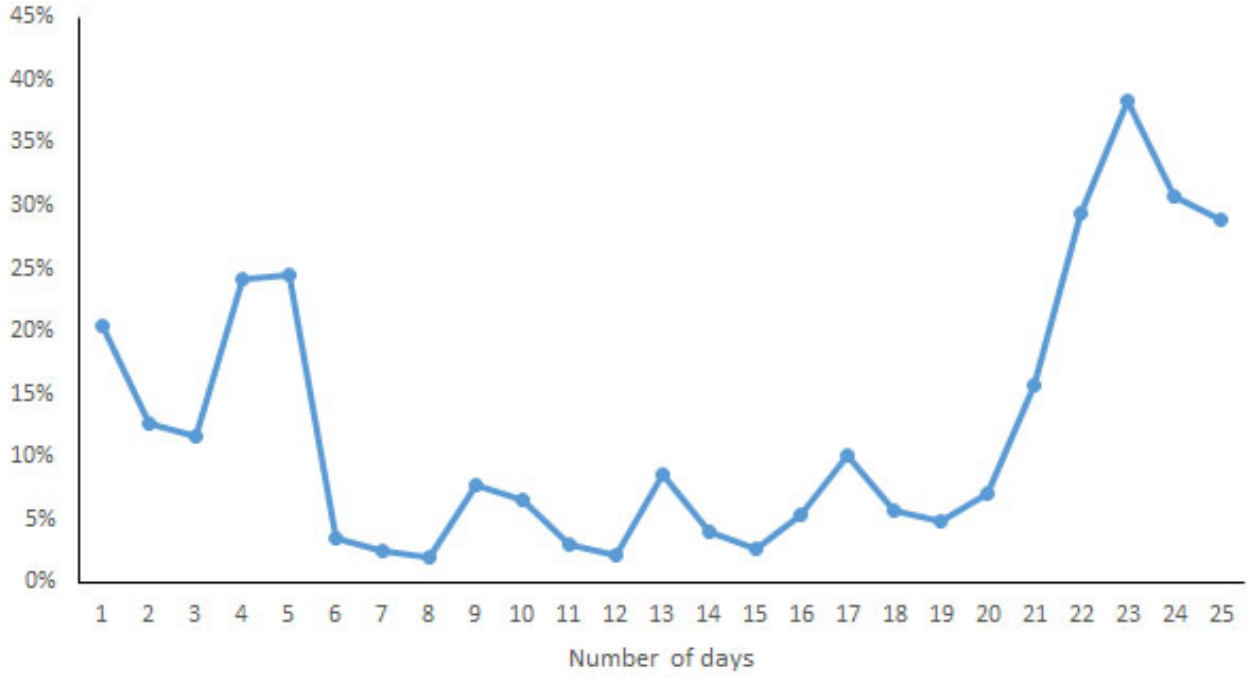


Figure 8. Coefficient of variation of daily measurements of distance by different devices.

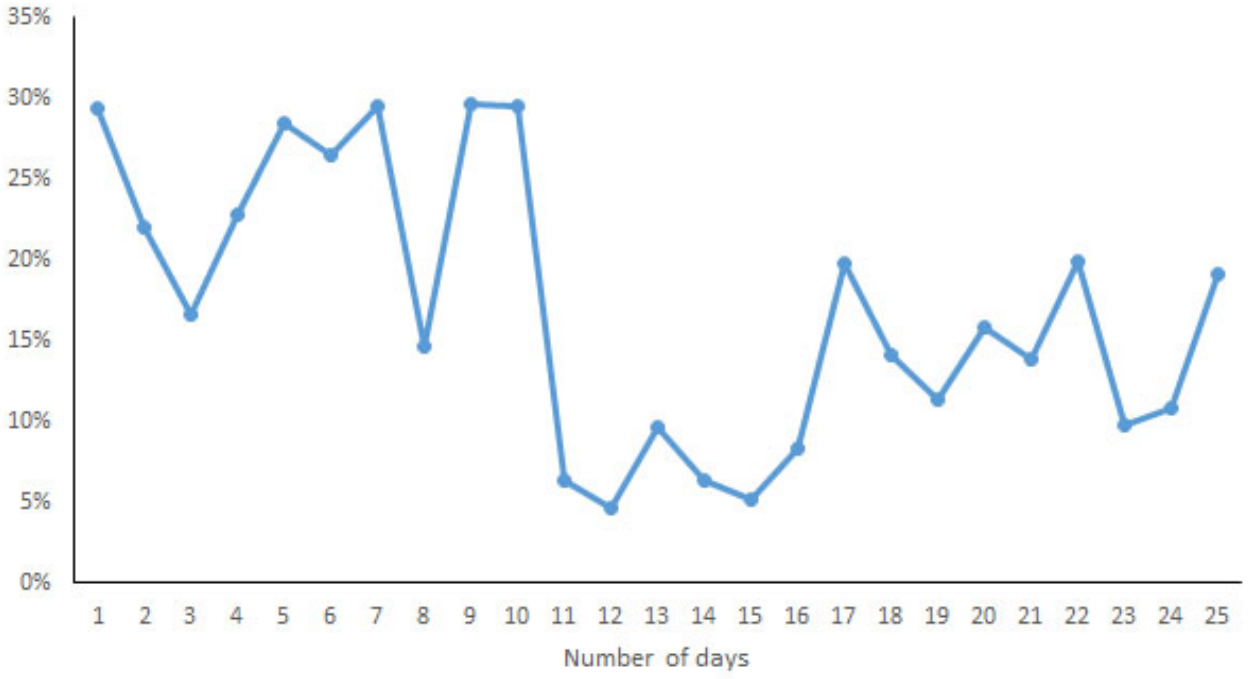


Figure 9. Coefficient of variation of daily measurements of active duration by different devices.

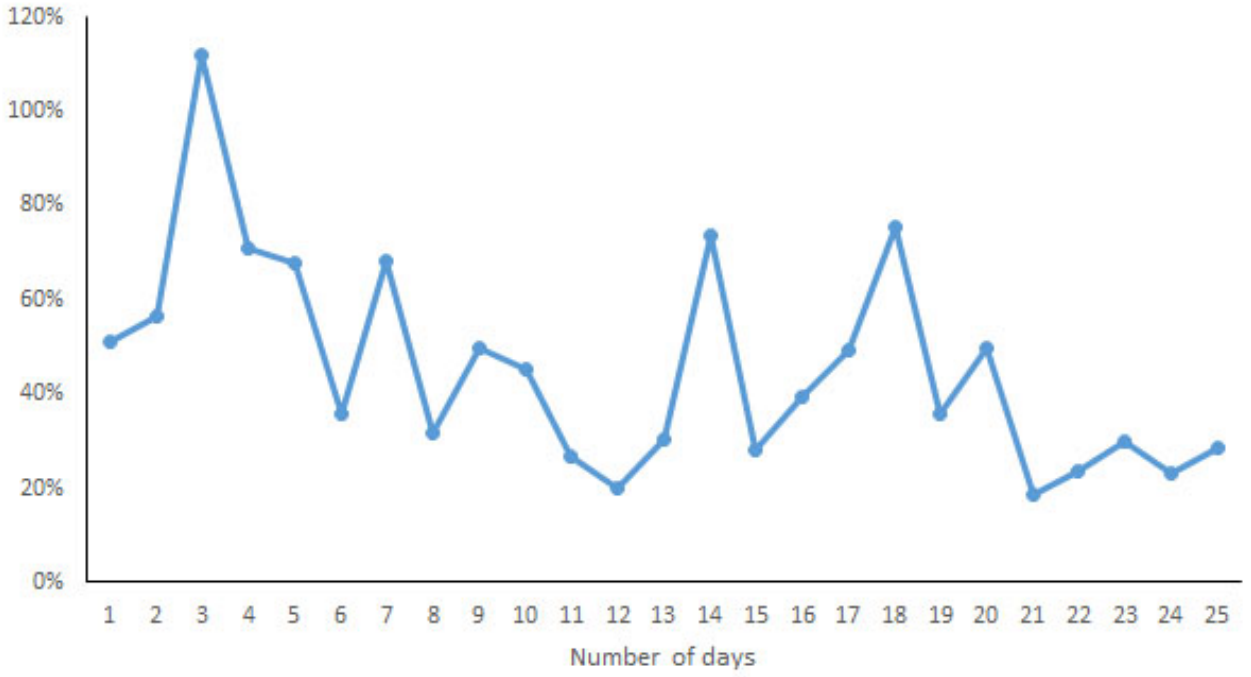


Figure 10. Coefficient of variation of daily measurements of total energy expenditure by different devices.

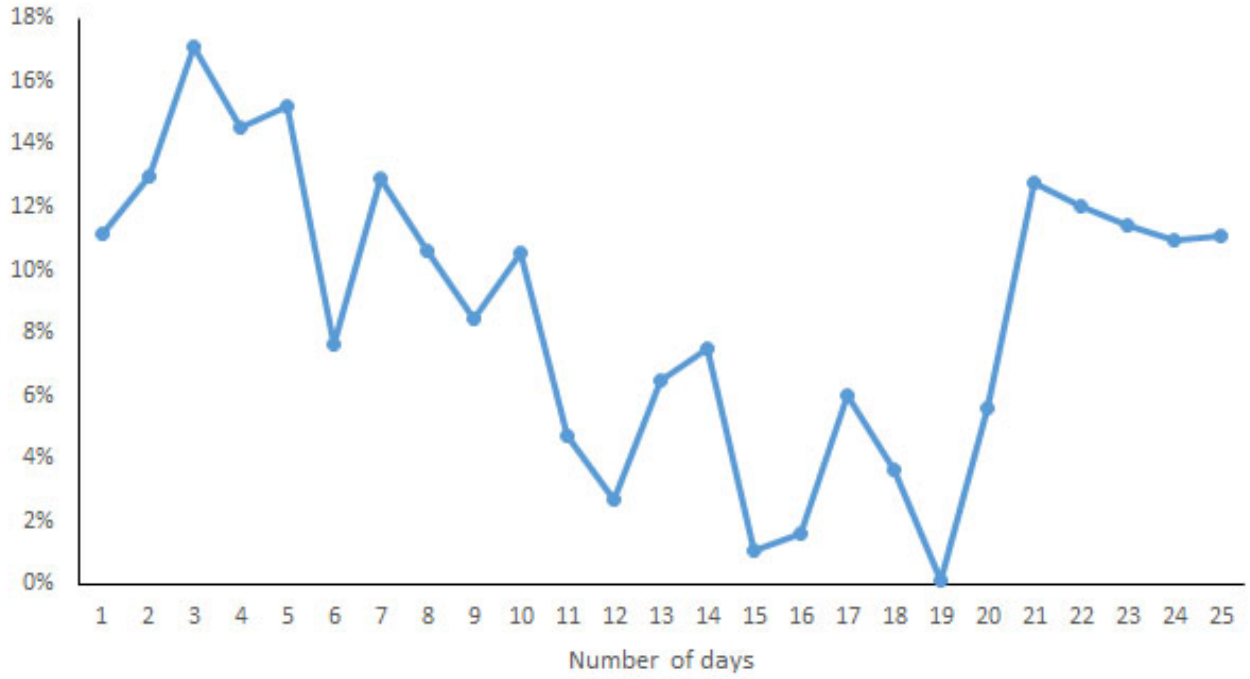


Figure 11. Coefficient of variation of daily measurements of activity energy expenditure by different devices.

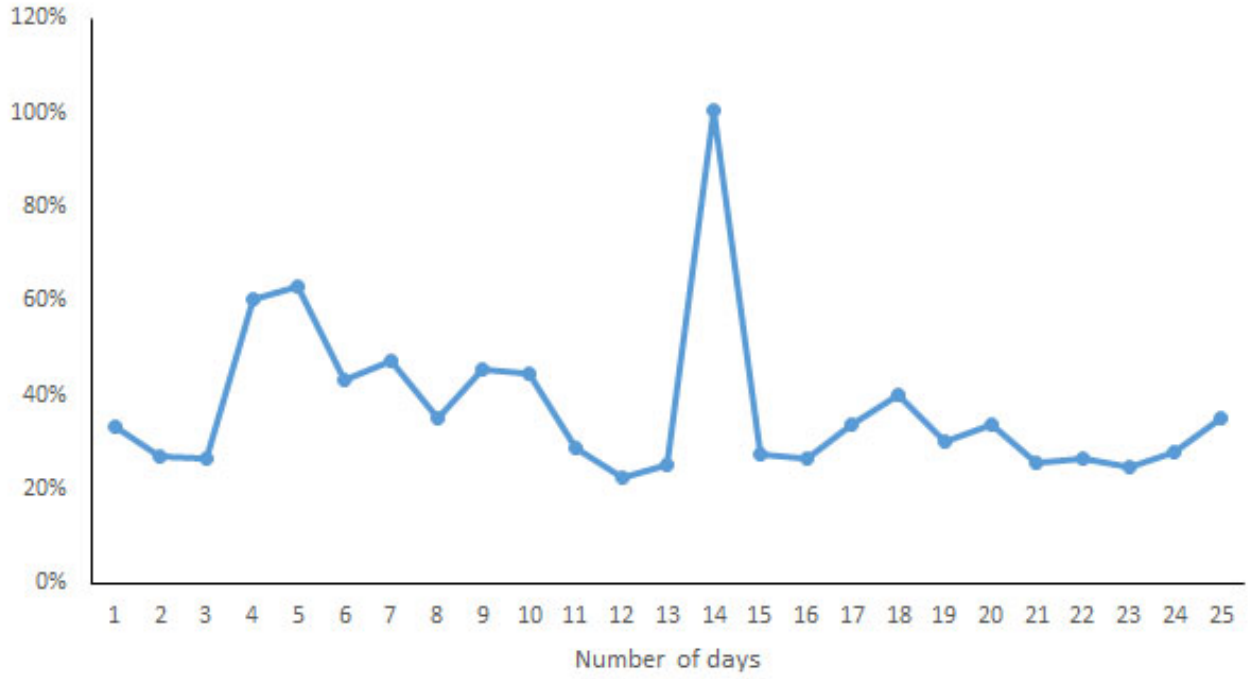


Figure 12. Coefficient of variation of daily measurements of sleep duration by different devices.

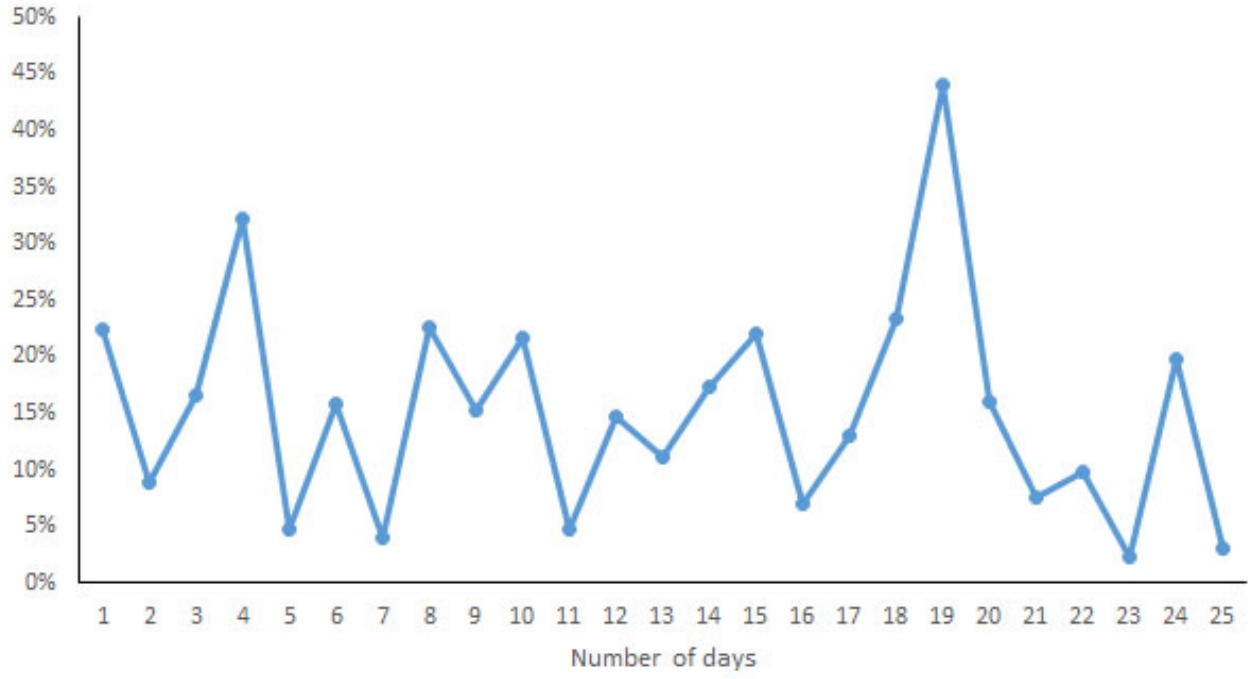
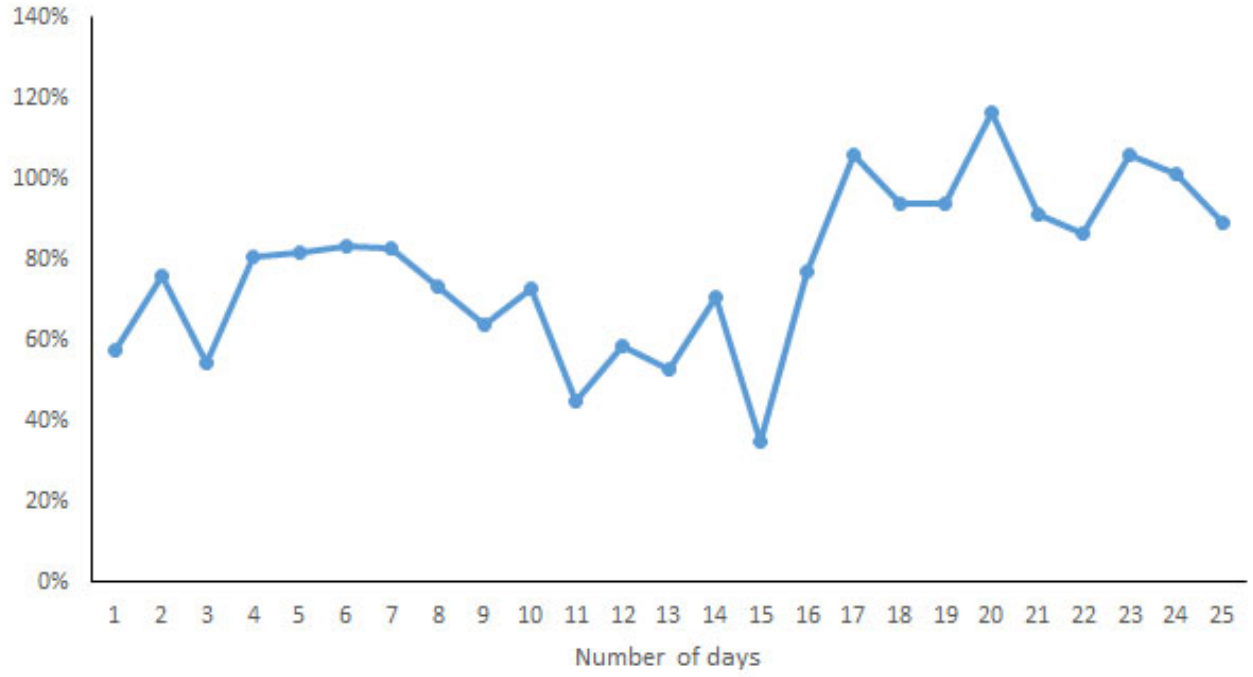


Figure 13. Coefficient of variation of daily measurements of deep sleep duration by different devices.



The Relationship Between Different Measurement Indicators

There were correlations between the number of steps and distance and between activity duration and EE. Given that most wearable devices measure activity data based on three-axis acceleration sensors and that the ICCs of the measurements on the number of steps and distance were close, as were the ICCs of the measurements on activity duration and EE, a scatter plot was generated for each device on the number of steps and distance (Figure 14). The results showed that except for Ledongli and Pacer, the measurements of the number of steps and distance by the other devices exhibited a strong positive linear correlation (R^2>.95), whereas the correlation between the two indicators measured by the third-party apps showed some differences. A scatter plot of activity duration and EE measurements by each device was generated and is shown in Figure 15. Except for Apple Watch, Fitbit Surge, and Huawei

Wrist bracelet, the measurements of activity duration and EE by the other devices showed a positive correlation (R^2>.7).

Figure 14. Scatter plot for each device showing measurements of the number of steps and distance.

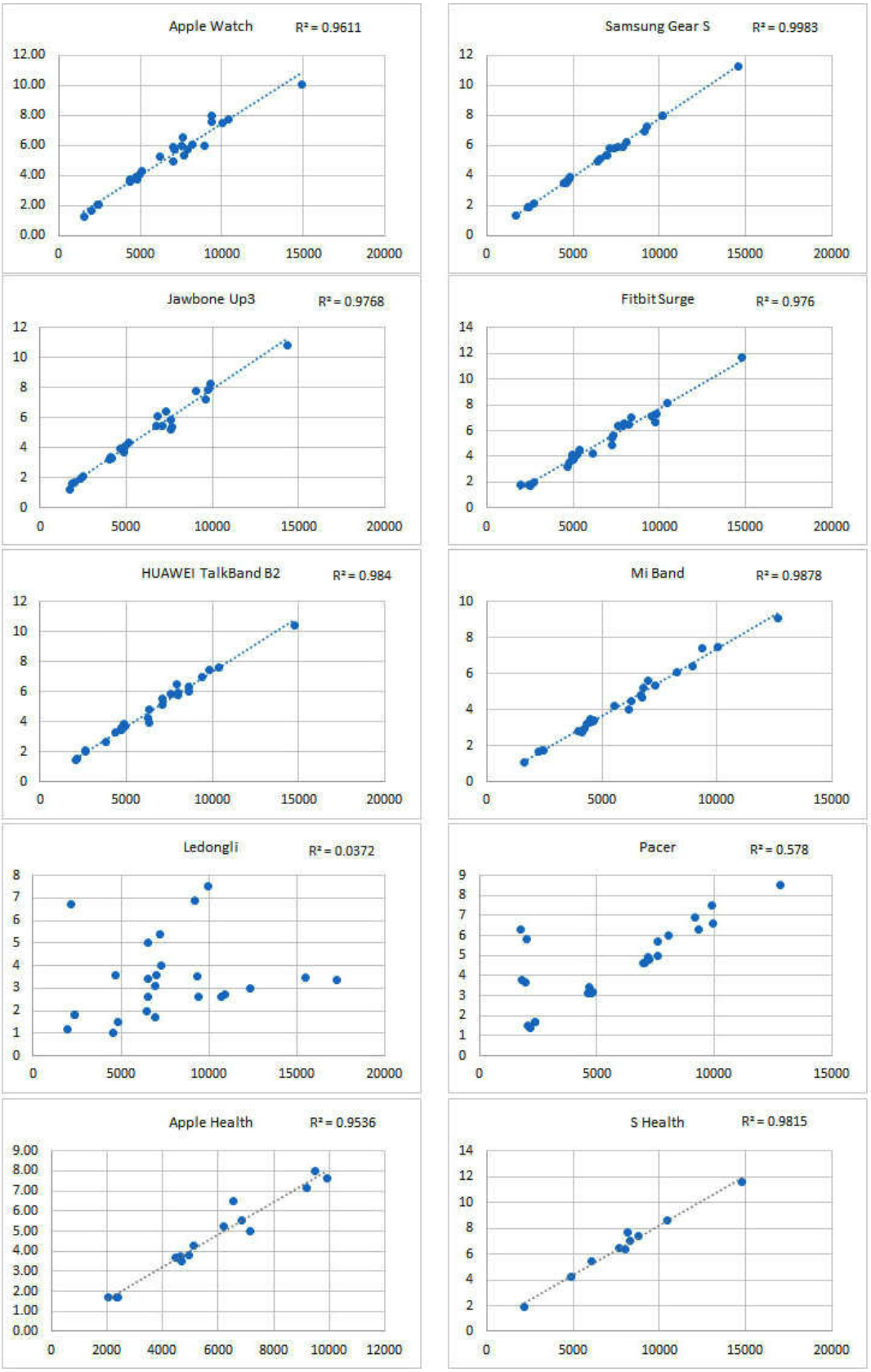
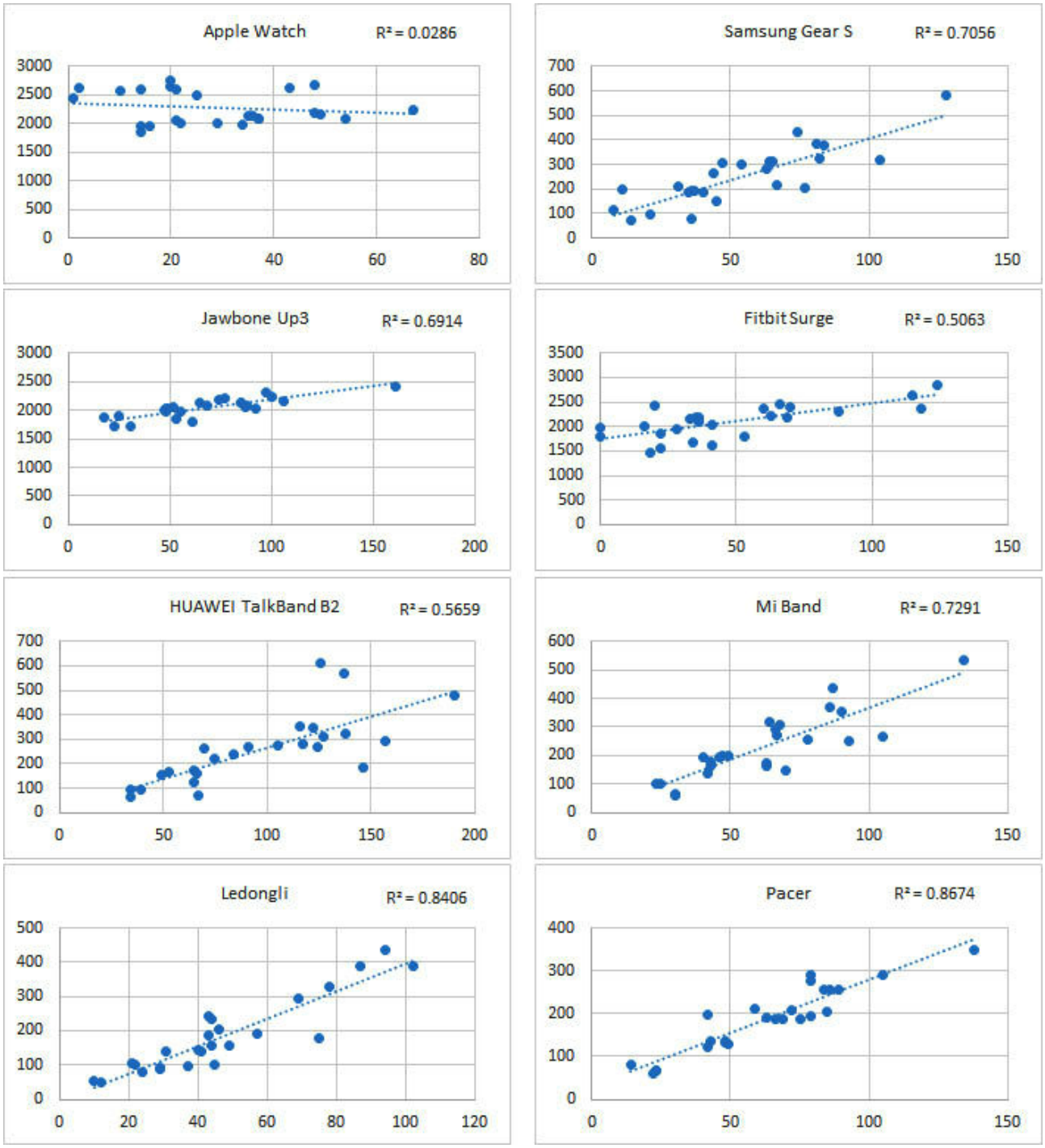


Figure 15. Scatter plot for each device showing the measurements of activity duration and energy expenditure.



Comparison of the Consistency of Measurements by Different Devices of the Same Brand

A strong correlation between different devices of the same brand was observed (r≥.88). The correlation coefficients of different measures for mobile phones and smart watches are showed in Table 3.

Table 3. Correlation coefficient on measurements by mobile phones and smart watches of the same brand.

Brand	Steps	Distance	EE[a]
Apple	.99	.99	.96
Samsung	.96	.91	.88

[a]EE: energy expenditure.

Discussion

Overview of This Study

In this study, the activity and sleep data of 5 subjects were collected on several consecutive days from the mainstream wearable devices and mobile phone apps in the state of everyday life activities to examine the consistency of the measurements by different wearable devices. The monitoring measurements by the mainstream wearable devices in the market showed a rather large daily variation on the measurements of the number of steps and distance, but the overall consistencies in the continuous measuring was excellent. Activity duration and EE displayed fair measurement consistencies, whereas the same for sleep duration and deep sleep duration were poor. In terms of the discrete level, the daily measurements on each indicator by different devices fluctuated to a certain degree. Furthermore, in all the wearable devices, there was a strong correlation between the number of steps and distance measured by the same device, and in some wearable devices and mobile apps, there was a rather strong correlation between measurements of activity duration and EE measured by the same device.

Compared with previous studies, this study has some unique innovations. First, in this study, the market representing smart watches, the internationally renowned smart bracelets, the Chinese-made smart bracelets, mobile phone health apps, and third-party health mobile phone apps were integrated in the same comparison system, which makes it possible to effectively compare the measurement differences derived from different devices under the same conditions, and one representative innovative result was that the number of steps and distance displayed excellent consistencies. The measurements by the third-party apps and the correlation between the 2 indicators were different from the values by the other devices. Second, in this study, activity and sleep data were collected from each subject for 5 days while participating in natural life activities, and all the subjects were members of the research group who were familiar with the content of the study. The continuous use of the wearable device ensured the subject to be familiar with sync rules of multiple devices, which effectively avoided the errors caused by inappropriate wearing of the device. Third, in this study, for the first time, the leading products in the market, smart watches and mobile phone apps of the same brand (Apple or Samsung), were compared in terms of measurement consistency. It is believed that products of the same manufacturer have similar technical optimization; however, it was found that the measurements by different devices of the same brand exhibited differences, likely due to factors such as hardware support and wear habits on the measurements.

However, there are 2 small limitations in this study. First, because the study was designed to monitor data in the subject's state of normal life activities, it was difficult to obtain the gold standard through the single research-level device for each indicator. Therefore, we intended to analyze the consistency of measurements on the same indicator by different wearable devices. If the gold standard control were available, the accuracy of each device would be obtained, which will be examined in our subsequent studies. In addition, because each subject was required to continuously wear 6 pieces of wearable devices and 1 mobile phone in the daily activities of a multiday duration while observing certain rules of use, it presented a certain burden of wear for the subject; thus, only the members of the close community of our research group were recruited under the principle of convenience sampling, in a very limited sample size. A series of measures were taken to reduce selection bias and ensure a relative complete normal activity cycle of an average person, such as the emphasis of maintaining original daily activities to the participants, as well as the addition of the weekends into the study duration during which more activities might be observed. Besides, the data analysis showed that the number of steps and distance had an excellent measurement consistency; moreover, the variance homogeneities of the number of steps (P=.96) and distance (P=.85) by different wearable devices were rather good, indicating that in terms of these 2 indicators, the degrees of discrepancy of the measurements by different wearable devices were not significantly different. Therefore, the data imbalance problem arising from the sampling was not significant, and the results were valid.

Consistency of Measurements for Wearable Devices

Previous studies have shown that the measurements on the number of steps and sleep duration were reliable, whereas for other indicators, the errors were high. Dontje et al [23] asked one subject to wear 10 pieces of Fitbit Ultra in the state of natural life for 8 consecutive days and showed that when comparing minutes, hours, and days, the reliabilities of the number of steps by multiple devices were excellent and the best when compared in terms of the number of steps per day (ICC=1). Ferguson et al [16] assessed the reliabilities of the number of steps, activity duration, sleep duration, and EE of 21 subjects wearing 7 devices for 48 h in a state of natural activity and found that the measurements for the number of steps (r=.94-.99) and sleep duration (r=.82-.92) were very reliable, whereas they were fairly reliable for total EE (r=.74-.81) and the duration of MVPA (r=.52-.91); however, the reliabilities of the measurements by different devices showed significant differences. Diaz et al [25] investigated 23 subjects with 4 pieces of Fitbit One worn on the waist and 2 pieces of Fitbit Flex worn on each wrist during the activities of 4 speed phases and showed that the measurements of the number of steps by all the devices were strongly correlated with the actual number and the measurement error was underestimated by 16.3% at the highest; moreover, the measurements of EE by all the devices were strongly correlated with the actual number, and the measurement error was overestimated by 52.4% at the highest; the measurements obtained by different devices were strongly correlated and highly reliable. Bai et al [27] asked 52 subjects to wear 5 pieces of consumer-level devices and 2 pieces of research-level devices and examined the accuracy of EE measurements during optional activities of various types and found that in general, the percentage of EE measurement error ranged from 15.3% to 30.4%, and the measurements by different devices had a generally high correlation with the gold standard (.71-.90). There were 3 devices that had a percentage of EE measurement error of less than 20% when resting; 2 devices had a percentage of EE measurement error less than 20% for

aerobic exercise, and for anaerobic exercise, the percentages of measurement error of all the devices were higher than 25%. Massimiliano et al [22] showed that the measurements for sleep duration, sleep efficiency, and wake-up duration by Jawbone Up exhibited excellent consistencies compared with the measurements obtained by Polysomnography (PSG), although there were still some errors and a certain degree of overestimation on sleep duration and efficiency. Meltzer et al [28] compared the accuracies of Fitbit Ultra and PSG on 63 subjects and found that in the normal mode, the sensitivity and accuracy of Fitbit Ultra were excellent but the specificity was rather poor; it significantly overestimated sleep duration and sleep efficiency. In the sensitive mode, the specificity was adequate but the sensitivity and accuracy were inadequate; it significantly underestimated sleep duration and sleep efficiency. The above studies show that the reliability of the measurement on the number of steps was rather high, whereas the measurements for activity duration and EE had a high error; the measurements for sleep were rather reliable, although in general, sleep duration was overestimated.

The conclusions of this study are consistent with the conclusions of previous studies but differ regarding sleep duration. This study found that the measurements for the number of steps and distance were reliable, whereas the measurements for the other indicators were not; however, previous studies found that the measurements on sleep duration were reliable. Further analysis of the CVs of the daily measurements on sleep duration by each device revealed that the average CV for all days was 17% and that of deep sleep duration, which also had a poor measurement consistency, was 74%. The data showed that the measurements on sleep duration by different devices varied little, leading to a low CV, which is consistent with the findings in previous reports that the errors in the measurement of sleep duration by different devices were low. However, the relative values of the measurements by different devices on different times fluctuated, leading to unstable performances of different devices in repeated measurements and poor measurement consistencies.

Significance of Measurements by Wearable Device

The number of steps and distance are reliable indicators of health evaluation for wearable devices. Studies have previously shown that daily activities such as walking and sitting showed a significant causal relationship with health and chronic condition rehabilitation [3-5]. In this study, we found that except for Ledongli, the consistencies of the measurements of number of steps and distance were excellent and could provide reliable judgment on the individuals' activity amount. Furthermore, except for Ledongli and Pacer, the measurements of the number of steps and distance showed a strong correlation, likely because distance was calculated from a linear function based on the total number of steps. However, in each device, there were discrete values that deviated from the regression line that were mainly derived from differences in the subjects' activity habits. The activity monitoring sensor of the devices included in this study was mainly the three-axis accelerometer, and it is believed that in addition to the users' initial base data that must be collected, the dynamic data on which the monitoring rely were generated from the real-time signal change of the acceleration. Moreover, the fitting function of each device was different, and the measurements by different devices differed due to factors such as the included variables, algorithm models, and so on.

Wearable devices are less reliable for measuring activity duration and EE. A variety of wearable devices do not specify activity duration. Academic studies often determine the presence of MVPA based on the metabolic equivalent of energy (EE at rest or sitting) [27,28]. Thus, moderate physical activity refers to 3-6 METs, requiring a moderate degree of motion and significant heart rate acceleration; vigorous physical activity refers to more than 6 METs, requiring a large amount of movement that lead to rapid breathing and a rapid increase in heart rate [29]. Due to their health benefits, MVPA are often used as a public health indicator to assess the level of activity in the population [27,28]. This study also found that activity duration and EE exhibited a strong correlation in the majority of devices, but the consistencies of the measurements on activity duration and EE by each device were rather poor, indicating that the 2 indicators are not suitable for the evaluation of activity. In addition, some of the devices differentiated resting EE and activity EE, and most of the devices only measured activity EE, whereas resting EE could actually reach 2000-3000 kcal, and the EE of everyday activity was less than 500 kcal. Basal metabolism may be affected by diet, temperature, endocrine factors, and so on; thus, resting EE is a dynamic value under the action of a variety of internal and external environmental factors. However, wearable devices mainly monitor the state of motion, so it is still doubtful whether the current wearable devices are able to provide information on resting EE.

Wearable devices are rather rudimentary on monitoring sleep. According to the Rechtschaffen and Kales classification, sleep can be divided into the rapid eye movement phase and the nonrapid eye movement phase; the fourth period of the nonrapid eye movement phase is the deep sleep stage, which has the high amplitude brain wave, mainly the delta-wave with a frequency of 1-2 times/s, and promotes physical and mental recovery [30]. The majority of the devices in this study could automatically determine the time points at which the user fell asleep and woke up; however, the measured sleep duration increased at varying degrees compared with the standard reference sleep duration, likely because of the level of activity on the bed during the time periods of falling asleep and waking up, which led to the device misreading the time periods as still being asleep. A few wearable devices were rather reliable in determining sleep duration, but the consistencies of the measurements by various devices were inadequate, and this function requires further calibration.

Conclusions

The consistencies of the number of steps and distance by wearable devices were excellent, and the 2 indicators can be used in health evaluations, whereas the consistencies of the measurements on activity duration, EE, sleep duration, and deep sleep duration were only fair or poor. These will directly affect consumers' acceptance of wearable devices and require the manufacturers' close attention and resolution as well.

Acknowledgments

This study was partly supported by the National Natural Science Foundation of China (NSFC) Grant #81471756.

Conflicts of Interest

None declared.

References

1. George ES, Rosenkranz RR, Kolt GS. Chronic disease and sitting time in middle-aged Australian males: findings from the 45 and Up Study. Int J Behav Nutr Phys Act 2013 Feb 08;10:20 [FREE Full text] [doi: 10.1186/1479-5868-10-20] [Medline: 23394382]
2. Chau JY, Grunseit AC, Chey T, Stamatakis E, Brown WJ, Matthews CE, et al. Daily sitting time and all-cause mortality: a meta-analysis. PLoS One 2013;8(11):e80000 [FREE Full text] [doi: 10.1371/journal.pone.0080000] [Medline: 24236168]
3. Tudor-Locke C, Bassett DR. How many steps/day are enough? Preliminary pedometer indices for public health. Sports Med 2004;34(1):1-8. [Medline: 14715035]
4. Dwyer T, Pezic A, Sun C, Cochrane J, Venn A, Srikanth V, et al. Objectively measured daily steps and subsequent long term all-cause mortality: the Tasped Prospective Cohort Study. PLoS One 2015;10:e0141274. [doi: 10.1371/journal.pone.0141274] [Medline: 26536618]
5. Dwyer T, Ponsonby AL, Ukoumunne OC, Pezic A, Venn A, Dunstan D, et al. Association of change in daily step count over five years with insulin sensitivity and adiposity: population based cohort study. BMJ 2011 Jan 13;342(jan13 1):c7249-c7249 [FREE Full text] [doi: 10.1136/bmj.c7249] [Medline: 21233153]
6. Vallanceff J, Eurich D, Gardiner P, Taylor L, Johnson S. Associations of daily pedometer steps and self-reported physical activity with health-related quality of life: results from the Alberta Older Adult Health Survey. J Aging Health 2016 Dec;28(4):661-674. [doi: 10.1177/0898264315609905] [Medline: 26486783]
7. Bravata DM, Smith-Spangler C, Sundaram V, Gienger AL, Lin N, Lewis R, et al. Using pedometers to increase physical activity and improve health: a systematic review. J Am Med Assoc 2007 Nov 21;298(19):2296-2304. [doi: 10.1001/jama.298.19.2296] [Medline: 18029834]
8. Tudor-Locke C, Bassett DR. How Many Steps/Day Are Enough? Sports Medicine 2004;34(1):1-8. [doi: 10.2165/00007256-200434010-00001] [Medline: 14715035]
9. Ridgers ND, McNarry MA, Mackintosh KA. Feasibility and effectiveness of using wearable activity trackers in youth: a systematic review. JMIR Mhealth Uhealth 2016 Nov 23;4(4):e129 [FREE Full text] [doi: 10.2196/mhealth.6540] [Medline: 27881359]
10. Sanders JP, Loveday A, Pearson N, Edwardson C, Yates T, Biddle SJ, et al. Devices for self-monitoring sedentary time or physical activity: a scoping review. J Med Internet Res 2016 May 04;18(5):e90 [FREE Full text] [doi: 10.2196/jmir.5373] [Medline: 27145905]
11. Neil Z. Analog. Full-featured pedometer design realized with 3-Axis digital accelerometer URL: http://www.analog.com/en/analog-dialogue/articles/pedometer-design-3-axis-digital-acceler.html [accessed 2016-10-23] [WebCite Cache ID 6lTDQrFBF]
12. Gualtieri L, Rosenbluth S, Phillips J. Can a free wearable activity tracker change behavior? The impact of trackers on adults in a physician-led wellness group. JMIR Res Protoc 2016 Nov 30;5(4):e237 [FREE Full text] [doi: 10.2196/resprot.6534] [Medline: 27903490]
13. Mercer K, Li M, Giangregorio L, Burns C, Grindrod K. Behavior change techniques present in wearable activity trackers: a critical analysis. JMIR Mhealth Uhealth 2016;4(2):e40 [FREE Full text] [doi: 10.2196/mhealth.4461] [Medline: 27122452]
14. Takacs J, Pollock CL, Guenther JR, Bahar M, Napier C, Hunt MA. Validation of the Fitbit One activity monitor device during treadmill walking. J Sci Med Sport 2014 Sep;17(5):496-500. [doi: 10.1016/j.jsams.2013.10.241] [Medline: 24268570]
15. Lee J, Kim Y, Welk GJ. Validity of consumer-based physical activity monitors. Med Sci Sports Exerc 2014 Sep;46(9):1840-1848. [doi: 10.1249/MSS.0000000000000287] [Medline: 24777201]
16. Ferguson T, Rowlands AV, Olds T, Maher C. The validity of consumer-level, activity monitors in healthy adults worn in free-living conditions: a cross-sectional study. Int J Behav Nutr Phys Act 2015;12:42 [FREE Full text] [doi: 10.1186/s12966-015-0201-9] [Medline: 25890168]
17. Rosenberger ME, Buman MP, Haskell WL, McConnell MV, Carstensen LL. 24 hours of sleep, sedentary behavior, and physical activity with nine wearable devices. Med Sci Sports Exerc 2015 Oct 17:457-465. [doi: 10.1249/MSS.0000000000000778] [Medline: 26484953]
18. Canalys. 2014. Fitbit accounted for nearly half of global wearable band shipments in Q1 2014 URL: https://www.canalys.com/newsroom/fitbit-accounted-nearly-half-global-wearable-band-shipments-q1-2014 [WebCite Cache ID 6oTu1yu8N]
19. Canalys. 2014. Fitbit accounted for nearly half of global wearable band shipments in Q1 URL: https://www.canalys.com/newsroom/media-alert-fitbit-maintains-leadership-share-wearable-band-market-apple-watch-entrance [accessed 2017-02-23] [WebCite Cache ID 6oTssqPie]

20. Mobihealthnews. 2013. Fitbit, Jawbone, Nike had 97 percent of fitness tracker retail sales URL: http://www.mobihealthnews.com/28825/fitbit-jawbone-nike-had-97-percent-of-fitness-tracker-retail-sales-in-2013/ [accessed 2016-10-23] [WebCite Cache ID 6lTEfStMh]
21. Dominick GM, Winfree KN, Pohlig RT, Papas MA. Physical activity assessment between consumer-and research-grade accelerometers: a comparative study in free-living conditions. JMIR Mhealth Uhealth 2016 Sep 19;4(3):e110 [FREE Full text] [doi: 10.2196/mhealth.6281] [Medline: 27644334]
22. de Zambotti M, Baker FC, Colrain IM. Validation of sleep-tracking technology compared with polysomnography in adolescents. Sleep 2015 Sep 01;38(9):1461-1468 [FREE Full text] [doi: 10.5665/sleep.4990] [Medline: 26158896]
23. Dontje ML, de Groot M, Lengton RR, van der Schans CP, Krijnen WP. Measuring steps with the Fitbit activity tracker: an inter-device reliability study. J Med Eng Technol 2015;39(5):286-290. [doi: 10.3109/03091902.2015.1050125] [Medline: 26017748]
24. Huang Y, Xu J, Yu B, Shull PB. Validity of FitBit, Jawbone UP, Nike+ and other wearable devices for level and stair walking. Gait Posture 2016 Jul;48:36-41. [doi: 10.1016/j.gaitpost.2016.04.025] [Medline: 27477705]
25. Diaz KM, Krupka DJ, Chang MJ, Peacock J, Ma Y, Goldsmith J, et al. Fitbit®: an accurate and reliable device for wireless physical activity tracking. Int J Cardiol 2015 Apr 15;185:138-140. [doi: 10.1016/j.ijcard.2015.03.038] [Medline: 25795203]
26. Montgomery-Downs HE, Insana SP, Bond JA. Movement toward a novel activity monitoring device. Sleep Breath 2012 Sep;16(3):913-917. [doi: 10.1007/s11325-011-0585-y] [Medline: 21971963]
27. Bai Y, Welk GJ, Nam YH, Lee JA, Lee J, Kim Y, et al. Comparison of consumer and research monitors under semistructured settings. Med Sci Sports Exerc 2015 Jul 6:151-158. [doi: 10.1249/MSS.0000000000000727] [Medline: 26154336]
28. Meltzer LJ, Hiruma LS, Avis K, Montgomery-Downs H, Valentin J. Comparison of a commercial accelerometer with polysomnography and actigraphy in children and adolescents. Sleep 2015 Aug;38(8):1323-1330 [FREE Full text] [doi: 10.5665/sleep.4918] [Medline: 26118555]
29. WHO. WHO. 2016. Global Strategy on Diet, Physical Activity and Health URL: http://www.who.int/dietphysicalactivity/physical_activity_intensity/en/ [accessed 2016-10-23] [WebCite Cache ID 6lTFMFb3j]
30. de Zambotti M, Claudatos S, Inkelis S, Colrain IM, Baker FC. Evaluation of a consumer fitness-tracking device to assess sleep in adults. Chronobiol Int 2015;32(7):1024-1028 [FREE Full text] [doi: 10.3109/07420528.2015.1054395] [Medline: 26158542]

Abbreviations

BMI: body mass index
CV: coefficient of variation
EE: energy expenditure
ICC: intraclass correlation coefficient
MET: metabolic equivalent of energy
MVPA: moderate to vigorous physical activity
PSG: polysomnography

Edited by G Eysenbach; submitted 23.10.16; peer-reviewed by JJ Mira, C Fernández; comments to author 24.11.16; revised version received 17.12.16; accepted 11.02.17; published 07.03.17.

Please cite as:
Wen D, Zhang X, Liu X, Lei J
Evaluating the Consistency of Current Mainstream Wearable Devices in Health Monitoring: A Comparison Under Free-Living Conditions
J Med Internet Res 2017;19(3):e68
URL: http://www.jmir.org/2017/3/e68/
doi:10.2196/jmir.6874
PMID:28270382

Original Paper

Ecological Momentary Assessment in Behavioral Research: Addressing Technological and Human Participant Challenges

Lora E Burke[1,2], MPH, PhD; Saul Shiffman[3], PhD; Edvin Music[1,3], MSIS, MBA; Mindi A Styn[1,3], PhD; Andrea Kriska[2], PhD; Asim Smailagic[4], PhD; Daniel Siewiorek[5], PhD; Linda J Ewing[3], PhD; Eileen Chasens[1], DNS; Brian French[6], MS; Juliet Mancino[3], MSc; Dara Mendez[2], PhD; Patrick Strollo[3], MD; Stephen L Rathbun[6], PhD

[1]Department of Health & Community Systems, University of Pittsburgh School of Nursing, Pittsburgh, PA, United States

[2]Department of Epidemiology, University of Pittsburgh Graduate School of Public Health, Pittsburgh, PA, United States

[3]University of Pittsburgh School of Medicine, Pittsburgh, PA, United States

[4]Institute for Complex Engineered Systems, Carnegie Mellon University, Pittsburgh, PA, United States

[5]Human Computer Interaction Institute, Carnegie Mellon University, Pittsburgh, PA, United States

[6]Department of Epidemiology and Biostatistics, University of Georgia College of Public Health, Athens, GA, United States

Corresponding Author:
Lora E Burke, MPH, PhD
Department of Health & Community Systems
University of Pittsburgh School of Nursing
415 Victoria Building
3500 Victoria Street
Pittsburgh, PA, 15261
United States
Phone: 1 412 624 2305
Fax: 1 412 383 7293
Email: lbu100@pitt.edu

Abstract

Background: Ecological momentary assessment (EMA) assesses individuals' current experiences, behaviors, and moods as they occur in real time and in their natural environment. EMA studies, particularly those of longer duration, are complex and require an infrastructure to support the data flow and monitoring of EMA completion.

Objective: Our objective is to provide a practical guide to developing and implementing an EMA study, with a focus on the methods and logistics of conducting such a study.

Methods: The EMPOWER study was a 12-month study that used EMA to examine the triggers of lapses and relapse following intentional weight loss. We report on several studies that informed the implementation of the EMPOWER study: (1) a series of pilot studies, (2) the EMPOWER study's infrastructure, (3) training of study participants in use of smartphones and the EMA protocol and, (4) strategies used to enhance adherence to completing EMA surveys.

Results: The study enrolled 151 adults and had 87.4% (132/151) retention rate at 12 months. Our learning experiences in the development of the infrastructure to support EMA assessments for the 12-month study spanned several topic areas. Included were the optimal frequency of EMA prompts to maximize data collection without overburdening participants; the timing and scheduling of EMA prompts; technological lessons to support a longitudinal study, such as proper communication between the Android smartphone, the Web server, and the database server; and use of a phone that provided access to the system's functionality for EMA data collection to avoid loss of data and minimize the impact of loss of network connectivity. These were especially important in a 1-year study with participants who might travel. It also protected the data collection from any server-side failure. Regular monitoring of participants' response to EMA prompts was critical, so we built in incentives to enhance completion of EMA surveys. During the first 6 months of the 12-month study interval, adherence to completing EMA surveys was high, with 88.3% (66,978/75,888) completion of random assessments and around 90% (23,411/25,929 and 23,343/26,010) completion of time-contingent assessments, despite the duration of EMA data collection and challenges with implementation.

Conclusions: This work informed us of the necessary preliminary steps to plan and prepare a longitudinal study using smartphone technology and the critical elements to ensure participant engagement in the potentially burdensome protocol, which spanned 12

months. While this was a technology-supported and -programmed study, it required close oversight to ensure all elements were functioning correctly, particularly once human participants became involved.

(J Med Internet Res 2017;19(3):e77) doi:10.2196/jmir.7138

KEYWORDS

ecological momentary assessment; relapse; obesity; smartphone; eating behavior; adherence

Introduction

Ecological momentary assessment (EMA) assesses individuals' current experiences, behaviors, and moods as they occur in real time and in their real-world settings [1]. Studies employing EMA methods have become more common in recent years, partially spurred by the ubiquitous availability of mobile devices and wearable sensors that provide access to individuals in their natural environment. This growth in EMA studies has expanded the array and scope of behaviors being studied. However, there continues to be an emphasis on substance abuse, particularly smoking [2-4]; other areas of focus include chronic pain [5-7], physical activity [8,9], lapses among dieters [10,11], and eating behaviors [11-14].

While a typical study may assess behaviors and emotions by retrospective questionnaires, this approach misses the specific situations that precipitate behavior changes. Most behaviors are dynamic and may change frequently throughout the day depending on context and social setting. The benefit of using EMA in these studies is that, through frequent assessments of ongoing life circumstances, EMA permits the estimation of risk of antecedents to the occurrence of a specific behavior. Unlike instruments or methods that are abstracted from the context in which events occur, EMA is able to capture environmental influences in a direct and immediate manner, which provides a direct examination of the mechanisms linking the immediate environment with risk, and, with aggregation, can also capture broader processes that may undermine behavior change [10,15]. EMA has made significant contributions to the behavioral sciences by describing dynamic changes in behaviors across contexts and environments in the everyday lives of study participants. However, despite its numerous benefits and the enhancement of this method by the technology available today, the burden that real-time, longitudinal data capture can place on the participant and the research team is significant.

We are using our weight loss relapse study as an exemplar to describe how we approached the EMA study. Relapse and weight regain are major issues in the treatment of obesity [16], yet knowledge of how relapses occur following intentional weight loss is rather limited. This is, in part, because of the dearth of prospective studies in the literature and the methodological limitations of the few studies reported [17,18]. To ensure the reader has an understanding of the major constructs discussed in our study of weight loss and relapse, we provide an operational definition of the terms that are the focus in several studies using EMA: temptation, lapse, and relapse. A *temptation* is a desire or sudden urge to eat something that was not in the individual's eating plan for that day or that would result in his or her exceeding the daily calorie or fat gram goal. A *lapse* means acting on that temptation and eating the food that fell outside the plan or goal, or that left a person feeling that he or she cheated on his or her diet. If the participant planned to eat a special dessert and budgeted his or her calories for this indulgence, it would not be a lapse, but if it occurred spontaneously without regard for their daily goal, it would be considered a lapse [11,19-21]. Lapses are contained, time-limited events. In contrast, a *relapse* is repeated episodes of lapsing, or what has been described as a return to previous behavior [19]. Marlatt and Gordon's [22] cognitive behavioral model of relapse underscored the processes by which an initial lapse could lead to a full-blown relapse; for example, eating a calorie-dense meal or dessert could lead to the person returning to their previous eating behaviors. The individual's cognitive and affective responses to a lapse are critical determining factors of whether the lapse will deteriorate into a relapse.

While the field of EMA studies has expanded rapidly, and the assessment of EMA has progressed from use of paper diaries to wireless devices, the literature is void of practical or procedural guidance on how to develop and implement an EMA study that assesses single or multiple behaviors over brief or extended periods. The purpose of this paper is to present practical steps and lessons learned in developing and implementing an EMA study, with a focus on the methods and logistics of conducting such a study. We also include strategies to ensure adequate adherence to EMA prompts; thus, we report data on adherence to daily EMA prompts and compare these data with those reported by studies of varying duration and frequency of EMA prompting. We include data from the EMPOWER study, which was a 12-month study that used EMA to examine the triggers of lapses and relapse following intentional weight loss. All participants were provided standard behavioral treatment for weight loss, which provided the background for the study of relapse. Standard behavioral treatment includes lifestyle modification, an approach that includes reduced energy intake, increased energy expenditure, and behavioral change strategies taught and practiced in groups [23]. The core behavioral change strategies are based on social cognitive theory and include goal setting, self-monitoring, cognitive restructuring, self-efficacy enhancement, and social support with feedback and guidance provided by interventionists to assist with development of problem-solving skills [23-28]. We prompted participants daily to complete EMA surveys at the beginning of the day (BOD) and end of the day (EOD) and also at random times during waking hours.

Methods

Given the theoretical and methodological reasons to collect EMA data using mobile devices, collection in this manner raises considerable practical considerations and challenges. The following section includes lessons learned from previous studies,

which were extremely important in informing the development of the EMPOWER study.

Part 1: Steps to Develop and Implement an EMA Study and Its Supporting Infrastructure

Defining an EMA Daily Data Collection Protocol

EMA employs 3 types of data collection protocols: *eventcontingent*; *signalcontingent* or " *random*"; and *timecontingent* [29]. The EMPOWER study used all 3 types of data collection protocols.

Individuals were instructed to initiate an *event-contingent* entry in the smartphone-based EMA app when some predefined event had occurred, such as a strong temptation to overeat, or a lapse (eg, having acted on a temptation and eaten some food inconsistent with a planned diet).

We used *signal-contingent* assessments scheduled at random times to obtain a representative sample of participants' moods and environments over the course of their study participation. These are "signal-contingent" assessments because the participant responds to a signal delivered at random, such as a beep from the device. Since signal-contingent assessments were delivered according to a known probability-based sampling design, we aggregated data to obtain unbiased estimates of the mean levels of the individuals' moods or levels of energy throughout the duration of the study. Moreover, we combined information from event-contingent assessments with that from the signal-contingent assessments to estimate risk of a salient event as a function of antecedents to the events, such as mood and environment [1,30,31].

Time-contingent assessment prompted the individual to make an entry at a fixed time, such as the beginning of each day, to assess the previous night's sleep. This process can be described as sampling a data collection event. Each of the 3 types of data collection protocols captures a data point on a target behavior, but any can be further sampled, such as additional questions posed regarding emotional state or environmental context. We provide more detail on architecting the assessment questions and response options, as part of finalizing the EMA protocol in the section on pilot studies.

Preliminary Studies With Electronic Diaries

Whatever the theoretical and methodological reasons to collect EMA data and to use mobile electronic devices to do so, such studies raise considerable practical considerations and challenges. The EMPOWER study was built on what our team learned from a clinical trial and pilot studies testing various devices and EMA sampling strategies. From 2004 to 2009, we conducted a 24-month clinical trial (Self-Monitoring And Recording using Technology [SMART]) that used personal digital assistants (PDAs), which required carrying the PDA and a cell phone [32]. We learned that carrying only 1 device was important to most people and that people had difficulty keeping 2 devices charged. Thus, our previous experience with technology and conducting 3 pilot studies testing various phones and EMA sampling strategies was extremely important in facilitating the success of the EMPOWER study.

Pilot Study 1: Test of Basic Infrastructure

The data collection architecture was a 3-tiered design with distributed Android (Google, Mountain View, CA, USA) app clients communicating with a public-facing Web server backed by an Oracle database (Oracle Corporation, Redwood Shores, CA, USA). To ensure that all pieces of technology (ie, smartphone, server, and database) worked in a synchronous and efficient way, we conducted thorough in-house testing. Research and management staff with varied technology skills and experiences used different makes and models of Android smartphones to identify as many bugs as possible related to differences in the Android operating system versions. We tested all aspects of end-to-end data flow, including sending interview data to the database, checking for updates to study scheduling parameters, and checking for app updates to resolve bugs. We also tested the functioning of the EMA algorithm focused mainly on correctness of the scheduling of EMA events; that is, that events were scheduled at the correct times and that the scheduled events actually occurred when intended.

Lessons learned: After a thorough 2-week in-house test, we learned more about the infrastructure needs to support data flow. In particular, we added reporting of more phone side events, such as phone restarts, and scheduled times of interviews to provide a more complete picture of the participants' interactions with the phones to aid in identifying when error conditions (such as scheduled interviews that never fired) occur. The multiple phone model testing allowed us to resolve a few model-specific bugs in the EMA app, but also let us identify which phone manufacturers and phone models behaved the most reliably. As many study participants had to upgrade to a compatible Android phone to join the study, we recommended they upgrade to one of the models that we found to be the most reliable, if possible [33].

Pilot Study 2: Test of EMA on Different Phone Models

We conducted a small pilot study with former participants (N=16) in weight loss studies conducted in our laboratory [34,35]. The purpose of the study was to evaluate acceptability and feasibility of using smartphones to collect EMA data. We provided the participants with a smartphone (Android; or iOS, Apple Inc, Cupertino, CA, USA) and training session on how to use the phone. We programmed the smartphones to prompt the participants 6 times per day at random times between 8:00 AM and 9:00 PM, and the participants were instructed to initiate a survey if they experienced a strong temptation to "go off their eating plan," or if they acted on this temptation and had a lapse.

Lessons learned: Initially, we sampled individuals 8 times per day with EMA prompts and received feedback from 10 of the 16 participants that this was too intensive. They felt that no more than 5 times per day was acceptable. Therefore, we reduced the prompts to 5 per day. We sampled participants on average for 21 days. Response to random prompts was approximately 71% and to end-of-day prompts, approximately 55%.

Acceptability: Overall, 100% of the participants strongly agreed or agreed that it was easy to use the smartphone; 75% (12/16) strongly agreed that 5 EMA prompts per day were adequate,

and 69% (11/16) said they would consider participating in a study lasting 6-12 months.

Lessons related to iOS: Since we did not have access to the full functionality of the iOS, we could not wake the phone and schedule EMA prompts; thus, the app was always running in the background and quickly drained the battery power. We learned that using phones with the iOS operating system required us to provide participants with 2 chargers, as the person needed to charge the phone during the day. In this pilot phase, we used first-generation iOS phones; however, because we did not have access to the operating system codes beyond the first generation of these phones, we could not program the EMA on later generations of this phone [36]. The Android operating system proved more accommodating; however, this may not be an issue today if the EMA programming is done on the server side.

Lessons learned related to personal use of phones: Overall, the pilot sample of 16 and the longer study period of 21 days permitted us to improve our backend systems to capture new issues more efficiently and act upon them in a timely manner, such as the impact of using other features or apps on the phone. We provided participants' phones with all features available for their use, unlike some studies that provided phones with everything locked except the EMA features. Participants tested the EMA app using study-provided phones, which helped us refine the questions and delivery of prompts, and also reinforced the notion that people who did not regularly use their mobile phones were more likely to forget about them, thus leaving them at home or letting the battery drain. Thus, we screened individuals for being a user of a mobile phone.

Pilot Study 3: Testing and Refining EMA Items

We conducted a pilot study through an anonymous Web-based questionnaire (N=133) with individuals who had previously participated in weight loss studies conducted in our laboratory [37]. The purpose of the study was to test and refine the item content for 4 types of EMA surveys: event-contingent, signal-contingent, and beginning-of-day and end-of-day surveys. We presented participants with draft assessments, and they provided feedback on the clarity of the instructions and assessment items.

Lessons related to testing prompts and refining EMA items: Participants thought the EMA items were acceptable but expressed concern about the volume of prompts they received and how to enter data to complete the EMA assessments. We learned that it was essential to train participants how to adequately complete the prompts and manage signals, such as using snooze mode to delay a reply. Figure 1 shows screenshots of EMA items.

Figure 1. Sample screenshots of ecological momentary assessment questions: (a) randomly scheduled, (b) self-initiated, (c) beginning of day, (d) end of day.

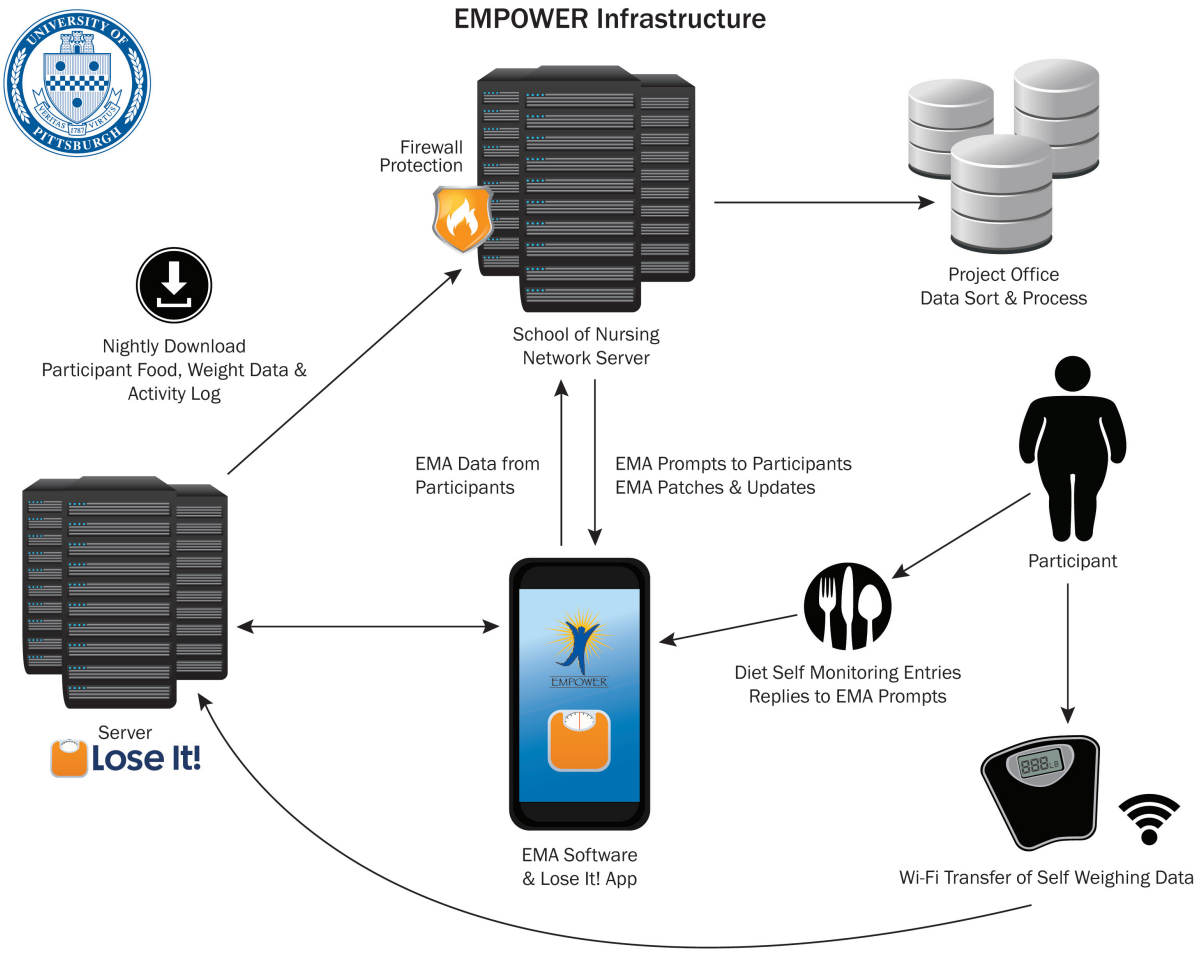
Summary of Pilot Studies

We learned a great deal in the series of pilot studies; most salient was the issue of the restrictive nature of Apple Inc to vet an app into iTunes. Thus, we decided not to pursue programming the EMA app on devices that had iOS. Concurrently, the Android-based smartphone was gaining in acceptability and, thus, we chose to use this model for the EMA programming. It also became clear that, regardless of the individual's technology expertise or skill, all participants would need training in the various aspects of providing data via a smartphone.

Part 2: EMPOWER Study: Applying the Lessons Learned in Developing the Infrastructure

The series of studies and lessons described above resulted in our developing the infrastructure that supported the EMPOWER study. The EMA data collection system comprised 3 primary components: an Android smartphone, a Web server, and a database server (Figure 2).

Figure 2. EMPOWER study infrastructure supporting collection of self-monitoring, weight, and ecological momentary assessment (EMA) data.



To support real-time data collection using an Android-based app, Wi–Fi-enabled weight scales, and self-monitoring third-party data through a commercially available self-monitoring app (Lose It!, FitNow, Inc, Boston, MA, USA), we went through the following steps in the development process: (1) determined the data flow to and from all sources, (2) identified storage and server processing solutions to support the data flow, (3) developed EMA survey apps and determined how to deliver updates to smartphones in real time, (4) acquired server solutions from the University of Pittsburgh and access to third-party data (Lose It!), (5) conducted thorough internal and pilot testing, and (6) developed proper oversight and troubleshooting plans for day-to-day operations and monitoring of the system.

Data Flow

It was critical for us to define the data flow from all sources that would later drive decisions needed on the number and type of servers. We identified that the data would come from 2 sources: participants' smartphones and the Lose It! servers that contained self-monitoring and weight data synchronized with participants' Lose It! account. Communication between servers and smartphones needed to be bidirectional so that the EMA app could be updated or patches released, as well as for downloading data from the smartphones and Lose It!. A thorough needs assessment demonstrated that we would need a Web server to process the data and a database server to store data. The administrative staff of FitNow, Inc agreed to have the self-monitoring data for the study participants downloaded from a special portal to our university server every night at midnight. We did this with a 24-hour delay so that if a participant did not record their dinner until the next morning, the data would be included in the download.

Web Server

The Web server was composed of scripts that were either called directly by the smartphone app with scripts or internally initiated data downloads on a scheduled basis. The smartphone scripts were called periodically by the participants' smartphones to send new survey and event data, or to retrieve updates for the app. Internally scripts were called once a day to collect calorie, exercise, and weight data from Lose It! and to verify that each participant received each of the scheduled EMA surveys.

Database Server

We used an Oracle server (version 10g, Oracle Corporation) as the backend database server. This relational database served as a repository for all EMA, Lose It!, and weight data, as well as a large amount of processing data. Each EMA assessment included a date and time stamp when the EMA interview was delivered, when a participant started and ended it, and its status (completed, missed, or abandoned).

Smartphone EMA Survey App

The smartphone EMA app developed for this study was based on an existing app that had been developed for another research group [38,39]. The pre-EMPOWER version of the app was designed to work as a stand-alone research tool with study-managed, locked-down Android devices. Android phones were chosen as the target technology for this app because they supported the full range of control the phone needed to support the study. These control points included reliable scheduling of EMA assessments, which was difficult to implement in iOS at the time our study was conducted. However, this specific problem has been resolved, as programming can be done on the server side instead of on the phone today, thus permitting use of different operating systems. Still, operating system compatibility is a technical concern that should always be considered early in development. To meet the study's requirements, the app had to be made robust enough to enable long-term running on participant-owned and -managed phones. Even so, having the existing infrastructure of the pre-EMPOWER app greatly reduced the development time needed to reach a stable, pilot-testable version of the EMPOWER EMA app.

Each participant had a copy of the app installed on their own smartphone and configured for the participant. The app provided most of the functionality of the system, including (1) scheduling all types of interview prompts (ie, BOD, which was set by the participant depending on when she or he woke up; EOD was also set by the participant and had to be at least 12 hours after BOD; and random), (2) providing the appropriate version of survey content as needed, and (3) transmitting recorded survey responses and important events (eg, survey scheduling, phone reboots, and update install times) at regular intervals to the Web server that logged these to the database, and polling the Web server for updates in parameters, any patches, and possible new EMA versions. Data were stored locally on the phone regardless and kept until transmitted to the server even if a few tries were required.

We decided to place the majority of the system's functionality for EMA data collection on the participants' phones to avoid loss of data when there was no network connectivity. Data would be transmitted to the server when connectivity was restored. This was especially relevant for a study with a 1-year duration. Participants were likely to have loss of network connectivity for hours (such as in certain campus buildings or travelling through rural areas) to days at a time (such as with travel on vacation or business). It also protected the data collection from any server-side failure or down time. By distributing the EMA data collection functionality among the participants' devices, no single system component failure could negatively affect data collection across all participants.

Access to Servers

The University of Pittsburgh's enterprise Web infrastructure offered a robust server hosting the university's community and research studies, which was the perfect fit for the study. After negotiation and ensuring the enterprise Web infrastructure programmers that the smartphones would not be using excessive bandwidth and that no sensitive or confidential information would pass between phones and servers, we gained access to the PHP (a server-side scripting language) Web server and Oracle database server described above. Every phone/participant had a unique ID, and all responses to questions were either integers or, in a few cases, free text, where participants had to further explain their selection. To access self-monitoring and weight data through Lose It!, we negotiated special access to those data for download to Oracle servers and accessed the data via a custom-made management portal for study interventionists to review participants' data through the Lose It! Web interface.

Day-to-Day Operations

With all participants in the field and with real-time data collection for 12 months, it was critical for us to ensure that participants were receiving EMA prompts when they were scheduled. To achieve this, we collected additional data to help us minimize down time in the case of a smartphone malfunctioning. Those data were stored in our database and consisted of all scheduled prompts recorded in a table along with boot and shutdown times when participants rebooted their phones. Every day, we ran a custom report that documented the last time a prompt was fired and when the next prompt was going to occur. If someone was scheduled to receive a prompt but did not, an email was sent to that participant to investigate if anything was wrong with the phone, if the app was still installed, or if there were problems with the Internet connection, such as no connection or a weak one. We found that, even if there was no Internet connection, participants were still receiving prompts and the data were queued for upload at the next connection to the Internet.

Data Management and Data Security

EMA data collected by smartphones were first processed by a Web server and then transferred to a secure database server in real time. The database and Web servers were hosted and maintained by the University of Pittsburgh Computing Services and Systems Development Network Operations Center that has 24/7 monitoring. The servers were behind the firewall with special permission given to us to pass the data. Security was reviewed by the university before we were granted permission to build the infrastructure. The dietary self-monitoring data were maintained on the Lose It! server by Fit, Inc. Overall, no personal health data were ever included in the data transmission. All participants were assigned random unique ID numbers and most of the data were integer values. The data were backed up nightly. While the servers were hosted and maintained by the University of Pittsburgh Network Operations Center, they were also overseen by the University of Pittsburgh School of Nursing's database administrator.

Part 3: Challenges Faced and Lessons Learned From the Conduct of the EMPOWER Study

The 1-year EMPOWER study infrastructure faced several challenges from which lessons may be learned regarding (1) variation among mobile phone models, manufacturers of the operating systems, and different versions of the same Android operating system, (2) phone operating system updates, and (3) conflict with nonstudy use of mobile phones. Previous experience and the study's 12-month duration drove the decision

to use participants' personal smartphones for data collection but introduced the technical challenge of supporting a wide range of smartphone models and operating system versions. While all of the Android smartphones broadly used the same operating system, and Android claims interoperability, there were significant and subtle differences in operating system behavior between smartphone manufacturers and also between operating system versions from a single manufacturer.

Phone Operating System Updates

In at least two cases, carrier-released operating system updates led to failures in the smartphone app that caused a loss of data. In both cases, the updates changed underlying operating system resources that caused the app to crash when attempting to prompt the user to complete a survey. In both cases, the operating system update-related error was only resolved after uninstalling and reinstalling all affected software. Tracking changes in the operating system update events on the participant's phone reduced the time to identify this potential source of failure. To address these issues, we recommend, in addition to automating tasks as much as possible to shorten troubleshooting turnaround time, having a system in place that collects useful data (eg, technical and user behavior) to stay informed. We also recommend doing thorough user training, always listening and responding to end users' observations, and collecting information on problems encountered. Each of these activities was helpful in addressing technical issues.

Conflict With Nonstudy Use of the Smartphone

The use of participants' personal smartphones for data collection introduced an additional source of variability that could lead to a range of problems. Nonstudy-related uses of the smartphone could affect the ability of the EMA app to function as intended, such as turning off the smartphone, and interactions with other installed apps led to scheduled EMA assessments failing to launch and prompts being missed. By looking at logs and talking to the participants, we could try to pinpoint what led to problems in the system; we could then educate the participants on how to avoid such problems.

Institutional Review Board: Preparing for Institutional Review Board Approval Today

This study was approved by the University of Pittsburgh institutional review board in 2010 without any concerns except the risk for participant burden related to use of the EMA or the dietary self-monitoring app. However, recently the institutional review board has increased its scrutiny of protocols using mobile devices. For protocols that use any form of apps since 2016, the investigator is required to answer an extensive list of questions and document details regarding the risk of a third party intercepting data, data plan expenses, data security, terms of agreement of the commercial app, and concern if a mobile app is deemed a mobile medical app, in which US Food and Drug Administration regulations apply. Investigators need to provide the answers to these concerns as well as the personnel to arrange for data exchange and security.

Training Participants in EMA and Ensuring Adherence to EMA Protocol

We used a single-group, observational study design to describe the microprocesses of lapse and relapse following intentional weight loss through the use of daily EMA surveys over 12 months. The 12-month duration was based on extensive data showing that individuals usually reach their peak weight loss at 6 months, which is often followed by partial or total regain of the weight that had been lost [40]. To provide the background for weight loss and regain, the study implemented a standard behavioral intervention for weight loss that was implemented through group sessions over the 12 months. The following paragraphs focus on the details of the EMPOWER study pertinent to the EMA protocol, such as training study participants in the use of a smartphone and completing the EMA data sampling, as well as strategies to enhance adherence to the EMA survey completion. Details of standard behavioral interventions for weight loss conducted in our laboratory are published elsewhere [32,35,41].

Once we verified that individuals met the standard eligibility criteria for behavioral treatment for weight loss, we confirmed whether they could participate in a study that required their response several times a day. This included a trained research staff person asking the person about their daily routine, the demands of their job, and whether they were permitted to have their phone nearby in the work environment. The staff person walked through the daily routine of EMA prompts and demonstrated the surveys on a phone and how to reply. Thus, potential participants were fully informed of the demands of the study and were told that they were vital partners in this research on identifying the triggers for relapse in real time. The innovative components of the study seemed to inspire some to want to participate and remain in the study. Developing this partnership at baseline and throughout the study and talking to participants in the treatment group sessions periodically about what we were aiming to learn in the study likely helped ensure the importance of their contributing. We have a track record of good retention in clinical trials, so we used standard strategies such as following up with an email or phone call if a person did not attend a treatment session or if we did not hear from them, or if we were not seeing self-monitoring data. We tried to be as flexible as possible within the limitations of a study protocol. However, while we may have permitted a participant to miss some sessions, we required participants to complete at least 60% of the daily EMA surveys to receive reimbursement for the data plan charges.

After determining participants' willingness to engage in the daily EMA surveys, we confirmed whether the individuals had a study-compatible smartphone. If they did not, we gave them written instructions on which phone to purchase and to bring in the receipt so the study could reimburse them for the expense. Once they had the phone, they were scheduled for a one-on-one enrollment session, during which the EMA app was installed on their phone and they were shown how to complete the EMA surveys and were required to give a return demonstration completing a survey to the staff person. As Figure 1 shows, the survey items were on the home screen of the phone, and replies were indicated by touching the screen or sliding a bar for a

Likert-scale response. Printed instructions with illustrations were also provided. If participants needed additional assistance, a staff member was available to troubleshoot by phone or in person. Participants were also provided with a Withings Wi–Fi scale (Withings, Inc, Cambridge, MA, USA) for daily self-weighing and printed instructions on how to set it up in their home. The scale transmitted the weight to the Lose It! server, which was visible when participants accessed Lose It! on their phone. We had instant access to information regarding participants' completion of EMA survey and use of the scale. If there appeared to be any problems, such as no weights being recorded or no responses to EMA prompts, the Data Manager alerted the interventionists, who contacted the participant to determine the problem and help resolve it. To ensure that participants knew what the terms *temptation*, *lapse*, and *relapse* meant for this study, we explained these terms as described above and illustrated them with several practical examples. We had several discussions with the participants about these terms and their related meanings and clarified the definitions as necessary until the participants were able to articulate the terms and definitions in their own words.

Incentives to Enhance Adherence to EMA

The pilot study increased our awareness of the risk of participant overburden inherent in the EMA study. To encourage responding to as many prompts as possible, we offered an incentive using a random, variable schedule based on a lottery principle, so that the more frequently the participants responded, the higher their likelihood of receiving an incentive (eg, a US $5.00 gift card). Participants also were told that they needed to complete 60% of the random prompt surveys in order to receive US $25 each month to compensate for a portion of the data plan fee. Others have used financial incentives for EMA completion and reported that it enhanced compliance [12]. Anecdotally, many of our participants reported that it did not matter, as they had family shared data plans, and some did not collect their incentives.

EMA Monitoring

Our EMA monitoring protocol was adapted from earlier studies conducted by Shiffman and colleagues [38,42]. Participants were instructed that they would be monitored for the duration of the 12-month study and that the frequency of prompts they received on their smartphones would vary over time. At random times throughout the waking hours of the day, the smartphone prompted the participants with an alarm tone to answer the EMA survey questions. If the individual was unable to respond immediately (eg, he or she was in a meeting), he or she could delay answering the prompt for up to 20 minutes, which was done by using a snooze function on the alarm for 5 minutes up to 4 times. The participant also could temporarily turn off the alarm tone when it would be disruptive (eg, sleeping, in church, or attending a performance). The number of items per prompted assessment varied, as did the questions, but there were usually 10-15 items, with some questions being asked every time. Questions and the response options were displayed on the screen; the app permitted touch screen responses to complete the assessment (see Figure 1). Skip patterns were used whenever appropriate to reduce participant burden. For example, participants could not observe others in their immediate environment eating if they were completely alone, and thus were not asked that question. As described earlier, there were 4 types of EMA prompting. Both the BOD and EOD surveys were delivered at a time indicated by the participant as acceptable. Use of the alarm mode on the smartphone also served to set limits on the times that random prompts could be delivered, but participants were required to have at least 12 hours between morning waking and going to sleep.

Duration of the Study

The typical weight change pattern observed in several reported studies revealed that weight loss often reaches its peak at approximately 6 months after initiation of treatment, followed by a 30% to 35% weight regain [43-45]. One of our previous studies demonstrated that 52.4% of the participants achieved significant weight loss at 6 months but began to regain in the months that followed [46]. Given these results, we selected 12 months of observation to capture the natural trajectory of weight loss and regain, specifically to capture in real time the triggers of lapses and relapses.

Results

Throughout this paper, we have addressed the challenges we encountered in pilot testing the EMA surveys and the mobile technology to facilitate the transmission of prompts and responses providing data. Through the pilot studies and the 12-month observation study that collected daily EMA data, we learned many lessons. Textbox 1 summarizes those key lessons.

Textbox 1. Summary of lessons learned regarding collection of ecological momentary assessment (EMA) data.

Participant-related preferences

- Carrying only 1 device was important to most people.
- People had difficulty keeping 2 devices charged.
- Random EMA prompts should be limited to a maximum of 5/day.
- Individuals who were not regular mobile phone users were not good candidates, as they might not keep a phone charged or nearby.
- Regardless of their technology experiences, it is essential to train all individuals in the use of a phone and in completing EMA prompts, and having staff available to assist with troubleshooting is advised.
- Providing the participant with a phone for their full use facilitated their keeping it charged and with them.
- To ensure compatibility with the needs of the study, a list of recommended phones was provided so that each participant could select their phone model.
- It is important to listen to end users to refine EMA questions and the assessment schedule.

Infrastructure-related needs

- Experience enabled the team to identify the most reliable phone models.
- The battery of phones with iOS drained quickly, requiring chargers at home and in the workplace.
- Phones with the Android operating system supported the full range of control needed to support the EMA study, permitted programming to be done on phone, rather than being Web based.
- To avoid loss of data due to loss of network connectivity, the majority of the system's functionality for EMA data collection was placed on the participants' phone (important for a person who travels).
- For a long-term study, it is important to track operating system updates that could potentially interfere with EMA; also, use of other apps installed by the participant might interfere with EMA prompts.

EMPOWER Study Sample

We completed enrollment in January 2014 with a sample of 151 enrolled over 2 years in 6 cohorts. No EMA data were collected from 1 participant who withdrew from the study immediately after baseline; data from this participant were not included in the following analyses. The sample (see Table 1) was predominantly female (136/150, 90.7%), white (121/150, 80.7%), employed full-time (124/150, 82.7%) and well educated, having completed a mean of 16 years of education. Participants were, on average, 51.09 (SD 10.19) years of age. The mean body mass index was 34.02 (SD 4.58) kg/m^2. The final cohort completed the study in March 2015 with an overall 87.4% retention rate. Reasons for participant withdrawal were pregnancy (n=3), development of diabetes (n=5), personal decision to withdraw (n=4), and lost to follow-up (n=7), resulting in a final sample of 132 participants (132/151, 87.4%) completing the final assessment.

Table 1. Sample characteristics at baseline (n=150).

Characteristics	n or mean	SD or %
Age in years, mean (SD)	51.09	10.19
Education in years, mean (SD)	16.41	2.81
Body mass index in kg/m^2, mean (SD)	34.02	4.58
Female sex, n (%)	136	90.7
White racial status, n (%)	121	80.7
Marital status, n (%)		
Married, or living with partner or significant other	93	62.0
Never married	25	16.7
Widowed, separated, or divorced	31	20.7
Employed full-time, n (%)	124	82.7
Household income in US $, n (%)		
<50,000	35	23.3
≥$50,000	107	71.3
Have intentionally lost 10-19 lbs, n (%)		
Never	8	5.3
1-2 times	60	40.0
3-5 times	55	36.7
6-10 times	17	11.3
Have intentionally lost 20-49 lbs, n (%)		
Never	51	34.0
1-2 times	70	46.7
3-5 times	16	10.7

Adherence to Answering and Completing the EMA Prompts

Table 2 presents the details of adherence by EMA prompt type. During the first 6 months of the 12-month study interval, participants completed 88.26% (66,978/75,888) of the daily random assessments and discontinued or abandoned less than 1% of the surveys that they started. Similarly, they completed around 90% of the BOD and EOD prompts over the first 6 months. With observations over 17,860 participant-days among the 150 participants, they initiated a survey to report an average of 1.48 lapses per week. When participants initiated a survey to report a temptation or lapse, they completed 98.44% (5055/5135) of those surveys, and they reported that they had a lapse in 43.80% (2214/5055) of the self-initiated surveys they completed. During the end-of-day surveys, participants stated that they did not report one or more temptations or lapses on 22.81% (5356/23,486) of the days.

Table 2. Adherence to ecological momentary assessment prompts by type of prompt during the first 6 and second 6 months of the study (N=150).

Prompt type		Completed	Abandoned[a]	Missed	Total
First 6 months					
	Random	66,978 (88.26%)	316 (0.42%)	8594 (11.32%)	75,888
	Event	5055 (98.44%)	80 (1.56%)	N/A[b]	5135
	BOD[c]	23,411 (90.29%)	143 (0.55%)	2375 (9.16%)	25,929
	EOD[d]	23,343 (89.75%)	132 (0.51%)	2535 (9.75%)	26,010
Second 6 months					
	Random	63,349 (85.10%)	244 (0.33%)	10,840 (14.56%)	74,443
	Event	2294 (98.75%)	29 (1.25%)	N/A	2323
	BOD	20,227 (85.61%)	131 (0.55%)	3269 (13.84%)	23,627
	EOD	20,308 (85.53%)	114 (0.48%)	3323 (13.99%)	23,745

[a]Abandoned means participant began a self-initiated report but discontinued before completing the items.

[b]N/A: not applicable.

[c]BOD: beginning of the day.

[d]EOD: end of the day.

As Table 3 reports, there was considerable variability among participants in the percentage completion of EMA prompts; note that differences in percentages reported in Table 3 are different from those reported in Table 2, since the means reported in Table 3 give equal weight to each participant, while the percentages in Table 2 give equal weight to each prompt. The medians are all greater than their respective means, since the distribution of percentage completed prompts is skewed to the left. While the range in completion rates for random assessments was quite wide, all but 1 participant completed half, and 80.7% (121/150) completed more than 80% of the random assessments. Variability in numbers of self-reports was especially high, with numbers of self-reports ranging from 0 to 567; only 1 participant never initiated a self-report. Participants were less adherent to daily weigh-ins and diet reporting using Lose It! than with the EMA assessments, and showed greater variability in the former than in the latter.

Table 3. Variability among participants in percentages of completed assessments by type and numbers of entered self-assessments (event contingent).

Assessment type	Mean	Median	SD	Range
Random	86.5%	88.7%	10.7%	16.7%-99.5%
Event	49.7	23	83.3	0-567
BOD[a]	87.6%	91.4%	12.4%	15.4%-100.0%
EOD[b]	87.3%	92.3%	13.0%	14.0%-100.0%
Weigh-ins	68.9%	76.9%	24.6%	4.1%-98.3%
Diet reporting	70.5%	78.3%	27.4%	4.4%-100.0%

[a]BOD: beginning of the day.

[b]EOD: end of the day.

The biggest challenge that we encountered was convincing participants to self-report episodes of temptation or lapses. This topic was raised with participants throughout the study. Some typical responses were “I am constantly tempted to eat what I shouldn’t so I would be initiating these surveys all day;” “When I am tempted, I need to walk away and forget it as soon as possible. The last thing I want to do is take time to report it;” or “I forget to report a temptation or lapse.”

Discussion

We engaged in multiple phases of preparation and implementation to conduct a longitudinal observational study employing daily EMA prompts for 12 months. At the time of study launch, we used what was considered novel technological, methodological, and statistical means to address the critical issue of lapse and relapse after intentional weight loss, a significant clinical and public health issue that has been problematic in health care for over 30 years.

A major part of our preliminary studies was focused on development of the supportive infrastructure and refinement of the EMA assessments. In this process we learned a great deal. Primarily, if one wishes to have the EMA assessment on the phone and prevent any issues with connectivity with cell phone towers and potential loss of data, it is not possible to use a phone with an iOS operating system. Other investigators have handled

this by providing their participant with a phone for EMA data collection, but the phone is locked for other uses [47]. Because of the duration of our study, and having learned that participants do not wish to carry 2 devices, we chose to provide participants a personal smartphone that accommodated the programming for EMA and that they could use for all purposes. This also prevented the loss of data when the person traveled, which likely has not been a major issue in the other EMA studies because of their shorter duration. However, today, many researchers are avoiding the issue of phone operating system compatibility by having the EMA program be Web based. Ehlers et al [48] demonstrated success in a 14-day pilot study with a low-cost approach that used participants' mobile phones, text messaging, and mobile Internet to explore daily relationships between self-worth and physical activity in middle-aged women. Fanning et al [49] used a similar approach in a 7-day study of college-aged adults examining mind wandering.

Other salient lessons learned pertained to the maximum frequency of random EMA prompts that participants tolerated. We delivered up to 5 random prompts per day plus the BOD and EOD, and also asked participant to initiate event-contingent surveys. Another investigator prompted participants for 6 semirandom prompts per day plus an EOD survey, but that study was 14 days in duration [12]. Event-contingent completion has not been reported often but is essential if one is studying a time-varying emotion such as urge to smoke or to eat [4,11,20]. Important future work needs to examine whether the frequency of random prompts affects adherence to completion of other assessments.

Finally, we knew in the pilot studies and in EMPOWER that it was critical to ensure that all participants were receiving the prompts and the data were coming into the server. The overarching lesson learned was that, while parts of the study were run by programmed algorithms, careful oversight and monitoring were necessary to ensure that all components were functioning as intended. Staff members who filled other roles in the study participated in this oversight, similar to overseeing adherence to the protocol in a randomized clinical trial. Also, staff members were prepared to assist participants with troubleshooting the phone or EMA, which likely improved participants' engagement, as they knew assistance was available and used it.

The length of previous EMA studies has ranged from 4 days [8] to 6 months [50]. As might be expected, retention is related to duration; for example, Stefano et al [51] reported on a study using smartphones to assess participants' body monitoring behavior for 5 days and reported that only 1 person withdrew, resulting in 95.6% retention. Other EMA studies were conducted for 7 days [52], 14 days [48], and 1 month [53] and reported retention rates of 90.7%, 75%, and 66%, respectively. Inada et al [50] reported a 77.8% retention rate in their 6-month study. Gidlow et al [54] reported on a 12-week EMA study that was focused on stress among 153 healthy working adults that achieved a retention rate of 89.5%, confirming that, similar to our study, high rates of retention can be achieved in longer studies. Several studies that ranged from 4 [55-57] to 8 [58] to 18 weeks [59] did not report retention rates.

Adherence to responding to the EMA surveys, crucial for gaining insight into the targeted behaviors, also varies across studies, modes of survey administration, and types of EMA sampling used. Goldschmidt et al [12] used handheld computers to collect EMA data after each binge-eating episode plus 6 semirandom prompts per day and a before bedtime survey for 14 days. Participants responded to 86% of the semirandom prompts, while 84% of the bedtime recordings were completed. Zenk et al [13] used a smartphone and reported that 68.9% of the 35 random signal-contingent surveys were completed in a 7-day study about snack food intake.

Many studies use event-contingent sampling, which entails asking a participant to initiate an EMA survey to report an event, such as a binge-eating episode, consumption of snack food, or, as in our study, a temptation to eat a food that was not part of their dietary plan or the actual consumption of the food, which represented a lapse. It is not possible to report adherence rates to these self-initiated reports, only the frequency. Participants in Goldschmidt and colleague's study reported 7.8 (6.5) binges and 11.1 (9.6) purges over the 12-day period [12]. McKee et al used only event-contingent sampling in their 7-day study of lapses among 80 adults who reported they were dieting and found that participants reported 898 instances of dietary temptation, or an average of 11.2 temptations per person over the 7 days [11]. These rates are significantly higher than those we observed in our study, which may be due to the 12-month duration of our study and that our participants were actively engaged in a weight loss intervention.

Our study was unique in its duration and 12-month retention rate of 87%. To date, no other study, to our knowledge, has assessed eating behaviors for as long as 12 months related to temptation, lapses, and relapse during a weight loss intervention. Our study contributes to the literature on EMA in describing the development and refinement of EMA items and the challenges and successes of developing the supporting infrastructure for a long-term EMA study. Additionally, our study adds information on strategies to enhance adherence in reporting episodes of slips and lapses through EMA.

Acknowledgments

Funding was provided by the National Institutes of Health (NIH), NHLBI R01HL107370, R01 HL107370-01 Research Supplements to Promote Diversity in Health-Related Research. We thank the participants of the study, who generously gave their time and effort to completing the numerous EMA queries.

Conflicts of Interest

None declared.

References

1. Shiffman S, Stone AA, Hufford MR. Ecological momentary assessment. Annu Rev Clin Psychol 2008;4:1-32. [Medline: 18509902]
2. Dunbar MS, Scharf D, Kirchner T, Shiffman S. Do smokers crave cigarettes in some smoking situations more than others? Situational correlates of craving when smoking. Nicotine Tob Res 2010 Mar;12(3):226-234 [FREE Full text] [doi: 10.1093/ntr/ntp198] [Medline: 20133379]
3. Shiffman S, Shumaker SA, Abrams DB, Cohen S, Garvey A, Grunberg NE, et al. Models of smoking relapse. Health Psychol 1986;5 Suppl:13-27. [Medline: 3582321]
4. Shiffman S, Read L, Jarvik ME. Smoking relapse situations: a preliminary typology. Int J Addict 1985 Feb;20(2):311-318. [Medline: 4008124]
5. Stone AA, Shiffman S, Schwartz JE, Broderick JE, Hufford MR. Patient non-compliance with paper diaries. BMJ 2002 May 18;324(7347):1193-1194 [FREE Full text] [Medline: 12016186]
6. Stone AA, Broderick JE, Schwartz JE, Shiffman S, Litcher-Kelly L, Calvanese P. Intensive momentary reporting of pain with an electronic diary: reactivity, compliance, and patient satisfaction. Pain 2003 Jul;104(1-2):343-351. [Medline: 12855344]
7. Stone AA, Shiffman S, Schwartz JE, Broderick JE, Hufford MR. Patient compliance with paper and electronic diaries. Control Clin Trials 2003 Apr;24(2):182-199. [Medline: 12689739]
8. Dunton GF, Liao Y, Kawabata K, Intille S. Momentary assessment of adults' physical activity and sedentary behavior: feasibility and validity. Front Psychol 2012;3:260 [FREE Full text] [doi: 10.3389/fpsyg.2012.00260] [Medline: 22866046]
9. Liao Y, Intille SS, Dunton GF. Using ecological momentary assessment to understand where and with whom adults' physical and sedentary activity occur. Int J Behav Med 2015 Feb;22(1):51-61. [doi: 10.1007/s12529-014-9400-z] [Medline: 24639067]
10. Shiffman S, Stone AA. Introduction to the special section: ecological momentary assessment in health psychology. Health Psychol 1998;17(1):3-5 [FREE Full text] [doi: 10.1037/h0092706]
11. McKee HC, Ntoumanis N, Taylor IM. An ecological momentary assessment of lapse occurrences in dieters. Ann Behav Med 2014 Dec;48(3):300-310. [doi: 10.1007/s12160-014-9594-y] [Medline: 24562984]
12. Goldschmidt AB, Wonderlich SA, Crosby RD, Engel SG, Lavender JM, Peterson CB, et al. Ecological momentary assessment of stressful events and negative affect in bulimia nervosa. J Consult Clin Psychol 2014;82(1):30-39. [doi: 10.1037/a0034974] [Medline: 24219182]
13. Zenk SN, Horoi I, McDonald A, Corte C, Riley B, Odoms-Young AM. Ecological momentary assessment of environmental and personal factors and snack food intake in African American women. Appetite 2014 Dec;83:333-341. [doi: 10.1016/j.appet.2014.09.008] [Medline: 25239402]
14. Schulz S, Laessle RG. Stress-induced laboratory eating behavior in obese women with binge eating disorder. Appetite 2012 Apr;58(2):457-461. [doi: 10.1016/j.appet.2011.12.007] [Medline: 22200410]
15. Shiffman S. Real-time self-report of momentary states in the natural environment: computerized ecological momentary assessment. In: Stone AA, Bachrach CA, Jobe JB, Kurtzman HS, Cain VS, editors. The Science of Self Report: Implication for Research and Practice. Erlbaum: Mahwah, NJ; 2000:277-296.
16. Wing RR, Phelan S. Long-term weight loss maintenance. Am J Clin Nutr 2005 Jul;82(1 Suppl):222S-225S [FREE Full text] [Medline: 16002825]
17. Grilo CM, Shiffman S, Wing RR. Relapse crises and coping among dieters. J Consult Clin Psychol 1989 Aug;57(4):488-495. [Medline: 2768608]
18. Grilo CM, Shiffman S, Wing RR. Coping with dietary relapse crises and their aftermath. Addict Behav 1993;18(1):89-102. [Medline: 8465681]
19. Witkiewitz K, Marlatt GA. Modeling the complexity of post-treatment drinking: it's a rocky road to relapse. Clin Psychol Rev 2007 Jul;27(6):724-738 [FREE Full text] [doi: 10.1016/j.cpr.2007.01.002] [Medline: 17355897]
20. Shiffman S, Balabanis MH, Gwaltney CJ, Paty JA, Gnys M, Kassel JD, et al. Prediction of lapse from associations between smoking and situational antecedents assessed by ecological momentary assessment. Drug Alcohol Depend 2007 Dec 1;91(2-3):159-168 [FREE Full text] [doi: 10.1016/j.drugalcdep.2007.05.017] [Medline: 17628353]
21. Shiffman S. Dynamic influences on smoking relapse process. J Pers 2005 Dec;73(6):1715-1748. [doi: 10.1111/j.0022-3506.2005.00364.x] [Medline: 16274451]
22. Marlatt G, Gordon J. Relapse Prevention: Maintenance Strategies in the Treatment of Addictive Behaviors. New York, NY: Guilford Press; 1985.
23. Wing R. Behavioral approaches to the treatment of obesity. In: Bray GA, Bourchard C, James WPT, editors. Handbook of Obesity: Clinical Applications. 2nd edition. New York, NY: Marcel Dekker; 2004:147-167.

24. Perri MG, Nezu AM, McKelvey WF, Shermer RL, Renjilian DA, Viegener BJ. Relapse prevention training and problem-solving therapy in the long-term management of obesity. J Consult Clin Psychol 2001 Aug;69(4):722-726. [Medline: 11550740]
25. Warziski MT, Sereika SM, Styn MA, Music E, Burke LE. Changes in self-efficacy and dietary adherence: the impact on weight loss in the PREFER study. J Behav Med 2008 Feb;31(1):81-92. [doi: 10.1007/s10865-007-9135-2] [Medline: 17963038]
26. Turk MW, Yang K, Hravnak M, Sereika SM, Ewing LJ, Burke LE. Randomized clinical trials of weight loss maintenance: a review. J Cardiovasc Nurs 2009;24(1):58-80 [FREE Full text] [doi: 10.1097/01.JCN.0000317471.58048.32] [Medline: 19114803]
27. Burke LE, Wang J, Sevick MA. Self-monitoring in weight loss: a systematic review of the literature. J Am Diet Assoc 2011 Jan;111(1):92-102 [FREE Full text] [doi: 10.1016/j.jada.2010.10.008] [Medline: 21185970]
28. Turk MW, Elci OU, Wang J, Sereika SM, Ewing LJ, Acharya SD, et al. Self-monitoring as a mediator of weight loss in the SMART randomized clinical trial. Int J Behav Med 2013 Dec;20(4):556-561 [FREE Full text] [doi: 10.1007/s12529-012-9259-9] [Medline: 22936524]
29. Wheeler L, Reis H. Self-recording of everyday life events: origins, types, and uses. J Pers 1991;59(3):339-354. [doi: 10.1111/j.1467-6494.1991.tb00252.x]
30. Rathbun SL, Shiffman S, Gwaltney CJ. Modelling the effects of partially observed covariates on Poisson process intensity. Biometrika 2007 Feb 07;94(1):153-165. [doi: 10.1093/biomet/asm009]
31. Neustifter B, Rathbun SL, Shiffman S. Mixed-Poisson point process with partially-observed covariates: ecological momentary assessment of smoking. J Appl Stat 2012;39(4):883-899 [FREE Full text] [doi: 10.1080/02664763.2011.626848] [Medline: 22544991]
32. Burke LE, Styn MA, Glanz K, Ewing LJ, Elci OU, Conroy MB, et al. SMART trial: a randomized clinical trial of self-monitoring in behavioral weight management-design and baseline findings. Contemp Clin Trials 2009 Nov;30(6):540-551 [FREE Full text] [doi: 10.1016/j.cct.2009.07.003] [Medline: 19665588]
33. Music E, Styn M, Burke L, French B, Siewiorek D, Smailagic A, et al. Development of technology infrastructure to support real-time data collection in behavioral research. [Poster presentation]. Ann Behav Med 2013;45(Suppl):s123.
34. Burke LE, Conroy MB, Sereika SM, Elci OU, Styn MA, Acharya SD, et al. The effect of electronic self-monitoring on weight loss and dietary intake: a randomized behavioral weight loss trial. Obesity (Silver Spring) 2011 Feb;19(2):338-344 [FREE Full text] [doi: 10.1038/oby.2010.208] [Medline: 20847736]
35. Burke LE, Ewing LJ, Ye L, Styn M, Zheng Y, Music E, et al. The SELF trial: a self-efficacy-based behavioral intervention trial for weight loss maintenance. Obesity (Silver Spring) 2015 Nov;23(11):2175-2182 [FREE Full text] [doi: 10.1002/oby.21238] [Medline: 26381151]
36. Burke L, Music E, Rathbun S, Olguin-Olguin D, French B, Styn M, et al. Feasibility of using smart phones for EMA data collection in a study of weight loss maintenance. Obesity 2010;18(Suppl. 2):s160.
37. Burke LE, Styn MA, Ewing LJ, Kriska AM, Strollo PJ, Shiffman S. Development of EMA items for study examining the triggers of relapse following intentional weight loss. 2012 Presented at: The Obesity Society 30th Annual Scientific Meeting; September 20-24, 2012; San Antonio, TX, USA.
38. Shiffman S, Hickcox M, Paty JA, Gnys M, Kassel JD, Richards TJ. Progression from a smoking lapse to relapse: prediction from abstinence violation effects, nicotine dependence, and lapse characteristics. J Consult Clin Psychol 1996 Oct;64(5):993-1002. [Medline: 8916628]
39. Shiffman S, Gwaltney CJ, Balabanis MH, Liu KS, Paty JA, Kassel JD, et al. Immediate antecedents of cigarette smoking: an analysis from ecological momentary assessment. J Abnorm Psychol 2002 Nov;111(4):531-545. [Medline: 12428767]
40. Wing RR, Tate DF, Gorin AA, Raynor HA, Fava JL. A self-regulation program for maintenance of weight loss. N Engl J Med 2006 Oct 12;355(15):1563-1571. [doi: 10.1056/NEJMoa061883] [Medline: 17035649]
41. Burke LE, Choo J, Music E, Warziski M, Styn MA, Kim Y, et al. PREFER study: a randomized clinical trial testing treatment preference and two dietary options in behavioral weight management--rationale, design and baseline characteristics. Contemp Clin Trials 2006 Feb;27(1):34-48. [doi: 10.1016/j.cct.2005.08.002] [Medline: 16233990]
42. Shiffman S, Waters AJ. Negative affect and smoking lapses: a prospective analysis. J Consult Clin Psychol 2004 Apr;72(2):192-201. [doi: 10.1037/0022-006X.72.2.192] [Medline: 15065954]
43. Butryn ML, Webb V, Wadden TA. Behavioral treatment of obesity. Psychiatr Clin North Am 2011 Dec;34(4):841-859 [FREE Full text] [doi: 10.1016/j.psc.2011.08.006] [Medline: 22098808]
44. Katan MB. Weight-loss diets for the prevention and treatment of obesity. N Engl J Med 2009 Feb 26;360(9):923-925. [doi: 10.1056/NEJMe0810291] [Medline: 19246365]
45. Weiss EC, Galuska DA, Kettel KL, Gillespie C, Serdula MK. Weight regain in U.S. adults who experienced substantial weight loss, 1999-2002. Am J Prev Med 2007 Jul;33(1):34-40. [doi: 10.1016/j.amepre.2007.02.040] [Medline: 17572309]
46. Zheng Y, Sereika S, Imes C, Ye L, Mattos M, Mancino J, et al. Trajectories of weight change and their association with barriers to healthy eating during 18-month weight loss treatment. Circulation 2014;129:A07.

47. Shrier LA, Shih M, Hacker L, de Moor C. A momentary sampling study of the affective experience following coital events in adolescents. J Adolesc Health 2007 Apr;40(4):357.e1-357.e8 [FREE Full text] [doi: 10.1016/j.jadohealth.2006.10.014] [Medline: 17367729]
48. Ehlers DK, Huberty J, Buman M, Hooker S, Todd M, de Vreede GJ. A novel inexpensive use of smartphone technology for ecological momentary assessment in middle-aged women. J Phys Act Health 2016 Mar;13(3):262-268. [doi: 10.1123/jpah.2015-0059] [Medline: 26284689]
49. Fanning J, Mackenzie M, Roberts S, Crato I, Ehlers D, McAuley E. Physical activity, mind wandering, affect, and sleep: an ecological momentary assessment. JMIR Mhealth Uhealth 2016 Aug 31;4(3):e104 [FREE Full text] [doi: 10.2196/mhealth.5855] [Medline: 27580673]
50. Inada S, Yoshiuchi K, Iizuka Y, Ohashi K, Kikuchi H, Yamamoto Y, et al. Pilot study for the development of a self-care system for type 2 diabetes patients using a personal digital assistant (PDA). Int J Behav Med 2016 Jun;23(3):295-299. [doi: 10.1007/s12529-016-9535-1] [Medline: 26780633]
51. Stefano EC, Hudson DL, Whisenhunt BL, Buchanan EM, Latner JD. Examination of body checking, body image dissatisfaction, and negative affect using ecological momentary assessment. Eat Behav 2016 Aug;22:51-54. [doi: 10.1016/j.eatbeh.2016.03.026] [Medline: 27086048]
52. Huh J, Cerrada CJ, Kirkpatrick MG, Dunton G, Leventhal AM. Social contexts of momentary craving to smoke among Korean American emerging adults. Addict Behav 2016 May;56:23-29. [doi: 10.1016/j.addbeh.2016.01.006] [Medline: 26802789]
53. Bromberg MH, Connelly M, Anthony KK, Gil KM, Schanberg LE. Prospective mediation models of sleep, pain, and daily function in children with arthritis using ecological momentary assessment. Clin J Pain 2016 Jun;32(6):471-477. [doi: 10.1097/AJP.0000000000000298] [Medline: 26340651]
54. Gidlow CJ, Randall J, Gillman J, Silk S, Jones MV. Hair cortisol and self-reported stress in healthy, working adults. Psychoneuroendocrinology 2016 Jan;63:163-169. [doi: 10.1016/j.psyneuen.2015.09.022] [Medline: 26447679]
55. Macias C, Panch T, Hicks YM, Scolnick JS, Weene DL, Öngür D, et al. Using smartphone apps to promote psychiatric and physical well-being. Psychiatr Q 2015 Jan 31:505-519. [doi: 10.1007/s11126-015-9337-7] [Medline: 25636496]
56. Odawara M, Hashizume M, Yoshiuchi K, Tsuboi K. Real-time assessment of the effect of biofeedback therapy with migraine: a pilot study. Int J Behav Med 2015 Dec;22(6):748-754. [doi: 10.1007/s12529-015-9469-z] [Medline: 25670026]
57. Cohn A, Brandon T, Armeli S, Ehlke S, Bowers M. Real-time patterns of smoking and alcohol use: an observational study protocol of risky-drinking smokers. BMJ Open 2015 Jan 06;5(1):e007046 [FREE Full text] [doi: 10.1136/bmjopen-2014-007046] [Medline: 25564150]
58. Shiffman S, Patten C, Gwaltney C, Paty J, Gnys M, Kassel J, et al. Natural history of nicotine withdrawal. Addiction 2006 Dec;101(12):1822-1832. [doi: 10.1111/j.1360-0443.2006.01635.x] [Medline: 17156182]
59. van Os J, Delespaul P, Barge D, Bakker RP. Testing an mHealth momentary assessment routine outcome monitoring application: a focus on restoration of daily life positive mood states. PLoS One 2014 Dec;9(12):e115254 [FREE Full text] [doi: 10.1371/journal.pone.0115254] [Medline: 25513813]

Abbreviations

BOD: beginning of the day
EMA: ecological momentary assessment
EOD: end of the day
PDA: personal digital assistant
SMART: Self-Monitoring And Recording using Technology

Edited by G Eysenbach; submitted 09.12.16; peer-reviewed by J Wang, D Ehlers, L Hu, P Ma; comments to author 03.01.17; revised version received 03.02.17; accepted 19.02.17; published 15.03.17.

Please cite as:
Burke LE, Shiffman S, Music E, Styn MA, Kriska A, Smailagic A, Siewiorek D, Ewing LJ, Chasens E, French B, Mancino J, Mendez D, Strollo P, Rathbun SL
Ecological Momentary Assessment in Behavioral Research: Addressing Technological and Human Participant Challenges
J Med Internet Res 2017;19(3):e77
URL: http://www.jmir.org/2017/3/e77/
doi:10.2196/jmir.7138
PMID:28298264

Original Paper

Comparison of Different Recruitment Methods for Sexual and Reproductive Health Research: Social Media–Based Versus Conventional Methods

Yoko Motoki[1], MD, PhD; Etsuko Miyagi[1], MD, PhD; Masataka Taguri[2], PhD; Mikiko Asai-Sato[1], MD, PhD; Takayuki Enomoto[3], MD, PhD; John Dennis Wark[4], MD, PhD; Suzanne Marie Garland[5,6,7], MD

[1]Department of Obstetrics and Gynecology, Yokohama City University School of Medicine, Yokohama, Kanagawa, Japan

[2]Department of Biostatistics, Yokohama City University School of Medicine, Yokohama, Kanagawa, Japan

[3]Department of Obstetrics and Gynecology, Niigata University Graduate School of Medical and Dental Science, Niigata, Niigata, Japan

[4]Department of Medicine and Bone and Mineral Medicine, Royal Melbourne Hospital, University of Melbourne, Parkville, Victoria, Australia

[5]Department of Microbiology and Infectious Diseases, Royal Women's Hospital, Parkville, Victoria, Australia

[6]Infection and Immunity, Murdoch Childrens Research Institute, Parkville, Victoria, Australia

[7]Department of Obstetrics and Gynaecology, University of Melbourne, Parkville, Victoria, Australia

Corresponding Author:
Etsuko Miyagi, MD, PhD
Department of Obstetrics and Gynecology
Yokohama City University School of Medicine
3-9 Fukuura, Kanazawa
Yokohama, Kanagawa, 236-0004
Japan
Phone: 81 45 787 2690
Fax: 81 45 701 3536
Email: emiyagi@yokohama-cu.ac.jp

Abstract

Background: Prior research about the sexual and reproductive health of young women has relied mostly on self-reported survey studies. Thus, participant recruitment using Web-based methods can improve sexual and reproductive health research about cervical cancer prevention. In our prior study, we reported that Facebook is a promising way to reach young women for sexual and reproductive health research. However, it remains unknown whether Web-based or other conventional recruitment methods (ie, face-to-face or flyer distribution) yield comparable survey responses from similar participants.

Objective: We conducted a survey to determine whether there was a difference in the sexual and reproductive health survey responses of young Japanese women based on recruitment methods: social media–based and conventional methods.

Methods: From July 2012 to March 2013 (9 months), we invited women of ages 16-35 years in Kanagawa, Japan, to complete a Web-based questionnaire. They were recruited through either a social media–based (social networking site, SNS, group) or by conventional methods (conventional group). All participants enrolled were required to fill out and submit their responses through a Web-based questionnaire about their sexual and reproductive health for cervical cancer prevention.

Results: Of the 243 participants, 52.3% (127/243) were recruited by SNS, whereas 47.7% (116/243) were recruited by conventional methods. We found no differences between recruitment methods in responses to behaviors and attitudes to sexual and reproductive health survey, although more participants from the conventional group (15%, 14/95) chose not to answer the age of first intercourse compared with those from the SNS group (5.2%, 6/116; *P*=.03).

Conclusions: No differences were found between recruitment methods in the responses of young Japanese women to a Web–based sexual and reproductive health survey.

(J Med Internet Res 2017;19(3):e73) doi:10.2196/jmir.7048

KEYWORDS

papillomavirus vaccines; reproductive health; sexual health; sexual behavior; Japan; uterine cervical neoplasms

Introduction

Prior research about sexual and reproductive health of young women has largely relied on self-reported survey studies. These surveys are challenging because of their low response rates and the potential bias of their participants [1].

Low survey response rates are caused by attitudes of young adults toward health issues and their personal lifestyles. Young adults have less interest in health research than older adults, as they have fewer health problems in general and do not perceive themselves to be at risk. Young adults are also extremely mobile during their late teens and 20s because of college, new jobs, or becoming married. Consequently, targeted mailings are not always received. Moreover, computerized assisted telephone interviews are less optimal today, as fewer homes have landlines. Biases can also occur in self-reported sexual and reproductive health surveys and include non-respondent bias and social disability bias [1]. Direct interviews and self-administered questionnaires may be biased in settings such as clinics, schools, and workplaces, even if they are submitted anonymously [1]. This could be because the participants felt nervous or uncomfortable with the teachers or parents learning about their sexual behavior.

Thus, social media–based recruitment to surveys can potentially improve the low response rate and biases in survey-based sexual and reproductive health research [2]. In a prior study, we reported that social media such as Facebook is a cost-efficient and a promising way to reach a general population of young women for sexual and reproductive health research [3-9]. In addition, Facebook is one of the popular social networking sites (SNSs) worldwide. Unlike the other SNSs such as LINE, mixi, and LinkedIn, Facebook can provide demographic-targeted advertisement (eg, gender, age, or place of residence) using the cost-per-click option, a low-cost system to charge per number of user clicks or views [9]. We have used Facebook previously in research to recruit general populations of young women, as it is one of the most popular SNSs that young women use to communicate. However, little is known whether social media–based or conventional recruitment methods (ie, face-to-face or flyer distribution) yield comparable survey responses from similar participants [2,10].

We conducted a survey to determine whether there was a difference in the sexual and reproductive health survey responses of young Japanese women based on recruitment methods: social media–based and conventional methods.

Methods

Study Oversight

This study was a component of a pilot study for the Yokohama-Kanagawa Cervical Cancer Prevention Project [3]. This study protocol was approved by the Institutional Ethics Committee of the Yokohama City University School of Medicine. The detailed methodology of this study was previously reported by Miyagi et al [3].

Targeted Participants

We recruited women aged 16-35 years in the Kanagawa Prefecture, Japan, with a population of 9 million, between July 2012 and March 2013 (9 months).

Recruitment Through Facebook

We defined participants recruited through Facebook as the “SNS group.” We recruited them by using a Facebook advertisement in the cost-per-click option.

Recruitment by Conventional Methods

We defined participants recruited through conventional methods as the “conventional group.” The conventional methods used included face-to-face recruitments in clinics and workplaces, and flyer distributions at schools, hospitals, and educational events, which accepted our concept of sexual and reproductive health research.

Enrollment

All participants were invited to access the Yokohama-Kanagawa Cervical Cancer Prevention Project website, through which their personal information was obtained. Subsequently, they were sent a link via email, where they were asked to complete a Web-based questionnaire (SurveyMonkey; SurveyMonkey, Inc).

Statistical Analysis

We asked the participants whether they felt embarrassed by the Web-based survey questions on a scale of 1-5 (1 = not embarrassing, 5 = extremely embarrassing). We also used a chi-square analysis of independence to compare the responses of the two different recruitment groups. We compared the scores between the two groups using Student’s *t* test. We analyzed the data by SPSS version 20 (IBM Corporation). A *P* value of < .05 was considered statistically significant.

Results

Attitudes to the Survey Between Recruitment Methods

During the 9-month study period, 394 participants expressed their interests and 264 consented to the study. Among the 243 who completed the Web-based questionnaire, the SNS group (52.3%, 127/243) and the conventional group (47.7%, 116/243) had similar attitudes for all questions regarding their sexual behavior, except for the question regarding the age when they became sexually active. Among those already sexually active (86.8%, 211/243), more participants in the conventional group (15%, 14/95) denied answering the age of their sexual debut, compared with those in the SNS group (5.2%, 6/116) (*P*=.03; Table 1).

Table 1. Characteristics of participants recruited through social media–based and other conventional methods.

Characteristics		Total (n=243)			SNS group (n=127)			Conventional group (n=116)			
		No	Rate (%)	95% CI	No	Rate (%)	95% CI	No	Rate (%)	95% CI	*P*
General											
Age Group											
	16-19	14	5.76	(2.8-8.7)	4	3.15	(0.1-6.2)	10	8.62	(3.5-13.7)	.10
	20-29	122	50.21	(43.9-56.5)	65	51.18	(42.5-59.9)	57	49.14	(40.0-58.2)	
	30-35	107	44.03	(37.8-50.3)	58	45.67	(37.0-54.3)	49	42.24	(33.3-51.2)	
District of residence											
	Yokohama City	143	58.85	(52.7-65.0)	75	59.06	(50.5-67.6)	68	58.62	(49.7-67.6)	>.99
	Others	100	41.15	(35.0-47.3)	52	40.94	(32.4-49.5)	48	41.38	(32.4-50.3)	
Educational level[a]											
	< High school graduate	5	2.07	(0.3-3.9)	2	1.57	(0.0-3.7)	3	2.61	(0.0-5.5)	.88
	High school graduate	47	19.42	(14.4-24.4)	26	20.47	(13.5-27.5)	21	18.26	(11.2-25.3)	
	> High school graduate	190	78.51	(73.3-83.7)	99	77.95	(70.7-85.2)	91	79.13	(71.7-86.6)	
Sexual and reproductive health											
Sexual experience											
	No answer	1	0.41	(0.3-0.5)	0	0.00	NA	1	0.86	(0.8-0.9)	.05
	Never	31	12.76	(8.9-17.4)	11	8.66	(3.8-13.6)	20	17.24	(11.1-25.1)	
	Yes	211	86.83	(82.6-91.1)	116	91.34	(86.4-96.2)	95	81.90	(74.9-88.9)	
Age at first intercourse[b]											
	No answer	20	9.48	(5.5-13.4)	6	5.17	(1.1-9.2)	14	14.74	(7.6-21.9)	.03
	12-15	14	6.64	(3.3-10.0)	9	7.76	(2.9-12.6)	5	5.26	(0.8-9.8)	
	16-18	71	33.65	(27.3-40.0)	46	39.66	(30.8-48.6)	25	26.32	(17.5-35.2)	
	19-24	98	46.45	(39.7-53.2)	52	44.83	(35.8-53.9)	46	48.42	(38.4-58.5)	
	25-30	8	3.79	(1.2-6.4)	3	2.59	(0.0-5.5)	5	5.26	(0.8-9.8)	
Condom use at first sex[b]											
	No answer	2	0.95	(0.0-2.3)	1	0.86	(0.0-2.5)	1	1.05	(0.0-3.1)	>.99
	Yes	171	81.04	(75.8-86.3)	94	81.03	(73.9-88.2)	77	81.05	(73.2-88.9)	
	No	38	18.01	(12.8-23.2)	21	18.10	(11.1-25.1)	17	17.89	(10.2-25.6)	

Characteristics		Total (n=243)			SNS group (n=127)			Conventional group (n=116)			
		No	Rate (%)	95% CI	No	Rate (%)	95% CI	No	Rate (%)	95% CI	*P*
Number of sexual partner within 12 months[b]											
	0-2	180	85.31	(80.5-90.1)	98	84.48	(77.9-91.1)	82	86.32	(79.4-93.2)	.85
	3+	18	8.53	(4.8-12.3)	12	10.34	(4.8-15.9)	6	6.32	(1.4-11.2)	
	No answer	13	6.16	(2.9-9.4)	6	5.17	(1.1-9.2)	7	7.37	(2.1-12.6)	
Chlamydia awareness											
	Yes	226	93.00	(89.8-96.2)	121	95.28	(91.6-99.0)	105	90.52	(85.2-95.8)	.21
	No	17	7.00	(3.8-10.2)	6	4.72	(1.0-8.4)	11	9.48	(4.2-14.8)	
Positive test history for chlamydia among whom they knew chlamydia											
	Yes	28	12.39	(8.1-16.7)	14	11.57	(5.9-17.3)	14	13.33	(6.8-19.8)	.84
	No	197	87.17	(82.8-91.5)	106	87.60	(81.7-93.5)	91	86.67	(80.2-93.2)	
	No answer	1	0.44	(0.0-1.3)	1	0.83	(0.0-2.4)	0	0.00	N.A.	
Awareness of HPV vaccines[b]											
	Yes	194	79.84	(74.8-84.9)	99	77.95	(70.7-85.2)	95	81.90	(74.9-88.9)	.52
	No/No answer	49	20.16	(15.1-25.2)	28	22.05	(14.8-29.3)	21	18.10	(11.1-25.1)	
Self-reported cervical cancer screening status among whom their age ≥ 20[c]											
	Yes	156	68.42	(62.1-74.2)	85	69.67	(61.5-77.8)	71	66.98	(58.0-75.9)	.67
	No	45	19.74	(14.5-24.8)	20	16.39	(9.8-23.0)	25	23.58	(15.5-31.7)	
	No answer	27	11.84	(8.0-16.5)	17	13.93	(7.8-20.1)	10	9.43	(3.9-15.0)	
Self-reported HPV vaccination status											
	Yes	30	12.35	(8.2-16.5)	11	8.66	(3.8-13.6)	19	16.38	(9.6-23.1)	.08
	No	206	84.77	(80.3-89.3)	113	88.98	(83.5-94.4)	93	80.17	(72.9-87.4)	
	Don’t know/No answer	7	2.88	(0.8-5.0)	3	2.36	(0.0-5.0)	4	3.45	(0.1-6.8)	

[a]One participant did not answer this question and was excluded.

[b]Question only for participants who answered 'yes' to having had sexual experiences.

[c]One who answered “I don't know” was excluded. The Japanese government invites women aged 20 years and older to cervical cancer screening.

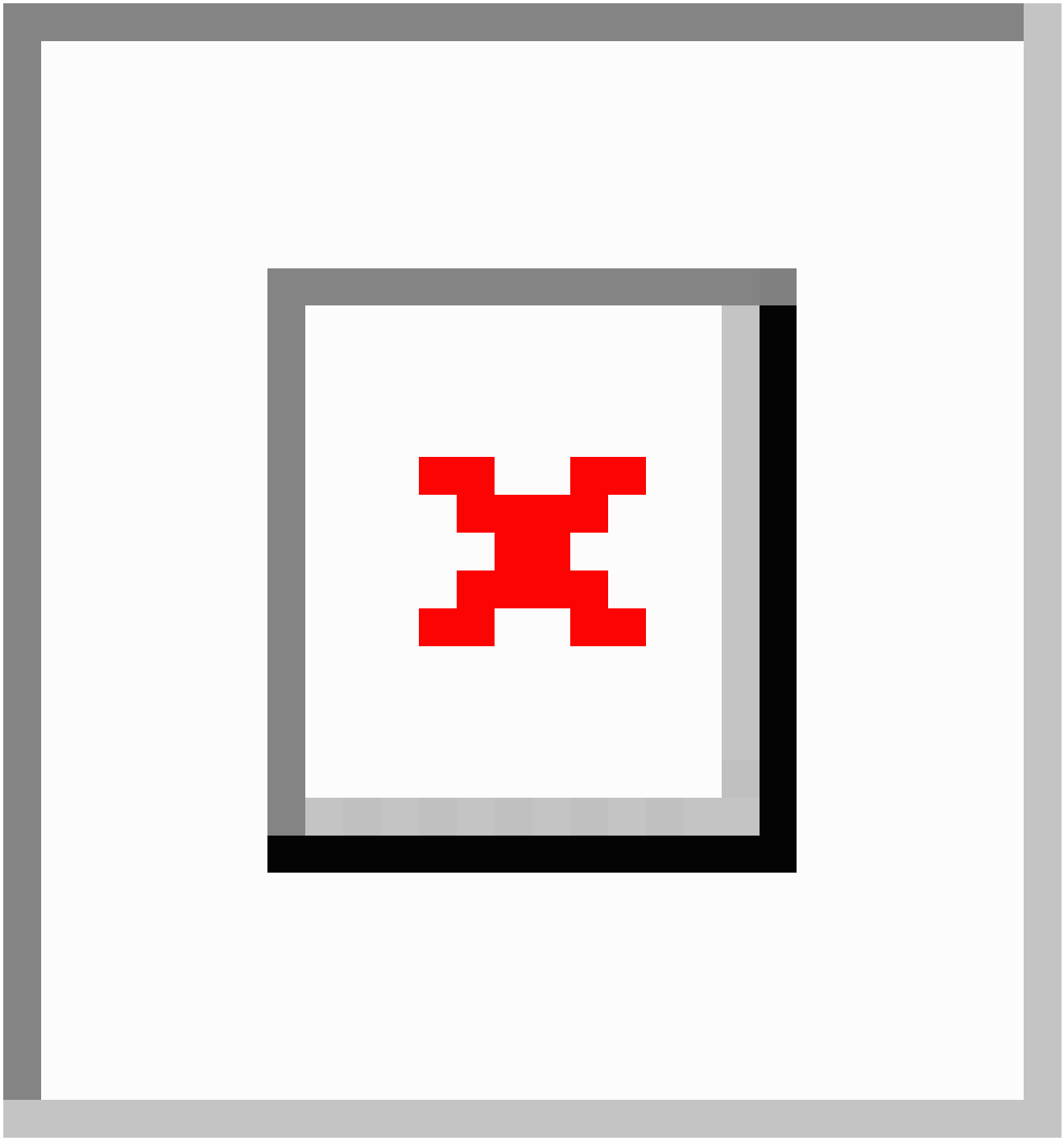
Among women who were sexually active, there was no significant statistical difference in the age of their sexual debut between the two groups (mean age, [SD]), SNS: 19.4 [2.8 years], Conventional: 18.9 [2.6 years], *P*=.16, by Student’s *t* test).

Differences in Responses to the Survey Between the Recruitment Methods

There was no statistical difference between the two groups in their responses to questions (embarrassing scores) about the

age when they became sexually active, number of sexual partners, regular condom use, sexually transmitted infection, and active use of oral contraceptives (Figure 1).

Figure 1. Mean embarrassing scores for Web-based questionnaires about sexual and reproductive behaviors among participants recruited through social media–based and other conventional methods.



Discussion

Principal Findings

Through this study, we found that Japanese women have similar responses to questions regarding sexual behavior and health, regardless of the way in which they were recruited.

We conducted this study in a similar manner to a prior Australian study by Fenner et al [9], but which did not incorporate conventional methods. We added the conventional methods because we were concerned that there were fewer social media users among Japanese women (72% in their teens and 64% in their 20s) than among Australian women (83% of 16-29-year-old women in 2010) [11].

Today in Japan (as of April 2016), 25 million people access their Facebook accounts monthly and 64% of them daily [12]. The use of smartphone has increased and encourages the frequent use of SNSs [13].

In our study, among those who were sexually active, more participants in the conventional group did not respond to the question regarding the age of sexual debut. We do not consider

that this difference means a conflict to similarity between recruitment methods. There are three reasons. First, there was no demographic difference between sexually active and nonactive (data not shown). Second, means and SDs of ages of sexual debut were similar between the two groups. Third, attitudes toward the Web-based survey were the same, irrespective of the recruitment method used (Figure 1). Therefore, we hypothesize that there might be a difference caused by anxiety about anonymity with our conventional methods of recruitment. This hypothesis should be verified by further research.

We can therefore assert that SNSs are promising and efficient tools for participant recruitment for sexual and reproductive health research.

Limitation

A limitation of this study was that we were able to recruit the SNS group as a randomized sample but not the conventional group. Some participants might have felt nervous about the possibility that their friends, teachers, or coworkers would learn about their sexual behavior, especially among participants in the conventional group. This might have resulted in a difference in the response rate to the question about age of sexual debut.

Conclusions

SNSs have made revolutionary changes in medical research. However, with respect to recruitment of study participants, these changes may be too large or rapid to permit validation of any differences between conventional and SNS-based recruitment methods. In this study, we confirmed similarities between these methods of recruitment for participation in a Web-based sexual and reproductive health survey. Cervical cancer prevention in Japan is declining. This is exemplified by low Papanicolaou cervical cytology screening rates which have resulted in increased cervical cancer mortality and incidence rates among younger Japanese women [14]. Moreover, the suspension of the human papillomavirus (HPV) vaccination promotion by the Japanese government in June 2013 has also decreased the coverage rate for the HPV vaccine from approximately 70% to almost zero [15]. Evidence-based measures must be taken to overcome the cervical cancer epidemic in Japan. This study confirmed that a nationwide SNS-based survey targeting young Japanese women would be a realistic way to spread information and awareness about cervical cancer prevention.

Acknowledgments

This study was funded by the Health and Research Grant awarded by the Ministry of Health, Labor and Welfare of Japan (H23-cancer research-general-001 and H26-innovative-cancer research-general-102) and a Japan Agency for Medical Research and Development Grant (15ck0106103h0102). The awarding agency had no role in the study design, data collection, analysis, interpretation of the data, and the decision to submit the manuscript for publication or the writing of the manuscript.

Authors' Contributions

YM and EM wrote the first draft of this manuscript. YM, EM, TE, JDW, and SMG contributed study design and data interpretation. YM and MT contributed data collection, statistical analysis, and quality control. YM, EM, TE, MA, JDW, and SMG contributed to critically reviewing the manuscript.

Conflicts of Interest

None declared.

References

1. McCallum EB, Peterson ZD. Investigating the impact of inquiry mode on self-reported sexual behavior: theoretical considerations and review of the literature. J Sex Res 2012 Mar 01;49(2-3):212-226. [doi: 10.1080/00224499.2012.658923] [Medline: 22380589]
2. Burkill S, Copas A, Couper MP, Clifton S, Prah P, Datta J, et al. Using the web to collect data on sensitive behaviours: a study looking at mode effects on the British National Survey of sexual attitudes and lifestyles. PLoS One 2016 Feb 11;11(2):e0147983 [FREE Full text] [doi: 10.1371/journal.pone.0147983] [Medline: 26866687]
3. Miyagi E, Motoki Y, Asai-Sato M, Taguri M, Morita S, Hirahara F, et al. Web-based recruiting for a survey on knowledge and awareness of cervical cancer prevention among young women living in Kanagawa prefecture, Japan. Int J Gynecol Cancer 2014 Sep;24(7):1347-1355 [FREE Full text] [doi: 10.1097/IGC.0000000000000220] [Medline: 25054449]
4. Heffernan KJ, Chang S, Maclean ST, Callegari ET, Garland SM, Reavley NJ, et al. Guidelines and recommendations for developing interactive eHealth apps for complex messaging in health promotion. JMIR Mhealth Uhealth 2016 Feb 09;4(1):e14 [FREE Full text] [doi: 10.2196/mhealth.4423] [Medline: 26860623]
5. Callegari ET, Reavley N, Garland SM, Gorelik A, Wark JD. Vitamin D status, bone mineral density and mental health in young Australian women: the Safe-D study. J Public Health Res 2015 Nov 30;4(3):594 [FREE Full text] [doi: 10.4081/jphr.2015.594] [Medline: 26753163]
6. Osborne SL, Tabrizi SN, Brotherton JML, Cornall AM, Wark JD, Wrede CD, et al. Assessing genital human papillomavirus genoprevalence in young Australian women following the introduction of a national vaccination program. Vaccine 2015 Jan 01;33(1):201-208. [doi: 10.1016/j.vaccine.2014.10.045] [Medline: 25444787]

7. Ahmed N, Jayasinghe Y, Wark JD, Fenner Y, Moore EE, Tabrizi SN, et al. Attitudes to Chlamydia screening elicited using the social networking site Facebook for subject recruitment. Sex Health 2013 Jul;10(3):224-228. [doi: 10.1071/SH12198] [Medline: 23702200]
8. Gunasekaran B, Jayasinghe Y, Fenner Y, Moore EE, Wark JD, Fletcher A, et al. Knowledge of human papillomavirus and cervical cancer among young women recruited using a social networking site. Sex Transm Infect 2013 Jun 01;89(4):327-329. [doi: 10.1136/sextrans-2012-050612] [Medline: 23047877]
9. Fenner Y, Garland SM, Moore EE, Jayasinghe Y, Fletcher A, Tabrizi SN, et al. Web-based recruiting for health research using a social networking site: an exploratory study. J Med Internet Res 2012 Feb 01;14(1):e20 [FREE Full text] [doi: 10.2196/jmir.1978] [Medline: 22297093]
10. Topolovec-Vranic J, Natarajan K. The use of social media in recruitment for medical research studies: a scoping review. J Med Internet Res 2016 Nov 7;18(11):e286 [FREE Full text] [doi: 10.2196/jmir.5698] [Medline: 27821383]
11. The Ministry of Internal Affairs and Communications, Japan. Soumu.go. 2011. Possibilities Engendered by a Symbiotic Networked Society URL: http://www.soumu.go.jp/johotsusintokei/whitepaper/eng/WP2011/chapter-3.pdf#page=1 [accessed 2016-11-25] [WebCite Cache ID 6mGxBGdgO]
12. Lifemedia, Inc. Research.lifemedia. 2015 Jun 17. Research of Japanese Facebook users June 2015 URL: http://research.lifemedia.jp/2015/06/150617_facebook.html [accessed 2016-11-25] [WebCite Cache ID 6mGyJTzEc]
13. The Ministry of Internal Affairs and Communications, Japan. Soumu.go. 2015 Jul 17. Press release for Internet utilization survey 2014 URL: http://www.soumu.go.jp/johotsusintokei/statistics/data/150717_1.pdf [accessed 2016-11-25] [WebCite Cache ID 6mGxYJiD4]
14. Motoki Y, Mizushima S, Taguri M, Takahashi K, Asano R, Kato H, et al. Increasing trends in cervical cancer mortality among young Japanese women below the age of 50 years: an analysis using the Kanagawa population-based Cancer Registry, 1975-2012. Cancer Epidemiol 2015 Oct;39(5):700-706. [doi: 10.1016/j.canep.2015.08.001] [Medline: 26277329]
15. Hanley JBS, Yoshioka E, Ito Y, Kishi R. HPV vaccination crisis in Japan. Lancet 2015 Jun 27;385(9987):2571. [doi: 10.1016/S0140-6736(15)61152-7] [Medline: 26122153]

Abbreviations

HPV: human papillomavirus
SD: standard deviation
SNS: social networking site

Edited by G Eysenbach; submitted 05.12.16; peer-reviewed by S Tao, K Takahashi; comments to author 25.01.17; revised version received 08.02.17; accepted 10.02.17; published 10.03.17.

Please cite as:
Motoki Y, Miyagi E, Taguri M, Asai-Sato M, Enomoto T, Wark JD, Garland SM
Comparison of Different Recruitment Methods for Sexual and Reproductive Health Research: Social Media–Based Versus Conventional Methods
J Med Internet Res 2017;19(3):e73
URL: http://www.jmir.org/2017/3/e73/
doi:10.2196/jmir.7048
PMID:28283466

Original Paper

Outcomes in Child Health: Exploring the Use of Social Media to Engage Parents in Patient-Centered Outcomes Research

Michele P Dyson[1], PhD; Kassi Shave[1], MSc; Ricardo M Fernandes[2,3], MD, PhD; Shannon D Scott[4], RN, PhD; Lisa Hartling[1], PhD

[1]Alberta Research Centre for Health Evidence, Department of Pediatrics, University of Alberta, Edmonton, AB, Canada

[2]Clinical Pharmacology Unit, Instituto de Medicina Molecular, University of Lisbon, Lisbon, Portugal

[3]Department of Pediatrics, Santa Maria Hospital, Lisbon, Portugal

[4]Faculty of Nursing, University of Alberta, Edmonton, AB, Canada

Corresponding Author:
Michele P Dyson, PhD
Alberta Research Centre for Health Evidence
Department of Pediatrics
University of Alberta
4-474 Edmonton Clinic Health Academy
11405 – 87 Avenue
Edmonton, AB, T6G 1C9
Canada
Phone: 1 780 492 1241
Fax: 1 780 248 5627
Email: mdyson@ualberta.ca

Abstract

Background: With the rapid growth of technology and its improved accessibility globally, social media is gaining an increasingly important role in health care. Patients are frequently engaging with social media to access information, share content, and interact with others in online health communities. However, the use of social media as a stakeholder engagement strategy has been minimally explored, and effective methods for involving participants in research on the identification of patient-centered outcomes remain unknown.

Objective: The aim of this study was to evaluate the process of using social media to engage parents in identifying patient-centered outcomes, using acute respiratory infections in children as an example to gauge feasibility.

Methods: We conducted a process evaluation of a two-phase Web-based strategy to engage parents in research on patient-centered outcomes. In the first phase, we developed a website and study-specific Facebook and Twitter accounts to recruit parents to complete a Web-based survey identifying patient-centered outcomes. In the second phase, we used Facebook to host discussion with parents based on the survey results. The reach of social media as an engagement strategy and the characteristics of the population recruited were assessed.

Results: During the first phase, there were 5027 visits to the survey site, 110 participants completed the survey, 553 unique users visited the study website (675 visits), the Facebook page received 104 likes, and the Twitter account gained 52 followers over the 14-week study period. Most survey respondents identified Facebook (51.8%, 57/110) or a friend (45.5%, 50/110) as their source of referral. It was found that 70.0% (77/110) of respondents resided in Canada, in urban centers (92.7%, 102/110), and 88.2% (97/110) had a college or university degree or higher. The median year of birth was 1978 and 90.0% (99/110) were female. Most survey responses (88.2%, 97/110) were completed during the first month of the study. In the second phase, 4 parents participated in the discussion, our Facebook page gained 43 followers, and our posts reached 5866 users. Mirroring the first phase, most followers were female (79%, 34/43), between 35 and 44 years (49%, 21/43), and resided in Canada (98%, 42/43). User engagement was variable, with no clear patterns emerging in timing, topic, or type of post.

Conclusions: Most participants were highly educated, urban dwelling, and female, limiting diversity and representativeness. The source of referral for most survey respondents was through Facebook or a friend, suggesting the importance of personal connections in the dissemination and uptake of information. User engagement with the Facebook-based discussion threads was inconsistent and difficult to initiate. Although there is substantial interest in the use of social media as a component of an effective

patient engagement strategy, challenges regarding participant engagement, representativeness, obtaining buy-in, and resources required must be carefully considered.

(J Med Internet Res 2017;19(3):e78) doi:10.2196/jmir.6655

KEYWORDS

social media; child health; patient outcome assessment; patient engagement

Introduction

Health Care

With major technological advances occurring in the past decade, the way patients seek out and engage in their health care is shifting. Nearly 3.5 billion people worldwide have access to a computer or mobile device at home, making a wealth of largely unvetted information widely available [1]. In health care, the Internet is gaining an increasingly important role, and 72% of American Internet users reported going online to seek out health information in 2012 [2]. Patients are using social networking websites to access health information from a wide variety of sources, post content about their personal health, and interact with others in online health communities [3]. This environment of open creation and exchange of user-generated content represents a significant evolution in the way information can be communicated and developed, and presents an opportunity for conducting research with an existing audience of engaged patients [4].

Along with its widespread use among health professionals and patients, the body of published literature examining the role of social media in health care is growing. Recent systematic review data have underscored the potential benefits, ethical issues, and unintended consequences of using social media in the health care environment [5-7]. For example, although online communities may represent an important source of social and emotional support for patients, participation in such communities may also result in diminished subjective well-being (eg, increased anxiety), and invoke concerns for privacy. Social media use may also impact, both positively and negatively, the relationships that patients have with their health care providers [8]. Despite the recent proliferation of literature with respect to the utility, benefits, and harms of social media use in health care, far less information exists with respect to utilizing social media to recruit and engage patients in health research.

The active engagement of patients in the research process is important, as it can improve on the credibility of the results and their relevance to the end users, thus reducing research waste [9]. Moreover, a number of studies have found that engaging patients in research improves rates of study enrollment and decreases losses to follow-up [9]. Within the context of patient engagement, patient-centered outcomes research seeks to elicit patients' unique perspectives on outcomes that are important to themselves and their families [10]. Patient-centered outcomes research, by definition, relies on engagement and input from patients themselves; however, participating in research may not be attractive for myriad reasons, including perceived and actual participant burden. Among the patient population, garnering interest in and excitement about research that will affect their own lives is crucial; however, knowledge of the optimal methods of engagement remains limited.

Social Media

Social media represents an innovative and potentially efficient method to attract attention, collect information, and generate dialogue from a widespread audience including patients [6,11]. Next to YouTube, Facebook and Twitter are the most accessed social media apps globally [12], thus representing opportune platforms to engage and recruit difficult-to-reach populations traditionally underrepresented in research [13-15]. Once engaged, effectively using social media to stimulate user-generated content can help to broaden the reach of a study through expanding personal and professional networks [16]. As participants engage in the study, people in their social networks are notified and may feel enticed to participate. Interactivity (eg, tweets, retweets, likes, comments, shares) may also build interest in and credibility of the study platform and subsequently increase followers.

In 2013 and 2016, we used social media to conduct two phases of recruitment for a mixed methods study (the OUTcomes in Child Health, OUTCH study) to understand parent perceptions of patient-centered outcomes for pediatric acute respiratory infections (ARIs), which represent a clinical area that is common among children and has a significant impact on the health care system, and patient and family well-being. Furthermore, ARIs are associated with significant morbidity and are one of the leading causes of illness and hospitalization in children [17]. Despite this burden of disease, there is a scarcity of research on families' perspectives of having a child with an ARI. Parents are a growing group of social media users [18,19], and given the breadth of our scope (ARIs; parents of children 0-17 years), represent a suitable audience to recruit through social media platforms.

In this paper, we evaluate the effectiveness of using social media as a strategy to recruit and engage with parents in determining patient-centered outcomes in a mixed methods study. Our specific objectives were to (1) systematically evaluate the process of using social media as a stakeholder engagement strategy and (2) examine the breadth and representativeness of the population engaged.

Methods

Study Design

Our mixed methods study, utilizing the sequential explanatory design [20], was conducted in two phases: an initial quantitative survey was conducted in which parents were asked to rate the importance of outcomes relevant to pediatric ARIs, followed by a qualitative follow-up phase in which we sought further

elaboration from parents on their rationale for their rating of outcomes for pediatric ARI. Ethics approval was granted for both phases by the University of Alberta Research Ethics Board; consent was implied through overt action on the part of participants. The methods for our social media engagement strategy for each phase of the study are described below.

Phase I: Utilizing Social Media to Recruit and Engage Parents in a Quantitative Survey

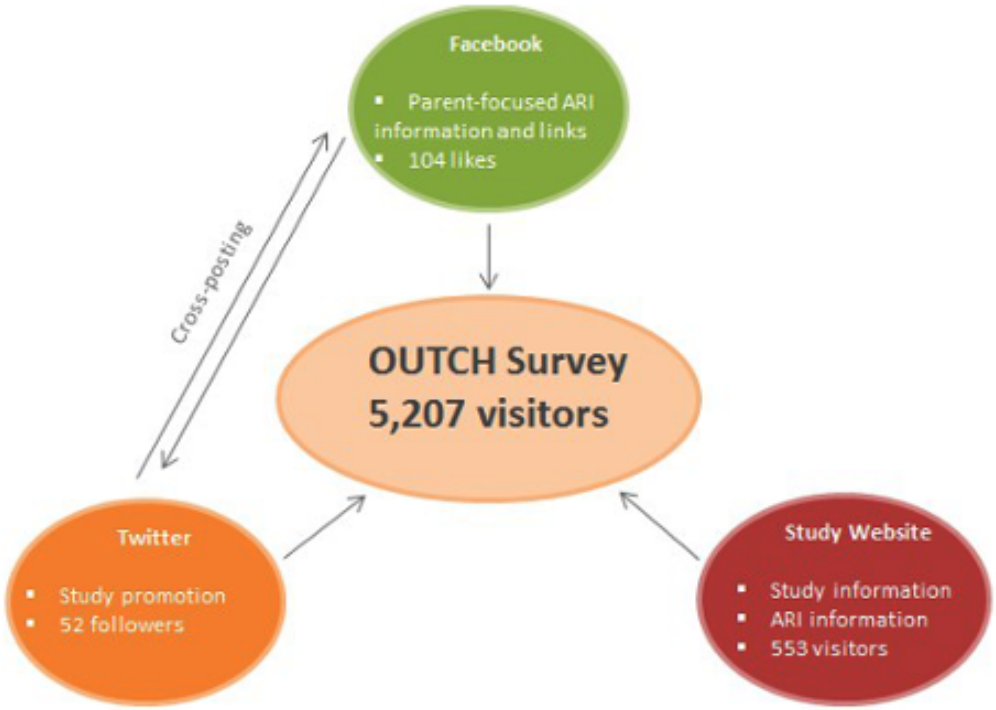
Our initial engagement strategy consisted of a suite of Web-based tools that we used as a hub to recruit potential participants and direct them to an open and voluntary Web-based survey on patient-important outcomes in pediatric ARI. We developed a Web presence with three components developed to promote the study and drive traffic to the survey: a study website [21] (see Multimedia Appendix 1) to host information about the study and conditions of interest; a Facebook page (OUTCH) [22] (see Multimedia Appendix 2) to post information and links related to pediatric ARI that would be of interest to parents; and a Twitter account (@OUTCH_Study) [23] (see Multimedia Appendix 3) to expand our network and cross-post content from our Facebook page. Although each of these platforms was used to present information in different formats, the primary goal of each was to engage parents with the aim of recruiting them as study participants, and direct them to the unique survey website (Figure 1).

Creating our own social media presence was important to establish a recognizable brand and to ensure credibility. We had direct control over the appearance and content on our website, which was necessary for ethical considerations, and employed consistent messaging throughout our channels. Although existing social media communities provide the advantage of an established group of followers, we sought to understand the opportunities and challenges associated with building a community of patients and families who share a common interest in informing patient-relevant research. Also, by relying on other channels to advertise one's study, control over the messaging is lost, increasing the potential for miscommunication, and increasing the complexity of tracking activity specific to our engagement strategy. Because social media, by its very nature, is interactive, we still expected to garner sufficient interest by engaging known, credible sources in our recruitment strategy. We also felt that having our own channels would facilitate continued active patient engagement moving forward.

We recruited parents through a snowball sampling technique [24], initially focusing on recruitment targets with the potential for a high yield of participants and then expanding through referrals and diffusion via social media. We targeted key local, national, and international English-speaking organizations with an interest in ARI and patient engagement, including TRanslating Emergency Knowledge for Kids (TREKK; research network with a Parent Advisory Group), the Cochrane Consumer Network (CCNet; organization that involves health care consumers in systematic reviews), the Alberta Centre for Child, Family & Community Research (ACCFCR; center that links government, academia, and the community), the Stollery Family Centered Care Network (network of families associated with a local children's hospital), and Mommy Connections (online parenting community) and asked them to promote the study to their audiences. We also built our audience by strategically seeking out followers (Twitter) or likes (Facebook) from individuals and organizations that emphasized parenting, health care, or local interest. Our social media platforms were intended to stimulate parent engagement and we posted daily content through both Facebook and Twitter. Facebook posts included information and links related to pediatric ARI that could be useful to parents. Posts on Twitter advertised the study and were linked to the Facebook page so that content was highlighted through both accounts. As an incentive, we included a raffle in which respondents could enter their names into a draw for an iPad mini, Kobo Touch, or Nike+ FuelBand.

We partnered with an online marketing company for the last 4 weeks of the 14-week survey period to evaluate the impact of paid advertising strategies on our recruitment efforts. We implemented a Facebook advertising campaign with standard ads (found along the right-hand side of the page) and page post ads (appear in the main News Feed), using an A/B testing format (head-to-head comparisons) to determine the highest performing ads. These evaluations were conducted over the course of the advertising campaign and the results were used to modify and target existing ads to maximize effectiveness.

Figure 1. Flow diagram outlining the online components of phase I of the OUTcomes in Child Health (OUTCH) study.



Phase II: Utilizing Social Media to Recruit and Engage Parents in Qualitative Follow-Up

In the second phase, our goal was to engage parents in discussion on our Facebook page, and our recruitment efforts focused almost exclusively on promotion through Facebook, which was chosen as our primary interface for interaction with parents based on our experiences and relative success with the platform over Twitter in phase I. We built in a 2-week promotional period in which all posts were focused on publicizing the upcoming study. During this phase, some of our affiliated research partners used Twitter to promote the study; however, all traffic was directed to our Facebook page. Over an 8-week study period, we used Facebook to host an online discussion forum in which we asked parents about the outcomes that matter to them when their child is ill with an ARI. Each week covered a new topic, with daily posts throughout the week, including discussion prompts and parent-friendly content related to respiratory infections in children. The discussion was moderated, with every comment receiving a follow-up response from the study team. To develop our network, we contacted several groups and organizations relevant to parents and child health in advance in an effort to obtain buy-in and encourage promotion of our study. These included online parenting communities (n=16), children's hospitals and associated foundations (n=14), and patient groups (n=3), both at the local and national levels. We also encouraged existing research networks and collaborators to help advertise our study (TREKK; Canadian Association of Paediatric Health Centres [CAPHC]; Knowledge Translation (KT) Canada; Alberta Research Centre for Health Evidence [ARCHE]). One of the parenting communities (Modern Mama) agreed to regularly share our posts for the duration of the study period for a nominal fee; any other engagement was informal and ad hoc.

Data Analysis

We measured user engagement with each of the online components through a series of metrics detailing reach and engagement from the study website (Google Analytics), survey website (Nooro), Facebook page (Facebook Insights), and Twitter account (followers). We also collected demographic information from our survey and available Facebook and Twitter data to examine the composition of our sample. We used descriptive statistics to analyze the data.

Results

The findings that describe the process evaluation of our social media engagement strategy are presented below, according to study phase. Data collected on patient-centered outcomes from the survey and qualitative follow-up are beyond the scope of this evaluation, and will be reported elsewhere.

Phase I: Utilizing Social Media to Recruit and Engage Parents in a Quantitative Survey

Our survey was live for 14 weeks between December 2013 and March 2014, and we received 110 responses. The study website received 675 visits from 553 unique visitors; 5207 users visited the survey website; the Facebook account received 104 "likes;" and the Twitter account gained 52 followers. The most common

sources of referral to the survey were Facebook (51.8%, 57/110), a friend (45.5%, 50/110), and Twitter (10.0%, 11/110).

The demographics of our sample are provided in Table 1. It was found that 90.0% (99/110) of respondents were female, the median year of birth was 1978 (range 1946-1995; age 38), and 88.2% (97/110) had a college or university degree or higher. Additionally, 70.0% (77/110) of respondents resided in Canada, and 92.7% (102/110) were based in urban centers.

Table 1. Characteristics of survey respondents in phase I (n=110).

Demographics		Response
Gender, n (%)		
	Female	99 (99.0)
	Male	11 (10.0)
Year of birth, median (range)		1978 (1946-1995)
Highest level of schooling completed, n (%)		
	Some high school	1 (0.9)
	High school graduate	5 (4.6)
	Some college or university	7 (6.4)
	College or university graduate	50 (45.5)
	Postgraduate education degree	47 (42.7)
Marital status, n (%)		
	Never married (single)	5 (4.6)
	Married or common-law	98 (89.1)
	Separated, divorced, or widowed	7 (6.4)
Annual household income in US $, n (%)		
	<30,000	5 (4.6)
	30,000-49,999	7 (6.4)
	50,000-69,999	15 (13.6)
	70,000-89,999	14 (12.7)
	>90,000	69 (62.7)
Country of residence, n (%)		
	Australia	2 (1.8)
	Canada	77 (70.0)
	England	8 (7.3)
	India	2 (1.8)
	United States	19 (17.3)
Type of community, n (%)		
	Urban (≥10,000 population)	102 (92.7)
	Rural (<10,000 population)	7 (6.4)
	Missing	1 (0.9)
Relationship to the children in the family, n (%)		
	Parent	106 (96.4)
	Step-parent	0 (0.0)
	Grandparent	4 (3.6)
	Other	2 (1.8)
Number of children living in the home, median (range)		2 (0-4)
What illness or illnesses that affect breathing has your child or children ever had? n (%)		
	Bronchiolitis	29 (26.4)
	Croup	49 (44.6)
	Strep throat or tonsillitis	36 (32.7)
	Sinusitis	32 (29.1)

Demographics		Response
	Wheezing	48 (43.6)
	Influenza	42 (38.2)
	Pneumonia	24 (21.8)
	Asthma	29 (26.4)
	Other	23 (20.9)
	Can’t remember	5 (4.6)
How did you hear about this survey? n (%)		
	Facebook	57 (51.8)
	Twitter	11 (10.0)
	Internet search	1 (0.9)
	Consumer group	4 (3.6)
	Friend	50 (45.5)
	Health care provider	3 (2.7)
	Research network	1 (0.9)
	Other	2 (1.8)

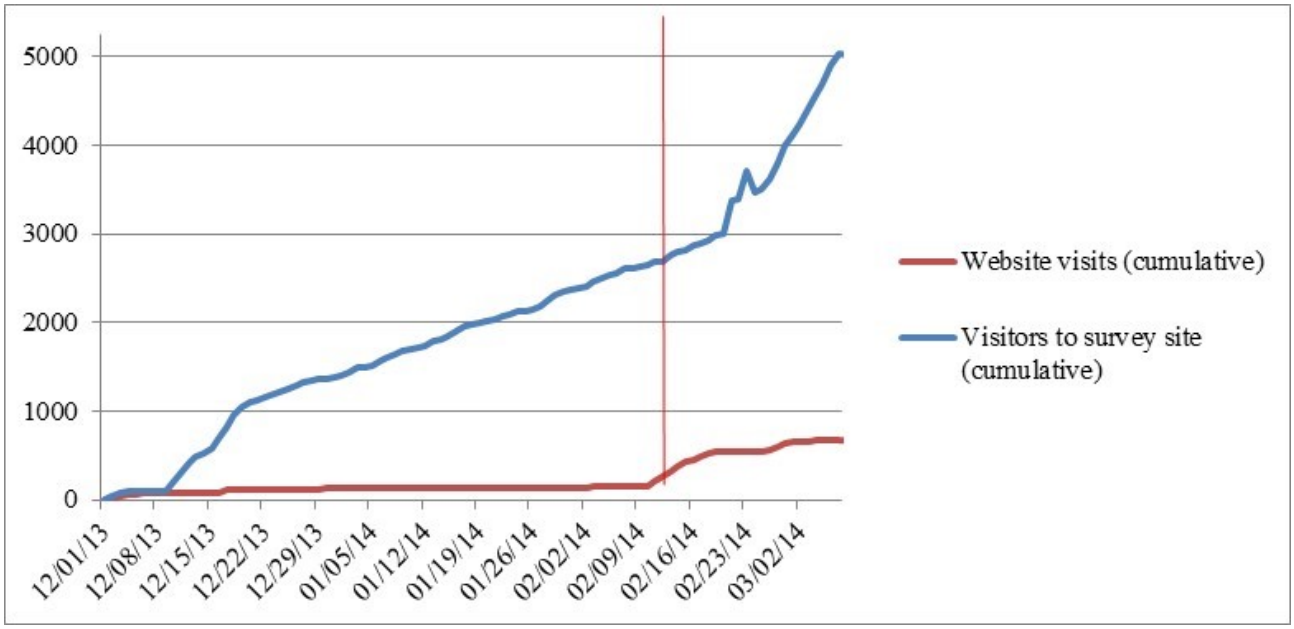
Study Website

On the study website, the home page received 88.2% (789/895) of all page views, with few users accessing the pages describing the study (6.6%, 59/895), the illnesses of interest (3.7%, 33/895), or study contact information (1.6%, 14/895). Engagement with the study website was typically brief, with 92.0% (621/675) of sessions resulting in drop-offs from the landing page, and 88.3% (596/675) of sessions lasting between 0 and 10 s. Most traffic (52.9%, 357/675) originated from the paid advertising campaign (described below), followed by 34.4% (232/675) from direct traffic (entering the specific Web address), and 11.1% (75/675) from social networks. Of the 75 visits that were derived from social media, 99% (74/75) were from Facebook and 1% (1/75) were from Twitter. Users accessed the website via a desktop computer in 44.6% (301/675) of cases, a mobile device in 37.6% (254/675), and a tablet in 17.8% (120/675). Of traffic originating from Facebook, 57% (42/74) of sessions were linked from the mobile site, and 43% (32/74) were from the main site.

Survey Website

Of 5027 visits to the survey website, 205 users opened the survey (4.08%, 205/5027). It was found that 37.1% (76/205) of surveys were blank, 9.3% (19/205) were partially completed, and 53.7% (110/205) were fully completed. On average, the 37-question survey took users 11.3 min (2 to 308.8 min) to complete. Of the surveys that were partially complete, all respondents completed the first page, and 16% (3/19) completed everything except the last section on demographics (18 questions). Data were not available on the origins of this traffic; however, traffic over time is depicted in Figure 2.

Figure 2. Web traffic to the study and survey websites (phase I). The red vertical line indicates the initiation of the advertising strategy.



Facebook

Over the course of the study period, the study Facebook page accumulated 93 lifetime total likes (104 total likes, minus 11 "unlikes" during the study period). The mean number of page views from unique logged-in users was 2 per day (median 1; interquartile range [IQR] 0-2; range 0-12). The daily total reach, encompassing the number of unique users who have seen any content associated with our page (eg, posts, likes, and comments), ranged from 0 to 78,869 people (mean 9111; median 20; IQR 9 to 3711), with substantial variation between the pre- and postadvertising phases (described below). Although we were unable to capture much information on the individuals who liked our page due to the use of privacy settings and the transient nature of followers (ie, can rapidly cycle between "liking" and "unliking" the page), we were able to obtain geographic locations for a sample of 63 users. Of these, 51% (32/63) were located in the United States, 46% (29/63) in Canada, and 3% (2/63) in Portugal; 87% (55/63) were in urban centers (≥10,000 population). In total, 10 followers out of all users (11%, 10/93) were directly linked to the study team (study investigators, center staff, and advertisers).

Twitter

We had 52 followers on Twitter. Over the 96-day study period, our tweets (n=168) received approximately 7200 impressions, indicating the number of times users saw a tweet on Twitter. Our tweets received 32 link clicks and 61 retweets. As with the Facebook profiles described above, we were limited in the amount of information we could extract on Twitter followers, but of a sample of 58 ever-followers, including 17 who unfollowed the account, users' profile descriptions primarily identified as a health care professional or researcher (n=22), an academic or professional organization (n=14), or a source of parenting or wellness information (n=7). Other categories included individual parents (n=5) and foundations or advocacy groups (n=2).

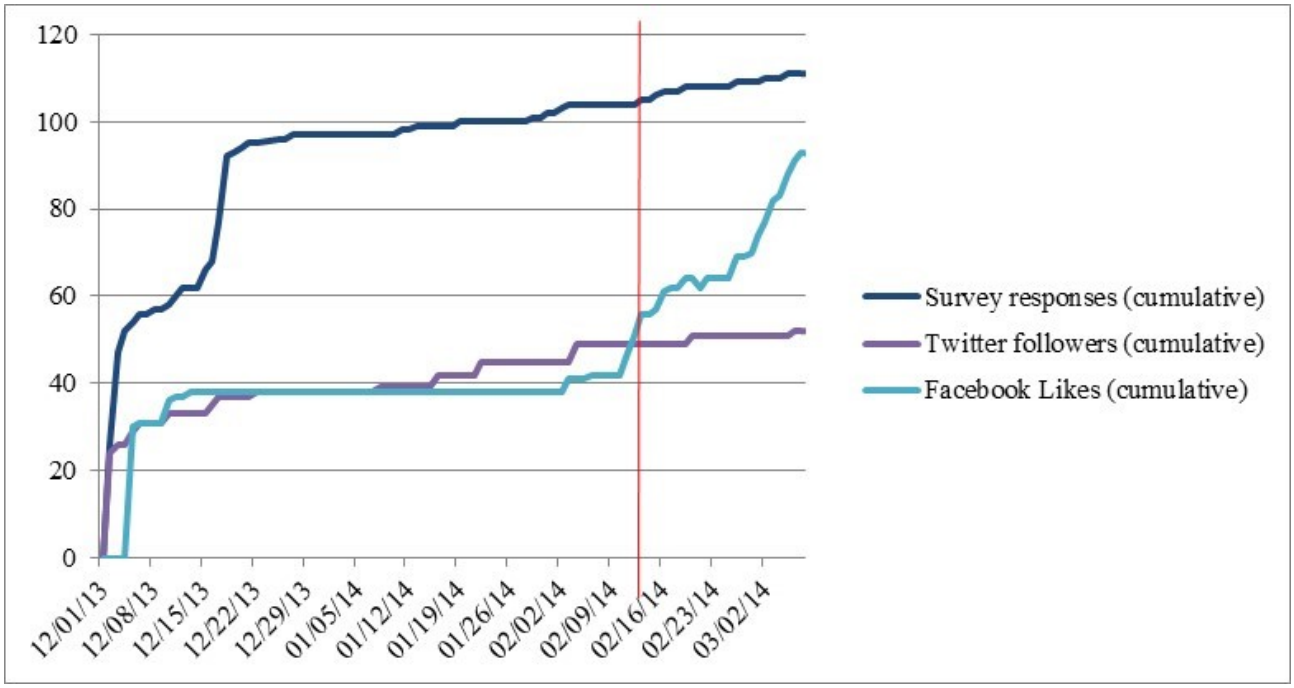
Impact of Advertising

Web traffic is depicted in Figures 2 and 3. In the 71 days before initiation of the paid advertising strategy, traffic to the study website was consistently low, with a mean of 2.3 visits per day (median 0; IQR 0-2; range 0-36). Our primary intention was to drive users to the survey website, and this was reflected with higher traffic, with a mean of 37.4 visits per day (median 30; IQR 18-46.5; range 0-133). Survey responses were highest in the first two days of the study period (42.7%; 47/110), with a mean response rate during the preadvertising phase of 1.5 per day. By the end of our initial study phase (preadvertising), 93.6% (103/110) of survey responses had been collected. Facebook likes and Twitter followers followed a similar pattern, with the sharpest rise at the beginning of the study, and activity leveling off after the first few days.

Our advertising strategy covered a 25-day period, in which 7 (6.4%; 7/110) respondents completed the questionnaire. The mean number of daily visits to the study website was 20.6 (median 10; IQR 3-5; range 0-61), and to the survey website was 94.8 (median 56; IQR 26-127; range 13-376). However, this encompasses a shift in emphasis. The first 10 days of advertisements were mistakenly driving traffic to the study website; after this period, clicks on the advertisements led directly to the survey website (Figure 2). Facebook likes increased during this phase (61 new likes), although engagement with the Twitter account remained static (3 new followers; Figure 3).

The total cost for our 4-week advertising campaign was CAD $1330 and reached an audience of an estimated 684,887 unique people. The advertisements received 1862 clicks that directed users to the study website or the survey. Overall, promoted posts were more effective than standard advertisements; and of the three highest performing advertisements (based on numbers of clicks and impressions), two were promoted posts and one was a standard advertisement. The top advertisement included content promoting our raffle.

Figure 3. Survey responses and social media engagement over time (phase I). The red vertical line indicates the initiation of the advertising strategy.



Phase II: Utilizing Social Media to Recruit and Engage Parents in Qualitative Follow-Up

We conducted our qualitative phase over a 2-week promotional and 8-week study period between January and March 2016. During this time, we published 52 posts, gained 43 Facebook followers (combined total for phases I and II: n=134), and attracted 4 participants to contribute to the discussion forum. Demographic data were collected from Facebook where possible; however, available data were limited by individuals' use of privacy settings (Table 2). Closely following the observed patterns from the survey phase, 79% (34/43) of Facebook followers were female, 49% (21/43) were 35-44 years of age, 98% (42/43) resided in Canada, and 100% (43/43) were based in urban centers.

Table 2. Characteristics of Facebook page followers gained in phase II (n=43).

Demographics		Response n (%)
Gender		
	Female	34 (79)
	Male	6 (14)
	Undisclosed	3 (7)
Age		
	18-24	2 (5)
	25-34	14 (33)
	35-44	21 (49)
	45-54	5 (12)
	55-64	1 (2)
	65+	0 (0)
Country of residence		
	Canada	42 (98)
	Egypt	1 (2)
Type of community		
	Urban (≥10,000 population)	22 (51)
	Rural (<10,000 population)	0 (0)
	Undisclosed geographic location	21 (49)

Facebook Post Activity

Over the course of 10 weeks, the 52 posts reached 5866 users (Table 3). Of these, 22 posts posed discussion questions to parents regarding the importance of certain outcomes related to ARI, 20 provided parent resources, and 10 were promotional. All contained images or videos, which were varied to include pictures of children, graphics, infographics, and other child-related content. We compared the three types of posts to determine whether patterns emerged in how users interacted with different content or types of engagement (Table 3). Although no strong trends surfaced, study discussion questions received the most attention, with the most comments (90%, 35/39), likes (46%, 17/37), and shares (67%, 48/72). However, on closer examination of our most successful posts (Table 4), we determined certain parent resources proved popular, with one post ranking second among the most liked posts by parents. Every Monday during the 8-week study period, we published a study discussion question that Modern Mama shared, increasing our reach for those posts. Two posts ranked among the five most successful in each category, each having a comparatively high reach and engagement through comments, likes, and shares. This is likely due to the fact that one was picked up and shared by the local children's hospital foundation, which has an established and engaged follower base, and the other promoted local research and was shared by users, who were aware of this study and associated parent resource. We did not observe changes in engagement over time, with individual posts instead seeming to stand on their own, without evidence of an increased following as the study progressed (Figures 4 and 5).

Table 3. Facebook post activity summary (phase II).

Period of study	Reach	Comments	Shares	Likes	Clicks	Impressions[a]
Promotional period						
Mean activity per post (median, range)	80 (55, 35-227)	0 (0, 0)	1 (1, 0-8)	1 (0, 0-9)	4 (0, 0-28)	131 (146, 64-182)
Mean activity by post type						
SDQ[b] (median, range)	-[c]	-	-	-	-	-
Parent resources (median, range)	59 (59, 56-62)	0 (0, 0)	0 (0, 0)	0 (0, 0)	0 (0, 0)	128 (128, 112-144)
Study promotion (median, range)	77 (49, 35-227)	0 (0, 0)	1 (0, 0-8)	2 (0, 0-9)	5 (0, 0-28)	119 (132, 64-164)
Total activity across all posts	800	0	10	12	40	1307
Study period						
Mean activity per post (median, range)	121 (72, 0-901)	1 (0, 0-15)	2 (1, 0-7)	1 (0, 0-4)	7 (3, 0-62)	119 (137, 10-259)
Mean activity by post type						
SDQ (median, range)	145 (89, 33-901)	2 (0, 0-15)	2 (0, 0-7)	1 (0, 0-3)	9 (4, 0-62)	122 (141, 14-259)
Parent resources (median, range)	97 (53, 0-600)	0 (0, 0-2)	1 (0, 0-7)	0 (0, 0-4)	5 (2, 0-36)	118 (138, 10-219)
Study promotion (median, range)	85 (55, 33-198)	0 (0, 0-1)	1 (0, 0-3)	1 (0, 0-3)	3 (3, 0-6)	134 (122, 58-235)
Total activity across all posts	5066	39 (17, 16, 6)[d]	62	25	293	5000
Summary						
Mean activity per post (median, range)	113 (66, 0-901)	1 (0, 0-15)	1 (0, 0-8)	1 (0, 0-9)	6 (2, 0-62)	121 (139, 10-259)
Total activity	5866	39 (17, 16, 6)[d]	72	37	333	6307

[a]An impression is the number of times a post is displayed to a Facebook user, with or without being clicked.

[b]SDQ: study discussion question.

[c]Hyphen indicates no posts of this type were published during the respective study period.

[d]Breakdown of study comments, participant comments, and untraceable comments.

Table 4. Summary of five most successful Facebook posts in each category by specific metrics (phase II).

Post timing and type	Post content	Reach (rank)	Comments[a] (rank)	Shares (rank)	Likes (rank)
Week 6 (Monday) SDQ[b]	Welcome to another week of the OUTCH[c] Study! Thanks for checking us out. This week, we're wondering about how your regular routine is disrupted when your child is sick. What happens in your family that's out of the ordinary? Tell us below!	900 (#1)	13 (#2)	5 (#5)	3 (#3)
Week 6 (Thursday) Parent resources	Coping with croup? Check out this great interactive Web page designed for parents. Learn more about croup, what you can do to help your little one, and when you need to see a doctor so you and your child can get back to your daily routines!	600 (#2)	2 (#4)	7 (#3)	4 (#2)
Week 1 (Wednesday) SDQ	Hey Parents! Have you had to take your child to the doctor or emergency department because they were having a hard time breathing? What was important to you?	352 (#3)	15 (#1)	4 (-)	3 (#3)
Promotion Week 2 (Monday) Study promotion	Starting next Monday: weekly discussion topics on acute respiratory infections. Tell us what matters most when your child is sick!	227 (#4)	0 (-)	8 (#1)	9 (#1)
Week 3 (Monday) SDQ	Welcome back parents! Here in Alberta we're just coming off of our Family Day long weekend, but we hope you had a relaxing weekend wherever you are! This week, we're talking about medical treatment for your kids. What was your experience when your child was sick and needed treatment to help their breathing? What went well? What could have been better? What worried you? What was reassuring? Please tell us in the comments below!	220 (#5)	0 (-)	4 (-)	2 (-)
Week 1 (Monday) SDQ	Thanks for checking out the OUTCH Study! We want to make sure that your child's health care is based on what's most important to you when their breathing is affected. This week, tell us what matters when your child needs to see the doctor.	198 (-)	0 (-)	3 (-)	3 (#3)
Promotion Week 2 (Thursday) Parent resources	Croup can be scary. Learn more about how to manage it at home and when it is time to see a doctor in this storybook!	167 (-)	0 (-)	2 (-)	3 (#3)
Week 5 (Monday) SDQ	Happy Monday, Moms and Dads! This week, we're talking about costs that might come up when your kids are sick. Tell us what has come up for you in the comments below!	141 (-)	0 (-)	7 (#2)	1 (-)
Week 2 (Wednesday) SDQ	If your child has needed tests like x-ray or blood tests when they're sick, what has mattered to you? Tell us in the comments below!	123 (-)	0 (-)	6 (#4)	0 (-)
Week 7 (Friday) SDQ	Hey Parents, has your child had any complications when they've been sick with a respiratory infection? Tell us about it in the comments below.	73 (-)	4 (#3)	1 (-)	0 (-)

[a]Only four posts have been ranked in this category because all others received no comments.

[b]SDQ: study discussion question.

[c]OUTCH: OUTcomes in Child Health.

Figure 4. Weekly reach of Facebook posts (phase II).

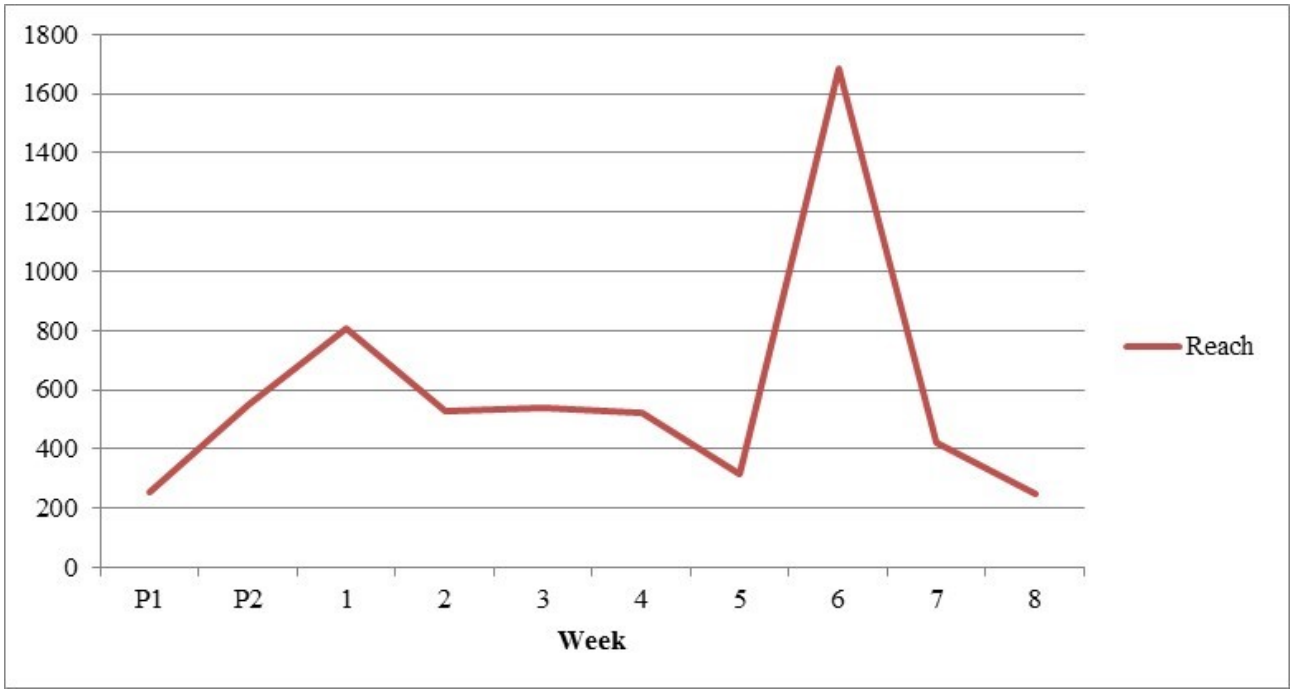
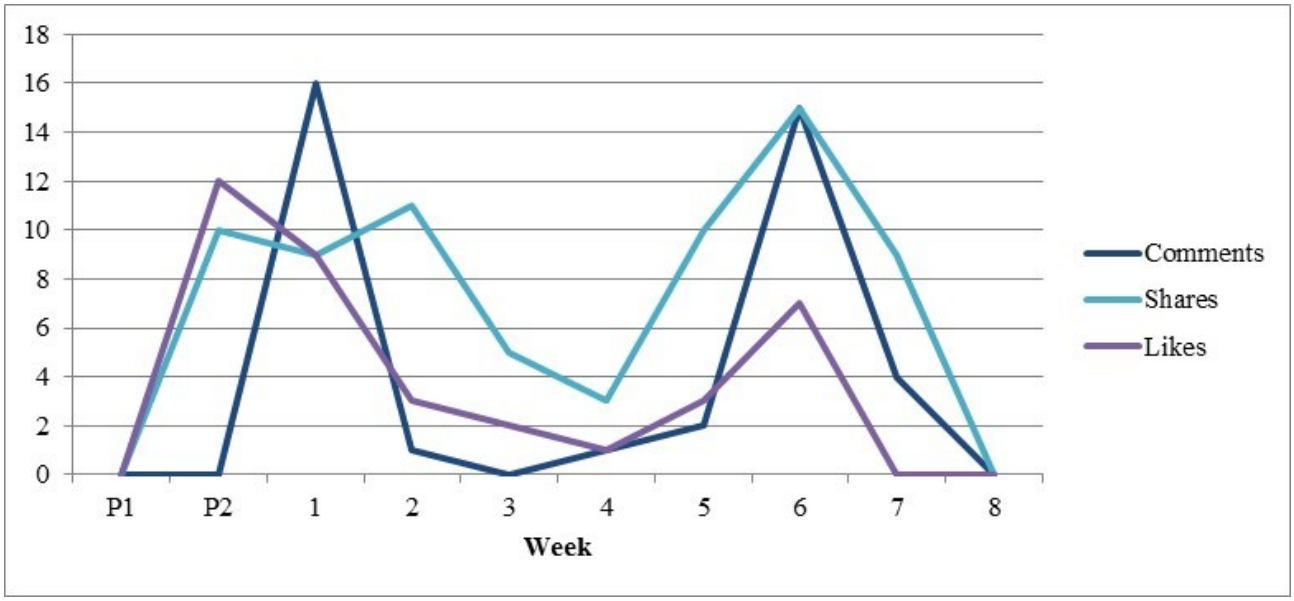


Figure 5. Weekly engagement with Facebook posts (phase II).



Stakeholder Engagement

We found that 4 parents contributed to the study discussion, of which 2 commented on more than one topic across study weeks, and 2 provided one comment each. Overall, our peak periods of activity were seen during the second promotional week and the first and sixth study weeks. Initially, members of the study team shared promotional material to encourage study involvement, driving activity levels in promotional Week 2. Week 1 encompassed an ongoing dialogue with a parent who had been recruited during the promotional period, resulting in our highest number of comments from the study. Week 6 benefitted from having one post shared by an influential organization, resulting in our highest reach and dialogue with two new study participants. The remaining weeks garnered relatively low levels of engagement. Stakeholder engagement over the study period is depicted in Multimedia Appendix 4.

Of the 33 organizations that we contacted in advance of the study, 5 provided support for promoting the study within their networks. With the exception of weekly shares by Modern Mama, each group advertised the study at one point in time. Other forms of collaboration from our research partners included tweets from TREKK and ARCHE, and a guest post published on the CAPHC blog (see Multimedia Appendix 4).

Discussion

Principal Findings

Social media is widely considered to be a promising set of tools in health research [25-34]. We evaluated a multicomponent

Web-based approach to engage parents in outcomes research that utilized a study website, Facebook page, and Twitter account to recruit participants to a survey and online discussion. Overall, this strategy was of limited effectiveness, resulting in 110 responses to the survey and recruitment of 4 parents to the discussion component.

On the basis of the limited demographic data available, we found that phases I (survey) and II (Facebook) of our study primarily reached a homogenous audience of highly educated urban-dwelling females. In part, this likely reflects that while parents of either gender are highly likely to use social media, mothers are significantly more likely to use Facebook than fathers and spend more time engaging with the platform [35]. However, the distribution of other demographic information suggests that our strategy was not successful in bridging other geographic and socioeconomic divides. As the majority of our audience was from Canada, we compared demographic data collected from our survey to recent data from Statistics Canada, highlighting this discrepancy. In terms of family structure, 89.1% (98/110) of our sample was married or in a common-law relationship, which is similar to national data (79.9% of children aged 14 years and under lived with married or common-law parents in 2011) [36]. Yet, while 42.7% (47/110) of our sample had completed a postgraduate degree, 9.4% of Canadians aged between 25 and 64 years old had done the same [37]. Additionally, we defined a rural community as having a population of less than 10,000 people, a classification which encompassed 6.4% (7/110) of our respondents. Statistics Canada uses a more restrictive definition of rural, with a population of less than 1000 people, and in 2006, 19.8% of the Canadian population lived in rural communities, and another 12.2% lived in small population centers (1000-29,999 population) [38]. In these respects, it does not appear that we engaged with a representative population. It may be argued that it is unnecessary or unrealistic to be representative of the general population because disparities remain in social media usage across groups. Among all American adults in 2015, 54% with a high school diploma or less used social media, compared with 70% with some college education and 76% with college or graduate degrees. Similarly, 58% of rural residents used social media, compared with 68% of suburban residents and 64% of urban residents [39]. However, pediatric ARIs are widespread across all demographics; therefore, we were unable to capture data on a representative sample of our target audience.

We observed a substantial gap between the number of visitors to the study and survey websites and those that proceeded to open and complete the survey. This may have been partly due to the design of the websites. Potential improvements to encourage increased participation may include optimizing the landing page to make it more appealing, having a prominent call to action, and reducing the number of clicks required to access the survey [40]. We had intentionally designed our online strategy to include multiple platforms that we could tailor according to their different functionalities. We wanted to maintain a study website that could provide detailed information about the study, as well as act as a resource for parents seeking information on ARI. However, landing on the study website necessitated an extra step for visitors wanting to access the survey, and this may have resulted in drop-offs due to the indirect pathway to the final website of interest [41]. Similarly, to meet the requirements of the approving ethics board, the landing page of the survey website required visitors to scroll through an information sheet before opening the survey, which may have turned some users away [42]. Although we had intended to use the best features of different online tools to emphasize different aspects of the study, we may have had more success with a consolidated strategy that could ease the demand on users.

Other potential contributors to the low response rate may have included initial uncertainty about eligibility, or being drawn in by the mention of a raffle but losing interest when asked to complete the survey. The length of the survey may also have posed an issue for some people [43]. Participants took a median time of 11.3 min (2 to 308.8 min) to complete the survey, which may have presented too onerous a task for some. Shortening the survey may have subsequently increased response rates [44,45]. Research has indicated that the longer a survey, the greater the perceived cost of participation and thus the lesser the likelihood that patients will be willing to participate [46]. As participants move through a survey, their responses also tend to diminish in quality such that the longer a survey, the lower the quality of responses as participants begin to fatigue [46]. Brief surveys (ie, one or two questions) can oftentimes be just as valid as longer ones, thus providing high quality data while also enticing more individuals to participate. In future patient engagement efforts, particular attention to survey length should be paid during the planning phase.

We initially used two different social media platforms in an effort to engage with parents: Facebook and Twitter. We put in equivalent efforts in promoting the study through both platforms, but the clear preference among this population was for Facebook. This is reflective of general social media usage trends, in which 71% of American online adults were on Facebook as of September 2014 and 23% used Twitter [47]. For the second phase, we drew from this experience and limited our efforts to engaging participants via Facebook. Despite a more focused strategy, it remained difficult to obtain buy-in, both at the levels of the target audience and of the key influencers we wished to involve, and we had little engagement with the account. Daily views and clicks were minimal, and were not proportional to the time and resources required to maintain a presence. The challenges we encountered in participant recruitment may, in part, be reflected by parents' perceived risk of participation. Minimally invasive research posing little individual risk is more appealing to potential study participants [48,49]. In light of this, obtaining buy-in using social media for research study participation may be difficult for certain health topics given the lack of confidentiality or anonymity associated with the interface. Using social media as a tool for parent engagement on scientific evidence or findings may prove less challenging, and would be an interesting topic for future research.

Recognizing the challenges associated with approaching members of the general population solely as researchers [50,51], we sought to increase credibility by partnering with health care organizations and parenting communities that were already known and trusted by our target audience; however, building

the necessary connections and relationships requires a significant investment of time, challenging feasibility within the budgetary and timeline constraints of grant funding. A further consideration is that many organizations will offer support for a fee, and therefore must be taken into account when budgeting for a project. Although, in principle, the possibility of messages spreading widely and organically via social media is appealing, without an established and well-connected social media presence, the reality is likely that far more often, a much more deliberate, and resource-intensive strategy is required. In future research, leveraging existing organizations and communities with access to the parent population early, consistently, and as frequently as budgeting permits throughout the duration of the study would likely improve buy-in and facilitate parent engagement and participation. Contacting organizational partners before the study start date, sending regular, brief reminders throughout the study period, and providing some degree of mutual benefit (eg, study progress updates, providing a summary of study results) to encourage their support may also prove beneficial.

During the first phase, we received far more traffic to the survey website from the paid advertising campaign than our organic strategy, although significantly fewer surveys were completed. Although we do not have data describing all visitors to the survey website, an interesting question that arises is when users determined whether or not they were eligible for the study. A potential explanation is that the Web traffic referred to the site was not actually representative of the target audience, in which case a more tailored recruitment strategy could be necessary. The majority of respondents indicated that the source of referral to the study was either through Facebook or a friend. Although we did not specifically target existing contacts, we noticed that many of the survey respondents belonged to the professional and social networks of the study team, suggesting not only that personal connections are an important facilitator, but also that our online reach was unable to spread much beyond our existing network. To maximize the utility of social media, it will be important to break this pattern. Of completed surveys, most were returned in the early days of the study period [52-54]. This is common to closed surveys with a set sampling frame; however, in theory, we could have had ongoing recruitment. This is likely reflective of our response pattern, in which many respondents had an association with the study team, and we were unable to expand our reach much beyond close contacts.

Strengths and Limitations

Patient-centered outcomes research is an emerging field, and the best methodologies for engaging stakeholders, including parents, have yet to be identified. Our study evaluated social media as a novel approach to engage parents in the research process, using a tailored and systematically measured strategy. Although our success in engaging participants was limited, an advantage of this project is that it highlights a number of the practical challenges associated with the use of social media for health research, providing guidance for future work in the area; specifically, the need to evaluate and compare purposive social media-based patient and family engagement strategies. Despite the fact that social media use is prevalent among our intended study population, there are nuances involved in navigating the dynamics and relationships that exist within these platforms that can make research challenging. Much of the existing evidence on the use of social media within health research has been based on ad hoc strategies and the assumption that the widespread use of social media will readily lend itself to research purposes; however, this does not capture the extent of the complexity. Within the field of patient-centered outcomes research, where stakeholder engagement is the driving component, identifying how patients and families perceive the use of social media in research, how they prefer to be involved, and whether different tools or approaches are more effective for different questions will be critical. The lessons learned from this study (Figure 6) will be of benefit to researchers seeking to engage patients in the identification of outcomes relevant to themselves and their families, as little guidance on using social media for this purpose currently exists.

The nature of social media allows for a transience and degree of control over privacy settings among users that does not permit comprehensive information to be collected on the demographic composition of the audience. Additionally, independent analytics programs were used to track metrics across each of our different online components, resulting in inconsistent data being collected from each source. Therefore, the information about our users and their activity was incomplete. A caveat to note, though, is that we used the free versions of each of these programs. Although upgrading would increase costs, it may permit a wider range of features that would allow a greater degree of harmony between data collected from different sources.

Our study was focused on parents of children with ARIs. This is a broad category, encompassing a range of conditions; therefore, we did not have a specific, well-defined end user or special interest group. Additionally, we concentrated on acute conditions. Our strategy may have had better success within a more defined condition or disease, or within a chronic condition, as it may have been more feasible to access a more readily available and already engaged network. However, respiratory conditions are common in children and patient-centered outcomes in this area are understudied, making evaluations important contributions to the evidence base.

Future research will be valuable in elucidating what makes a social media strategy effective in engaging stakeholders. In our study, we created and utilized new social media channels with no prior following. Using established social media channels to recruit and engage participants may have yielded different results (eg, increased engagement) and would be a worthwhile avenue for further exploration. The enthusiasm surrounding the potential uses of social media platforms can overshadow real constraints, including resources required, challenges in creating a network among the target audience, and navigating new territory for ethical approval. Many of the conventional ethics requirements are difficult to translate to research using social media given the vast potential reach. We did not have a defined sample that we were trying to access; therefore, issues related to sample size and participant demographics were not easily delineated upfront. Additionally, we were asked by the ethics board to provide Facebook and Twitter posts in advance, which runs counter to the evolving and interactive nature of social media platforms. Gaining exposure and continuing to evaluate

aspects of social media use will help to inform the development of efficient and exciting strategies to connect with patients, their families, and the public.

Figure 6. Lessons learned.

- Select audience appropriate social media channels based on current usage in the target population
- Begin promoting the study early, prior to data collection to garner interest
- Leverage the connections of existing organizations and communities early, consistently, and as frequently as budgeting permits for the duration of the study
- Send brief, regular reminders to organizational partners throughout the duration of the study encouraging support through their social media channels, and consider how to make their support mutually beneficial
- Closely consider participant burden, balancing information needs with the cost of participation for patients

Conclusions

Our social media engagement strategy resulted in a sample of survey respondents and discussion forum participants that were highly educated, urban dwelling, and female. The source of referral for most survey respondents was through Facebook or a friend, suggesting the importance of personal connections in the dissemination and uptake of information; however, user engagement with the Facebook-based discussion threads was inconsistent and difficult to initiate. Interaction with our online presence was limited relative to the resources required to deliver the approach. The use of social media as a component of a patient engagement strategy for outcomes research is associated with several challenges, and requires thoughtful design and implementation, dedicated resources, and methods in place to ensure representativeness. Future work to evaluate effective components of social media strategies, as well as patient and family preferences for engagement using this modality, will advance the science of patient-centered outcomes research.

Acknowledgments

The authors would like to thank Sarah Oberik-Olivieri for her graphic design work for this study, and Allison Gates for reviewing and providing feedback on the manuscript. This work was supported by KT Canada, grant number CIHR 87776 (SG-1) and Alberta Innovates—Health Solutions, grant number 201400561. LH is supported in part by a Canadian Institutes of Health Research New Investigator Award. SS is a Canada Research Chair (Tier II) for Knowledge Translation in Child Health and is also supported by an Alberta Innovates Health Solutions Population Health Investigator Award.

Authors' Contributions

MPD and LH designed the study. Data collection and analysis were conducted by MPD and KS with input from LH, RMF, and SDS. MPD oversaw all aspects of the study's implementation. All authors contributed to manuscript preparation and approved the final version. The funders had no role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Conflicts of Interest

None declared.

Multimedia Appendix 1

A static snapshot of the study website.

[JPG File, 195KB - jmir_v19i3e78_app1.jpg]

Multimedia Appendix 2

A static snapshot of the study Facebook page.

[JPG File, 244KB - jmir_v19i3e78_app2.jpg]

Multimedia Appendix 3

A static snapshot of the study Twitter page.

[JPG File, 171KB - jmir_v19i3e78_app3.jpg]

Multimedia Appendix 4

A network diagram of stakeholder engagement.

[PNG File, 121KB - jmir_v19i3e78_app4.png]

References

1. Internet Live Stats. Internet Live Stats. 2016. Internet users URL: http://www.internetlivestats.com/internet-users/ [accessed 2017-02-02] [WebCite Cache ID 6kBFdpYhE]
2. Fox S, Duggan M. Pew Research Center. 2013. Pew Internet & American Life Project URL: http://www.pewinternet.org/files/old-media/Files/Reports/PIP_HealthOnline.pdf [accessed 2017-03-02] [WebCite Cache ID 6kBIFnDhQ]
3. Fernández-Luque L, Bau T. Health and social media: perfect storm of information. Healthc Inform Res 2015 Apr;21(2):67-73 [FREE Full text] [doi: 10.4258/hir.2015.21.2.67] [Medline: 25995958]
4. Meyer ET, Schroeder R. The world wide web of research and access to knowledge. Knowl Manage Res Pract 2009 Aug 10;7(3):218-233. [doi: 10.1057/kmrp.2009.13]
5. Moorhead SA, Hazlett DE, Harrison L, Carroll JK, Irwin A, Hoving C. A new dimension of health care: systematic review of the uses, benefits, and limitations of social media for health communication. J Med Internet Res 2013;15(4):e85 [FREE Full text] [doi: 10.2196/jmir.1933] [Medline: 23615206]
6. Hamm MP, Chisholm A, Shulhan J, Milne A, Scott SD, Klassen TP, et al. Social media use by health care professionals and trainees: a scoping review. Acad Med 2013 Sep;88(9):1376-1383. [doi: 10.1097/ACM.0b013e31829eb91c] [Medline: 23887004]
7. Hors-Fraile S, Atique S, Mayer MA, Denecke K, Merolli M, Househ M. The unintended consequences of social media in healthcare: new problems and new solutions. Yearb Med Inform 2016 Nov 10(1):47-52. [doi: 10.15265/IY-2016-009] [Medline: 27830230]
8. Smailhodzic E, Hooijsma W, Boonstra A, Langley DJ. Social media use in healthcare: a systematic review of effects on patients and on their relationship with healthcare professionals. BMC Health Serv Res 2016 Aug 26;16:442 [FREE Full text] [doi: 10.1186/s12913-016-1691-0] [Medline: 27562728]
9. Domecq JP, Prutsky G, Elraiyah T, Wang Z, Nabhan M, Shippee N, et al. Patient engagement in research: a systematic review. BMC Health Serv Res 2014;14:89 [FREE Full text] [doi: 10.1186/1472-6963-14-89] [Medline: 24568690]
10. Frank L, Basch E, Selby JV, Patient-Centered Outcomes Research Institute. The PCORI perspective on patient-centered outcomes research. J Am Med Assoc 2014 Oct 15;312(15):1513-1514. [doi: 10.1001/jama.2014.11100] [Medline: 25167382]
11. Hamm MP, Shulhan J, Williams G, Milne A, Scott SD, Hartling L. A systematic review of the use and effectiveness of social media in child health. BMC Pediatr 2014;14:138 [FREE Full text] [doi: 10.1186/1471-2431-14-138] [Medline: 24886048]
12. Alexa. Alexa. 2016. The top 500 sites on the web URL: http://www.alexa.com/topsites [accessed 2017-03-02] [WebCite Cache ID 6kBGgVx86]
13. Quach S, Pereira JA, Russell ML, Wormsbecker AE, Ramsay H, Crowe L, et al. The good, bad, and ugly of online recruitment of parents for health-related focus groups: lessons learned. J Med Internet Res 2013 Jan;15(11):e250 [FREE Full text] [doi: 10.2196/jmir.2829] [Medline: 24231040]
14. O'Connor A, Jackson L, Goldsmith L, Skirton H. Can I get a retweet please? Health research recruitment and the Twittersphere. J Adv Nurs 2014 Mar;70(3):599-609. [doi: 10.1111/jan.12222] [Medline: 23909740]
15. Amon KL, Campbell AJ, Hawke C, Steinbeck K. Facebook as a recruitment tool for adolescent health research: a systematic review. Acad Pediatr 2014;14(5):439-47.e4. [doi: 10.1016/j.acap.2014.05.049] [Medline: 25169155]
16. Syred J, Naidoo C, Woodhall SC, Baraitser P. Would you tell everyone this? Facebook conversations as health promotion interventions. J Med Internet Res 2014;16(4):e108 [FREE Full text] [doi: 10.2196/jmir.3231] [Medline: 24727742]
17. Mulholland K. Global burden of acute respiratory infections in children: implications for interventions. Pediatr Pulmonol 2003 Dec;36(6):469-474. [doi: 10.1002/ppul.10344] [Medline: 14618637]
18. Masson H, Balfe M, Hackett S, Phillips J. Lost without a trace? Social networking and social research with a hard-to-reach population. Brit J Soc Work 2011 Nov 29;43(1):24-40. [doi: 10.1093/bjsw/bcr168]
19. Drozd F, Raeder S, Kraft P, Bjørkli CA. Multilevel growth curve analyses of treatment effects of a Web-based intervention for stress reduction: randomized controlled trial. J Med Internet Res 2013;15(4):e84 [FREE Full text] [doi: 10.2196/jmir.2570] [Medline: 23607962]
20. Creswell J, Clark V. Designing and Conducting Mixed Methods Research. Thousand Oaks, CA: Sage Publications, Inc; 2007.

21. Outcomes in Child Health. OUTCH. 2016. What is the OUTCH study? URL: http://www.outch-study.com/ [accessed 2017-03-02] [WebCite Cache ID 6kBGmVP0K]
22. Outcomes in Child Health. Facebook. 2016. OUTCH URL: http://on.fb.me/1MRsjJd [accessed 2017-03-01] [WebCite Cache ID 6kBGslz6H]
23. Outcomes in Child Health. Twitter. 2016. OUTCH URL: https://twitter.com/outch_study [accessed 2017-03-02] [WebCite Cache ID 6ke6Oqu5l]
24. Miles M, Huberman A. Qualitative data analysis: an expanded sourcebook. Thousand Oaks, CA: Sage Publications; 1994.
25. Coiera E. Social networks, social media, and social diseases. Br Med J 2013;346:f3007. [Medline: 23697672]
26. Swan M. Crowdsourced health research studies: an important emerging complement to clinical trials in the public health research ecosystem. J Med Internet Res 2012;14(2):e46 [FREE Full text] [doi: 10.2196/jmir.1988] [Medline: 22397809]
27. Burke-Garcia A, Scally G. Trending now: future directions in digital media for the public health sector. J Public Health (Oxf) 2014 Dec;36(4):527-534 [FREE Full text] [doi: 10.1093/pubmed/fdt125] [Medline: 24468840]
28. Allison M. Can web 2.0 reboot clinical trials? Nat Biotechnol 2009 Oct;27(10):895-902. [doi: 10.1038/nbt1009-895] [Medline: 19816442]
29. Bonilla-Warford N. Many social media options exist for optometrists. Optometry 2010 Nov;81(11):613-614. [Medline: 21504079]
30. Brownstein CA, Brownstein JS, Williams DS, Wicks P, Heywood JA. The power of social networking in medicine. Nat Biotechnol 2009 Oct;27(10):888-890. [doi: 10.1038/nbt1009-888] [Medline: 19816437]
31. Eytan T, Benabio J, Golla V, Parikh R, Stein S. Social media and the health system. Perm J 2011;15(1):71-74 [FREE Full text] [Medline: 21505625]
32. Hawn C. Take two aspirin and tweet me in the morning: how Twitter, Facebook, and other social media are reshaping health care. Health Aff (Millwood) 2009;28(2):361-368 [FREE Full text] [doi: 10.1377/hlthaff.28.2.361] [Medline: 19275991]
33. Spallek H, O'Donnell J, Clayton M, Anderson P, Krueger A. Paradigm shift or annoying distraction: emerging implications of web 2.0 for clinical practice. Appl Clin Inform 2010;1(2):96-115 [FREE Full text] [doi: 10.4338/ACI-2010-01-CR-0003] [Medline: 23616830]
34. Morris K. Tweet, post, share--a new school of health communication. Lancet Infect Dis 2011 Jul;11(7):500-501. [Medline: 21842546]
35. Duggan M, Lenhart A, Lampe C, Ellison N. Pew Research Center. 2015 Jul 16. Parents and Social Media URL: http://www.pewinternet.org/2015/07/16/parents-and-social-media/ [accessed 2017-03-01] [WebCite Cache ID 6kBH2IGjE]
36. Milan A, Bohnert N. Statistics Canada. 2012 Sep. Portrait of families and living arrangements in Canada: Families, households and marital status URL: http://www12.statcan.gc.ca/census-recensement/2011/as-sa/98-312-x/98-312-x2011001-eng.pdf [accessed 2017-03-02] [WebCite Cache ID 6kBITDLHx]
37. Statistics Canada. 2015 Dec 23. Education in Canada: attainment, field of study and location of study URL: http://www12.statcan.gc.ca/nhs-enm/2011/as-sa/99-012-x/2011001/tbl/tbl01-eng.cfm [accessed 2017-03-02] [WebCite Cache ID 6kBHDDlqf]
38. Statistics Canada. 2011. Standard Geographical Classification (SGC) URL: http://www.statcan.gc.ca/eng/subjects/standard/sgc/notice/sgc-06 [accessed 2016-08-31] [WebCite Cache ID 6kBHHpxsC]
39. Perrin A. Pew Research Center. 2015 Oct 08. Social Media Usage URL: http://www.pewinternet.org/2015/10/08/social-networking-usage-2005-2015/ [accessed 2017-03-02] [WebCite Cache ID 6kBHP4Yxj]
40. Edwards PJ, Roberts I, Clarke MJ, Diguiseppi C, Wentz R, Kwan I, et al. Methods to increase response to postal and electronic questionnaires. Cochrane Database Syst Rev 2009(3):MR000008. [doi: 10.1002/14651858.MR000008.pub4] [Medline: 19588449]
41. Porter J. User Interface Engineering. 2003. Testing the Three-Click Rule URL: https://www.uie.com/articles/three_click_rule/ [accessed 2017-03-02] [WebCite Cache ID 6kBHTmuV9]
42. Angus VC, Entwistle VA, Emslie MJ, Walker KA, Andrew JE. The requirement for prior consent to participate on survey response rates: a population-based survey in Grampian. BMC Health Serv Res 2003 Nov 18;3(1):21 [FREE Full text] [doi: 10.1186/1472-6963-3-21] [Medline: 14622444]
43. Sauermann H, Roach M. Increasing web survey response rates in innovation research: An experimental study of static and dynamic contact design features. Res Policy 2013 Feb;42(1):273-286. [doi: 10.1016/j.respol.2012.05.003]
44. Fan W, Yan Z. Factors affecting response rates of the web survey: A systematic review. Comput Human Behav 2010 Mar;26(2):132-139. [doi: 10.1016/j.chb.2009.10.015]
45. McPeake J, Bateson M, O'Neill A. Electronic surveys: how to maximise success. Nurse Res 2014 Jan;21(3):24-26. [doi: 10.7748/nr2014.01.21.3.24.e1205] [Medline: 24460562]
46. Galesic M, Bosnjak M. Effects of Questionnaire Length on Participation and Indicators of Response Quality in a Web Survey. Public Opin Q 2009 May 28;73(2):349-360. [doi: 10.1093/poq/nfp031]
47. Pew Research Center. 2014. Social Networking Fact Sheet URL: http://www.pewinternet.org/fact-sheets/social-networking-fact-sheet/ [accessed 2016-08-31] [WebCite Cache ID 6kBHa8Nz6]
48. Cobb EM, Singer DC, Davis MM. Public interest in medical research participation: differences by volunteer status and study type. Clin Transl Sci 2014 Apr;7(2):145-149 [FREE Full text] [doi: 10.1111/cts.12142] [Medline: 24456538]

49. Cottler LB, McCloskey DJ, Aguilar-Gaxiola S, Bennett NM, Strelnick H, Dwyer-White M, et al. Community needs, concerns, and perceptions about health research: findings from the clinical and translational science award sentinel network. Am J Public Health 2013 Sep;103(9):1685-1692 [FREE Full text] [doi: 10.2105/AJPH.2012.300941] [Medline: 23409875]
50. Corbie-Smith G, Thomas SB, Williams MV, Moody-Ayers S. Attitudes and beliefs of African Americans toward participation in medical research. J Gen Intern Med 1999 Sep;14(9):537-546 [FREE Full text] [Medline: 10491242]
51. Mills EJ, Seely D, Rachlis B, Griffith L, Wu P, Wilson K, et al. Barriers to participation in clinical trials of cancer: a meta-analysis and systematic review of patient-reported factors. Lancet Oncol 2006 Feb;7(2):141-148. [doi: 10.1016/S1470-2045(06)70576-9] [Medline: 16455478]
52. Sheehan K, McMillan S. Response variation in e-mail surveys: An exploration. J Advert Res 1999;39(4):45-54.
53. Zheng J. Survey Monkey. 2011 Jun 08. How Many Days Does it Take for Respondents to Respond to Your Survey? URL: https://www.surveymonkey.com/blog/2011/06/08/time-to-respond/ [accessed 2017-03-02] [WebCite Cache ID 6kBHsCWVp]
54. PeoplePulse. PeoplePulse. 2016. Survey Response Rates: Tips on How to Increase Your Survey Response Rates URL: http://www.peoplepulse.com/resources/useful-articles/survey-response-rates/ [accessed 2017-03-02] [WebCite Cache ID 6kBHeLrji]

Abbreviations

ACCFCR: Alberta Centre for Child, Family & Community Research
ARCHE: Alberta Research Centre for Health Evidence
ARI: acute respiratory infection
CAPHC: Canadian Association of Paediatric Health Centres
CCNet: Cochrane Consumer Network
IQR: interquartile range
KT Canada: Knowledge Translation Canada
OUTCH: OUTcomes in Child Health
TREKK: TRanslating Emergency Knowledge for Kids

Edited by G Eysenbach; submitted 19.09.16; peer-reviewed by S Gilbert, M Barwick; comments to author 17.11.16; revised version received 23.12.16; accepted 11.02.17; published 16.03.17.

Please cite as:
Dyson MP, Shave K, Fernandes RM, Scott SD, Hartling L
Outcomes in Child Health: Exploring the Use of Social Media to Engage Parents in Patient-Centered Outcomes Research
J Med Internet Res 2017;19(3):e78
URL: http://www.jmir.org/2017/3/e78/
doi:10.2196/jmir.6655
PMID:28302593

Original Paper

"Fitspiration" on Social Media: A Content Analysis of Gendered Images

Elise Rose Carrotte[1], BPsych (Hons); Ivanka Prichard[2], BBSc (Hons in Psychology), PhD; Megan Su Cheng Lim[1,3,4], BBiomedSc(Hons), PhD

[1]Burnet Institute, Melbourne, Australia

[2]School of Health Sciences, Flinders University, Adelaide, Australia

[3]School of Public Health and Preventive Medicine, Monash University, Melbourne, Australia

[4]Melbourne School of Population and Global Health, The University of Melbourne, Melbourne, Australia

Corresponding Author:
Elise Rose Carrotte, BPsych (Hons)
Burnet Institute
85 Commercial Road
Commercial Road
Melbourne, 3004
Australia
Phone: 61 385062365
Fax: 61 3 9282 2100
Email: elise.carrotte@burnet.edu.au

Abstract

Background: "Fitspiration" (also known as "fitspo") aims to inspire individuals to exercise and be healthy, but emerging research indicates exposure can negatively impact female body image. Fitspiration is frequently accessed on social media; however, it is currently unclear the degree to which messages about body image and exercise differ by gender of the subject.

Objective: The aim of our study was to conduct a content analysis to identify the characteristics of fitspiration content posted across social media and whether this differs according to subject gender.

Methods: Content tagged with #fitspo across Instagram, Facebook, Twitter, and Tumblr was extracted over a composite 30-minute period. All posts were analyzed by 2 independent coders according to a codebook.

Results: Of the 415/476 (87.2%) relevant posts extracted, most posts were on Instagram (360/415, 86.8%). Most posts (308/415, 74.2%) related thematically to exercise, and 81/415 (19.6%) related thematically to food. In total, 151 (36.4%) posts depicted only female subjects and 114/415 (27.5%) depicted only male subjects. Female subjects were typically thin but toned; male subjects were often muscular or hypermuscular. Within the images, female subjects were significantly more likely to be aged under 25 years (P<.001) than the male subjects, to have their full body visible (P=.001), and to have their buttocks emphasized (P<.001). Male subjects were more likely to have their face visible in the post (P=.005) than the female subjects. Female subjects were more likely to be sexualized than the male subjects (P=.002).

Conclusions: Female #fitspo subjects typically adhered to the thin or athletic ideal, and male subjects typically adhered to the muscular ideal. Future research and interventional efforts should consider the potential objectifying messages in fitspiration, as it relates to both female and male body image.

(J Med Internet Res 2017;19(3):e95) doi:10.2196/jmir.6368

KEYWORDS

social media; physical fitness; women's health; men's health; body image

Introduction

Overview of Fitspiration

An estimated 79% of young people use some form of social media daily [1]. Research suggests that young people are increasingly turning to social media for information about health and health behavior norms [2]. In recent years, a new fitness trend has emerged, providing Web-based and social media–based content designed to inspire individuals to exercise and be healthy. "Fitspiration," commonly shortened to "fitspo,"

is the broad term used to describe this "fitness inspiration" [3]. Fitspiration on social media allows users to view exercise-related images and videos and communicate with like-minded individuals. It often contains exercise tips, recipes, and photographs of food or people (including professional photographs, self-portraits ("selfies") and "before and after" images to highlight changes in weight or muscle). It is conceptually different from "thinspiration" (#thinspo), a form of media that deliberately promotes weight loss and thinness, and glorifies aspects of disordered eating behavior [4]. Social media users may follow dedicated fitspiration pages and profiles so that related content appears in their newsfeeds. Fitspiration-related social media posts are often tagged using hashtags—short words or phrases preceded by the hash or number symbol (#)—such as "#fitspo," allowing social media users to easily search for posts related to this topic. A recent cross-sectional survey estimated that 31% of young Australians like or follow fitspiration on social media, with young women more than twice as likely as young men to like these pages [5].

Impact of Fitspiration

Fitspiration is perceived to "model" ideas about health and fitness, shaping health beliefs and encouraging a "moral obligation" to achieve a particular body type among young women [6]. Experimental research has demonstrated negative effects from acute exposure to fitspiration among women, including increased negative mood and body dissatisfaction [7]. Acute exposure to fitspiration-style athletic ideal images (which depict a thin yet toned or muscular female body [8]) and exposure to "thinspiration"-style thin ideal images [9] predict body dissatisfaction and compulsive exercising among women at similar rates [10-12].

Meanwhile, contemporary men also commonly experience body dissatisfaction, particularly muscle dissatisfaction [13]. Exposure to traditional media (eg, print or magazine) muscular ideal images is associated with increased drive for muscularity and depression [14] and lowered muscle satisfaction among men [15]. Furthermore, exposure to images of male models actively engaging in sport has been shown to decrease men's satisfaction with their fitness levels and overall appearance [16]. However, the impact of male fitspiration images on men remains unknown.

Objectification Theory

Despite the focus on fitness, fitspiration images are argued to focus heavily on the appearance of the body and emphasize looks rather than body functionality [7]. According to objectification theory [17], this treatment of the body as an object ("objectification") is common in Western society, occurs in many forms, and disproportionately affects women. A common form of objectification is sexual objectification, where the body is treated and visually inspected as a collection of sexually appealing body parts [17]. Past research has shown that over 50% of the time, traditional media presents the female form as a "sex object," using a woman's sexuality to sell a product via facial expression, the amount of skin shown, and sexually suggestive camera angles. This figure rises to approximately 76% when examining images of women in men's magazines [18]. Furthermore, objectification may also be evident through facial prominence in images. Research suggests that images of men tend to focus on the head and face, whereas for women, the emphasis is on the whole body; this is known as "face-ism" [19]. A comprehensive analysis of objectification in fitspiration images would identify gender differences in these areas and provide avenues for future research.

Fitspiration Content Analyses

Content analyses have important implications for policy, as they can identify which types of content are potentially beneficial or harmful, and the degree to which fitspiration may be targeting different demographics. Two recent studies analyzed fitspiration websites; finding that these websites contain similar content to thinspiration websites, mostly depict women (>90% of images), and include objectified and sexualized women who are objectively thin [3,20]. However, it is unclear whether any social media–based fitspiration was included in these website analyses. In addition, gendered comparisons were absent from these studies [3,20], presumably as the majority of website-based content depicted women. Due to the popularity of fitspiration on social media, and its potential to facilitate peer-based body comparisons and reinforce social norms around health and fitness [5], it is also important to also study social media–based fitspiration. Recently, Tiggemann and Zaccardo [21] examined fitspiration on the image-based social media platform Instagram, finding that although most of these images (67%) depicted women, around 29% of images depicted men. Using categorical measurements, they found that most female subjects were thin and visibly muscular, whereas most male subjects were of medium build with a high level of muscularity. A quarter of subjects were engaged in some exercise activity and the majority of both men and women were objectified.

Although these previous content analyses have provided a broad overview of the common appearance-based messages of fitspiration, they do not provide any inferential statistics related to potential differences in male and female images. Tiggemann and Zaccardo [21] provided limited description of body type, activity engagement, and objectification between male and female subjects; however, this was only on one media platform (Instagram). It is currently unclear whether social media fitspiration content is most commonly posted to Instagram or to another social media platform, and whether these patterns exist across different social media platforms such as Facebook, Twitter, and Tumblr. Instagram and Tumblr are highly visual platforms compared with Twitter and Facebook, which have more of a mix of text- and image-based content. Furthermore, Instagram and Tumblr have more female users than male users [22].

Although women are more likely than men to access social media–based fitspiration [5], since over a quarter of Instagram content appears to be aimed at men [21], it is also important to examine messages aimed at men. Prior studies of promuscularity websites indicate that these websites contain messages about rigid exercise and dietary practices and aspirations toward an "ideal" muscular body [23]. Such messages may contribute toward rates of body dissatisfaction, disordered eating, and compulsive exercise behaviors in men. Studying these gendered differences in social media fitspiration content by using inferential statistics to clearly identify significant differences,

could help identify patterns of fitspiration messages, inform future experimental research, and help to develop and refine interventions aimed at men and women. Furthermore, no prior content analyses have explored the age of fitspiration subjects. It is important to determine whether fitspiration is targeting particular age groups. Of note, body image concerns tend to commence earlier in girls than boys [24], and younger people disproportionately access fitspiration [4].

This Study

This study aimed to describe and identify the characteristics of fitspiration content posted across social media (ie, Instagram, Facebook, Twitter, and Tumblr) via a public hashtag with regards to body image messages, food or dieting messages, and exercise messages by gender of subjects (see Multimedia Appendix 1 for variable description). In accordance with objectification theory and media trends, it was hypothesized that (1) posts would more frequently appear on Instagram and Tumblr than Twitter and Facebook; (2) more posts would depict images of women than images of men; (3) women would be thinner whereas men would be more muscular; (4) posts depicting women would be more likely to demonstrate objectification through sexualized imagery, emphasis placed on the look of the full body, emphasis placed on specific body parts (stomach or buttocks), and fewer depictions of the face, and (5) women would appear younger than men.

Methods

Selection of Content

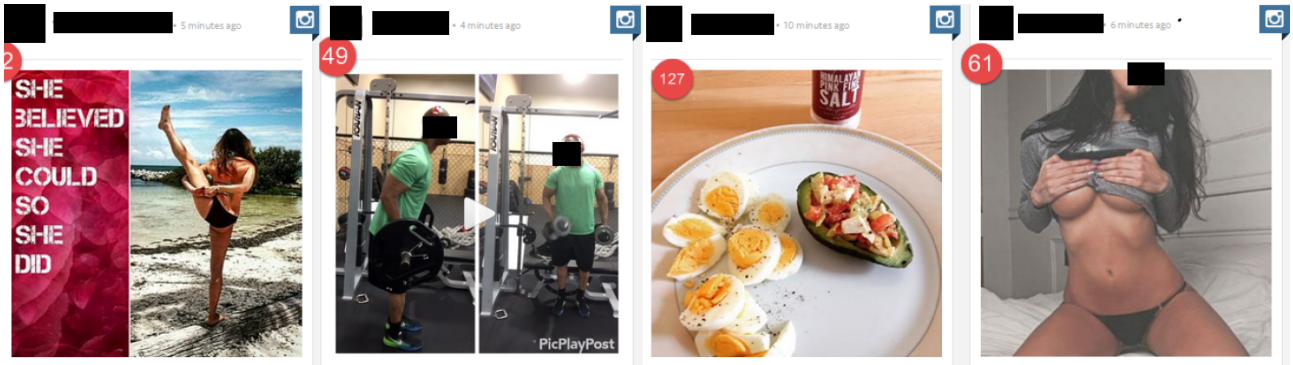
This content analysis involved analyzing recent posts with the "#fitspo" hashtag across 4 social media platforms: Instagram, Facebook, Twitter, and Tumblr. At 3 randomly generated timeslots, #fitspo was searched across Instagram, Tumblr, Facebook, and Twitter and all posts were extracted using screenshot methods over the next 10 minutes. No best practice tools are available for systematically searching social media, and various websites' default search algorithms do not allow systematic searching. At the time of data collection, Facebook and Pinterest did not allow accurate searching of "most recent" tagged posts, instead showing a combination of recent and "popular" posts when searching a tag. Instagram allows searching of recent posts for public, but not private profiles, but does not allow searching of popular posts. To address these issues, the website tagboard.com was used, which tracks recent, public posts with hashtags across Instagram, Facebook, and Twitter. Tumblr allows searching of both recent and popular posts; all profiles are public, so this platform was searched directly. The authors were unable to search for recent posts on Pinterest, another platform of interest, so this platform was not included in the analysis.

Coding Strategy

Given the emergent nature of research into this field, posts were analyzed with a codebook developed specifically for this project by the authors (see Figure 1 for example images and Multimedia Appendix 1 for codebook). This codebook was expanded from Boepple and Thompson's [3], and Tiggemann and Zaccardo's [21] analyses. Variables were also informed by objectification theory [17]. Twenty-eight variables were chosen by the authors. These included variables related to the relevance and structure of the post (eg, social media platform type, presence of a caption, presence of a photo or video), the theme of the post (exercise or food), number of people present and their apparent gender and age (in categories), whether the post depicted the subject's face or body, whether the post emphasized the subject's stomach or buttocks, and whether the post encouraged healthy eating or included a person actively exercising. For variables relating to the subject of the post, if multiple people of the same gender were present, coders were instructed to code variables related to the main subject of the post. This was determined by the structure of the post highlighting 1 subject (eg, 1 subject in the foreground). If this was unclear, the coder analyzed the subject on the left of the image (if viewing the image from the left to the right, this would be the first person viewed). For ease of interpretation, age was later collapsed into a binary variable of <25 or 25+ years for the gendered analysis.

Thinness and muscularity of the people in the images were rated using line drawing figure scales: the Figure Rating Scale (also known as Stunkard scale) for male and female thinness [25] and modified versions of this scale for male and female muscularity [26,27]. These scales use simple drawings of male and female figures to assess thinness and muscularity of men and women on scales of 1 (very thin or very little visible muscle) to 9 (very overweight or very muscular). Body parts emphasized were determined by visual cues such as proximity to the camera, cropping, and captions which drew attention to particular body parts.

Figure 1. Examples of posts included in the content analysis. These posts demonstrate the following variables, among others: (L-R) full body, active exerciser, healthy eating, and sexualisation.



Each post was analyzed by 2 independent coders. Both coders were trained in using the codebook before the analysis process using example posts not included in the analysis. Coders were instructed to assess the main messages or clear implications of the post, and to use any visible captions, hashtags, and comments to give context to the post. Coders were instructed to view videos, if possible, using links embedded in screenshots taken with tagboard.com. If this was not possible, coders rated the single frame video preview. Reliability was analyzed between the 2 coders using Cohen kappa for categorical variables (see Multimedia Appendix 1). Categorical variables included in the analysis exceeded the recommendation for a minimum value of κ=.60 [28] and percentage agreements for these variables ranged from 87.8% (for "Full Body") to 99.5% (for "Platform"). Five categorical variables were removed due to low reliability. Reliability was analyzed for ordinal and interval-level variables using Spearman rho; these values were highly correlated, ranging from with ρ=.74 for thinness to ρ=.96 for number of women present. A third reviewer, the lead author, independently analyzed any discrepancies in accordance to the codebook with consideration of the 2 primary coders' data, making the final decision and forming the final dataset.

Gender Analysis

In posts containing men only and those containing women only, gender differences in mean thinness and muscularity ratings were analyzed using independent *t* tests, and chi-square or Fisher exact tests were used for analyzing categorical variables by gender of subjects. Cohen *d* or the phi coefficient was used to calculate effect sizes. Analyses were conducted in Stata version 13 (StataCorp LLC).

Results

Description of Content

Across the 30-minute composite study period, 476 social media posts tagged with "#fitspo" were retrieved across the 4 platforms. After removal of 61 irrelevant posts (κ=.73; see Multimedia Appendix 1), 415 of 476 posts (87.2%) were coded (Table 1). The most content was posted to Instagram (360/415, 86.8%), followed by Tumblr (39/415, 9.4%), Facebook (12/415, 2.9%), and Twitter (4/415, 1.0%) (κ=.98). Due to the relatively small number of posts across Tumblr, Facebook, and Twitter, variables were not compared by platform.

Most #fitspo posts contained a photograph or another image, and around 1 in 10 posts contained a video. More women appeared in posts than men, but men still appeared in nearly one third of posts; 40.5% (168/415) depicted at least one woman and 31.6% (131/415) depicted at least one man. In general, subjects of posts appeared to be young adults. Most posts were thematically related to exercise or fitness, although around 1 in 6 was thematically related to food (Table 1). Most messages about food (64/68, 94%) were found in posts without any people present, whereas most messages about exercise (245/295, 83.1%) were found in posts with at least one person present.

Table 1. Description of social media posts tagged with #fitspo (N=415).

Type of variable	Category	Level	Posts n (%)
Structure	Text	Motivational or inspirational quote or text	67 (16.1)
	Caption	Caption present	330 (79.5)
	Photo	Photo present	332 (80.0)
	Video	Video present	41 (9.9)
	Selfie	Post contains a selfie	112 (27.0)
Subjects	Person[a]	At least one person present	283 (68.2)
	Age (years)	<18	7 (1.7)
		18-24	104 (25.1)
		25-34	81 (19.5)
		35-44	12 (2.9)
		45+	1 (0.2)
		Multiple people of different ages	10 (2.4)
		Unclear	68 (16.4)
		Not applicable; no people present	132 (31.8)
	Gender[b]	Women only	151 (36.4)
		Men only	114 (27.5)
		Women and men	17 (4.1)
		Unclear	1 (0.2)
		Not applicable; no people present	132 (31.8)
Theme	Exercise	Thematically relates to exercise or fitness only	295 (70.1)
	Food	Thematically relates to food or eating only	68 (16.4)
	Both	Thematically relates to both food or exercise	13 (3.2)
	Neither	Thematically relates to neither food nor exercise	39 (9.4)

[a]Mean number of people present per post: 1.4 (SD 1.7, when at least one person present).

[b]Mean number of men present: 1.1 (SD 0.7); mean number of women present: 1.4 (SD 2.0, when at least one man or one woman were present, respectively).

Gendered Analysis

Gendered analyses of posts depicting only men and posts depicting only women (265/415, 63.9%) were conducted. On the Figure Rating Scales, female subjects had significantly lower mean thinness scores (Mean 3.0, SD 1.2) than male subjects (Mean 4.4, SD 1.1), P<.001, Cohen d=−1.11, large effect. Female subjects also had significantly lower muscularity scores (Mean 4.1, SD 1.7) than male subjects (Mean 6.0, SD 1.2), P<.001, Cohen d=−1.30, large effect. Of note, these thinness and muscularity analyses were conducted only among posts where reviewers were able to assess thinness (n=126 for female subjects, n=89 for male subjects) and muscularity (n=116 for female subjects, n=94 for male subjects).

Categorical variables are presented in Table 2 (variable descriptions available in Multimedia Appendix 1). Posts containing women only were significantly more likely to display the subject's full body than posts containing men only. Posts containing men only were significantly more likely to have their subject's face visible than posts containing women only. Nearly half of posts emphasized the subject's stomach or contained an active exerciser, regardless of gender. Women were sexualized significantly more than men, although sexualization was common for both men and women. Posts containing women only emphasized the subject's buttocks significantly more frequently than posts containing men only. Before and after images, and messages about food and healthy eating, were relatively rare regardless of gender. Messages about food and healthy eating were generally depicted in posts without any people present.

Table 2. Content of social media #fitspo posts by gender of subject when only one gender was present in the post.

Type of variable	Category	Female subjects only (n=151) n (%)	Male subjects only (n=114) n (%)	*P* value	Phi (ϕ)	Effect size
Age of subjects	<25 years	78 (70.9)	31 (37.8)	<.001	.33	Medium
Structure of post						
	Motivational text	11 (7.3)	11 (9.7)	.49		
	Caption	118 (78.2)	86 (75.4)	.60		
	Photo	135 (89.4)	98 (86.0)	.40		
	Video	18 (11.9)	21 (18.4)	.14		
	Selfie	66 (43.7)	38 (33.3)	.09		
Theme of post						
	Food only[a]	1 (0.7)	2 (1.8)	.58		
	Exercise only	126 (83.4)	104 (91.2)	.06		
Objectification						
	Face visible	76 (50.3)	77 (67.5)	.005	.17	Small
	Full body or nearly full body visible	98 (64.9)	51 (44.7)	.001	.20	Small
	Emphasis on stomach	69 (45.7)	47 (41.2)	.47		
	Emphasis on buttocks[a]	38 (25.2)	2 (1.8)	<.001	.71	Large
	Sexualization	72 (47.7)	33 (29.0)	.002	.19	Small
Other variables						
	Before or after[a]	3 (2.0)	1 (0.9)	.64		
	Healthy eating[a]	4 (2.7)	5 (4.4)	.51		
	Active exerciser	80 (53.0)	61 (53.5)	.93		

[a]Chi-square tests only performed with minimum frequency >5 per cell. Fisher exact test was used if cell frequency was 5 or fewer.

Discussion

Principal Findings

This study aimed to analyze the content of posts with the #fitspo hashtag over a composite 30-minute period across the social media platforms Instagram, Facebook, Twitter, and Tumblr, comparing messages of fitspiration by gender of the subject. As hypothesized, the vast majority of posts in this time period were posted to Instagram, followed by Tumblr, probably due to their highly visual nature. Similar to the findings of Tiggemann and Zaccardo [21], female subjects were more frequently depicted in fitspiration than male subjects; however, a third of posts depicted at least one male subject. This indicates that, in contrast to the female focus of previous fitspiration research (eg, [3,7]), it is likely that fitspiration is reaching men and may influence the body image, exercise, and health behaviors of male followers as well as female followers. As such, future research should investigate the potential impact of these images on men.

Posts depicting women, which are likely aimed at female social media users, typically depicted young adult women meeting either the thin ideal or the athletic ideal. Posts depicting very muscular women were also common, and may indicate an increase in the popularity of body building among women. Female subjects were frequently objectified and sexualized, with emphasis on the idealized look of their full body and body parts such as the stomach and buttocks. Considering the popularity of fitspiration among young women [5], and the detrimental effects on female body image observed previously [7,10], it is fair to assume that young women viewing this content are frequently exposed to images of thin and athletic ideal bodies, often sexualized, and that this content has the power to influence their body image and encourage exercise to alter their appearance. Furthermore, only half of the images of women actually depicted women's faces (face-ism [19]), following a similar trend to that observed by Tiggemann and Zaccardo [21]. This trend may also be related to greater self-objectification, particularly in the context of social media, where the profile pictures of male users are more likely to depict their faces compared with female social media users [29].

Interestingly, men were frequently objectified, albeit in different ways to women. Although men had their faces in images more often than women, posts depicting men (presumably aimed at male social media users) depicted subjects that were highly muscular—significantly more muscular than female subjects. Images of men focused on stomachs at a similar frequency to women, but unlike women, men's stomachs nearly always had

visible abdominal muscles. Although not included in this analysis, images of men often emphasized visible biceps and pectoral muscles. Compared with posts with female subjects, in the images of men there was little emphasis on thinness or weight loss. This is aligned with current body ideals for men [13] and a general trend seen in the media of increased muscle mass in images of male bodies over time [30]. However, previous research indicates that young men access fitspiration on social media less frequently than young women [5]. There may be a small group of social media users posting content to the #fitspo hashtag who are very active and have a small but dedicated male audience. It is also possible that male social media users are accessing similar content but not on social media or do not consider this content to be fitspiration, and use different hashtags to label the fitness material they post (eg, #swoll which refers to swollen muscles).

Despite these concerns, it is possible that social media users viewing fitspiration are inspired to exercise, and that they view fitspiration in a positive manner. Qualitative research indicates that young female fitspiration users appreciate the ease of access to health and fitness information provided by fitspiration and generally choose to follow normalized and dominant health discourses [6]. Many posts were identified with balanced approaches to health and fitness. However, only half of the posts contained a subject actively exercising. A large number of posts focused on users' bodies, including flat stomachs and muscles, often in a posed and sexualized manner. The implication of these findings is that a subset of fitspiration is focused on appearance. Previous research has noted that exercising for appearance-based purposes is associated with increased body image concerns and disordered eating symptomatology [31]. This content also suggests that being fit and healthy is equivalent to fitting in with current masculine and feminine body ideals; in many posts, fitness and beauty were depicted as being essentially the same concept.

More experimental and longitudinal research is needed to identify the impacts of fitspiration on both body image and exercise behavior, particularly for male users. However, in the context of the small body of experimental research [7] which indicates short-term harms of fitspiration in women, and qualitative research which indicates that some young female social media users have internalized messages about idealized bodies depicted in fitspiration [6], there may be a need to develop interventions to prevent such harms in the long term. It has been suggested that focusing on the benefits of fitness without also emphasizing thinness may offer promising results regarding body image and physical activity for young women [32]. Such an approach may also be worthwhile for young men, emphasizing the benefits of exercise without emphasizing muscularity.

The results of the content analysis in this study suggest that interventions to reduce potential harms of fitspiration could focus on critically analyzing objectifying messages in fitspiration, and other content aiming to inspire people to exercise and be healthy. Interventions aimed at women could aim to reduce pressures to be both thin and muscular and deconstruct the relationship between fitness, sexuality, and beauty; interventions aimed at men could focus on reducing the pressure to be muscular (particularly in the upper body). Further research could aim to develop these interventions for both general communities (eg, social media users) and clinical populations (eg, young people experiencing eating disorders). Furthermore, fitspiration communities appear to be most active on Instagram, suggesting that any social media–based interventions should focus primarily on Instagram, with a secondary emphasis on Tumblr and other platforms. However, due to the diverse nature of tagging on different social media sites, it is possible that fitspiration communities are very active on other social media platforms. Furthermore, fitspiration websites [20] should not be ignored when developing interventions.

Of note, coders also identified several themes that were not included in the analysis and have not been included in previous content analyses. These include: presence of professional fitness models; suspected instances of image manipulation; emphasis on back muscles; depiction of protein powder or supplements; tie-ins with particular products, companies, diets, or trends (eg, "clean eating," "bikini bodies"); or depictions of particular subcommunities such as "fit mothers" and people tracking "fitness journeys." Future research should consider these areas and the potential impact that they might have on male and female body image.

Limitations

The authors acknowledge the limitations of this study. Although the study utilized 3 independent coders, the majority of variables included in the framework were subjective. Certain variables were nonspecific and only allowed simple yes or no coding. It is possible that our composite half hour did not accurately represent social media fitspiration; social media users may follow dedicated fitspiration blogs and pages rather than use hashtag-based searching, especially since searching for recent hashtags is difficult on some platforms. Fitspiration profiles may also be private and some posts would not have been extracted in our analysis. We only analyzed 1 hashtag ("#fitspo") due to time resources; this hashtag was chosen over #fitspiration (which was used by Tiggemann and Zaccardo [21]) as #fitspo returned approximately 4 times the number of results. Although many posts analyzed also contained the #fitspiration hashtag, our analyses can only be interpreted as relating to this 1 hashtag. Little detail could be provided about videos as not all were able to be viewed; some of these videos were coded as single still images. Furthermore, in order to analyze similar styles of posts across platforms, only "most recent" posts were viewed. Although representative of posts uploaded and tagged within a particular time frame, we are unable to provide information about the "most popular" form of fitspiration.

Conclusions

Overall, this study indicates that fitspiration on social media often encourages exercise in order to reach an appearance aligned with gendered body image ideals [8,13]. Our results also indicate that fitspiration imagery features men nearly as often as it features women and that men are just as likely to be objectified as women, albeit in different ways. These findings suggest a need to experimentally examine the impact of fitspiration-style posts on body image and exercise behaviors,

including fitspiration and other fitness media aimed at men. If such research identifies harms of fitspiration, further research is also required to determine the best strategies to minimize potential harms. Interventions should focus on Instagram and Tumblr and consider their highly visual nature and mobile formats, and gendered body image messages.

Acknowledgments

We gratefully thank Paige Kernebone and Michelle Motteram for their contribution to coding the content, and Alyce Vella and Cassandra Wright for assisting with the development of the codebook. The authors also acknowledge the contribution to this work of the Victorian Operational Infrastructure Support Program received by the Burnet Institute. ML is supported by the Jim and Margaret Beever Fellowship from the Burnet Institute.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Content Analysis Codebook.

[PDF File (Adobe PDF File), 44KB - jmir_v19i3e95_app1.pdf]

References

1. AIMIA: The Digital Industry Association of Australia. Sensis. 2015. Sensis Social Media Report May2015 URL: https://www.sensis.com.au/asset/PDFdirectory/Sensis_Social_Media_Report_2015.pdf [accessed 2017-02-19] [WebCite Cache ID 6oP1l6hpP]
2. Vaterlaus JM, Patten EV, Roche C, Young JA. #Gettinghealthy: The perceived influence of social media on young adult health behaviors. Comput Human Behav 2015 Apr;45:151-157. [doi: 10.1016/j.chb.2014.12.013]
3. Boepple L, Thompson JK. A content analytic comparison of fitspiration and thinspiration websites. Int J Eat Disord 2015 Mar 16:98-101. [doi: 10.1002/eat.22403] [Medline: 25778714]
4. Ghaznavi J, Taylor LD. Bones, body parts, and sex appeal: an analysis of #thinspiration images on popular social media. Body Image 2015 Jun;14:54-61. [doi: 10.1016/j.bodyim.2015.03.006] [Medline: 25880783]
5. Carrotte ER, Vella AM, Lim MS. Predictors of "liking" three types of health and fitness-related content on social media: a cross-sectional study. J Med Internet Res 2015 Aug 21;17(8):e205 [FREE Full text] [doi: 10.2196/jmir.4803] [Medline: 26297689]
6. Jong ST, Drummond MJ. Exploring online fitness culture and young females. Leisure Studies 2016 May 08:758-770. [doi: 10.1080/02614367.2016.1182202]
7. Tiggemann M, Zaccardo M. "Exercise to be fit, not skinny": the effect of fitspiration imagery on women's body image. Body Image 2015 Sep;15:61-67. [doi: 10.1016/j.bodyim.2015.06.003] [Medline: 26176993]
8. Markula P, Yiannakis A, Melnick M. Firm but shapely, fit but sexy, strong but thin: the postmodern aerobicizing female bodies. Contemporary Issues in Sociology of Sport 2001;12(4):424-453. [doi: 10.1123/ssj.12.4.424]
9. Knobloch-Westerwick S. Thinspiration: self-improvement versus self-evaluation social comparisons with thin-ideal media portrayals. Health Commun 2015;30(11):1089-1101. [doi: 10.1080/10410236.2014.921270] [Medline: 25317922]
10. Prichard I, Drummond C, Drummond M, Tiggemann M. Images of inspiration? The effects of viewing different types of #Fitspo images on mood, body image and exercise behaviour. 2015 Presented at: Youth Health Conference; 2015; Melbourne, Australia.
11. Benton C, Karazsia BT. The effect of thin and muscular images on women's body satisfaction. Body Image 2015 Mar;13:22-27. [doi: 10.1016/j.bodyim.2014.11.001] [Medline: 25528369]
12. Homan K. Athletic-ideal and thin-ideal internalization as prospective predictors of body dissatisfaction, dieting, and compulsive exercise. Body Image 2010 Jun;7(3):240-245. [doi: 10.1016/j.bodyim.2010.02.004] [Medline: 20226748]
13. Olivardia R, Pope Jr HG, Borowiecki III JJ, Cohane G. Biceps and body image: the relationship between muscularity and self-esteem, depression, and eating disorder symptoms. Psychol Men Masc 2004;5(2):112-120. [doi: 10.1037/1524-9220.5.2.112]
14. McCray J. The effects of exposure to images of the male muscular ideal on body image and muscularity concerns in men. Dissertation. Grand Forks: The University of North Dakota; 2005:1-262.
15. Hargreaves DA, Tiggemann M. Muscular ideal media images and men's body image: social comparison processing and individual vulnerability. Psychol Men Masc 2009;10(2):109-119. [doi: 10.1037/a0014691]
16. Mulgrew KE, Johnson LM, Lane BR, Katsikitis M. The effect of aesthetic versus process images on men's body satisfaction. Psychol Men Masc 2014;15(4):452-459. [doi: 10.1037/a0034684]

17. Fredrickson BL, Roberts T. Objectification theory: toward understanding women's lived experiences and mental health risks. Psychol of Women Q 1997 Jun;21(2):173-206. [doi: 10.1111/j.1471-6402.1997.tb00108.x]
18. Stankiewicz JM, Rosselli F. Women as sex objects and victims in print advertisements. Sex Roles 2008 Jan 15;58(7-8):579-589. [doi: 10.1007/s11199-007-9359-1]
19. Archer D, Iritani B, Kimes DD, Barrios M. Face-ism: Five studies of sex differences in facial prominence. J Pers Soc Psychol 1983;45(4):725-735. [doi: 10.1037/0022-3514.45.4.725]
20. Boepple L, Ata RN, Rum R, Thompson JK. Strong is the new skinny: a content analysis of fitspiration websites. Body Image 2016 Jun;17:132-135. [doi: 10.1016/j.bodyim.2016.03.001] [Medline: 27045871]
21. Tiggemann M, Zaccardo M. 'Strong is the new skinny': a content analysis of #fitspiration on Instagram. J Health Psychol 2016:1-9. [doi: 10.1016/j.bodyim.2016.03.001] [Medline: 27045871]
22. Womseo. Neilson 2012: State of the media? The social media report 2012 URL: https://womseo.com/wp-content/uploads/2012/12/The-Social-Media-Marketing-Report-2012.pdf [accessed 2017-02-19] [WebCite Cache ID 6oP33NLb4]
23. Murray SB, Griffiths S, Hazery L, Shen T, Wooldridge T, Mond JM. Go big or go home: a thematic content analysis of pro-muscularity websites. Body Image 2016 Mar;16:17-20. [doi: 10.1016/j.bodyim.2015.10.002] [Medline: 26523689]
24. Etcoff N, Orbach S, Scott J, D'Agostino H. Beperkthoudbaar. 2006 Beyond stereotypes: Rebuilding the foundation of beauty beliefs: Findings of the 2005 global study URL: http://www.beperkthoudbaar.info/upload/documents/dove/DoveBeyondStereotypesWhitePaper.pdf [accessed 2017-02-19] [WebCite Cache ID 6oP39ZdTW]
25. Stunkard AJ, Sørensen T, Schulsinger F. Use of the Danish adoption register for the study of obesity and thinness. Res Publ Assoc Res Nerv Ment Dis 1983;60:115-120. [Medline: 6823524]
26. Lynch S, Zellner D. Figure preferences in two generations of men: the use of figure drawings illustrating differences in muscle mass. Sex Roles 1999;40(9-10):833-843. [doi: 10.1023/A:1018868904239]
27. Grossbard JR, Neighbors C, Larimer ME. Perceived norms for thinness and muscularity among college students: what do men and women really want? Eat Behav 2011 Aug;12(3):192-199 [FREE Full text] [doi: 10.1016/j.eatbeh.2011.04.005] [Medline: 21741017]
28. McHugh ML. Interrater reliability: the kappa statistic. Biochem Med (Zagreb) 2012;22(3):276-282 [FREE Full text] [Medline: 23092060]
29. Smith LR, Cooley SC. International faces: an analysis of self-inflicted face-ism in online profile pictures. J Intercult Commun Res 2012 Nov;41(3):279-296. [doi: 10.1080/17475759.2012.728771]
30. Law C, Labre MP. Cultural standards of attractiveness: a thirty-year look at changes in male images in magazines. Journal Mass Commun Q 2002 Sep 01;79(3):697-711. [doi: 10.1177/107769900207900310]
31. Prichard I, Tiggemann M. Relations among exercise type, self-objectification, and body image in the fitness centre environment: the role of reasons for exercise. Psychol Sport Exerc 2008;9(6):855-866 [FREE Full text] [doi: 10.1016/j.psychsport.2007.10.005]
32. Homan K, McHugh E, Wells D, Watson C, King C. The effect of viewing ultra-fit images on college women's body dissatisfaction. Body Image 2012 Jan;9(1):50-56. [doi: 10.1016/j.bodyim.2011.07.006] [Medline: 21880562]

Edited by G Eysenbach; submitted 17.07.16; peer-reviewed by E Holland, M Moreno, C Fernández; comments to author 14.11.16; revised version received 11.12.16; accepted 15.02.17; published 29.03.17.

Please cite as:
Carrotte ER, Prichard I, Lim MSC
"Fitspiration" on Social Media: A Content Analysis of Gendered Images
J Med Internet Res 2017;19(3):e95
URL: http://www.jmir.org/2017/3/e95/
doi:10.2196/jmir.6368
PMID:28356239

Original Paper

The Efficacy of Internet-Based Mindfulness Training and Cognitive-Behavioral Training With Telephone Support in the Enhancement of Mental Health Among College Students and Young Working Adults: Randomized Controlled Trial

Winnie WS Mak[1], PhD; Floria HN Chio[1], MPhil; Amy TY Chan[1], PhD; Wacy WS Lui[2], MSSc; Ellery KY Wu[1], MSSc

[1]Department of Psychology, The Chinese University of Hong Kong, Hong Kong, China (Hong Kong)

[2]Center for Personal Growth and Crisis Intervention of the Corporate Clinical Psychology Services, Hospital Authority, Hong Kong, China (Hong Kong)

Corresponding Author:
Winnie WS Mak, PhD
Department of Psychology
The Chinese University of Hong Kong
Shatin, NT
Hong Kong,
China (Hong Kong)
Phone: 852 39436577
Fax: 852 26035019
Email: wwsmak@psy.cuhk.edu.hk

Abstract

Background: College students and working adults are particularly vulnerable to stress and other mental health problems, and mental health promotion and prevention are needed to promote their mental health. In recent decades, mindfulness-based training has demonstrated to be efficacious in treating physical and psychological conditions.

Objective: The aim of our study was to examine the efficacy of an Internet-based mindfulness training program (iMIND) in comparison with the well-established Internet-based cognitive-behavioral training program (iCBT) in promoting mental health among college students and young working adults.

Methods: This study was a 2-arm, unblinded, randomized controlled trial comparing iMIND with iCBT. Participants were recruited online and offline via mass emails, advertisements in newspapers and magazines, announcement and leaflets in primary care clinics, and social networking sites. Eligible participants were randomized into either the iMIND (n=604) or the iCBT (n=651) condition. Participants received 8 Web-based sessions with information and exercises related to mindfulness or cognitive-behavioral principles. Telephone or email support was provided by trained first tier supporters who were supervised by the study's research team. Primary outcomes included mental and physical health-related measures, which were self-assessed online at preprogram, postprogram, and 3-month follow-up.

Results: Among the 1255 study participants, 213 and 127 completed the post- and 3-month follow-up assessment, respectively. Missing data were treated using restricted maximum likelihood estimation. Both iMIND (n=604) and iCBT (n=651) were efficacious in improving mental health, psychological distress, life satisfaction, sleep disturbance, and energy level.

Conclusions: Both Internet-based mental health programs showed potential in improving the mental health from pre- to postassessment, and such improvement was sustained at the 3-month follow-up. The high attrition rate in this study suggests the need for refinement in future technology-based psychological programs. Mental health professionals need to team up with experts in information technology to increase personalization of Web-based interventions to enhance adherence.

Trial Registration: Chinese Clinical Trial Registry (ChiCTR): ChiCTR-TRC-12002623; https://www2.ccrb.cuhk.edu.hk/registry/public/191 (Archived by WebCite at http://www.webcitation.org/6kxt8DjM4).

(J Med Internet Res 2017;19(3):e84) doi:10.2196/jmir.6737

KEYWORDS

mental health promotion; Internet-based interventions; mindfulness-based training; cognitive-behavioral training; randomized controlled trial

Introduction

According to the World Health Organization (WHO) [1], mental health is an essential part of health that contributes to the overall well-being of every individual. However, approximately 450 million people suffer from mental health problems worldwide [2]. Working adults are particularly vulnerable in Hong Kong. For instance, a survey found that 25% of the 1031 employees interviewed reported feeling down, depressed, or hopeless in the previous month, and 90% of them reported needing more mental health support at work [3]. Another survey conducted among 1207 employees also showed that 82.5% and 27.6 % of the employees suffered from stress and depression because of work, respectively [4]. Work stress can be detrimental to mental health and it is associated with the onset of depression and anxiety among the young working adults [5]. The Hong Kong Mental Morbidity Survey (HKMMS), which was the first territory-wide epidemiological study in Hong Kong, recruited 5719 Chinese participants in the general population of Hong Kong and found that 13.3% of adults have common mental disorder, with individuals aged 26-35 years having significantly higher weighted prevalence (16.5%) than the general adult population [6]. In view of their vulnerability to mental health problems, prevention programs are urgently needed to promote their mental health. As the WHO has suggested, mental health is not the mere absence of disease or infirmity but it also includes positive functioning and state of mind [1]. It is therefore important to provide a prevention program that does not focus only on the reduction of mental health problems or psychological distress but also promotes their positive functioning and state of mind.

In addition to working adults, emerging adults such as college students also experience high levels of stress. They are in the midst of identity exploration [7] and the instability involved in their transitional stage of becoming full-fledged adults may expose them to higher risks of psychological disturbance. Based on a survey conducted in 2006 covering 10 tertiary education institutions in Hong Kong, 21%, 41%, and 27 % of the 7915 first-year students reported moderate severity or above on depression, anxiety, and stress symptoms, respectively [8]. Another study also found high prevalence of depressive symptoms among college students, with 43.9% of the students reporting a score of 16 or above on the Center for Epidemiologic Studies Depression Scale, which is suggestive of depressive symptoms [9]. Thus, prevention and mental health promotion are also needed among college students.

Besides mental health, according to the WHO, physical health is also one of the components that is intimately related to mental health [1]. In Hong Kong, it has been found that pain and sleep disturbance are prevalent, with 80.3% of the 1051 Hong Kong Chinese adults who were interviewed indicating some pain over the past year [10] and 68.6% of the 529 Hong Kong college students interviewed reporting symptoms of insomnia [11]. Furthermore, pain, fatigue, and sleep disturbance are highly interconnected and the presence of all 3 physical problems is associated with poorer physical and mental health [12]. Since mental and physical health are highly interdependent [1], a mental health promotion program can potentially have benefits on physical health as well.

Although effective treatments are available, it is noted that two-thirds of people who suffered from mental disorder did not seek help due to the stigma in seeking mental health services [13]. In the HKMMS, only 26% of those with common mental disorders have sought mental health services in the past year, with only 3.9% having sought help from a psychologist [6]. Internet-based interventions provide an alternative to face-to-face therapy in enhancing mental health for people who may not seek help due to stigma or other reasons. It is anonymous, self-paced, and easily accessible. The high scalability and penetration of Internet-based interventions also offers advantages over face-to-face treatment.

Increasing evidence has shown the efficacy of Internet-based interventions in the treatment of anxiety and depression, as well as the promotion of mental health in the general public. Meta-analysis found the efficacy of Internet-based cognitive behavioral intervention in the treatment of anxiety and depression [14,15]. In addition to cognitive behavioral therapy, mindfulness-based interventions provide another means in enhancing mental and physical health. Mindfulness is a nonjudgmental awareness of the present moment with curiosity and openness. [16,17]. Among its many training programs, mindfulness-based stress reduction (MBSR) and mindfulness-based cognitive therapy (MBCT) are the most widely applied and are found to be efficacious in the treatment of depression and anxiety, as well as improving the physical and mental health conditions in the clinical and nonclinical populations [18-23]. Meta-analyses have also showed that mindfulness-based training is efficacious in the reduction of stress and anxiety among working adults and college students in clinical and community settings [24,25].

Although the face-to-face mindfulness-based interventions are efficacious, few have tested the efficacy when delivered through the Web. Two previous feasibility and pilot studies showed preliminary evidence of Internet-based mindfulness programs in improving stress in nonclinical population [26,27] and another randomized controlled trial showed the efficacy of Internet-based mindfulness training in enhancing quality of life among people in the clinical population [28]. In Hong Kong, one study showed that an 8-week Internet-based mindfulness training was efficacious in enhancing mental well-being among college students at postprogram and 3-month follow-up compared with the waitlist control [29]. Taken together, the preliminary evidence supported the feasibility of Internet-based mindfulness training in promoting mental health.

Although much work has been done on mindfulness training and Internet-based cognitive behavioral training, few have tested these Internet-based interventions in Asia. Also, the efficacy of

Internet-based mindfulness training in promoting mental health is at its early stage. With the risks and prevalence of depression and anxiety observed among the college students and working adults, this study aimed to test the efficacy of an Internet-based mindfulness training for the prevention and promotion of their physical and mental health, compared with the well-established Internet-based cognitive behavioral training in a randomized controlled trial. We hypothesized that both training could enhance the physical and mental health at postprogram and 3-month follow-up.

Methods

Trial Design

This study was a 2-arm, randomized, open-label, parallel positive-control trial involving two Internet-based interventions: a mindfulness training program named iMIND versus a cognitive-behavioral training program named iCBT. Clinical ethics approval was obtained from the principal investigator's institution (Joint Chinese University of Hong Kong–New Territories East Cluster Clinical Research Ethics Committee) as well as from the Hospital Authority Kowloon Central or East Cluster and the Department of Health of Hong Kong.

Participants

The study targeted college students and young working adults and recruitment was done through (1) sending mass emails to students, teachers, and staff at different universities in Hong Kong; (2) distributing announcements to the staff of the Hospital Authority; (3) placing leaflets and posters in civil servant primary care clinics under Hong Kong Department of Health; and (4) posting advertisements in local libraries, newspapers, magazines, and social networking site Facebook.

Individuals who were interested in participating in the study visited our website where they were screened by completing Web-based questionnaires on mental health and demographics. Inclusion criteria included (1) age 18 years or above, (2) ability to read and understand Chinese, (3) computer literacy, and (4) consistent access to the Internet. Exclusion criteria included (1) an indication of suicidality by a score of 1-4 (out of 6) in item 16, 21, or 28 of the Mental Health Inventory (MHI) [30]; (2) currently receiving professional mental health services; and (3) currently taking psychotropic medication. If participants indicated suicidality in the screening questionnaire, they were to be given a list of resources and hotline on mental health services in the community.

Eligible individuals were given detailed information about the study aims, length of the program, participant involvement, and the assignment of intervention through randomization. They were also informed that the study was conducted by the Department of Psychology at the Chinese University of Hong Kong. Participants provided informed consent by clicking the "I agree" button at the bottom of the study description page. From there, participants received an activation link via email and were then randomly assigned to 1 of the 2 conditions by computer-generated numbers. The pre-, post-, and follow-up assessments were completed by the eligible participants on the Web, instead of through supporters, so that the assessment could be free from assessors' biases from knowing the participants' assigned conditions. Individuals who did not meet the eligibility criteria received an on-screen message and email with a thank you note and a list of resources on mental health services in the community.

Interventions

iMIND and iCBT were administered via the Internet on 2 separate Web pages that were in the Chinese language. Functional tests were conducted before the release of the website. Each program consisted of 8 30- to 45-minute sessions. Both programs lasted for 8 weeks. The delivery format of iMIND involved didactic readings (eg, nature of human suffering according to the Buddhist perspective), experiential learning (eg, guided meditation), and daily life applications (eg, developing awareness on how letting go of one's attachment could lead to inner peace). To enhance the user experiences, we made improvement on iMIND based on its predecessor [29] by making the content more interactive (eg, weekly well-being tracking, built-in multimedia within each lesson, dynamic content display) and more aesthetically appealing (eg, color coordinated and theme-consistent graphics with easy-to-use navigation). Recently, scholars have started to raise concerns about contemporary mindfulness teachings for their over-simplification and deviation from its traditional Buddhist root [31-33]. In response to this, our current iMIND program incorporated core notions in traditional Buddhism including discernment, compassion, impermanence, interdependence of all beings, and nonattachment [34]. By contextualizing our mindfulness training within the traditional Buddhist foundation, the training program aimed to facilitate participants to develop their own rationale behind practices. Such intentions would set the foundation for continuous and regular practices, and could potentially affect practice outcomes [35,36].

The content of iCBT was organized based on MacDonald and O'Hara's 10 elements of mental health [37], with mental health promotion resources from the WHO and government reports from the United Kingdom and Australia. At the end of each session, participants were provided with homework assignments to practice what was learned and apply the skills in their daily lives. In the iMIND program, videos of stretching and audios of body scan and sitting meditation were provided to the participants to guide them through their exercises. In the iCBT program, worksheets including mood diary, cognitive restructuring, and healthy lifestyle plan were provided for participants to record their responses. All contents were developed by the research team members who were clinical psychologists and mindfulness practitioners. The iMIND and iCBT content was turned into the Web page by eLearningPro Limited. A brief overview of the session content is shown in Table 1. No further revision on the content and the Web page was made after the trial was launched. Screenshots of how the interventions appeared in the Web page are shown in Multimedia Appendix 1.

Table 1. Overview of session content.

Session	Content (iMIND[a])	Content (iCBT[b])
1	Introduction on mindfulness	Introduction on mental health
2	Observing thoughts, feelings, and sensations as they are	Stress, body reactions, and emotion regulations
3	Mindful attitudes and nature of suffering	Cognitive distortions and strategies to cope with stress
4	Being in the present moment	Emotion regulation
5	Letting go in times of difficulties	Resilience in times of adversities
6	Ways to stay mindful	Ways to increase self-esteem
7	Mindful communications	Effective communication skills
8	Review and applications	Review and applications

[a]iMIND: Internet-based mindfulness training program.

[b]iCBT: Internet-based cognitive-behavioral training program.

Previous research has shown that (1) guided self-help has higher completion rates than unguided self-help [38], (2) programs with weekly telephone reminders are more efficacious than those without [39], and (3) technician-assisted telephone or email support for Internet-based interventions is as efficacious as clinician-assisted telephone or email support [40,41]. Given these findings, for the duration of the 8-session program in both conditions, trained first tier supporters contacted each participant weekly via telephone and email to (1) acknowledge their time spent on the program, (2) ensure their understanding of course-related instructions, (3) encourage them to continue participating, and (4) provide guidelines for homework activities. Scripted guidelines and training were provided to the supporters. Participants were instructed to call and/or email our research assistant for clarification in case of questions or problems during the course of the intervention. When a participant fell below a score of 13 or answered 0 or 1 on any of the items on the Well-Being Index (WBI) [42], first tier supporters would refer them to second tier supporters (who were clinical psychologists) to evaluate their mental health status, address their questions, and make referrals for more intensive treatments as needed. Each participant was monitored through weekly self-report measure (ie, the WBI) as well as their first and second tier supporters. The first tier supporters also contacted the participants in both conditions once a month after the end of the program to maintain contact and interest in completing postprogram evaluations. The CONSORT e-health checklist is shown in Multimedia Appendix 2.

Measures

Baseline Measures

At baseline, participants provided demographic and background information including age, gender, education level, income, marital status, religion, and previous experience with systematic mindfulness training (ie, mindfulness-based stress reduction therapy, MBSR, or mindfulness-based cognitive therapy, MBCT), regular meditation practices, cognitive-behavioral training, and yoga. To assess the route of participation, participants also indicated how they learned about this study.

Mental Health Measures

Mental Well-Being

The WHO 5-item WBI [43] was used to measure overall mental well-being. Each item was rated on a 6-point Likert scale from 0 (never) to 5 (all of the time). The scale has been used among the Chinese with an internal consistency of .90 [36]. In this study, the Cronbach alphas of the WBI were .92, .93, and .94 at baseline, postprogram, and 3-month follow-up, respectively.

Psychological Distress

The 18-item MHI was used to assess psychological distress [30]. Each item was rated on a 6-point Likert scale from 1 (all of the time) to 6 (none of the time). Previous research showed that the MHI's internal consistency (Cronbach alphas) ranged from .81 to .91, with stability coefficients ranging from .60 to .76 over a 1-year interval [44]; its validity has also been supported among the Chinese. In this study, its Cronbach alphas were .93, .94, and .95, at baseline, postprogram, and 3-month follow-up, respectively.

Life Satisfaction

Life satisfaction was assessed by the 5-item Satisfaction with Life Scale (SWLS) [45]. Participants rated the extent to which they endorsed each item on a 6-point Likert scale from 1 (strongly disagree) to 6 (strongly agree). Its reliability has been substantiated (eg, test-retest reliability of .84 over a 1-month interval) and its convergent validity has been demonstrated by its high correlations with other life satisfaction measures [46]. The scale has been used extensively among the Chinese and its validity has also been supported among Hong Kong university students [47]. In this study, the Cronbach alphas of the SWLS were .91, .91, and .90 at baseline, postprogram, and 3-month follow-up, respectively.

Physical Health Measures

Energy

Average level of energy was measured by the visual analogue scale (VAS) [48]. Participants rated their average daily energy level on a 100mm long line from 0 (no energy) to 100 (a lot of energy). Its validity is supported by its usage in measuring energy or fatigue level among patients with diagnoses of fatigue-related medical conditions [49].

Sleep Disturbance

The 4-item sleep disturbance subscale of the Medical Outcomes Study (MOS) Sleep Scale [50] was used to assess how well participants slept without tapping into other sleep-related medical conditions. Three items related to sleep disturbance were rated on a 6-point Likert scale from 1 (all of the time) to 6 (none of the time) and 1 item related to the time needed to fall asleep was assessed on a 5-point Likert scale from 1 (less than 15 minutes) to 5 (more than 60 minutes). Scores were converted to an index that ranged from 0-100 with higher scores indicating a higher level of sleep disturbance. Research has demonstrated its acceptable level of internal consistency reliability (>.70) and its responsiveness to change [50]. It has also been validated among community adults and has been used among the Chinese population [51]. The Cronbach alphas of MOS sleep scale were .83, .70, and .70 at baseline, postprogram, and 3-month follow-up, respectively.

Pain

Average level of pain was measured by VAS [52]. Participants rated their average daily pain level on a 100mm long line from 0 (no pain) to 100 (very severe pain). It has been used in a variety of settings and is sensitive to treatment effects [53]. It has been shown to be reliable in pain assessment when compared with other subjective pain measuring methods [54]. It also showed good reliability and validity among Chinese adults [55].

Usage and Satisfaction Measures

Usage is defined as the time (in minutes) spent in the previous week on browsing the website and practicing the assigned homework. Participants reported these figures at the beginning of every session. At the end of the 8-week program, attitude toward and satisfaction with the Internet-based interventions were assessed using the Chinese version of the 8-item Client Satisfaction Questionnaire (CSQ) [56]. Each item was rated on a 4-point Likert scale from 1 to 4 and response options differed for different items. The Cronbach alpha of CSQ was .91 in this study.

Credibility and Expectancy

At baseline, participants completed the 6-item Credibility or Expectancy Questionnaire (CEQ) that aimed to examine if expectancies or perception of treatment credibility were related to outcomes. Five items were rated on a 9-point Likert scale from 1 (not at all) to 9 (very much) and 1 item was rated on an 11-point Likert scale ranging from 0 (0%) to 11 (100%). The CEQ comprises 2 factors: cognitively based credibility and affectively based expectancy. It was shown to have a total item correlation of .78 [57] and the scale has been used among Chinese patients [58]. Standardized scores were computed for the 2 subscales. In this study, the Cronbach alphas for credibility and expectancy were .82 and .84, respectively.

Analysis

All analyses were conducted using SPSS version 20.0 (IBM Corp) . Linear mixed models were conducted to test if both conditions showed improvements in all outcomes over time. Compound symmetry covariance was used and missing data were treated using restricted maximum likelihood estimation. Model for each outcome variable consisted of the time effect, condition effect, and the interaction effect of time by condition. When the main effect of time was significant, follow-up analyses were conducted to compare the outcomes in postprogram and follow-up program with the preprogram, and results were adjusted with Bonferroni correction. *T* tests were also conducted to test for equivalence in treatment expectancy and satisfaction about course content across the 2 conditions.

Results

Recruitment and Participant Characteristics

Participants were recruited between July 2013 and March 2015. A total of 4215 registrants were screened for eligibility. Among those who registered, 932 (22.11%, 932/4215) registrants were deemed ineligible, 1202 (28.52%, 1202/4215) eligible registrants did not activate their accounts, whereas 2081 (49.37%, 2081/4215) eligible registrants proceeded with account activation followed by randomization. Our sample consisted of those who, after randomization, completed the presurvey and received course materials (N=1255). About one-fifth of the participants (n=253) completed the entire 8-session program, 16.97% (213/1255) completed the postprogram survey, and 10.12% (127/1255) completed the 3-month follow-up (see Figure 1 for the flow diagram). No adverse events were reported during the course of the study.

Participants learned about the study from a variety of avenues: work institutions or universities (36.65%, 460/1255), Facebook (28.21%, 354/1255), family or relatives or friends (21.27%, 267/1255), other means such as posters and leaflets (10.92%, 137/1255), and primary care clinics (2.95%, 37/1255). Tables 2 and 3 display the baseline characteristics of the participants in both conditions. Overall, participants had a mean age of 32.62 years (SD 12.54), were predominantly female (74.34%, 933/1255), with half of them being college graduates (52.51%, 659/1255). About one-thirds (34.6%, 434/1255) were college students and about half (51.07%, 641/1255) were working full-time (see Table 1). Both conditions reported similar treatment expectancy and credibility (t<0.57, P>.30). Findings showed that both conditions expressed similar CSQ usage satisfaction, t_{211} =−0.07, P=.94. In terms of utilization, iMIND condition (Mean 189.89, SD 501.00) spent more time browsing the course content than their iCBT counterparts (Mean 135.98, SD 347.08), $t_{1063.9}$ =−2.20, P=.03). However, iCBT condition (Mean 240.63 minutes, SD 578.52) spent more time on homework assignment than iMIND condition (Mean 118.42, SD 401.42), $t_{1162.3}$=4.37, P<.001.

To investigate the potential causes of attrition, we compared the baseline attributes between the attrition group (did not complete postprogram assessment; n=1042) and the retention group (n=213). No significant differences in their demographic characteristics were found, except for yoga experience. A slightly higher percentage of participants reported having had yoga experience in the retention group (28.6%) than in the attrition group (21.8%), χ^2_1 =4.5, P<.05. In terms of their psychological and stress profile, the attrition group was lower in mental well-being (WBI: Mean 2.04, SD 1.05; MHI: Mean

3.89, SD 0.84; SWLS: Mean 3.86, SD 1.40, t>2.32, P<.05), energy (Mean 53.11, SD 26.15; t_{1253}=−2.01; P=.05), and treatment expectancy (credibility: Mean −0.05, SD 0.86; expectancy: Mean −0.03, SD 0.88) at preprogram than those who completed the postprogram assessment (WBI: Mean 2.25, SD 1.06; MHI: Mean 4.03, SD 0.80; SWLS: Mean 4.19, SD 1.35; energy: Mean 57.04, SD 25.65; credibility: Mean 0.23, SD 0.81; expectancy: Mean=0.16, SD=0.83).

Table 2. Baseline characteristics across conditions.

Characteristics		iCBT[a] (n=651)	iMIND[b] (n=604)
Age in years, mean (SD)		32.52 (12.41)	32.73 (12.68)
Gender, n (%)			
	Male	173 (26.6)	149 (24.7)
	Female	478 (73.4)	455 (75.3)
Education, n (%)			
	Primary or below	2 (0.3)	1 (0.2)
	Secondary	125 (19.2)	118 (19.5)
	Bachelor or diploma	346 (53.1)	313 (51.8)
	Master or above	178 (27.3)	172 (28.5)
Employment, n (%)			
	Student	226 (34.7)	208 (34.4)
	Full-time	331 (50.8)	310 (51.3)
	Part-time or freelance	29 (4.5)	28 (4.7)
	Others	65 (10)	58 (9.6)
Religion, n (%)			
	No religion	392 (60.2)	382 (63.1)
	Christianity	178 (27.3)	153(25.3)
	Catholicism	28 (4.3)	26 (4.3)
	Buddhism	41 (6.3)	34 (5.6)
	Others	12(1.9)	10 (1.7)
Systematic mindfulness training, n (%)			
	Yes	36 (5.5)	40 (6.6)
	No	615 (94.5)	564 (93.4)
Regular meditation practices, n (%)			
	Yes	60 (9.2)	57 (9.4)
	No	591 (90.8)	547 (90.6)
Yoga experience, n (%)			
	Yes	149 (22.9)	139 (23.0)
	No	502 (77.1)	465 (77.0)
Cognitive behavioral, n (%) therapy experience, n (%)			
	Yes	16 (2.5)	18 (3.0)
	No	635 (97.5)	586 (97.0)

[a]iCBT: Internet-based cognitive behavioral training program.

[b]iMIND: Internet-based mindfulness training program.

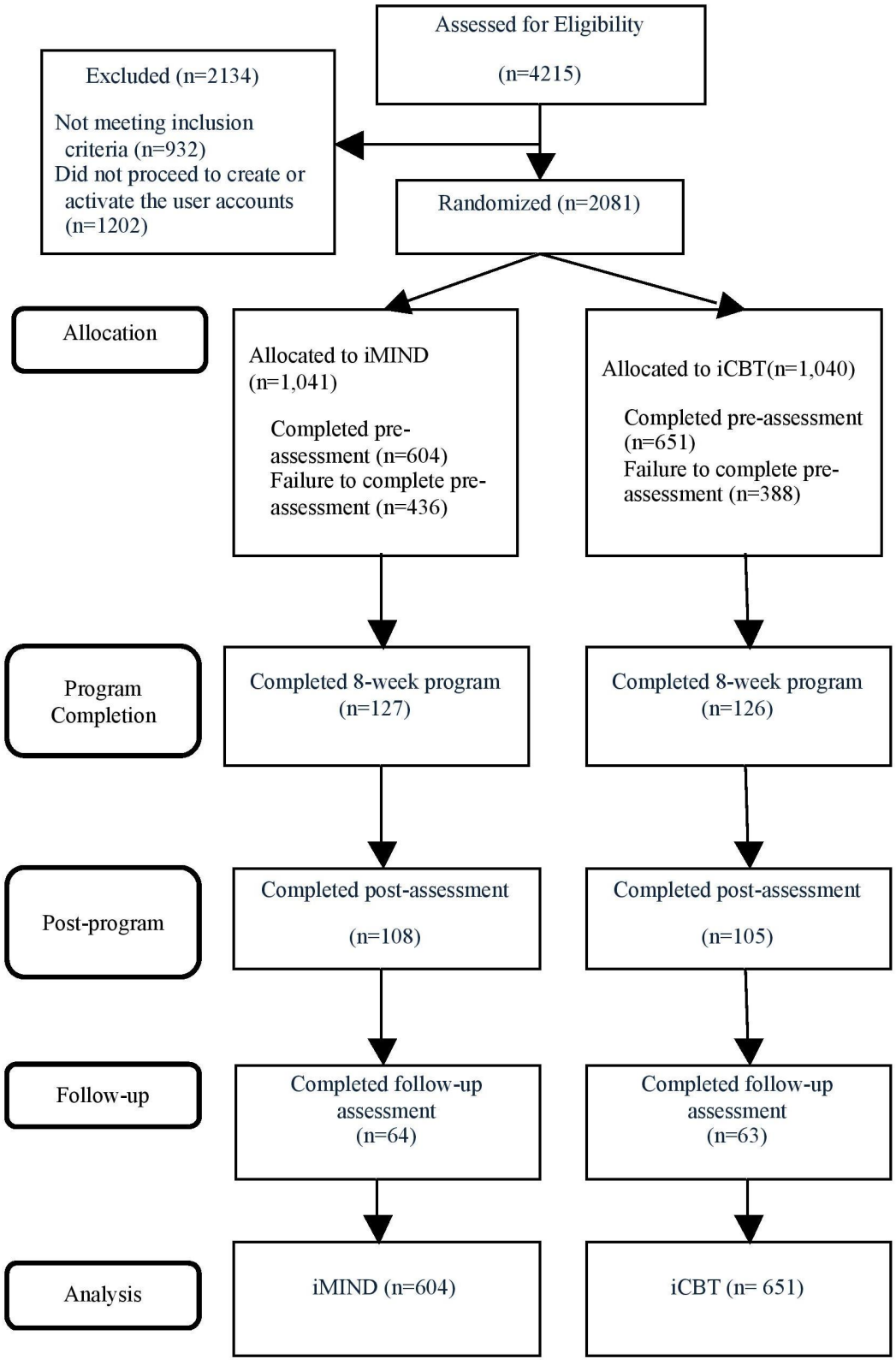
Table 3. Baseline characteristics across conditions.

Measures		iCBT[a] (n=651), Mean (SD)	iMIND[b] (n=604), Mean (SD)
Well-being index		2.02 (1.05)	2.14 (1.06)
Mental health inventory		3.90 (0.83)	3.93 (0.83)
Life satisfaction scale		3.90 (1.36)	3.94 (1.43)
Sleep disturbance		26.92 (20.24)	26.03 (20.69)
Pain		26.14 (25.75)	26.31 (25.62)
Energy		52.01 (26.41)	55.67 (25.64)
Credibility or expectancy questionnaire			
	Credibility	−0.02 (.86)	0.03 (0.86)
	Expectancy	0.01 (0.86)	−0.01 (0.88)

[a]iCBT: Internet-based cognitive behavioral training program.

[b]iMIND: Internet-based mindfulness training program.

Figure 1. Flow diagram of this study.



Mental Health Measures

Well-Being Index

Results from the linear mixed model indicated a significant time effect (P<.001). Mental well-being significantly increased from baseline to postprogram (mean difference=−0.83, 95% CI −0.996 to −0.66, P<.001), and this increase was maintained at 3-month follow-up (mean difference=−0.73, 95% CI −0.94 to −0.51, P<.001) in both iMIND and iCBT. WBI was not significantly different between iMIND and iCBT (mean

difference=−0.03, 95% CI −0.20 to 0.14, *P*=.75), and the 2 conditions did not differ in their improvements over time (*P*=.56)

Mental Health Inventory

The results indicated that there was a significant time effect (*P*<.001). It significantly increased from baseline to postprogram in both iMIND and iCBT (mean difference=−0.46, 95% CI −0.59 to −0.33, *P*<.001) and was maintained at 3-month follow-up (mean difference=−0.25, 95% CI −0.41 to −0.09, *P*=.001). MHI was not significantly different between iMIND and iCBT (mean difference=0.08, 95% CI −0.05 to 0.22, *P*=.22). The interaction effect of time x condition (*P*=.18) was also not significant, indicating that iMIND and iCBT showed similar improvement over time.

Life Satisfaction

The results indicated a significant time effect (*P*<.001), with life satisfaction significantly increased from baseline to postprogram in both iMIND and iCBT (mean difference=−0.77, 95% CI −0.96 to −0.59, *P*<.001), and this increase was maintained at 3-month follow-up (mean difference=−0.85, 95% CI −1.08 to −0.62, *P*<.001). The improvement was not significantly different between iMIND and iCBT (mean difference=−0.05, 95% CI −0.26 to 0.16, *P*=.62), and the time x condition interaction (*P*=.88), was not significant.

Physical Health Measures

Energy

Results showed that energy improved over time (*P*<.001). The improvement was significant at postprogram, (mean difference=−12.60, 95% CI −16.54 to −8.67), *P*<.001), and was maintained at 3-month follow-up in both iMIND and iCBT, (mean difference=−13.42, 95% CI −18.37 to 8.47, *P*<.001). The effect did not differ between iMIND and iCBT (mean difference=−2.01, 95% CI −6.07 to 2.05, *P*=.33), and no significant interaction effect of time x condition (*P*=.67) was found.

Sleep Disturbance

Improvement was shown over time (*P*<.001). The improvement was significant at postprogram (mean difference=8.12, 95% CI 5.66-10.58, *P*<.001), and 3-month follow-up in both iMIND and iCBT (mean difference=7.46, 95% CI 4.39-10.53, *P*<.001). The improvement was not significantly different between iMIND and iCBT (mean difference=0.16, 95% CI −2.79 to 3.12, *P*<.001), and the interaction effect of time x condition (*P*=.91), was not significant.

Pain

Results showed that pain significantly improved over time (*P*=.01). The improvement was shown in postprogram (mean difference=3.95, 95% CI 0.27-7.63, *P*=.03). No significant effect was found for the condition effect (mean difference=−0.31, 95% CI −4.29 to 3.67, *P*=.88), and the time by condition effect (*P*=.58) were not significant. Tables 4 and 5 show a summary of the means, standard errors, effect sizes, and time effects of the outcome measures across conditions.

Table 4. Means and standard errors across conditions.

Measures[a]		iCBT[b] (n=651) Mean (SE[d])			iMIND[c] (n=604) Mean (SE)		
		Pre	Post	Follow-up	Pre	Post	Follow-up
Mental health measures							
	WBI[e]	2.02 (0.04)	2.90 (0.10)	2.84 (0.12)	2.14 (0.04)	2.92 (0.10)	2.78 (0.12)
	MHI[f]	3.90 (0.03)	4.43 (0.07)	4.25 (0.09)	3.93 (0.03)	4.31 (0.08)	4.08 (0.09)
	SWLS[g]	3.90 (0.05)	4.69 (0.11)	4.71 (0.14)	3.94 (0.06)	4.69 (0.11)	4.82 (0.14)
Physical health measures							
	Energy	52.01 (0.99)	66.00 (2.26)	66.52 (2.87)	55.67 (1.03)	66.89 (2.29)	68.00 (2.89)
	Sleep disturbance	26.92 (0.77)	19.13 (1.51)	18.03 (1.85)	26.03 (0.81)	17.58 (1.53)	19.99 (1.88)
	Pain	26.14 (0.99)	21.04 (2.15)	23.12 (2.71)	26.31 (1.03)	23.50 (2.19)	21.43 (2.73)

[a]Significant time effects were shown for all measures (*P* s<.05). All post and follow-up scores were significantly improved compared with the prescores, except that pain did not show any improvement at follow-up compared with prescore. Interaction effect of time x condition were all nonsignificant, indicating that the improvements over time were similar across conditions.

[b]iCBT: Internet-based cognitive behavioral training program.

[c]iMIND: Internet-based mindfulness training program.

[d]SE: standard error.

[e]WBI: well-being index.

[f]MHI:Mental Health Inventory.

[g]SWLS: Satisfaction with Life Scale.

Table 5. Overall time effects and effect sizes across conditions.

Measures	Scales	iCBT[a] (n=651) Cohen's d[c]		iMIND[b] (n=604) Cohen's d		Overall time effect			
		Post versus pre	Follow-up versus pre	Post versus pre	Follow-up versus pre	Post versus pre mean difference (95% CI)	*P* value	Follow-up versus pre mean difference (95% CI)	*P* value
Mental health measures									
	WBI[d]	0.86	0.81	0.79	0.65	−0.83 (−0.996 to −0.66)	<.001	−0.73 (−0.94 to 0.51)	<.001
	MHI[e]	0.70	0.46	0.51	0.20	−0.46 (−0.59 to −0.33)	<.001	−0.25 (−0.41 to −0.09)	.001
	SWLS[f]	0.55	0.64	0.52	0.61	−0.77 (−0.96 to −0.59)	<.001	−0.85 (−1.08 to −0.62)	<.001
Physical health measures									
	Energy	0.56	0.58	0.45	0.49	−12.60 (16.54-8.67)	<.001	−13.42 (−18.37 to −8.47)	<.001
	Sleep disturbance	0.41	0.46	0.44	0.31	8.12 (5.66-10.58)	<.001	7.46 (4.39-10.53)	<.001
	Pain	0.21	0.12	0.11	0.20	3.95 (0.27-7.63)	.03	3.95 (−0.67 to 8.57)	.12

[a]iCBT: Internet-based cognitive behavioral training program.

[b]iMIND: Internet-based mindfulness training program.

[c]Cohen's d was computed from postprogram or 3-month follow-up score minus preprogram score divided by the pooled standard deviation.[d]WBI: Well-Being Index.[e]MHI: Mental Health Inventory.[f]SWLS: Satisfaction with Life Scale.

Discussion

Principal Findings

This study developed and evaluated the efficacy of the Internet-based mindfulness training in comparison with an Internet-based cognitive-behavioral training on college students and young working adults in Hong Kong. Results showed that the Internet-based mindfulness training was as efficacious as the widely supported Internet cognitive-behavioral training in improving mental well-being, psychological distress, life satisfaction, energy level, sleep disturbance, and pain at the end of the 8-week program. Furthermore, users' perceived credibility, expectancy, and satisfaction of both programs were similar. The results are encouraging as both Internet-based programs received support for their utility, satisfaction, and efficacy in mental health promotion. Given the weight of mental illness disease burden in our communities, this study shows that Internet-based mindfulness and cognitive-behavioral training programs with minimal guided support can be a highly scalable and convenient way for prevention and promotion of mental and physical health among college students and young working adults

In Hong Kong, the majority of individuals who seek help for mental health issues do not receive psychiatric and clinical psychological services in primary and secondary care settings until their problems have become severe. In comparison with face-to-face interventions, Internet-based interventions are more easily accessible and affordable and have the potential to fulfill the need for mental health promotion and prevention in community settings. This study provided empirical support for the efficacy of Internet-based cognitive-behavioral and mindfulness training programs, which can be easily incorporated into existing service provision portfolios that promote mental health and reduce psychological distress among the college students and young working adult population in Hong Kong.

In terms of service management, these developed Internet-based interventions are highly sustainable. In Hong Kong, the number of mobile phone customers reached over 8 million in June 2016, and the amount of mobile data usage has been 10-folded from 2006 to 2016, demonstrating the rapid increase of mobile phone and mobile Internet usage [59]. With the high penetration of Internet-based programs and the increasing prevalence of mobile phone and tablet device utilization, Internet-based programs meet the public mental health goal of reaching the general public for mental health promotion and prevention under the stepped care model, especially if the current programs can be converted into mobile phone apps in the future. Although attrition may be high, given it is an easily accessible public health tool, if its dissemination in the population is wide, such Internet-based mental health promotional tools can still be an important augmentation to face-to-face interventions in filling the role of mental health promotion and illness prevention that is lacking in the current mental health services system in Hong Kong [60]. One size does not fit all; we need myriad approaches of varying dosage and penetration rates to reach a wider population in order to prevent the tremendous mental illness burden [61,62]. Thus, we believe Internet-based programs still have their merits by reaching out to many more individuals at a much shorter period of time. Even if only a tenth of the thousands improved over the course of the programs, it is still worthwhile to make this accessible to individuals who may not have access or do not prefer to have face-to-face interventions.

Future research should explore methods for enhancing adherence of Internet-based health solutions in order to harness the expanding proliferation of technology among the public. For instance, recent studies have begun to incorporate ecological momentary intervention components into Internet-based programs so that interventions can be directed to real-time events and be more personalized [63,64]. Emerging evidence has suggested efficacy of ecological momentary interventions in the enhancement of a variety of health behaviors [65]. To leverage the power of technology in the promotion of mental health, interventions that incorporate these methods may reduce attrition rate and further maximize the efficacy of the Internet-based programs.

Although improvements in outcomes were observed at postprogram, the improvements for pain were not maintained at 3-month follow-up. This could be the result of reduced practice or application of skills learned on the websites. In addition, this might also due to the low level of pain observed within this group of population. The floor effect might have limited the possibility in detecting improvement in pain at postprogram and 3-month follow-up.

The two Internet-based interventions in this study yielded similar results. Future studies can explore how individual differences may affect intervention benefits. It may be possible that cognitive styles can play a role in the receptivity of iMIND and iCBT and matching their styles with the treatment approach may maximize the outcome.

Limitations

This study has several limitations. First, our target population was college students and young working adults. By nature, our sample is skewed toward those who were educated or were employed. As our programs were Internet-based, it is possible that they appealed to a selective group in the population who were more comfortable in accessing interventions over the Internet with their personal computers. They might have higher mental health literacy and be more willing to participate in Internet-based mental health programs. These biases in our sample limit the generalizability of our findings to all segments of the population (eg, less educated individuals, older adults). It is possible that the delivery of mental health materials over the Internet may only be appropriate for specific segments of the populations, rather than the entire population. Future studies should focus on how Internet-based interventions can cater to different segments of the populations through various adaptations.

Second, the attrition rate of our study is high. High attrition rate has been a perennial problem for Internet-based interventions. Similarly high attrition rates have been reported in other Internet-based mental health programs. For example, Christensen and colleagues [66] reported an attrition rate of 74% in their Internet-based cognitive behavioral therapy (CBT)

program for depression. Another study reported an attrition rate of 98.8% in an Internet-based CBT for panic disorder [67]. A study based on the MoodGym had an attrition rate of 73.9% for trial participants and 99.2% for public registrants [68]. In a more recent systematic review of Internet-based interventions for anxiety and depression, the completion of protocol rates for depression sites ranged from 43% to 99% [69]. Another systematic review on Internet-based interventions for psychological disorders also found that there was an additional 0% to 18% of participants who would further dropout from post to follow-up assessment [70]. Moreover, during the inception of our project, mobile phones and tablets were not as omnipresent as they are today. Because of that, our Web-based modules were designed using Adobe Flash and thus only catered for desktop viewing, which may deter usage.

Third, we did not include a waitlist control group in this study. As this study aimed to compare Internet-based mindfulness training with a well-established Internet-based cognitive behavioral training, and previous study has found Internet-based mindfulness training to have significant improvements in mental health than waitlist control [29], we decided not to have a waitlist control in this study so as not to withhold intervention from our participants. Fourth, we did not ask whether our participants received any other psychological intervention during the study period. Thus, the findings may potentially be attributed to additional intervention that the participants have received. Finally, this study found that those who quit the programs scored lower on mental health measures, energy level, mindful awareness, and treatment expectancy at the outset. To ensure interventions are catered to those most in need, future studies should explore the reasons behind attrition and identify corresponding remedies. As suggested by existing research, utilization can potentially be promoted via built-in incentives, personalized feedback, and user collaboration [71].

Conclusions

In sum, this study showed that both Internet-based mindfulness training and Internet-based cognitive-behavioral training were efficacious in improving mental and physical health indicators among college students and young working adults in a convenient fashion. To leverage the power of technology in reducing mental illness burden, it is paramount for mental health professionals to work in tandem with professionals in other disciplines (eg, designers, computer scientists) in creating user-friendly programs that enable seamless integration into users' daily lives.

Acknowledgments

We would like to acknowledge the Health and Health Services Research Fund (Ref. No. 09100711) for funding this project.

Conflicts of Interest

The study was supported by the Health and Health Services Research Fund (Ref. No. 09100711). The first author of the study, Winnie Mak, is one of the developers of the content of the trials but does not own the source code of the website.

Multimedia Appendix 1

Screenshots of the iMIND and iCBT.

[PDF File (Adobe PDF File), 1MB - jmir_v19i3e84_app1.pdf]

Multimedia Appendix 2

CONSORT EHEALTH checklist.

[PDF File (Adobe PDF File), 349KB - jmir_v19i3e84_app2.pdf]

References

1. World Health Organization. Promoting mental health: concepts, emerging evidence, practice: summary report. Geneva: World Health Organization; 2004.
2. World Health Organisation. The World Health Report 2001: Mental Health : New Understanding, New Hope (World Health Report). Geneva: World Health Organization; 2001.
3. Tse SSK, Wong PWC. SOCSC. 2014. Press Conference Presentation: Knowledge, attitude, practices about mental health of the workforce URL: http://www.socsc.hku.hk/events/press_information/press/pdf/2014/20140917/Working%20Population_%20Press%20Conference%20English_Final.pdf [accessed 2016-06-03] [WebCite Cache ID 6jzHfPeiz]
4. Welford R. Csr360gpn. Work life balance in Hong Kong: Survey results URL: http://www.csr360gpn.org/uploads/files/resources/WLB_2008_Final.pdf [WebCite Cache ID 6jzIBcHht]
5. Melchior M, Caspi A, Milne BJ, Danese A, Poulton R, Moffitt TE. Work stress precipitates depression and anxiety in young, working women and men. Psychol Med 2007 Aug;37(8):1119-1129 [FREE Full text] [doi: 10.1017/S0033291707000414] [Medline: 17407618]

6. Lam LC, Wong CS, Wang M, Chan W, Chen EY, Ng RM, et al. Prevalence, psychosocial correlates and service utilization of depressive and anxiety disorders in Hong Kong: the Hong Kong Mental Morbidity Survey (HKMMS). Soc Psychiatry Psychiatr Epidemiol 2015 Sep;50(9):1379-1388. [doi: 10.1007/s00127-015-1014-5] [Medline: 25660760]
7. Arnett JJ. Emerging adulthood: a theory of development from the late teens through the twenties. Am Psychol 2000 May;55(5):469-480. [Medline: 10842426]
8. Wong JG, Cheung EP, Chan KK, Ma KK, Tang SW. Web-based survey of depression, anxiety and stress in first-year tertiary education students in Hong Kong. Aust N Z J Psychiatry 2006 Sep;40(9):777-782. [doi: 10.1111/j.1440-1614.2006.01883.x] [Medline: 16911753]
9. Song Y, Huang Y, Liu D, Kwan JS, Zhang F, Sham PC, et al. Depression in college: depressive symptoms and personality factors in Beijing and Hong Kong college freshmen. Compr Psychiatry 2008;49(5):496-502. [doi: 10.1016/j.comppsych.2008.02.005] [Medline: 18702936]
10. Ng KF, Tsui SL, Chan WS. Prevalence of common chronic pain in Hong Kong adults. Clin J Pain 2002;18(5):275-281. [Medline: 12218497]
11. Sing CY, Wong WS. Prevalence of insomnia and its psychosocial correlates among college students in Hong Kong. J Am Coll Health 2010;59(3):174-182. [doi: 10.1080/07448481.2010.497829] [Medline: 21186447]
12. Wong WS, Fielding R. The co-morbidity of chronic pain, insomnia, and fatigue in the general adult population of Hong Kong: prevalence and associated factors. J Psychosom Res 2012 Jul;73(1):28-34. [doi: 10.1016/j.jpsychores.2012.04.011] [Medline: 22691556]
13. World Health Organization. WHO. Mental disorders affect one in four people: Treatment available but not being used URL: http://www.who.int/whr/2001/media_centre/press_release/en/ [WebCite Cache ID 6jzJW8HRb]
14. Reger MA, Gahm GA. A meta-analysis of the effects of internet- and computer-based cognitive-behavioral treatments for anxiety. J Clin Psychol 2009 Jan;65(1):53-75. [doi: 10.1002/jclp.20536] [Medline: 19051274]
15. Spek V, Cuijpers P, Nyklícek I, Riper H, Keyzer J, Pop V. Internet-based cognitive behaviour therapy for symptoms of depression and anxiety: a meta-analysis. Psychol Med 2007 Mar;37(3):319-328. [doi: 10.1017/S0033291706008944] [Medline: 17112400]
16. Kabat-Zinn J. Full catastrophe living: Using the wisdom of your mind and body to face stress, pain, and illness. New York: Delacorte; 1990.
17. Bishop SR, Lau M, Shapiro S, Carlson L, Anderson ND, Carmody J, et al. Mindfulness: a proposed operational definition. Clin Psycholci and Pract 2004;11:230-241. [doi: 10.1093/clipsy.bph077]
18. Grossman P, Niemann L, Schmidt S, Walach H. Mindfulness-based stress reduction and health benefits: a meta-analysis. J Psychosom Res 2004 Jul;57(1):35-43. [doi: 10.1016/S0022-3999(03)00573-7] [Medline: 15256293]
19. Baer RA. Mindfulness training as a clinical intervention: a conceptual empirical review. Clin Psychol: Sci and Pract 2003;10:125-143. [doi: 10.1093/clipsy/bpg015]
20. Miller JJ, Fletcher K, Kabat-Zinn J. Three-year follow-up and clinical implications of a mindfulness meditation-based stress reduction intervention in the treatment of anxiety disorders. Gen Hosp Psychiatry 1995 May;17(3):192-200. [Medline: 7649463]
21. Speca M, Carlson LE, Goodey E, Angen M. A randomized, wait-list controlled clinical trial: the effect of a mindfulness meditation-based stress reduction program on mood and symptoms of stress in cancer outpatients. Psychosom Med 2000;62(5):613-622. [Medline: 11020090]
22. Smith BW, Shelley BM, Dalen J, Wiggins K, Tooley E, Bernard J. A pilot study comparing the effects of mindfulness-based and cognitive-behavioral stress reduction. J Altern Complement Med 2008 Apr;14(3):251-258. [doi: 10.1089/act.2008.14505] [Medline: 18370583]
23. Shapiro SL, Schwartz GE, Bonner G. Effects of mindfulness-based stress reduction on medical and premedical students. J Behav Med 1998 Dec;21(6):581-599. [Medline: 9891256]
24. Virgili M. Mindfulness-based interventions reduce psychological distress in working adults: a meta-analysis of intervention studies. Mindfulness 2013 Dec 13;6(2):326-337. [doi: 10.1007/s12671-013-0264-0]
25. Regehr C, Glancy D, Pitts A. Interventions to reduce stress in university students: a review and meta-analysis. J Affect Disord 2013 May 15;148(1):1-11. [doi: 10.1016/j.jad.2012.11.026] [Medline: 23246209]
26. Glück TM, Maercker A. A randomized controlled pilot study of a brief web-based mindfulness training. BMC Psychiatry 2011;11:175 [FREE Full text] [doi: 10.1186/1471-244X-11-175] [Medline: 22067058]
27. Krusche A, Cyhlarova E, King S, Williams JM. Mindfulness online: a preliminary evaluation of the feasibility of a web-based mindfulness course and the impact on stress. BMJ Open 2012;2(3):e000803 [FREE Full text] [doi: 10.1136/bmjopen-2011-000803] [Medline: 22614170]
28. Boettcher J, Aström V, Påhlsson D, Schenström O, Andersson G, Carlbring P. Internet-based mindfulness treatment for anxiety disorders: a randomized controlled trial. Behav Ther 2014 Mar;45(2):241-253 [FREE Full text] [doi: 10.1016/j.beth.2013.11.003] [Medline: 24491199]
29. Mak WW, Chan AT, Cheung EY, Lin CL, Ngai KC. Enhancing Web-based mindfulness training for mental health promotion with the health action process approach: randomized controlled trial. J Med Internet Res 2015;17(1):e8 [FREE Full text] [doi: 10.2196/jmir.3746] [Medline: 25599904]

30. Veit CT, Ware JE. The structure of psychological distress and well-being in general populations. J Consult Clin Psychol 1983 Oct;51(5):730-742. [Medline: 6630688]
31. Rosch E. More than mindfulness: When you have a tiger by the tail, let it eat you. Psychol Inq 2007 Oct 19;18(4):258-264. [doi: 10.1080/10478400701598371]
32. Stanley S. 'Things said or done long ago are recalled and remembered': the ethics of mindfulness in early Buddhism, psychotherapy and clinical psychology. Eur J Psychother Couns 2013 Jun;15(2):151-162. [doi: 10.1080/13642537.2013.795338]
33. Monteiro LM, Musten R, Compson J. Traditional and contemporary mindfulness: finding the middle path in the tangle of concerns. Mindfulness 2014 Apr 29;6(1):1-13. [doi: 10.1007/s12671-014-0301-7]
34. Sumedho B. The four noble truths. Hertfordshire: Amaravati Publications; 1992.
35. Shapiro SL, Carlson LE, Astin JA, Freedman B. Mechanisms of mindfulness. J Clin Psychol 2006 Mar;62(3):373-386. [doi: 10.1002/jclp.20237] [Medline: 16385481]
36. Shapiro DH. A preliminary study of long-term meditators: goals, effects, religious orientation, cognitions. Journal of Transpersonal Psychology 1992;24:23.
37. MacDonald G, O'Hara K. Ten elements of mental health, its promotion and demotion: Implications for practice. London: Society of Health Education and Health Promotion Specialists; 1998.
38. Titov N, Andrews G, Choi I, Schwencke G, Mahoney A. Shyness 3: randomized controlled trial of guided versus unguided Internet-based CBT for social phobia. Aust N Z J Psychiatry 2008 Dec;42(12):1030-1040. [doi: 10.1080/00048670802512107] [Medline: 19016091]
39. Titov N, Andrews G, Choi I, Schwencke G, Johnston L. Randomized controlled trial of web-based treatment of social phobia without clinician guidance. Aust NZ J Psychiatry 2009 Jan;43(10):913-919. [doi: 10.1080/00048670903179160]
40. Titov N, Andrews G, Davies M, McIntyre K, Robinson E, Solley K. Internet treatment for depression: a randomized controlled trial comparing clinician vs. technician assistance. PLoS One 2010;5(6):e10939 [FREE Full text] [doi: 10.1371/journal.pone.0010939] [Medline: 20544030]
41. Robinson E, Titov N, Andrews G, McIntyre K, Schwencke G, Solley K. Internet treatment for generalized anxiety disorder: a randomized controlled trial comparing clinician vs. technician assistance. PLoS One 2010;5(6):e10942 [FREE Full text] [doi: 10.1371/journal.pone.0010942] [Medline: 20532167]
42. Psychiatric Research Unit. Psykiatri-regionh. Interpretation of the Items of the WHO-Five Well-being Index (WHO-5) Questionnaire URL: https://www.psykiatri-regionh.dk/who-5/Documents/WHO5_Chinese_PR.pdf [WebCite Cache ID 6jzL3H1oR]
43. Psychiatric Research Unit. Psykiatri-regionh. 1998. WHO (Five) Well-Being Index (1998 version) URL: https://www.psykiatri-regionh.dk/who-5/who-5-questionnaires/Pages/default.aspx [WebCite Cache ID 6jzLVtH2d]
44. Ware JEJ, Johnston SA, Davies AR, Brook RH. Conceptualization and measurement of health for adults in the health insurance study (volIII): Mental health. Santa Monica. Santa Monica, CA: Rand Corporation; 1979.
45. Diener E, Emmons RA, Larsen RJ, Griffin S. The satisfaction with life scale. J Pers Assess 1985 Feb;49(1):71-75. [doi: 10.1207/s15327752jpa4901_13] [Medline: 16367493]
46. Pavot W, Diener E, Colvin CR, Sandvik E. Further validation of the Satisfaction with Life Scale: evidence for the cross-method convergence of well-being measures. J Pers Assess 1991 Aug;57(1):149-161. [doi: 10.1207/s15327752jpa5701_17] [Medline: 1920028]
47. Sachs J. Validation of the Satisfaction with Life Scale in a sample of Hong Kong University students. Psychologia 2012;41(6):E41-E50. [doi: 10.2117/psysoc.2003.225]
48. Folstein MF, Luria R. Reliability, validity, and clinical application of the visual analogue mood scale. Psychol Med 1973 Nov;3(4):479-486. [Medline: 4762224]
49. O'Connor PJ. Mental energy: assessing the mood dimension. Nutr Rev 2006 Jul;64(7 Pt 2):S7-S9. [Medline: 16910215]
50. Hays RD, Martin SA, Sesti AM, Spritzer KL. Psychometric properties of the Medical Outcomes Study Sleep measure. Sleep Med 2005 Jan;6(1):41-44. [doi: 10.1016/j.sleep.2004.07.006] [Medline: 15680294]
51. Kwan P, Yu E, Leung H, Leon T, Mychaskiw MA. Association of subjective anxiety, depression, and sleep disturbance with quality-of-life ratings in adults with epilepsy. Epilepsia 2009 May;50(5):1059-1066 [FREE Full text] [doi: 10.1111/j.1528-1167.2008.01938.x] [Medline: 19170734]
52. Wallerstein SL. Scaling clinical pain and pain relief. In: Bromm B, editor. Pain Measurement in Man: Neurophysiological Correlates of Pain. New York: Elsevier; 1984:389-396.
53. Joyce CR, Zutshi DW, Hrubes V, Mason RM. Comparison of fixed interval and visual analogue scales for rating chronic pain. Eur J Clin Pharmacol 1975 Aug 14;8(6):415-420. [Medline: 1233242]
54. Lundeberg T, Lund I, Dahlin L, Borg E, Gustafsson C, Sandin L, et al. Reliability and responsiveness of three different pain assessments. J Rehabil Med 2001 Nov;33(6):279-283. [Medline: 11766958]
55. Li L, Liu X, Herr K. Postoperative pain intensity assessment: a comparison of four scales in Chinese adults. Pain Med 2007 Apr;8(3):223-234 [FREE Full text] [doi: 10.1111/j.1526-4637.2007.00296.x] [Medline: 17371409]
56. Larsen DL, Attkisson CC, Hargreaves WA, Nguyen TD. Assessment of client/patient satisfaction: development of a general scale. Eval Program Plann 1979;2(3):197-207. [Medline: 10245370]

57. Devilly GJ, Borkovec TD. Psychometric properties of the credibility/expectancy questionnaire. J Behav Ther Exp Psychiatry 2000 Jun;31(2):73-86. [Medline: 11132119]
58. Kim J, Lee MS, Jung S, Choi J, Lee S, Ko J, et al. Acupuncture for persistent allergic rhinitis: a multi-centre, randomised, controlled trial protocol. Trials 2009;10:54 [FREE Full text] [doi: 10.1186/1745-6215-10-54] [Medline: 19602250]
59. OFCA. 2016. Key statistics for telecommunications in Hong Kong URL: http://www.ofca.gov.hk/filemanager/ofca/en/content_108/wireless_en.pdf [WebCite Cache ID 6nXBC3tfw]
60. Kazdin AE, Blase SL. Rebooting psychotherapy research and practice to reduce the burden of mental illness. Perspect Psychol Sci 2011 Jan;6(1):21-37. [doi: 10.1177/1745691610393527] [Medline: 26162113]
61. Kazdin AE, Rabbitt SM. Novel models for delivering mental health services and reducing the burdens of mental illness. Clin Psychol Sci 2013 Jan 23;1(2):170-191. [doi: 10.1177/2167702612463566]
62. Onken LS, Carroll KM, Shoham V, Cuthbert BN, Riddle M. Reenvisioning clinical science unifying the discipline to improve the public health. Clin Psychol Sci 2014 Jan 1;2(1):22-34 [FREE Full text] [doi: 10.1177/2167702613497932] [Medline: 25821658]
63. Kelly J, Gooding P, Pratt D, Ainsworth J, Welford M, Tarrier N. Intelligent real-time therapy: harnessing the power of machine learning to optimise the delivery of momentary cognitive-behavioural interventions. J Ment Health 2012 Aug;21(4):404-414. [doi: 10.3109/09638237.2011.638001] [Medline: 22251028]
64. Burns MN, Begale M, Duffecy J, Gergle D, Karr CJ, Giangrande E, et al. Harnessing context sensing to develop a mobile intervention for depression. J Med Internet Res 2011;13(3):e55 [FREE Full text] [doi: 10.2196/jmir.1838] [Medline: 21840837]
65. Heron KE, Smyth JM. Ecological momentary interventions: incorporating mobile technology into psychosocial and health behaviour treatments. Br J Health Psychol 2010 Feb;15(Pt 1):1-39 [FREE Full text] [doi: 10.1348/135910709X466063] [Medline: 19646331]
66. Christensen H, Griffiths KM, Mackinnon AJ, Brittliffe K. Online randomized controlled trial of brief and full cognitive behaviour therapy for depression. Psychol Med 2006 Dec;36(12):1737-1746. [doi: 10.1017/S0033291706008695] [Medline: 16938144]
67. Farvolden P, Denisoff E, Selby P, Bagby RM, Rudy L. Usage and longitudinal effectiveness of a Web-based self-help cognitive behavioral therapy program for panic disorder. J Med Internet Res 2005;7(1):e7 [FREE Full text] [doi: 10.2196/jmir.7.1.e7] [Medline: 15829479]
68. Christensen H, Griffiths KM, Korten AE, Brittliffe K, Groves C. A comparison of changes in anxiety and depression symptoms of spontaneous users and trial participants of a cognitive behavior therapy website. J Med Internet Res 2004 Dec 22;6(4):e46 [FREE Full text] [doi: 10.2196/jmir.6.4.e46] [Medline: 15631970]
69. Christensen H, Griffiths KM, Farrer L. Adherence in internet interventions for anxiety and depression. J Med Internet Res 2009;11(2):e13 [FREE Full text] [doi: 10.2196/jmir.1194] [Medline: 19403466]
70. Melville KM, Casey LM, Kavanagh DJ. Dropout from Internet-based treatment for psychological disorders. Br J Clin Psychol 2010 Nov;49(Pt 4):455-471. [doi: 10.1348/014466509X472138] [Medline: 19799804]
71. Mohr DC, Burns MN, Schueller SM, Clarke G, Klinkman M. Behavioral intervention technologies: evidence review and recommendations for future research in mental health. Gen Hosp Psychiatry 2013;35(4):332-338 [FREE Full text] [doi: 10.1016/j.genhosppsych.2013.03.008] [Medline: 23664503]

Abbreviations

CBT: cognitive behavioral therapy
CEQ: Credibility or Expectancy Questionnaire
CSQ: Client Satisfaction Questionnaire
HKMMS: Hong Kong Mental Morbidity Survey
MBSR: mindfulness-based stress reduction
MBCT: mindfulness-based cognitive therapy
MHI: Mental Health Inventory
SWLS: Satisfaction with Life Scale
VAS: visual analogue scale
WBI: Well-Being Index
WHO: World Health Organization

Edited by G Eysenbach; submitted 04.10.16; peer-reviewed by M Harper Shehadeh, A Feinstein; comments to author 03.11.16; revised version received 16.01.17; accepted 04.02.17; published 22.03.17.

Please cite as:
Mak WWS, Chio FHN, Chan ATY, Lui WWS, Wu EKY
The Efficacy of Internet-Based Mindfulness Training and Cognitive-Behavioral Training With Telephone Support in the Enhancement of Mental Health Among College Students and Young Working Adults: Randomized Controlled Trial
J Med Internet Res 2017;19(3):e84
URL: http://www.jmir.org/2017/3/e84/
doi:10.2196/jmir.6737
PMID:28330831

Original Paper

Assessing Feasibility and Acceptability of Web-Based Enhanced Relapse Prevention for Bipolar Disorder (ERPonline): A Randomized Controlled Trial

Fiona Lobban[1], PhD; Alyson L Dodd[1,2], PhD; Adam P Sawczuk[1,3], Pg Cert; Ozgur Asar[4,5], PhD; Dave Dagnan[6], PhD; Peter J Diggle[5], PhD; Martin Griffiths[1], BSc Hons; Mahsa Honary[1], PhD; Dawn Knowles[1], MSc; Rita Long[1]; Richard Morriss[7], MD; Rob Parker[1], MSc; Steven Jones[1], PhD

[1]Spectrum Centre, Faculty of Health and Medicine, Lancaster University, Lancaster, United Kingdom

[2]Department of Psychology, Northumbria University, Newcastle Upon Tyne, United Kingdom

[3]Institute of Psychology, Health and Society, University of Liverpool, Liverpool, United Kingdom

[4]Department of Bio-statistics and Medical Informatics, Acibadem University, Istanbul, Turkey

[5]CHICAS, Faculty of Health and Medicine, Lancaster University, Lancaster, United Kingdom

[6]Cumbria Partnership NHS Foundation Trust, Penrith, Cumbria, United Kingdom

[7]Department of Psychiatry and Applied Psychology, Institute of Mental Health, University of Nottingham, Nottingham, United Kingdom

Corresponding Author:
Fiona Lobban, PhD
Spectrum Centre
Faculty of Health and Medicine
Lancaster University
Bailrigg Campus
Lancaster, LA14YT
United Kingdom
Phone: 44 01524593752
Fax: 44 0000000000
Email: f.lobban@lancaster.ac.uk

Abstract

Background: Interventions that teach people with bipolar disorder (BD) to recognize and respond to early warning signs (EWS) of relapse are recommended but implementation in clinical practice is poor.

Objectives: The objective of this study was to test the feasibility and acceptability of a randomized controlled trial (RCT) to evaluate a Web-based enhanced relapse prevention intervention (ERPonline) and to report preliminary evidence of effectiveness.

Methods: A single-blind, parallel, primarily online RCT (n=96) over 48 weeks comparing ERPonline plus usual treatment with "waitlist (WL) control" plus usual treatment for people with BD recruited through National Health Services (NHSs), voluntary organizations, and media. Randomization was independent, minimized on number of previous episodes (<8, 8-20, 21+). Primary outcomes were recruitment and retention rates, levels of intervention use, adverse events, and participant feedback. Process and clinical outcomes were assessed by telephone and Web and compared using linear models with intention-to-treat analysis.

Results: A total of 280 people registered interest online, from which 96 met inclusion criteria, consented, and were randomized (49 to WL, 47 to ERPonline) over 17 months, with 80% retention in telephone and online follow-up at all time points, except at week 48 (76%). Acceptability was high for both ERPonline and trial methods. ERPonline cost approximately £19,340 to create, and £2176 per year to host and maintain the site. Qualitative data highlighted the importance of the relationship that the users have with Web-based interventions. Differences between the group means suggested that access to ERPonline was associated with: a more positive model of BD at 24 weeks (10.70, 95% CI 0.90 to 20.5) and 48 weeks (13.1, 95% CI 2.44 to 23.93); increased monitoring of EWS of depression at 48 weeks (−1.39, 95% CI −2.61 to −0.163) and of hypomania at 24 weeks (−1.72, 95% CI −2.98 to −0.47) and 48 weeks (−1.61, 95% CI −2.92 to −0.30), compared with WL. There was no evidence of impact of ERPonline on clinical outcomes or medication adherence, but relapse rates across both arms were low (15%) and the sample remained high functioning throughout. One person died by suicide before randomization and 5 people in ERPonline and 6 in WL reported ideas of suicide or self-harm. None were deemed study related by an independent Trial Steering Committee (TSC).

Conclusions: ERPonline offers a cheap accessible option for people seeking ongoing support following successful treatment. However, given high functioning and low relapse rates in this study, testing clinical effectiveness for this population would require very large sample sizes. Building in human support to use ERPonline should be considered.

Trial registration: International Standard Randomized Controlled Trial Number (ISRCTN): 56908625; http://www.isrctn.com/ISRCTN56908625 (Archived by WebCite at http://www.webcitation.org/6of1ON2S0)

(J Med Internet Res 2017;19(3):e85) doi:10.2196/jmir.7008

KEYWORDS

Internet; randomized controlled trial; feasibility studies; bipolar disorder

Introduction

Bipolar Disorder

Bipolar disorder (BD) is a lifelong mental health condition characterized by extreme fluctuating mood including recurrent episodes of depression and mania, which generally starts in adolescence and affects approximately 1-1.5% of adults worldwide [1]. The impact of BD on employment and relationships can be devastating, and the condition has high financial costs, estimated at £5.2 billion annually in England alone [2]. Preventing relapse is a key goal of most interventions for BD. Interventions that teach people to recognize and respond to early warning signs (EWS) are recommended by clinical guidelines worldwide [3-5] but implementation in routine clinical practice is poor [6]. Enhanced Relapse Prevention (ERP), a structured manualized intervention for frontline care staff, has shown significant benefit and is well received by patients and staff [7]. However, delivered face-to-face it will only ever be available to a small percentage of people with BD due to low rates of psychological intervention provision even among those who remain in secondary care services. In this study, we test the feasibility and acceptability of a Web-based version of ERP: ERPonline. Web-based interventions in mental health offer the potential to broaden access, reduce waiting times, delivery costs and stigma, and improve quality through standardized delivery [8,9]. There is growing evidence for short-term benefits of Internet-delivered psychological treatments for depression and anxiety disorders compared with waitlist (WL) controls [10], although understanding their implementation into real-world services is still in its infancy [11]. In BD, the evidence, while promising, is at an earlier stage, comprising small-scale feasibility studies [12-18]. These studies, along with results from an international multisite survey [19], suggest that people with BD can use, and are interested in further using, Web-based mental health support. However, detailed evidence is lacking on what kinds of psychosocial support can be Web-based, the best ways to deliver these, who accesses Web-based interventions, what processes and outcomes are impacted on, and how to best design rigorous trials to evaluate them on the Web. This information is essential to inform definitive clinical and cost-effectiveness trials. This study addresses these issues in a novel randomized controlled trial (RCT) to assess feasibility and acceptability of ERPonline with all recruitment and assessments of outcome performed remotely.

Objectives

The objectives of this study were as follows:

First, to assess the feasibility of (1) creating a Web-based version of enhanced relapse prevention for BD (ERPonline) and (2) an RCT design using Web-based and telephone data collection to evaluate effectiveness.

Second, to determine the acceptability of ERPonline for people with BD via (1) ERPonline website usage, (2) number and type of adverse events associated with site use, and (3) detailed feedback from participants about their experiences of ERPonline to inform future developments.

Third, to determine the feasibility and acceptability of data collection via the Internet and telephone measured by recruitment and retention rates, data completion, and direct feedback from participants.

Fourth, to test the impact of the intervention on hypothesized mechanisms of change to understand processes underlying any impact.

Finally, to estimate the likely effect size of the intervention on a range of outcomes, particularly noting any negative impacts.

Methods

Design

A single-blind RCT with nested qualitative study comparing ERPonline plus usual treatment with a “waitlist control” arm with delayed access to ERPonline plus usual treatment. Primary outcomes were feasibility and acceptability. Process and clinical outcomes were assessed to identify measures sensitive to change collected remotely and to explore potential positive and negative impacts of the intervention. Remote data collection and online recruitment increased the external validity of the trial by encouraging participation from those unable or unwilling to engage in face-to-face clinical trials, who are also more likely to be those people unable or unwilling to engage in face-to-face clinical support for whatever reason. The study was not powered to test statistically significant impact. The trial was preregistered and full protocol published [16]. Ethics approval was given by UK National Research Ethics Service (NRES) Committee North West (Ref 12/NW/0594).

Participants

We aimed to recruit 125 participants, anticipating a dropout rate of up to 35% (based on retention rates from previous trials of Web-based interventions for BD [12-18]) providing 40 people per arm, sufficient to meet the aims of our study to assess feasibility and acceptability. Participants were aged ≥18 years, residents of the United Kingdom, with a confirmed diagnosis

of BD (1 or 2), at risk of relapse (≥3 previous episodes, ≥1 in the preceding 2 years), and with access to the Internet. We excluded people in current episode (within previous 4 weeks), currently under Mental Health Act section and therefore likely to be in current episode or at high risk of harm to self or others, or unable to understand English sufficiently to engage with the study.

Recruitment Strategy

The study was presented to clinical teams in 8 NHS Mental Health Trusts in England, and staff were reminded in monthly team meetings to direct service users to the Web-based registration site. An advert was placed in a UK charity newsletter (Bipolar UK), and on a charity website (Bipolar Scotland). A link to ERPonline was put in NHS Choices, and British Broadcasting Corporation (BBC) health online presented a short article that linked to the website. The research team regularly tweeted about the study, and our service user lead was interviewed on local radio about the study.

People were invited to visit the site which explained the study, allowed them to check eligibility, and to register an interest in participating. Participants read online participant information sheet and completed an online consent form. Consent and capacity were reassessed at each assessment point.

Intervention

ERPonline was developed with extensive input from a reference group of 8 adults with BD to adapt the original ERP manual to a Web-based format. Input (online and face-to-face) occurred throughout the study, but was more extensive during the initial development of the ERPonline site and included feedback on content of draft modules, user testing of the ERPonline website, and providing video and case material of lived experience which are integral parts of the intervention site. The aim of the intervention is to help people develop a coherent working model of their mood changes, recognize and manage triggers and EWS, and develop coping strategies to manage these effectively. Key modules are summarized in Table 1 with more detail in the protocol paper [20].

Table 1. Key intervention modules in Web-based enhanced relapse prevention intervention (ERPonline).

Section	Module title	Module description	ERPonline (n=47) average number of module views per person Mean (SD[a]) Median[b] (min-max)	ERPonline (n=47) average time spent per module per person (min) Mean (SD[a]) Median[b] (min-max)
Getting started	How to use the site	Ways to navigate the site to get the best from the available modules	7.43 (5.77) 7 (0-25)	8.61 (8.70) 7.5 (0-46.5)
	Introduction	Explains what ERPonline is, rationale for this approach, why it might be useful, and how to involve a relative or friend if desired	4.28 (5.46) 3 (0-30)	5.88 (8.99) 2 (0-40)
	What is bipolar?	Background information about what bipolar disorder is, theories about causes, common consequences, and an overview of available treatments	8.00 (7.34) 7 (0-30)	11.95 (13.51) 7 (0-52.5)
Key Modules	Mood charting	How to use a Web-based tool to monitor mood on a daily basis to help recognize normal mood fluctuation and pick up early signs of a mood episode	138.38(445.54) 13 (0-2519)	122.89 (386.43) 14.5 (0-2150.5)
	Life charting	Complete a chart of past mood episodes, identifying potential triggers and coping strategies for future mood changes	49.09 (147.55) 11 (0-990)	47.34 (114.45) 8.00 (0-744)
	Identifying triggers	Detailed analysis of triggers of previous mood episodes, followed by a personalized plan of how to manage triggers	11.34 (25.57) 1 (0-148)	15.07 (30.09) 0.5 (0-140.5)
Specific moods	Early warning signs (EWS) -high mood	Detailed analysis of EWS of high mood to develop a relapse signature for (hypo) mania	14.47 (26.72) 0 (0-90)	17.36 (32.31) 0 (0-114)
	Coping strategies -high mood	Review of current strategies to manage high mood and introduction to new strategies that may be helpful	5.68 (10.62) 0 (0-44)	10.0 (19.78) 0 (0-71.5)
	Early warning signs (EWS) -low mood	Detailed analysis of EWS of low mood to develop a relapse signature for depression	9.26 (22.47) 0 (0-94)	9.38 (21.87) 0 (0-84.5)
	Coping strategies -low mood	Review of current strategies to manage low mood and introduction to new strategies that may be helpful	3.83 (11.26) 0 (0-59)	3.95 (11.23) 0 (0-52.5)

Section	Module title	Module description	ERPonline (n=47) average number of module views per person Mean (SD[a]) Median[b] (min-max)	ERPonline (n=47) average time spent per module per person (min) Mean (SD[a]) Median[b] (min-max)
Wrapping things up	Staying well strategies	Identifying and managing stress levels Understanding the importance of social rhythms and how to regulate these to manage mood How relationships with other people impact on mood	4.06 (6.11) 0 (0-25)	4.39 (8.30) 0 (0-37)
	Your staying well plan	An individualized summary of staying well strategies, early warning signs to look out for, and coping strategies to regulate mood	3.09 (5.94) 0 (0-30)	2.35 (4.60) 0 (0-18.5)

[a]SD: standard deviation.

[b]The median value 0 indicates that at least half the sample did not visit this module.

Each module included information, suggested strategies, and case examples. Users interacted with the site to input personal information relevant to their own triggers, EWS, and coping strategies. These informed an individualized staying well plan. The site also provided signposting to additional formal and informal support. Participants were free to choose the order they visited modules (although they were listed in logical order), and were invited to involve a supporter of their choosing. Each module included recommendations of how the supporter could be involved in relapse prevention. All participants continued to receive any other treatment as usual throughout the study. The home page is shown in Figures 1 and 2 for illustration.

Figure 1. The online enhanced relapse prevention intervention (ERPonline) home page - top of page.

Figure 2. The online enhanced relapse prevention intervention (ERPonline) home page - bottom of the page.

Procedure: Randomization and Masking

Diagnostic eligibility was confirmed using Structured Clinical Interview for Diagnostic and Statistical Manual of Mental Disorders-IV (DSM-IV) [21] administered by telephone by trained research assistants. Training consisted of scoring training videos, conducting clinical sensitive interviews on the telephone, and recording an interview with someone with BD who provided experiential feedback and which was then rated by a supervising clinical academic. Training continued until ratings were reliable and clinical style was of high quality. Monthly supervision to ensure reliability in scoring of telephone interviews continued throughout the study. Following baseline assessments (telephone and Web-based) participants were randomly allocated by an independent clinical trials unit (CTU) using 1:1 ratio, minimized on number of previous episodes (<8, 8-20, 21+) and including a random element to minimize predictability of allocation. Those in the ERPonline arm received an email containing a weblink and instructions of how to log-on to the site using a unique username and password. Control participants received an email or telephone call informing them of the allocation and emphasizing the importance of continued participation throughout the trial. All communication with CTU and participants regarding randomization was conducted by the trial manager (unblinded). All communication with participants reiterated the importance of not telling the researcher carrying out the follow-up interviews which group they were in, and why this was important. All assessments were conducted by blind researchers. Blindness was further maintained using restricted file access to any data showing randomization, and prefacing each follow-up interview with a reminder about why it was important not to say anything about which arm they were in.

During the trial, participants were sent an additional email inviting them to provide qualitative feedback. The ERPonline group was asked to complete a Web-based survey about their views of the ERPonline site and any improvements they would recommend. They were also given the option (between 2 and 12 months following randomization) to take part in a telephone interview about their experiences of using ERPonline. Given the relatively novel primarily Web-based trial design, the WL control group was sent a survey before accessing the site about their reasons for engaging in the trial, and their experience of taking part in the trial.

A reflective log detailed our experiences throughout the trial.

Measures

Proposed mechanisms of change, were assessed at baseline, 24 and 48 weeks including frequency of EWS monitoring (EWS checklist for relapse in depression and mania [22] Likert scale 1=never to 4=very regularly), adapted Brief Illness Perception Questionnaire (BIPQ score 0-110: higher score=more negative beliefs) [23], and the Medication Adherence Rating Scale (MARS 0-10: higher score=higher compliance) [24]. These

measures were either designed for use with people with BD (EWS checklists and BIPQ), or have been successfully used with this population (MARS). They are all self-report, have high face validity, and have been shown to be valid and reliable measures, making them highly applicable to Web-based use.

Interviewer-rated outcome measures were administered by telephone by two trained research assistants, at baseline, 12-, 24-, 36-, and 48-week follow-up. These included SCID-LIFE [25] providing a retrospective weekly rating of depression (1-6) and mania (1-6) (scores of 5 or 6 indicate major mood episode); Hamilton Depression Rating Scale (HAM-D scores above 7 indicate mild depression, above 13 moderate, and above 18 severe) [26], and Mania Rating Scale (MRS scores of 11 and above indicate hypomania) [27]; the Personal and Social Performance Scale (PSP scores 70-80 indicate mild difficulties, above 80 is good functioning) [28]; and the Multidimensional Scale of Independent Functioning (MSIF Likert scale 1=normal functioning, to 7=total disability) [29]. Self-report outcome measures were collected at baseline, 24, and 48 weeks and included the Work and Social Adjustment Scale (WSAS less than 10=subclinical; 10-20 some functional impairment; above 20 moderate psychopathology) [30]; Quality of Life in Bipolar Disorder (QoLBD range 48-240 with high score=higher quality of life) [31]; and the Bipolar Recovery Questionnaire (BRQ score 0-3600, high score=higher recovery) [32]. Web-based versions of the EQ5D5L [33] and the Client Service Receipt Inventory (CSRI) [34] were piloted to assess the feasibility of collecting this data on the Web and to test the sensitivity to change in this population as neither have been previously used in this format. A checklist to record current treatment was developed for the study to define usual treatment.

The only change to the published protocol was to record only the frequency of monitoring of EWS for hypomania and depression, as early feedback from participants indicated the full checklist was too long. All serious adverse events (SAEs) were recorded and reported to the Trial Steering Committee (TSC). All participants were given a £10 shopping voucher on completion of measures at each assessment point.

Analysis

Descriptive statistics report the characteristics of the sample recruited; use of the website; and rates of recruitment, retention, and data completion in each arm of the trial. The impact on repeated process and outcome measures was tested using linear models with correlated errors, which allow for correlation between repeated measures from the same participant. For ordinal data, we used generalized linear mixed models. We report both unadjusted analyses, and those adjusting for any differences in baseline demographic (age, gender, ethnicity, employment, education) and clinical variables (number of previous episodes, and whether or not prescribed a mood stabilizer).

Incomplete records from participants were retained, and analyses used maximum likelihood estimation for all model parameters. Statistical comparison of outcomes was made between the two trial arms at 24 and 48 weeks follow-up.

Weekly ratings of depression and mania from the SCID-LIFE were used to analyze time to first relapse (any and separately for depression [requiring 2 consecutive weeks], or mania [1 week]) and the proportion of time spent in episode (defined as SCID-LIFE rating of 5 or 6), or in subsyndromal state (SCID-LIFE rating of 3 or 4) or euthymic (SCID-LIFE rating of 1 or 2). To analyze the impact of the intervention on time to first relapse, we used a Cox's proportional hazards regression model. Beta regression was used to compare the proportion of time spent in episode or subsyndromal or euthymic in each arm.

The study is not powered to test for statistically significant impact and therefore we do not specify a primary outcome, or set a level of statistical significance for interpreting analyses. All analyses were run from R open-source computing environment version 3.3.1 (R Foundation for Statistical Computing).

Content analysis of qualitative survey data highlighted the individual points made and these were grouped into key themes. Interview transcripts were analyzed in depth using indexing and charting methods inspired by Framework Analysis [35]. All transcripts were independently coded by the interviewer (MG) and a second member of the research team. Codes were compiled into a tentative coding frame with thematic headings. Narrative summaries were created from each of the conceptual themes across all cases. This data will be reported in full elsewhere but here we present key data relevant to the feasibility and acceptability aims of the trial.

Results

Quality Assurance

Only two unblindings occurred. In both instances, the participant inadvertently indicated their group during a telephone assessment (one WL control, one ERPonline). Subsequent assessments were completed by a blind Research assistant. At each follow-up, 10% of the SCID-LIFE interviews were rated by both researchers and kappa statistic calculated to assess interrater reliability. These ranged from acceptable (κ=.54, 95% CI 0.39-0.69, at 36 weeks based on 142 weekly ratings for 6 participants) to high (κ=.90, 95% CI 0.82-0.98, at 12 weeks based on 192 weekly ratings for 8 people).

Feasibility and Acceptability of Trial Design

Participant flow is detailed in Figure 3, including recruitment, over half of which came via online sources. A total of 96 people were randomized (49 to WL, 47 to ERPonline) over a 17-month period, with 80% retention in telephone and online follow-up at all time points, except week 48 (76%). Attrition was 11% lower in WL arm.

Participants were predominantly diagnosed with BD1 (90%) and had a chronic relapsing course (67% had over 21 relapses; see Table 2). Despite this, the group was currently high functioning and had a positive attitude to recovery. The vast majority were taking and adherent to medication to manage their mood and over half had previously received psychological treatment for BD (where specified, this was most commonly described as cognitive behavior therapy [CBT]).

Table 2. Key characteristics of participant sample at baseline.

Participant characteristics		Wait list (n=49)	ERPonline (n=47)
Baseline demographic and clinical variables			
Age, mean (SD[a])		43.8 (11.45)	42 (12.23)
Gender, n (%)			
	Female	32 (65)	27 (57)
Ethnicity, n (%)			
	White British	44 (90)	38 (81)
	Any other white	2 (4)	5 (11)
	Black British	-	1 (2)
	Caribbean	1 (2)	-
	Asian British	1 (2)	-
	Indian	-	1 (2)
	Any other mixed	-	1 (2)
	Missing	1 (2)	1 (2)
Occupational status, n (%)			
	Full-time paid or self	16 (33)	21 (45)
	Part-time paid or self	13 (27)	6 (13)
	Voluntary	3 (6)	4 (9)
	Not employed	10 (20)	6 (13)
	Student	3 (6)	2 (4)
	Housewife or househusband	-	3 (6)
	Retired	4 (8)	5 (11)
Education, n (%)			
	No formal qualifications	1 (2)	-
	CSE[b] or O Level or GCSE[c]	5 (10)	4 (9)
	A Level	7 (14)	7 (15)
	Degree	17 (35)	16 (34)
	PG[d] Diploma or qualification	13 (27)	13 (28)
	Doctorate or PhD	3 (6)	7 (15)
Total number of past episodes (baseline), n (%)			
	<7	6 (12)	2 (4.)
	8-20	12 (25)	12 (26)
	21+	31 (63)	33 (70)
Taking mood stabilizer (baseline), n (%)			
	Not on any medication (so item rated not applicable)	3 (6)	2 (4)
	No mood stabilizer	15 (31)	11 (23)
	Lithium	12 (25)	11 (23)
	Sodium valproate	5 (10)	14 (30)
	Carbamazepine	3 (6)	1 (2)
	Lamotrigine	11 (22)	8 (17)
Treatment history (baseline), n (%)			

Participant characteristics		Wait list (n=49)	ERPonline (n=47)
	Ever used mental health services	46 (94)	42 (89)
	Clinical diagnosis of BD1[e] (vs BD2)	44 (90)	44 (94)
	Ever seen a psychiatrist	46 (94)	43 (92)
	Ever prescribed medication for BD	49 (100)	46 (98)
	Currently taking medication for bipolar disorder	40 (82)	40 (85)
	Ever received therapy or psychosocial intervention for bipolar disorder	26 (53)	27 (57)
	Currently receiving therapy for bipolar disorder	8 (16)	10 (21)
Process measures			
Early warning signs monitoring frequency— depression, n (%)			
	Never	3 (6)	3 (6)
	Occasionally	18 (37)	11 (23)
	Fairly regularly	8 (37)	20 (43)
	Very regularly	10 (20)	13 (28)
Early warning signs monitoring frequency— hypomania, n (%)			
	Never	6 (12)	5 (11)
	Occasionally	21 (43)	14 (30)
	Fairly regularly	16 (33)	18 (38)
	Very regularly	6 (12)	10 (21)
Brief Illness Perception Questionnaire—total (high score=more negative model), mean (SD)		60.6 (10.5)	60.7 (9.9)
Medication Adherence Rating Scale, mean (SD)		6.9 (2.2)	7.0 (2.1)
Outcome measures			
Hamilton Depression Rating Scale, mean (SD)		4.5 (5.3)	3.5 (4.2)
Mania Rating Scale, mean (SD)		1.3 (2.4)	1.0 (1.7)
Personal and Social Performance Scale, mean (SD)		79.7 (11.2)	79.1 (13.1)
Multidimensional Scale of Independent Functioning, n (%)		25 (51)	21 (45)
		14 (29)	16 (34)
		8 (16)	6 (13)
		2 (4)	2 (4)
		0 (0)	2 (4)
		0 (0)	0 (0)
Work and Social Adjustment Scale, mean (SD)		11.4 (9.4)	12.8 (8.7)
Quality of Life in Bipolar Disorder, mean (SD)		162.7 (33.5)	162.5 (22.8)
Bipolar Recovery Questionnaire, mean (SD)		2332 (394)	2342 (383)

[a]SD: standard deviation.

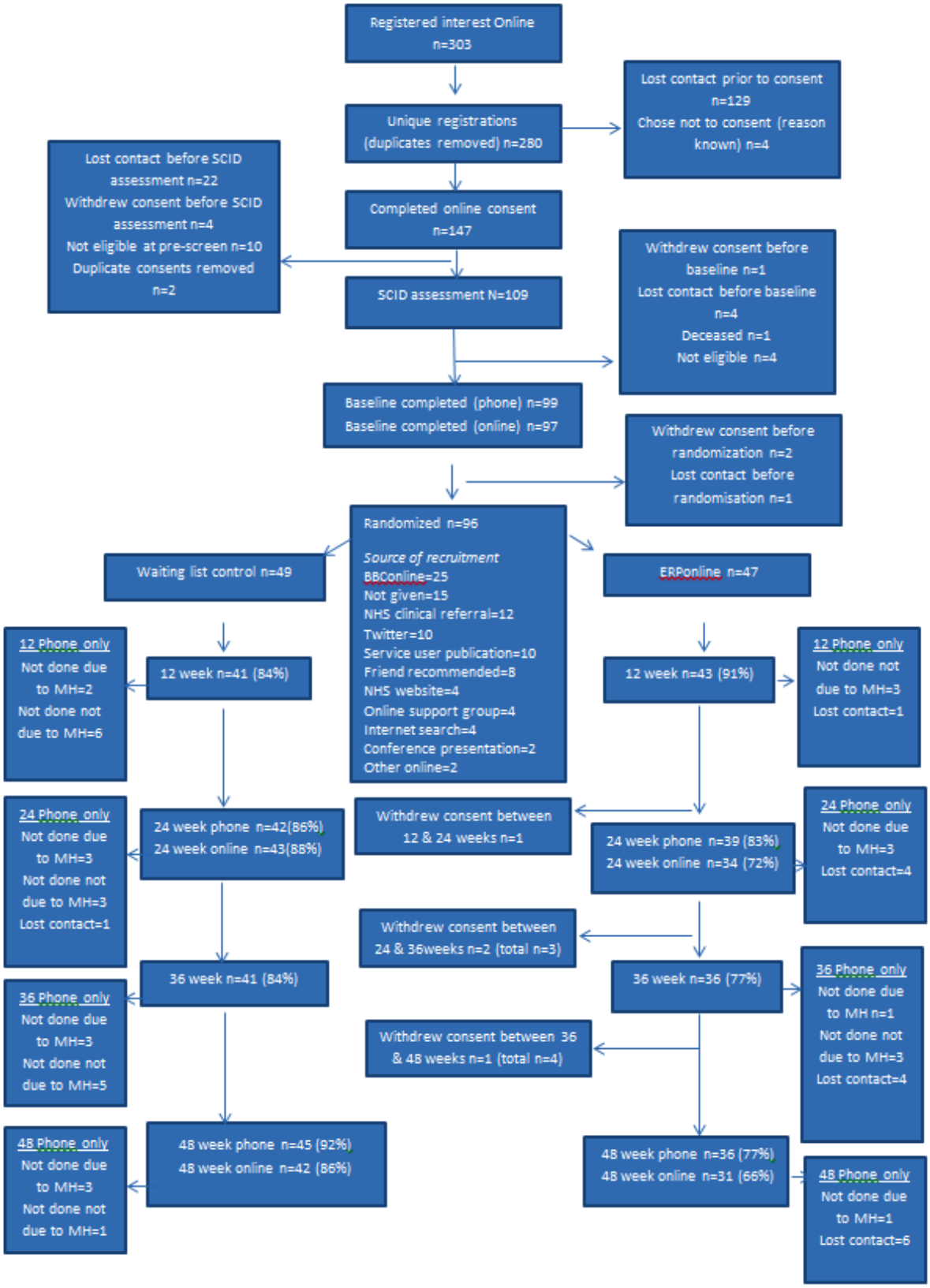
[b]CSE: Certificate of Secondary Education.

[c]GCSE: General Certificate of Secondary Education.

[d]PG: postgraduate.

[e]BD: bipolar disorder.

Figure 3. The online enhanced relapse prevention intervention (ERPonline) consort diagram.



Feasibility and Acceptability of Data Collection

Telephone and Web-based data collection procedures were generally acceptable to participants based on information from 41 participants (WL survey, n=22; ERPonline interviews, n=19 reported separately). Survey data indicated factors that encouraged people to take part including opportunity to improve their own resilience and self-management, wanting to help others, and recognizing the importance of research on improving Web-based interventions, due to a perceived gap in face-to-face

services and some existing websites feeling unsafe. Many of these factors were also cited as facilitating retention in the trial, as well as factors such as text reminders about follow-ups, and viewing the research team as sensitive, polite, and nonintrusive. Participants reported that the research process was well-managed, clearly explained, straightforward, and flexible, and they liked the shopping vouchers. Some participants believed completing the measures had changed their thinking about their mood and diagnosis. Barriers to retention included procedural difficulties, such as remembering follow-up times and rescheduling missed appointments, feeling "weird" to have interviews only on the phone and difficulties finding private space for the phone calls; issues with measures, some of which were too long (CSRI) and could be tiring, distressing, and required recall over long periods of time; and technical difficulties with Web-based questionnaires. Some reported feeling disappointed to be in the WL control arm, although had remained in the trial.

Additional key data collection lessons we learnt included the importance of checking electronic communication is received (some of our reminder emails were initially going into junk folders), the need to accommodate the high demand for evening and weekend telephone appointments, and the importance of text reminders for telephone appointments.

Feasibility and Acceptability of ERPonline Intervention

ERPonline is low cost at an estimated £19,340 to create, and approximately £2176 per year to host and maintain the site. Development costs included time to adapt content from the ERP manual; discussion and feedback with coauthors; Web developer time to build the site; filming and producing videos; and costs for the Service User Reference Group to feedback on early iterations. Hosting costs include software updates and technical issues (estimated at 2 h per week) and space on a server. To keep costs low, ERPonline was delivered unsupported and without a Web-based moderated forum, despite these being part of the intended design.

Activity levels were highly skewed. Two people allocated to ERPonline never visited the site. Mean number of page views per person was 259 (SD 577), median was 85 (range 0-3203). Participants spent a mean of 259 min (SD 509), median 76 min (range 0-2770) accessing ERPonline throughout the 48-week intervention period. The most frequently viewed modules were "Life Charting" (median views 11; range 0-990 per person) and "Mood Charting" (median views 13; range 1-2519 per person), which is unsurprising as they offered an ongoing monitoring function. "Coping strategies for Low Mood" (median views 0: range 0-44 per person) and "Your Staying Well Plan" (median views 0; range 0-30 per person) were the least frequently visited, but also occurred toward the end of the listed modules (see Table 1 for number of visits and time spent on all modules).

A total of 17 ERPonline participants (36%) responded to the Web-based survey. Overall these participants were satisfied (13/17 people [76%] somewhat or very satisfied), found it somewhat or very helpful (12/17, 71%), and very or extremely relevant (13/17, 76%). Only one person said they would not recommend it to a friend. Most useful features were recognizing EWS of relapse, shared experiences through videos, mood monitoring, ability to revisit and refresh skills, improved knowledge and self-management of BD, ease of use, and being able to use the site with the family. The key recommendation for improvement was additional support with working through the materials. The sample of questionnaire respondents described themselves as confident (n=16, 94% very or extremely confident) and regular (n=12, 71% at least daily) Internet users.

A total of 19 people took part in qualitative interviews about their experience of the trial and use of the ERPonline site. The key finding was the importance of relationships that the individual developed with the ERPonline team in determining retention into the study and use of the site:

> *I think the sort of general thoroughness and kindness of the people I dealt with that certainly contributed to, you know, me sort of staying in the study. Everybody's been really upbeat, very positive, very accommodating* [P10]

This is particularly interesting when we consider there was no face-to-face contact. Sole direct contact was by telephone at three monthly interviews for SCID-LIFE interviews, which for some was preferable to face-to-face:

> *…I think it being over the 'phone makes it a bit easier. If it's face-to-face I would have probably not been quite so comfortable answering. But yeah over the 'phone was definitely not so bad.* [P13]

Crucial to the strength of the alliance was the perceived trustworthiness of the team and being made aware of the extensive user involvement in design and content of the site:

> *It's always available and also the information's on there has been put together by the people who do know what they're doing.* [P7]

> *I suppose the prospect of the online study kept me quite interested and the fact that it was developed by other people with Bipolar and that was something that I was definitely interested in…* [P17]

A key recommendation to improve the ERPonline site was to integrate human support to facilitate ongoing use of the site. This is consistent with previous studies in which adherence was higher in groups receiving a Web intervention plus support as compared with Web intervention only [31]. Participants felt websites should be used to support interventions delivered by real people and not as a cheaper replacement:

> *The cash strapped health service will rely heavily on these sort of techniques which I think only fill one part of the market. I think they only really deal with, you know, and a comparatively narrow field of potential patients. I think they're very useful but I do think the gold standard involves some sort of face-to-face psychological therapy. And I think the clinical literature bears that out so, I want more jobs for psychologists basically* [Henry: 23.893-23.898]

Table 3. Descriptive statistics on outcome and process measures at 24- and 48-week follow-ups.

Variable	Group	Baseline, n WL[a]=49(I[c]); 49(O[d]) ERP[b]=47(I); 47(O)	12 weeks WL=41(I) ERP=43 (I)	24 weeks WL=42(I); 43(O) ERP=39(I); 34(O)	36 weeks WL=41(I) ERP=36(I)	48 weeks WL=45(I); 42(O) ERP=36(I); 31(O)
Process measures						
Early warning signs monitoring frequency—depression (O), n (%)	never WL	3 (6)		8 (19)		4 (10)
	Occasionally	18 (37)		13 (30)		11 (26)
	Fairly regularly	18 (37)		17 (40)		19 (45)
	Very regularly	10 (20)		5 (12)		6 (14)
	never ERP	3 (6)		1 (3)		1 (3)
	Occasionally	11 (23)		11 (32)		6 (19)
	Fairly regularly	20 (43)		15 (44)		10 (32)
	Very regularly	13 (28)		7 (21)		14 (45)
Early warning signs monitoring frequency—hypomania (O), n (%)	never WL	6 (12)		11 (26)		4 (10)
	Occasionally	21 (43)		17 (40)		18 (43)
	Fairly regularly	16 (33)		9 (21)		11 (26)
	Very regularly	6 (12)		6 (14)		7 (17)
	never ERP	5 (11)		2 (6)		1 (3)
	Occasionally	14 (30)		10 (29)		8 (26)
	Fairly regularly	18 (38)		15 (44)		11 (35)
	Very regularly	10 (21)		7 (21)		11 (35)
BIPQ[e]—total (O), mean (SD[f])	WL	60.6 (10.5)		49.4 (22.6)		49.4 (24.6)
	ERP	60.7 (9.9)		39.3 (27.0)		36.2 (29.1)
MARS[g] (O), mean (SD)	WL	6.9 (2.2)		6.7 (2.6)		6.6 (2.5)
	ERP	7.0 (2.1)		6.8 (2.7)		7.0 (2.2)
Outcome measures						
HAM-D[h] (I), mean (SD)	WL	4.5 (5.3)	7.5 (7.4)	7.3 (8.6)	6.8 (8.6)	8.2 (9.0)
	ERP	3.5 (4.2)	6.6 (6.7)	6.9 (8.0)	6.0 (8.3)	7.1 (9.3)
MRS[i] (I), mean (SD)	WL	1.3 (2.4)	2.7 (3.7)	2.2 (4.1)	1.7 (2.9)	1.7 (2.2)
	ERP	1.0 (1.7)	2.5 (4.4)	2.4 (3.9)	2.0 (4.0)	1.4 (2.4)
PSP[j] (I), mean (SD)	WL	79.7 (11.2)	75.0 (15.1)	76 (16.4)	79.8 (14.8)	78.4 (15.6)
	ERP	79.1 (13.1)	77.8 (15.1)	76.7 (15.4)	77.8 (16.1)	80.7 (16.1)
MSIF[k]-global (frequencies for scores categories 1, 2, 3, 4, 5, 6) (I), n (%)	WL					

Variable	Group	Baseline, n WL[a]=49(I[c]); 49(O[d]) ERP[b]=47(I); 47(O)	12 weeks WL=41(I) ERP=43 (I)	24 weeks WL=42(I); 43(O) ERP=39(I); 34(O)	36 weeks WL=41(I) ERP=36(I)	48 weeks WL=45(I); 42(O) ERP=36(I); 31(O)
	1	25 (51)	24 (59)	20 (48)	25 (61)	28 (62)
	2	14 (29)	8 (20)	13 (31)	10 (24)	9 (20)
	3	8 (16)	4(10)	5 (12)	3 (7)	6 (13)
	4	2 (4)	5 (12)	3 (7)	1 (2)	1 (2)
	5	0 (0)	0 (0)	1 (2)	2 (5)	0 (0)
	6	0 (0)	0 (0)	0 (0)	0 (0)	1 (2)
	ERP					
	1	21 (45)	26 (60)	18 (46)	20 (56)	22 (61)
	2	16 (34)	10 (23)	12 (31)	9 (25)	9 (25)
	3	6 (13)	3 (7)	6 (15)	5 (14)	4 (11)
	4	2 (4)	3 (7)	2 (5)	0 (0)	1 (3)
	5	2 (4)	1 (2)	1 (3)	2 (6)	0 (0)
	6	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)
WSAS[l] (I), mean (SD)	WL	11.4 (9.4)		12.4 (10)		12.9 (10.6)
	ERP	12.8 (8.7)		14.3 (9.1)		14.8 (10.3)
QoLBD[m] (O), mean (SD)	WL	162.7 (33.5)		161.2 (39.7)		154.9 (36.1)
	ERP	162.5 (22.8)		156.5 (33.4)		151.8 (41.7)
BRQ[n] (O) mean (SD)	WL	2332 (394)		2309 (504)		2336 (468)
	ERP	2342 (383)		2451 (430)		2414 (577)

[a]WL: waitlist.

[b]ERP: enhanced relapse prevention.

[c]I: interviewer rated by telephone.

[d]O: completed online.

[e]BIPQ: Brief Illness Perception Questionnaire (score 0-110 higher score=more negative beliefs).

[f]SD: standard deviation.

[g]MARS: Medication Adherence Rating Scale (0-10 higher score=higher compliance).

[h]HAM-D: Hamilton Depression Rating Scale (scores above 7 indicate mild depression, above 13 moderate, and above 18 severe).

[i]MRS: Mania Rating Scale (scores of 11 and above indicate hypomania).

[j]PSP: Personal and Social Performance Scale (scores 70-80 indicate mild difficulties, above 80 is good functioning).

[k]MSIF: Multidimensional Scale of Independent Functioning (Likert scale 1=normal functioning, to 7=total disability).

[l]WSAS: Work and Social Adjustment Scale (less than 10=subclinical; 10-20 some functional impairment; above 20 moderate psychopathology).

[m]QoLBD: Quality of Life in Bipolar Disorder (range 48-240 with high score=higher quality of life).

[n]BRQ: Bipolar Recovery Questionnaire (score 0-3600 high score=higher recovery).

Table 4. Comparison of linear models with correlated errors to test for differences between waitlist (WL) and Web-based enhanced relapse prevention intervention (ERPonline) on outcome and process measures at 12-, 24-, and 48-week follow-ups. Unadjusted model showing estimates of difference between beta estimates at each time point.

Variable	Model 1—Unadjusted analysis							
	12-week follow-up estimate	95% CI (*P* value)	24-week follow-up estimate	95% CI (*P* value)	36-week follow-up estimate	95% CI (*P* value)	48-week follow-up estimate	95% CI (*P* value)
Early warning signs monitoring frequency—depression			0.73	−1.86 to 0.40 (.20)			−1.39	−2.61 to −.163 (.03)
Early warning signs monitoring frequency—hypomania			−1.72	−2.98 to −.47 (.01)			−1.61	−2.92 to −.30 (.02)
Brief Illness Perception Questionnaire[a]—total			10.70	0.90 to 20.5 (.03)			13.18	2.44 to 23.93 (.02)
Medication Adherence Rating Scale			−0.102	−1.07 to 0.87 (.84)			−0.327	−1.34 to 0.685 (.53)
Personal and Social Performance Scale	−2.91	−9.19 to 3.37 (.36)	−0.198	−6.77 to 6.38 (.95)	0.77	−5.87 to 7.41 (.82)	−2.87	−9.27 to 3.52 (.38)
Multidimensional Scale of Independent Functioning—global (frequencies for scores categories 1, 2, 3, 4, 5, 6)	0.169	−0.867 to 1.20 (.75)	.029	−0.959 to 1.02 (.95)	−0.074	−1.15 to 1.00 (.89)	.281	−0.785 to 1.35 (.61)
Work and Social Adjustment Scale			−1.61	−5.67 to 1.46 (.30)			−0.53	−3.90 to 2.83 (.76)
Quality of Life in Bipolar Disorder			6.67	−5.73 to 19.1 (.29)			3.99	−9.72 to 17.7 (.57)
Bipolar Recovery Questionnaire			−52.3	−195 to 90.7 (.47)			−38.09	−208.58 to 132.41 (.66)

[a]Brief Illness Perception Questionnaire total measures how negative a model the person has—so high score=more negative model.

Table 5. Comparison of linear models with correlated errors to test for differences between waitlist (WL) and Web-based enhanced relapse prevention intervention (ERPonline) on outcome and process measures at 12-, 24-, and 48-week follow-ups. Adjusted for any differences in baseline demographic (age, gender, ethnicity, employment, education) and clinical variables (number of previous episodes, and whether or not prescribed a mood stabilizer).

Variable	Model 2—Adjusted analysis							
	12-week follow-up estimate	95% CI (*P* value)	24-week follow-up estimate	95% CI (*P* value)	36-week follow-up estimate	95% CI (*P* value)	48-week follow-up estimate	95% CI (*P* value)
Early warning signs monitoring frequency—depression			−0.721	−1.85 to 0.413 (.21)			−1.38	−2.61 to −0.153 (.03)
Early warning signs monitoring frequency—hypomania			−1.70	−2.96 to −0.452 (.01)			−1.60	−2.91 to −0.291 (.02)
Brief Illness Perception Questionnaire[a]—total			11.06	1.25 to 20.9 (.03)			13.6	2.69 to 24.6 (.02)
Medication Adherence Rating Scale			−0.128	−1.11 to 0.855 (.80)			−0 .356	−1.39 to 0.674 (.50)
Personal and Social Performance Scale	−2.99	−8.84 to 2.863 (.32)	−0.430	−7.02 to 6.16 (.90)	.704	−6.06 to 7.47 (.84)	−3.21	−9.69 to 3.26 (.33)
Multidimensional Scale of Independent Functioning—global (frequencies for scores categories 1, 2, 3, 4, 5, 6)	0.092	−0.907 to 1.09 (.86)	−0.075	−1.03 to 0.878 (.88)	−0.084	−1.12 to 0.95 (.87)	0.275	−0.743 to 1.29 (.60)
Work and Social Adjustment Scale			−1.46	−4.53 to 1.64 (.36)			−0.40	−3.84 to 3.03 (.82)
Quality of Life in Bipolar Disorder			6.23	−6.13 to 18.6 (.32)			3.58	−10.4 to 17.5 (.62)
Bipolar Recovery Questionnaire			−63.29	−206.47 to 79.88 (.39)			−35.84	−209.78 to 138.10 (.69)

[a]Brief Illness Perception Questionnaire total measures how negative a model the person has—so high score=more negative model.

Estimate of Impact on Outcome and Process Measures

Descriptive statistics on process and outcome measures at each time point are shown in Table 3. Comparison between WL and ERPonline on process and outcome measures at 12-, 24-, and 48-week follow-ups are shown in Table 4. Models adjusting for baseline demographic and clinical variables are also shown in Table 5.

Process Measures

ERPonline increased the frequency of monitoring early signs of mood change (EWS—depression and EWS—hypomania), evident for hypomania at 24 weeks (−1.72, 95% CI −2.98 to −0.47), and for both at 48 weeks (depression −1.39, 95% CI −2.61 to −0.163; hypomania −1.61, 95% CI −2.92 to −0.30), and improved working model of mood changes (BIPQ—high score indicates more negative model) at both 24 weeks (10.70, 95% CI 0.90-20.5) and 48 weeks follow-ups (13.18, 95% CI 2.44-23.93; Table 4). All differences are robust to adjustments in model 2 for baseline differences between the groups (Table 5). Medication adherence was high (indicated by high score on the MARS) throughout the study and did not differ between groups.

Mood and Functioning

Depression and hypomania were low at all time points suggesting a generally stable and euthymic group. Similarly, functioning on WSAS, PSP, and MSIF at baselines were suggestive of very mild impairment in work and social performance and remained so throughout. Time spent in euthymic, subsyndromal, and relapse mood states respectively in WL were 93% (SD 8%), 5% (SD 7%), and 3% (SD 15%), and in ERPonline: 95% (SD 8%), 4% (SD 6%), and 2% (SD 4%). There were no notable differences between the two groups in any of the outcome measures at any of the time-points.

Relapse

Of the 96 participants, one provided no SCID-LIFE data at follow-up. Only 15 (16%) participants experienced a relapse over the 48-week follow-up; 11 (11%) depressive and 7 (7%) mania-type. There were no notable differences between groups on time to any relapse (unadjusted hazard ratio [HR] 1.67, 95% CI 0.60-4.71), *P*=.33; time to depressive episode (HR 1.53, 95% CI 0.49-4.83), *P*=.47; time to mania type episode (HR 2.87, 95% CI 0.56-14.8), *P*=.12. Given the low level relapses, no Kaplan-Meier curves are presented.

Adverse Events

During the trial, 1 participant completed suicide before randomization, 11 participants (11% of those randomized) reported suicidality and self-harm, and 1 made a suicide attempt (before withdrawing from the study). It was found that 6 were

in the WL arm, and 5 were receiving ERPonline. None of the SAEs were deemed study related by an independent TSC.

Discussion

Principal Findings

ERPonline is a novel Web intervention for improving relapse prevention and providing National Institute for Health and Care Excellence (NICE) congruent information to people with BD. This study indicates that the development and evaluation of this type of approach in a rigorous RCT using telephone and Web-based assessments is both feasible and acceptable. Important lessons were learnt relevant to each of our study aims, but which also have relevance to the wider development and evaluation of remote-access approaches for other health problems.

With the help of our Service User Reference Group, we were able to develop a Web-based version of an existing ERP intervention for people with BD at a very low cost that received largely positive feedback, and led to no evident adverse events. Activity was highly skewed but over 90% of our sample visited the site more than once, which can be compared with MyRecoveryPlan [17] that reported site returns for 71% in a coached group, and only 44% in an unsupported group. On the basis of levels of use of the different modules and direct participant feedback, engagement could be enhanced by making the intervention more interactive and providing support to use it.

Recruitment, retention, and data completion strategies were largely successful. Retention was higher than demonstrated in previous Web-based trials with people who are affected with BD [14,15,17]. Key features of the trial design that facilitated this included payment for completing assessments, text and email reminders, a WL control design, and a friendly flexible research team who were willing to offer telephone appointments at times to suit participants including out-of-office hours. However, to reach the sample size required for large scale clinical and cost effectiveness trials, paying for advertising through popular websites such as Google and Facebook may be necessary [36].

Feedback about the experience of taking part in a primarily Web-based trial was mixed. Some participants reported difficulties finding a private space to take telephone calls or finding Web-based measures difficult, tiring or distressing, whereas others valued the flexibility, convenience, and felt more able to be open about the problems they had experienced than in a face-to-face interview. This suggests that trials which offer a choice of data collection options may be most effective in achieving recruitment and retention targets.

However, further work is needed to test the validity and reliability of these data collection approaches. Our data showed that while the hypothesized increase in EWS monitoring and development of more positive beliefs about mood swings did occur in those receiving ERPonline compared with WL control group, we did not see any benefit of ERPonline on any of the clinical outcome measures. This was largely due to the ceiling effect on our outcome measures. Only 16% of the total sample experienced any relapse, compared with expected levels of 50-70% [37].

This ceiling effect was consistent across all outcome measures and all assessors. Therefore, the most likely explanations are either that the method of data collection is leading to underreporting of problems, or that the participating sample reflect a different population from those taking part in more traditionally designed face-to-face clinical studies.

With regards to the first possibility, whereas we did not directly test the reliability of the data compared with a face-to-face interview, other studies have done this comparing telephone and face-to-face interview data of SCID assessments found high levels of agreement [38]. Our team have also carried out a parallel Web-based RCT which included the same Web-based and telephone assessments, delivered through researchers trained by the same methods, and which will report relapse rates of 47% which are akin to those expected from previous research data and much higher than in this study [39].

The second possible explanation can be explored by examining the characteristics of our participants. Compared with bipolar samples recruited to other face-to-face trials [37] including one evaluating clinician delivered ERP [7], and samples in other Web-based trials which all show higher relapse rates [14,15,17], our sample are more euthymic, highly educated, likely to be in employment, and have had surprisingly high levels of access to previous psychological therapy. Further work is needed to better understand how using a primarily Web-based trial design may impact on sample characteristics, and the information they provide.

The Future for ERPonline

ERPonline offers a cheap and easily accessible option for people who are seeking ongoing support following successful treatment, which is currently unavailable. However, given the high functioning and low relapse rates evident in this study, testing the clinical effectiveness of ERPonline for this population would require very large sample sizes. Alternatively, ERPonline could target people at an earlier stage of treatment, who have had not yet received more expensive face-to-face psychological therapy, and need support to understand their mood swings, consider the pros and cons of medication use, and explore the usefulness of monitoring and managing EWS of relapse. For this group, ERPonline may offer a way to reduce the need for expensive individual therapy. Consistent with participant recommendations and previous research, we also need to consider how best to integrate support mechanisms to facilitate use of the intervention, either by integrating the Web-based resource with clinician delivered relapse prevention, or through online peer support as described in other Web-based interventions for BD [10,11,13]. This study highlights the importance of the relationship that the users have with Web-based interventions and how this develops as an extension of the relationship with the humans perceived as offering and supporting its use. Web-based interventions offered in isolation in this context seem unlikely to engage people in the same way and may be perceived negatively as attempts to save money rather than improve care. Our study has explored the feasibility and acceptability of a specific Web-based intervention (ERPonline),

but does not address the broader social issue of how acceptable the increasing use of digital health technology is to people with mental health problems [40].

Strengths of Study

Extensive user involvement improved the content of the ERPonline website, identified recruitment sources, and ensured the measures were appropriate and not too burdensome. The sample was sufficiently large to be able to comment on patterns in the data likely to be indicative of effects on process and outcome measures in a larger trial. Independent randomization, trained blind assessors, and the use of well-established outcome and process measures ensured that the data are reliable and valid. Extensive reflection and learning around feasibility was built into the design process using face-to-face meetings and an online reflection log.

Limitations

Despite 280 unique site registrations, only 145 people consented, and due to ineligibility and drop out, only 96 were randomized. We have no data on why nearly half the sample registering an interest, then chose not to take part, though for some it may have been delay between prestart expression of interest and randomization. During the trial, we had higher dropout in the ERPonline arm, which is common in trials with a WL control arm and is likely due to the perceived reward of the intervention retaining people through WL. Survey responses were incomplete for feedback on trial participation (22/49 in WL group, 45%) and for feedback on the ERPonline intervention (17/47 in ERPonline arm, 36%). The bias in responders is likely to skew the nature of the feedback which on the whole was very positive.

In summary, we were able to successfully adapt and deliver online a relapse prevention intervention for BD previously used face-to-face. The intervention was successfully evaluated against a WL control group using a RCT design with high levels of retention and data completeness over 48 weeks. Participants had high rates of previous bipolar episodes but had accessed previous psychosocial interventions (where specified, most commonly described as CBT) for BD. Web-based interventions may prove an important cheap, feasible, and acceptable step forward in creating a choice of evidence-based interventions for people with BD at different stages of recovery, but may be more appropriately designed with built-in support and targeted at those with less prior experience of effective care.

Acknowledgments

The authors would like to thanks all the participants, the Service User Reference Group, and the Trial Steering Committee including Professor Danny Smith (Chair), Professor Gillian Haddock, Dr Kim Wright, Mr Tim Rawcliffe, and Dr Chris Sutton. This paper presents independent research commissioned by the National Institute of Health Research (NIHR) under the Research for Patient Benefit (RfPB) Programme (Grant Reference Number PB-PG-0211-10001). The views expressed are those of the authors and not necessarily those of the NHS, the NIHR or the Department of Health. The funder played no role in the study design, conduct or interpretation of the data. The corresponding author had full access to all the data in the study and had final responsibility for the decision to submit for publication. RM is funded by the Collaboration for Leadership in Applied Health Research and Care and East Midlands and the NIHR MindTech Health Technology Co-operative.

Authors' Contributions

FL was the study Chief Investigator, leading the design of the trial and ERPonline intervention, study conduct, and write up of the study. AD managed day-to-day management of the trial and data collection. OA and PD did the statistical analysis. AS, MG, and DK collected the data. MH analyzed the Web usage data. RL Chaired the SURG, ensuring continued service user input throughout the study. RP developed the ERPonline website. RM and SJ contributed to the design analysis and interpretation of data, and edited drafts of the paper. DD facilitated recruitment through NHS Trusts. All authors were involved in drafting and final approval of this paper.

Conflicts of Interest

ERPonline was developed by the authors and therefore this is not an independent evaluation. Prof Morriss is the Mood Disorders Theme Lead for the NIHR MindTech Health Technology Co-operative. He also chaired the NICE Clinical Guideline Development Group for bipolar disorder that recommended psychological treatment for long-term maintenance of bipolar disorder.

References

1. Whiteford HA, Degenhardt L, Rehm J, Baxter AJ, Ferrari AJ, Erskine HE, et al. Global burden of disease attributable to mental and substance use disorders: findings from the Global Burden of Disease Study 2010. Lancet 2013 Nov 09;382(9904):1575-1586. [doi: 10.1016/S0140-6736(13)61611-6] [Medline: 23993280]
2. McCrone P. Paying the Price: The Cost of Mental Health Care in England to 2026. London: King's Fund Publishing; 2008.
3. National Institute for HealthClinical Excellence. NICE. London: National Institute for Health and Care Excellence; 2014 Nov 24. Bipolar disorder: assessment and management (clinical guideline 185) URL: https://www.nice.org.uk/guidance/cg185/resources/bipolar-disorder-assessment-and-management-35109814379461 [accessed 2017-03-02] [WebCite Cache ID 6ofJKc9Fl]

4. Malhi GS, Bassett D, Boyce P, Bryant R, Fitzgerald PB, Fritz K, et al. Royal Australian and New Zealand college of psychiatrists clinical practice guidelines for mood disorders. Aust N Z J Psychiatry 2015 Dec;49(12):1087-1206. [doi: 10.1177/0004867415617657] [Medline: 26643054]
5. Lehman AF, Lieberman JA, Dixon LB, McGlashan TH, Miller AL, Perkins DO, American Psychiatric Association, Steering Committee on Practice Guidelines. Practice guideline for the treatment of patients with schizophrenia, second edition. Am J Psychiatry 2004 Feb;161(2 Suppl):1-56. [Medline: 15000267]
6. Merikangas KR, Jin R, He J, Kessler RC, Lee S, Sampson NA, et al. Prevalence and correlates of bipolar spectrum disorder in the world mental health survey initiative. Arch Gen Psychiatry 2011 Mar;68(3):241-251 [FREE Full text] [doi: 10.1001/archgenpsychiatry.2011.12]
7. Lobban F, Taylor L, Chandler C, Tyler E, Kinderman P, Kolamunnage-Dona R, et al. Enhanced relapse prevention for bipolar disorder by community mental health teams: cluster feasibility randomised trial. Br J Psychiatry 2010 Jan;196(1):59-63 [FREE Full text] [doi: 10.1192/bjp.bp.109.065524] [Medline: 20044662]
8. Hollis C, Falconer C, Martin J, Whittington C, Stockton S, Glazebrook C, et al. J Child Psychol Psychiatr. 2016 Dec 10. Annual research review: digital health interventions for children and young people with mental health problems: a systematic and meta-review URL: http://onlinelibrary.wiley.com/doi/10.1111/jcpp.12663/full [accessed 2017-03-02] [WebCite Cache ID 6ofKCi0xv]
9. Wallin EE, Mattsson S, Olsson EM. The preference for internet-based psychological interventions by individuals without past or current use of mental health treatment delivered online: a survey study with mixed-methods analysis. JMIR Ment Health 2016 Jun 14;3(2):e25 [FREE Full text] [doi: 10.2196/mental.5324] [Medline: 27302200]
10. Arnberg FK, Linton SJ, Hultcrantz M, Heintz E, Jonsson U. Internet-delivered psychological treatments for mood and anxiety disorders: a systematic review of their efficacy, safety, and cost-effectiveness. PLoS One 2014;9(5):e98118 [FREE Full text] [doi: 10.1371/journal.pone.0098118] [Medline: 24844847]
11. Drozd F, Vaskinn L, Bergsund H, Haga S, Slinning K, Bjørkli CA. The implementation of internet interventions for depression: a scoping review. J Med Internet Res 2016 Sep 08;18(9):e236 [FREE Full text] [doi: 10.2196/jmir.5670] [Medline: 27608548]
12. Smith DJ, Griffiths E, Poole R, di FA, Barnes E, Kelly MJ, et al. Beating bipolar: exploratory trial of a novel internet-based psychoeducational treatment for bipolar disorder. Bipolar Disord 2011;13(5-6):571-577. [doi: 10.1111/j.1399-5618.2011.00949.x] [Medline: 22017225]
13. Murray G, Leitan ND, Berk M, Thomas N, Michalak E, Berk L, et al. Online mindfulness-based intervention for late-stage bipolar disorder: pilot evidence for feasibility and effectiveness. J Affect Disord 2015 Jun 01;178:46-51. [doi: 10.1016/j.jad.2015.02.024] [Medline: 25795535]
14. Lauder S, Chester A, Castle D, Dodd S, Gliddon E, Berk L, et al. A randomized head to head trial of MoodSwings.net.au: an Internet based self-help program for bipolar disorder. J Affect Disord 2015 Jan 15;171:13-21. [doi: 10.1016/j.jad.2014.08.008] [Medline: 25282145]
15. Proudfoot J, Parker G, Manicavasagar V, Hadzi-Pavlovic D, Whitton A, Nicholas J, et al. Effects of adjunctive peer support on perceptions of illness control and understanding in an online psychoeducation program for bipolar disorder: a randomised controlled trial. J Affect Disord 2012 Dec 15;142(1-3):98-105. [doi: 10.1016/j.jad.2012.04.007] [Medline: 22858215]
16. Barnes C, Hadzi-Pavlovic D, Wilhelm K, Mitchell P. A web-based preventive intervention program for bipolar disorder: outcome of a 12-months randomized controlled trial. J Affect Disord 2015 Mar 15;174:485-492. [doi: 10.1016/j.jad.2014.11.038] [Medline: 25554993]
17. Simon GE, Ludman EJ, Goodale LC, Dykstra DM, Stone E, Cutsogeorge D, et al. An online recovery plan program: can peer coaching increase participation? Psychiatr Serv 2011 Jun;62(6):666-669 [FREE Full text] [doi: 10.1176/ps.62.6.pss6206_0666] [Medline: 21632737]
18. Todd NJ, Jones SH, Hart A, Lobban FA. A web-based self-management intervention for Bipolar Disorder 'living with bipolar': a feasibility randomised controlled trial. J Affect Disord 2014 Dec;169:21-29. [doi: 10.1016/j.jad.2014.07.027] [Medline: 25129531]
19. Bauer R, Conell J, Glenn T, Alda M, Ardau R, Baune BT, et al. Internet use by patients with bipolar disorder: results from an international multisite survey. Psychiatry Res 2016 Aug 30;242:388-394. [doi: 10.1016/j.psychres.2016.05.055] [Medline: 27391371]
20. Lobban F, Dodd AL, Dagnan D, Diggle PJ, Griffiths M, Hollingsworth B, et al. Feasibility and acceptability of web-based enhanced relapse prevention for bipolar disorder (ERPonline): trial protocol. Contemp Clin Trials 2015 Mar;41:100-109 [FREE Full text] [doi: 10.1016/j.cct.2015.01.004] [Medline: 25602581]
21. First M. User's Guide for the Structured Clinical Interview for DSM-IV Axis I Disorders SCID-I: Clinician Version. Washington, DC: American Psychiatric Press; 1997.
22. Lobban F, Solis-Trapala I, Symes W, Morriss R, ERP Group, University of Liverpool. Early warning signs checklists for relapse in bipolar depression and mania: utility, reliability and validity. J Affect Disord 2011 Oct;133(3):413-422. [doi: 10.1016/j.jad.2011.04.026] [Medline: 21640385]
23. Lobban F, Solis-Trapala I, Tyler E, Chandler C, Morriss RK. The role of beliefs about mood swings in determining outcome in bipolar disorder. Cogn Ther Res 2012 Mar 29;37(1):51-60. [doi: 10.1007/s10608-012-9452-9]

24. Thompson K, Kulkarni J, Sergejew AA. Reliability and validity of a new medication adherence rating scale (MARS) for the psychoses. Schizophr Res 2000 May 05;42(3):241-247. [Medline: 10785582]
25. Keller MB, Lavori PW, Friedman B, Nielsen E, Endicott J, McDonald-Scott P, et al. The longitudinal interval follow-up evaluation. a comprehensive method for assessing outcome in prospective longitudinal studies. Arch Gen Psychiatry 1987 Jun;44(6):540-548. [Medline: 3579500]
26. Hamilton M. A rating scale for depression. J Neurol Neurosurg Psychiatry 1960 Feb;23:56-62 [FREE Full text] [Medline: 14399272]
27. Bech P, Rafaelsen OJ, Kramp P, Bolwig TG. The mania rating scale: scale construction and inter-observer agreement. Neuropharmacology 1978 Jun;17(6):430-431. [Medline: 673161]
28. Morosini PL, Magliano L, Brambilla L, Ugolini S, Pioli R. Development, reliability and acceptability of a new version of the DSM-IV social and occupational functioning assessment scale (SOFAS) to assess routine social functioning. Acta Psychiatr Scand 2000 Apr;101(4):323-329. [Medline: 10782554]
29. Jaeger J, Berns SM, Czobor P. The multidimensional scale of independent functioning: a new instrument for measuring functional disability in psychiatric populations. Schizophr Bull 2003;29(1):153-168. [Medline: 12908671]
30. Mundt JC, Marks IM, Shear MK, Greist JH. The work and social adjustment scale: a simple measure of impairment in functioning. Br J Psychiatry 2002 May;180:461-464 [FREE Full text] [Medline: 11983645]
31. Michalak EE, Murray G, Collaborative RESearch Team to Study Psychosocial Issues in Bipolar Disorder (CREST.BD). Development of the QoL.BD: a disorder-specific scale to assess quality of life in bipolar disorder. Bipolar Disord 2010 Nov;12(7):727-740. [doi: 10.1111/j.1399-5618.2010.00865.x] [Medline: 21040290]
32. Jones S, Mulligan LD, Higginson S, Dunn G, Morrison AP. The bipolar recovery questionnaire: psychometric properties of a quantitative measure of recovery experiences in bipolar disorder. J Affect Disord 2013 May;147(1-3):34-43. [doi: 10.1016/j.jad.2012.10.003] [Medline: 23182591]
33. Herdman M, Gudex C, Lloyd A, Janssen M, Kind P, Parkin D, et al. Development and preliminary testing of the new five-level version of EQ-5D (EQ-5D-5L). Qual Life Res 2011 Dec;20(10):1727-1736 [FREE Full text] [doi: 10.1007/s11136-011-9903-x] [Medline: 21479777]
34. Chisholm D, Knapp MR, Knudsen HC, Amaddeo F, Gaite L, van WB. Client socio-demographic and service receipt inventory--European version: development of an instrument for international research. Epsilon study 5. European psychiatric services: inputs linked to outcome domains and needs. Br J Psychiatry Suppl 2000(39):s28-s33. [Medline: 10945075]
35. Ritchie J, Lewis J. Qualitative Research Practice: A Guide for Social Science Students and Researchers. London: Sage Publications; 2003.
36. Morgan AJ, Jorm AF, Mackinnon AJ. Internet-based recruitment to a depression prevention intervention: lessons from the Mood Memos study. J Med Internet Res 2013 Feb 12;15(2):e31 [FREE Full text] [doi: 10.2196/jmir.2262] [Medline: 23403043]
37. Oud M, Mayo-Wilson E, Braidwood R, Schulte P, Jones SH, Morriss R, et al. Psychological interventions for adults with bipolar disorder: systematic review and meta-analysis. Br J Psychiatry 2016 Mar;208(3):213-222. [doi: 10.1192/bjp.bp.114.157123] [Medline: 26932483]
38. Hajebi A, Motevalian A, Amin-Esmaeili M, Hefazi M, Radgoodarzi R, Rahimi-Movaghar A, et al. Telephone versus face-to-face administration of the structured clinical interview for diagnostic and statistical manual of mental disorders, fourth edition, for diagnosis of psychotic disorders. Compr Psychiatry 2012 Jul;53(5):579-583. [doi: 10.1016/j.comppsych.2011.06.001] [Medline: 21820651]
39. Jones S, Wainwright LD, Jovanoska J, Vincent H, Diggle PJ, Calam R, et al. An exploratory randomised controlled trial of a web-based integrated bipolar parenting intervention (IBPI) for bipolar parents of young children (aged 3-10). BMC Psychiatry 2015 Jun 06;15:122 [FREE Full text] [doi: 10.1186/s12888-015-0505-y] [Medline: 26047808]
40. Poder TG, Bellemare CA, Bédard SK, Lemieux R. Social acceptance and population confidence in telehealth in Quebec. BMC Health Serv Res 2015;15:72 [FREE Full text] [doi: 10.1186/s12913-015-0727-1] [Medline: 25889230]

Abbreviations

BBC: British Broadcasting Corporation
BD: bipolar disorder
BIPQ: Brief Illness Perception Questionnaire
BRQ: Bipolar Recovery Questionnair
CBT: cognitive behavior therapy
CSE: Certificate of Secondary Education
CSRI: Client Service Receipt Inventory
CTU: clinical trials unit
DSM-IV: Diagnostic and Statistical Manual of Mental Disorders-IV
ERP: enhanced relapse prevention
EWS: early warning signs

GCSE: General Certificate of Secondary Education
HAM-D: Hamilton Depression Rating Scale
HR: hazard ratio
MARS: Medication Adherence Rating Scale
MRS: Mania Rating Scale
MSIF: Multidimensional Scale of Independent Functioning
NHS: National Health Service
NICE: National Institute for Health and Care Excellence
PG: postgraduate
PSP: Personal and Social Performance Scale
QoLBD: Quality of Life in Bipolar Disorder
RA: Research Associate
RCT: randomized controlled trial
SAE: severe adverse events
SCID: Structured Clinical Interview for Diagnostic and Statistical Manual of Mental Disorders-IV (DSM-IV)
TSC: Trial Steering Committee
WL: waitlist
WSAS: Work and Social Adjustment Scale

Edited by G Eysenbach; submitted 17.11.16; peer-reviewed by A Sharma, T Poder; comments to author 22.12.16; revised version received 18.01.17; accepted 11.02.17; published 24.03.17.

Please cite as:
Lobban F, Dodd AL, Sawczuk AP, Asar O, Dagnan D, Diggle PJ, Griffiths M, Honary M, Knowles D, Long R, Morriss R, Parker R, Jones S
Assessing Feasibility and Acceptability of Web-Based Enhanced Relapse Prevention for Bipolar Disorder (ERPonline): A Randomized Controlled Trial
J Med Internet Res 2017;19(3):e85
URL: http://www.jmir.org/2017/3/e85/
doi:10.2196/jmir.7008
PMID:28341619

Original Paper

Methods for Coding Tobacco-Related Twitter Data: A Systematic Review

Brianna A Lienemann[1], PhD; Jennifer B Unger[1], PhD; Tess Boley Cruz[1], MPH, PhD; Kar-Hai Chu[1], PhD

Department of Preventive Medicine, Keck School of Medicine, University of Southern California, Los Angeles, CA, United States

Corresponding Author:
Brianna A Lienemann, PhD
Department of Preventive Medicine
Keck School of Medicine
University of Southern California
2001 N Soto St
Los Angeles, CA, 90032
United States
Phone: 1 3109241711
Fax: 1 3109241711
Email: blienema@usc.edu

Abstract

Background: As Twitter has grown in popularity to 313 million monthly active users, researchers have increasingly been using it as a data source for tobacco-related research.

Objective: The objective of this systematic review was to assess the methodological approaches of categorically coded tobacco Twitter data and make recommendations for future studies.

Methods: Data sources included PsycINFO, Web of Science, PubMed, ABI/INFORM, Communication Source, and Tobacco Regulatory Science. Searches were limited to peer-reviewed journals and conference proceedings in English from January 2006 to July 2016. The initial search identified 274 articles using a Twitter keyword and a tobacco keyword. One coder reviewed all abstracts and identified 27 articles that met the following inclusion criteria: (1) original research, (2) focused on tobacco or a tobacco product, (3) analyzed Twitter data, and (4) coded Twitter data categorically. One coder extracted data collection and coding methods.

Results: E-cigarettes were the most common type of Twitter data analyzed, followed by specific tobacco campaigns. The most prevalent data sources were Gnip and Twitter's Streaming application programming interface (API). The primary methods of coding were hand-coding and machine learning. The studies predominantly coded for relevance, sentiment, theme, user or account, and location of user.

Conclusions: Standards for data collection and coding should be developed to be able to more easily compare and replicate tobacco-related Twitter results. Additional recommendations include the following: sample Twitter's databases multiple times, make a distinction between message attitude and emotional tone for sentiment, code images and URLs, and analyze user profiles. Being relatively novel and widely used among adolescents and black and Hispanic individuals, Twitter could provide a rich source of tobacco surveillance data among vulnerable populations.

(J Med Internet Res 2017;19(3):e91) doi:10.2196/jmir.7022

KEYWORDS

tobacco; Internet; social marketing; review

Introduction

As Twitter has grown in popularity to 313 million monthly active users [1], researchers have increasingly been using it as a data source for tobacco-related research. Twitter is a microblogging platform where users have 140 characters to share thoughts, jokes, information, images, and URLs (ie, Web addresses). Twitter posts (ie, tweets) are in real time and often public, with the potential to reach a wide audience. Users can retweet or share tweets with others, which can cause tweets to spread to large numbers of users ("go viral"). Posts can be seen when users follow each other or search for specific terms (eg, #vape). Hashtags signify a topic for users to participate in the

conversation. Antitobacco campaigns may use a hashtag to start a conversation about the harms of tobacco, for example, the Truth Initiative account, @truthinitiative, promotes the use of #tobaccofreegen in the user description [2]. Similarly, the tobacco industry and independent manufacturers can use Twitter to advertise their products. Imperial Brands uses such an approach when it promotes the electronic cigarette (e-cigarette) blu through its account, @blucigs, with the hashtag #JustYouAndblu in the user description and messages to engage with Twitter users [3]. Furthermore, social media can be used to counter antitobacco campaigns. When the anti–e-cigarette campaign Still Blowing Smoke [4] was released by the State of California on television and Facebook, the pro–e-cigarette campaign Not Blowing Smoke [5], developed by pro-vaping groups, countered by taking the Twitter handles @StillBlwngSmoke [6] and @NotBlwngSmoke [7] to challenge the California campaign's messages [8].

Tobacco-related tweets can reach a relatively young and ethnically diverse audience. Pew Research Center studies have found that, in the United States, a third of teenagers use Twitter [9], while 20% of adults have accounts with the majority being younger than 50 years [10]. Twitter is especially popular among girls aged 15 to 17 years with 49% having accounts [9]. Similarly, 45% of black, 34% of Hispanic, and 31% of white adolescents use Twitter [9]. Comparably, a larger proportion of black (28%) and Hispanic (28%) than white (20%) adult *Internet users* also use Twitter [10]. Therefore, tobacco conversations on Twitter, whether pro or anti, may be particularly likely to reach these populations.

Research utilizing Twitter data is fairly novel without established standards across studies. Thus, it could be advantageous to establish what methods are being used and their strengths and weaknesses. Standards for reporting social media data are needed to be able to compare methods and results across studies [11]. This review focuses specifically on the methodology of tobacco-related studies that code Twitter data categorically by examining data collection methods, coding methods, and coding categories. It addresses the questions, "What methodologies are used to categorically code tobacco-related Twitter data" and "What recommendations can be made for future studies?"

Methods

Data Sources

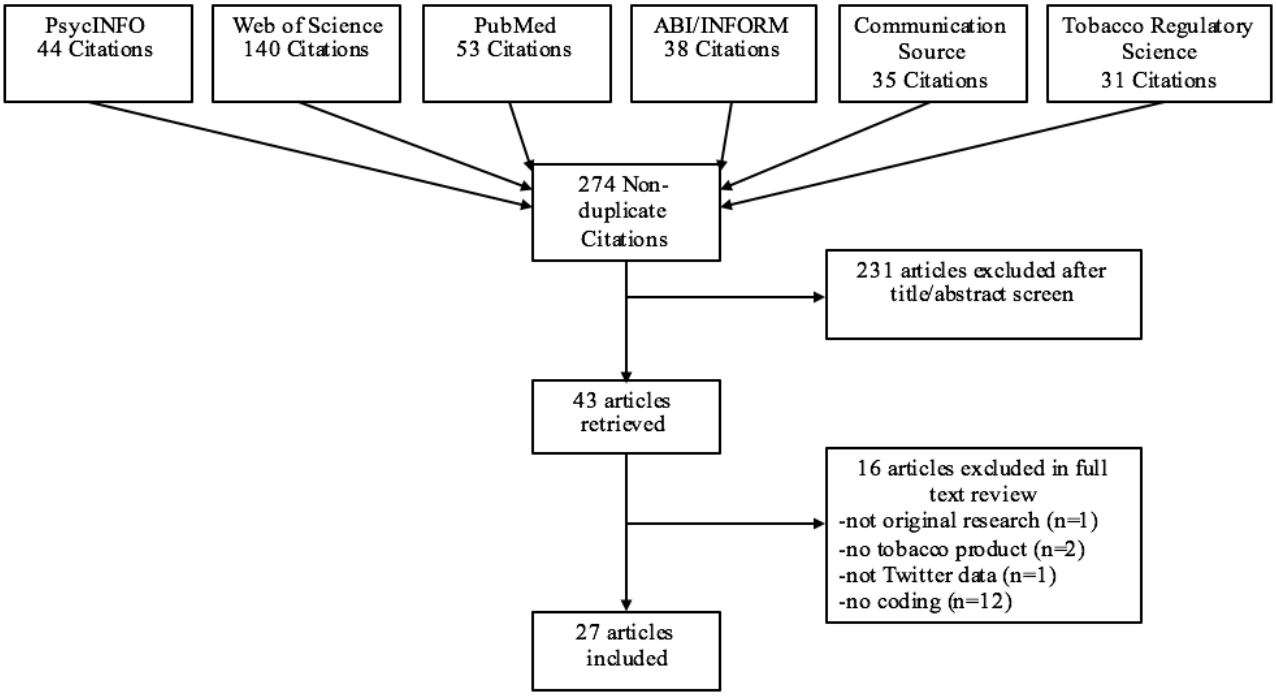
A literature search was conducted in July 2016 using the databases PsycINFO, Web of Science, PubMed, ABI/INFORM, Communication Source, and the journal Tobacco Regulatory Science. Searches included a Twitter term and a tobacco term: (Twitter OR tweet) AND (tobacco OR nicotine OR...) (Table 1). Tobacco terms were selected based on an article on noncigarette tobacco products [12] and the US Food and Drug Administration's (FDA) article, *Recognize Tobacco in Its Many Forms* [13]. Searches were limited to peer-reviewed journals in English published from January 2006 to those available in July 2016. The beginning date was selected because Twitter was launched in 2006. The initial search produced 274 nonduplicate articles (Figure 1).

Table 1. Tobacco search terms.

Search term[a]	Tobacco products covered by search term
tobacco	Tobacco, smokeless tobacco, chewing tobacco, dissolvable tobacco
nicotine	Nicotine, electronic nicotine delivery system
cig*	Cigarette, cigar, little cigar, large cigar, cigarillo, electronic cigarette, e-cigarette, e-cig
pipe	Pipe, waterpipe
bidi	Bidi
kretek	Kretek
shisha	Shisha
hookah	Hookah, e-hookah, hookah pen
narghile	Narghile
argileh	Argileh
cheroot	Cheroot
smok*	Smoke, smokeless tobacco, smoking, smoker
chew	Chew, chewing tobacco
snuff	Snuff, dry snuff, moist snuff
snus	Snus
betel quid	Betel quid
gutkha	Gutkha
zarda	Zarda
toombak	Toombak
dissolvable	Dissolvable, dissolvable tobacco
ENDS	ENDS (electronic nicotine delivery system)
vap*	Vape, vaper, vape pen, vaping, vapor

[a]Asterisk (*) represents stemmed words; for example, cig* would capture all words beginning with cig.

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) diagram of articles included in the systematic review.



Study Selection

One coder independently reviewed all titles and abstracts and selected 43 articles from the initial 274 that referenced any of the tobacco products and Twitter. Full text of the articles was then reviewed. Explicit inclusion criteria were determined a priori to reduce coder bias when selecting articles. Articles were included in the review if they met the following criteria: (1) original research in a peer-reviewed journal, (2) focused on tobacco, (3) analyzed Twitter data, and (4) coded data categorically (eg, sentiment, theme). If an article was excluded for failing to meet multiple inclusion criteria, it was counted in the higher-order criteria for exclusion. A total of 16 articles were excluded, so 27 articles were included in the review (Figure 1). We excluded 1 article because it was a narrative review rather than original research, 2 articles because they did not address a tobacco product, 1 article because it analyzed Web search results and tobacco control websites rather than Twitter data, and 12 articles because they did not categorically code tobacco-related Twitter data. For example, a study may have tracked changes in the number of tweets over time but not coded the tweets categorically. The 27 included articles ranged from 2011 to 2016. Although there is risk for bias in studies, this review considered all studies that met the inclusion criteria to evaluate the strengths and weakness of all methodological approaches within this domain. The primary focus of this review is the methodology of qualitative studies rather than the outcomes of quantitative studies. Therefore, the risk of publication bias of studies with significant results and selective reporting of significant results is minimal for this type of review compared with a meta-analysis of quantitative outcomes.

Data Extraction

One author read each article to extract methodological information on data collection (data sources, date collected, tobacco topics, search keyword selection), coding methods (type of coding method, number of coders, number of tweets coded, coded retweets, number of Twitter accounts, followed URLs, coding agreement), and coded categories. This information is presented in detail in Tables 2, 3, and 4 and is summarized in the results.

Results

Data Collection Methods

Data Sources

In total, 22 of the 27 articles discussed the sources they used for their sample of Twitter messages, while 5 articles did not. Twitter provides 3 primary sources of data: Twitter's Search application programming interface (API), Twitter's Streaming API, and Twitter's Firehose. An additional method is Twitter's REST API, which allows tracking specific users by their username. One study collected data using Twitter's REST API.

Twitter's Search API is free to use and provides a maximum of 3200 past tweets (ie, published in the past 7 days, so it is not in real time) with a limit of 180 searches every 15 minutes [14,15]. Twitter's Search API was used in 2 studies (Table 2). There are programs developed to interact between Twitter's Search API and specific analyses programs. For example, twitteR package for R [16], NCapture for NVivo [17], and Social Network Importer for NodeXL, the free add-on for Excel [18], provide access to Twitter's Search API. A small subset of the studies used these programs: twitteR package for R (n=1), NCapture (n=1), and Social Network Importer for NodeXL (n=1).

More useful to researchers is Twitter's Streaming API, which provides all tweets related to the search terms up to a limit of 1% of the Twitter database for that time period. If the selected search terms are infrequently used across Twitter (eg, the name of a local tobacco campaign), all tweets related to the search terms will be available. However, if the selected search terms are commonly used, tweets related to those terms will be available up to a limit of 1% of the current Twitter database. Twitter's Streaming API is free, publicly available data [19]. A total of 5 studies used Twitter's Streaming API (Table 2). Twitonomy is an analytics tool that accesses Twitter's Streaming API and offers both free and premium packages for a small fee [20]. One article used Twitonomy.

Providing the greatest access to data, Twitter's Firehose has real-time access to 100% of Twitter content. Twitter's Firehose formerly was handled by multiple data providers (eg, Gnip, DataSift, and Topsy). After the acquisition of Gnip in 2014, Twitter transitioned to only allowing access to Twitter's Firehose through Gnip for a fee from August 2015 onward [21]. A total of 6 articles used Gnip, making it the most common method to collect data (Table 2). There are social listening programs that buy Twitter data from data providers such as Gnip. Radian6 [22], Simply Measured [23], and Sysomos Heartbeat [24] collect data from across social networking sites, blogs, forums, and news sites for a fee. A few articles used these social listening programs to collect data: Radian6 (n=2), Simply Measured (n=1), and Sysomos Heartbeat (n=1).

Table 2. Data collection methods.

Article	Date collected	Type of tweets or accounts	Keyword selection[a,b]	Data source	Retrieval precision	Retrieval recall
[8]	March 22 to June 27, 2015	Tweets about the California Department of Public Health "Still Blowing Smoke" media campaign about the harms of e-cigarettes and the pro–e-cigarette campaign "Not Blowing Smoke"	#stillblowingsmoke, stillblowingsmoke, "still blowing smoke", stillblngsmoke, "still blng smoke", #notblowingsmoke, notblowingsmoke, "not blowing smoke", notblngsmoke, "not blng smoke", @CAPublicHealth	Gnip	97.5%	NR[c]
[25]	May 1, 2013, to May 1, 2014	E-cigarettes	vaping, vape, vaper, vapers, vapin, vaped, evape, vaporing, e-cig*, ecig*, e-pen, epen, e-juice, ejuice, e-liquid, eliquid, cloud chasing, cloudchasing, deeming AND regulation, deeming AND FDA, deemed AND FDA, deem* AND FDA	Gnip	59.23%	NR
[26]	July 1, 2008, to February 28, 2013	E-cigarettes	*55 keywords (only examples reported)*: *general e-cigarette terms* (*eg*, electronic cigarette, eCig), *specific brand names* (*eg*, blu, NJoy, green smoke), *and terms about e-cigarette use* (*eg*, vaping) *Excluded words related to tobacco or drugs (only examples reported)*: marijuana, hookah	Radian6	91%	93%
[27]	December 5, 2011, to July 17, 2012 (15-day intervals)	Tobacco	cig*, nicotine, smok*, tobacco; hookah, shisha, waterpipe, e-juice, e-liquid, vape *and* vaping	Twitter's Streaming API[d]	57.25%	95%-99%
[28]	January 1 to December 31, 2014	E-cigarettes and smoking cessation	*E-cigarette keywords*: vaping, vaper, vapes, vapers, e-cigarette, e-cig, mod, eGo, mod, RBA, RDA, vape, "vape pen", "e-hookah", "e-pipe", "e-shisha", "hookah pen", "vape pipe", #vape #vapelife #vapor #vapeporn #vapenation #vapestars #vaperazzi #vapingstyle #vaperevolution #vapeswag #vapesirens #vaperscommunity #vapepics #vapesociety #socialvapers #vapefamily #vapefreedom #vapelove #vapers #vapstagram #vapelyfe #vapeshop #vapeon #vapestrong #girlswhovape #alldayvape #adv #vapersoul #VGOD #ecig #ecigarette *Smoking cessation keywords*: quit, stop, "quit smoking", "stop smoking", "quit cigarettes", "smoke less", "smoking less", help, NRT, patch, lozenge, spray, gum, nicorette, nicotine, #quit #quitsmoking #quitsmokingcigarettes #Cessationnation *Exclusion keywords*: marijuana, weed, pot, dank, trees, green, cheeba, THC, cannabis, sativa, indica, bud, marihuana, MJ, "mary jane"	Sysomos Heartbeat	NR	NR
[29]	February 1 to April 30, 2014	Blu e-cigarettes' tweets and retweets	@blucigs	Twitter REST API	NR	NR
[30]	April 12 to May 10, 2014	Hookah or shisha	hookah, #hookah, shisha, #shisha, hooka, #hooka, sheesha, #sheesha	Simply Measured	99.56%	NR
[31]	November 1, 2011, to August 31, 2013	Hookah, cigarettes, and cigars	cigar, cigars, cigarette, cigarettes, hookah, waterpipe, water pipe, shisha, sheesha	Twitter's Streaming API	NR	NR
[32]	October 4 to November 3, 2010	Tobacco	Smoking, tobacco, cigarette, cigar, hookah, hooka	Twitter's Search API	NR	NR
[33]	May 1, 2012, to June 30, 2012	E-cigarettes	*Keywords*: e-cigarette, ecigarette, e-cig, ecig *Additional keywords AND "cig" or "cigarette"*: electronic, blu, njoy	Gnip	>99% of a random sample of 500 tweets	NR

Article	Date collected	Type of tweets or accounts	Keyword selection[a,b]	Data source	Retrieval precision	Retrieval recall
[34]	December 6, 2012, to June 20, 2013	Tobacco or cessation price promotion	*Tobacco-related*: cig(s), cigarette(s), smoking, tobacco, blu cigarette, njoy cigarette, ecig, e-cig, @blucig, e-cigarette, ecigarette, from:blucigs*, ecigs, e-cigs, ecigarettes, e-cigarettes, "green smoke", "south beach smoke", cartomizer, (atomizer OR atomizers) –perfume*, ehookah OR e-hookah, ejuice OR ejuices OR e-juice OR e-juices, eliquid OR eliquids OR e-liquid OR e-liquids, e-smoke OR e-smokes, (esmoke OR esmokes), eversmoke, "joye 510", joye510, lavatube OR lavatubes, logicecig OR logicecigs, smartsmoker, smokestik OR smokestiks, "v2 cig" OR "v2 cigs" OR v2cig OR v2cigs, vaper OR vapers OR vaping, zerocig OR zerocigs, cartomizers, Vuse, MarkTen *Price-related*: Coupon(s), Promo(s), Promotions(s), Promotional, Discount(s)(ed), Save, Code(s)	Gnip	56.94%	NR
[35]	July 2014	Slogans for the Dutch health campaign "Smoking is so outdated" (Roken kan echt niet meer)	#rokenkanechtnietmeer [#smokingissooutdated]	Twitter's Search API	NR	NR
[36]	December 2013	Little cigars	Swisher Sweets, Black & Milds	Twitonomy	67.50%	NR
[37]	September 2012 and January to May 2013	Genetic information on smoking	genetic, smoking	NR	49.1%	NR
[38]	August 2010	Smoking cessation accounts	*Searched for smoking cessation accounts using the following terms*: "quit or stop smoking" or "smoking cessation"	NR	NR	NR
[39]	January 8-15, 2014	Tweets about Chicago Department of Public Health's e-cigarette Twitter campaign	@ChiPublicHealth	twitteR package for R and NodeXL	NR	NR
[40]	January 2010 to January 2015	E-cigarettes	vape, ecig, ecigarette, vaping, ejuice, vapers, drip AND tip, dripping, eliquid AND flavor, e AND juice, e AND liquid, smoke AND free, off AND cigarettes, ex AND smoker, no AND analogs, I AND quit	NR	NR	NR
[41]	January 2012 to December 2014	E-cigarettes	e(-)cig, e(-)cigarette, electronic cigarette, *etc*	Twitter's Streaming API	81% to 90.8% for 4 groups of 500 randomly sampled automated tweets	NR
[42]	September to December 2013 and March 2015	E-cigarettes	Electronic-cigarette, e-cig, e-cigarette, e-juice, e-liquid, vape-juice, vape-liquid	Twitter's Streaming API and Twitter's Firehose	97.21%	86.63%
[43]	April 21 to October 20, 2014	Blu and V2 e-cigarettes' tweets and retweets	@blucigs, @v2cigs	NR	100%	NR

Article	Date collected	Type of tweets or accounts	Keyword selection[a,b]	Data source	Retrieval precision	Retrieval recall
[44]	July 7 to 21, 2014	Tweets about the Centers for Disease Control and Prevention's (CDC) Tips From Former Smokers campaign	#cdctips, CDC AND smoking	Social Network Importer for NodeXL	81.70%	NR
[45]	May 1, 2013, to May 1, 2014	E-cigarettes	vaping, vape, vaper, vapers, vapin, vaped, evape, vaporing, e-cig*, ecig*, e-pen, epen, e-juice, ejuice, e-liquid, eliquid, cloud chasing, cloudchasing, deeming AND regulation, deeming AND FDA, deemed AND FDA, deem* AND FDA	Gnip	59.23%	NR
[46]	March to June 2013	Tobacco control program tweets during the months that the national CDC Tips smoking cessation campaign aired	*Google search for tobacco control programs using the terms* " tobacco program " *and* " quitline ." *If the site included a link to a Twitter account, that account was included.*	Radian6	NR	NR
[47]	March 15 to June 9, 2012	Tweets about the CDC's Tips campaign	*Tobacco behavior*: cig(s), cigarette(s), nicotine, smoke(s), smoker, smoking, tobacco *Tobacco policy*: @cdcgov, @cdctobaccofree, @drfriedencdc, @fdatobacco, @smokefreegov, antitobacco, antismoking, CDC, quitline, quitnow, secondhand+smoke, smokefree, smokefree.gov, tobaccofree *Ad specific*: #cdctips, amputation, amputee, Buerger's+Disease, heart+attack, hole+neck, hole+throat, lung+cancer, stoma, stroke, throat+cancer *Engagement*: ad, commercial, campaign, PSA	Gnip	78.87%	94%
[48]	February 5-12, 2014	CVS Health-related tweets surrounding the announcement of ending tobacco sales	#cvs, #cvsquits	Twitter's Streaming API	72.38%	NR
[49]	50 most recent tweets from July 18, 2012	Smoking cessation accounts	*Searched for smoking cessation accounts using the terms*: "quit smoking" *and* "smoking cessation"	NR	NR	NR
[50]	February 23 to April 9, 2015	Exposure to secondhand e-cigarette aerosol	"secondhand vape" OR "secondhand vaping" OR "second-hand vape" OR "second-hand vaping" OR "vape smoke" OR "ecig smoke" OR "e-cig smoke" OR "e-cigarette smoke" OR "vape shs" OR "ecig shs" OR "vape secondhand smoke" OR "vape second-hand smoke" OR "esmoke" OR "e-smoke"	NCapture	NR	NR

[a]Asterisk (*) represents stemmed words; for example, cig* would capture all words beginning with cig.

[b]Words in italics were not keywords used for searches.

[c]NR: not reported.

[d]API: application programming interface.

Date Collected

Twitter data were collected across the studies from 2008 to 2015. The time span of Twitter data collected within an individual study ranged from 1 day to 5 years with a median of 14 weeks (Table 2).

Tobacco Topics

Among the 27 studies, 41% (n=11) analyzed messages related to e-cigarettes, 19% (n=5) related to other tobacco products, and 22% (n=6) about specific tobacco campaigns (Table 2). The remaining topics included smoking cessation accounts (n=2), tobacco or cessation price promotion (n=1), genetic information on smoking (n=1), and ending tobacco sales at CVS Health (n=1).

Search Keyword Selection

Kim and colleagues [11] proposed a framework of three steps to develop and validate search filters. This framework was selected because it provided a standard in which to compare

studies. Most of the articles partially fulfilled these steps within the framework. The first step is to develop a search filter. All the articles generated a list of keywords presumably based on expert knowledge of the topic or a systematic search of language related to the topic of interest. However, only 4 articles discussed the process of discarding keywords that returned a high proportion of irrelevant results and adding new keywords as new terms appeared in the returned literature [25-28]. It should be noted that for some studies this process of developing a search filter may be irrelevant because they are coding all tweets from a specific account (eg, coding tweets from the blu e-cigarette account, @blucigs) [29].

Search keyword selection was tailored to the studies' specific topics (Table 2). To search for e-cigarette Twitter data, variations on e-cigarette (eg, e-cig), vape (eg, vaping), e-liquid (eg, eliquid), and e-juice (eg, ejuice) were common. Some e-cigarette studies also included major e-cigarette companies or brands as key search terms (eg, Njoy). Studies that analyzed Twitter data on specific tobacco campaigns used a variety of tactics such as searching for variations on the campaign name (eg, still blowing smoke), the source of the campaign (eg, CDC), specific features of the campaign ads (eg, lung+cancer), tobacco products and behavior (eg, smoke), and general campaign terms (eg, PSA). Some studies also used a combination of searching for terms with and without hashtags (eg, hookah, #hookah). In 2 studies, marijuana terms (eg, weed) were used as exclusion keywords [26,28].

The second step of the framework is to apply the search filter and split data into retrieved and unretrieved sets. The third step is to assess the search filter on its ability to distinguish between relevant and irrelevant messages [11]. Precision refers to how much retrieved data are relevant, whereas recall refers to how much relevant data are retrieved. Recall is similar to measures of sensitivity. Precision is much less difficult to accurately estimate than recall because recall requires conclusions to be drawn about the tweets that were not retrieved. A precision score of 100% means that all retrieved data were relevant, while a recall score of 100% means that all relevant data were retrieved. However, precision and recall are inversely related. As a search filter expands to collect more data, the proportion of relevant data decreases. Publications can present a precision-recall curve to show the extent of this trade-off for their search filter. A good search filter will maintain a relatively high level of precision as recall increases. A total of 15 articles reported precision, which ranged from 49.1% to 100% (Table 2). A total of 4 articles reported recall, which ranged from 86.6% to 99% (Table 2).

Coding Methods

Coding methods for the studies included hand-coding, machine learning, or a combination of the two. Hand-coding involves one or more human coders categorizing data. When 2 or more coders independently code data, a coding agreement score (eg, kappa) between the coders can be calculated. In contrast, machine learning uses an algorithm for a computer to learn how to code data. However, human-coding is used for an initial subset of data to help refine the algorithm to improve its accuracy. Coding categories may be determined a priori based on prior research or they may be developed inductively through the process of coding. Studies used hand-coding only (n=17), machine learning + hand-coding (n=8), and machine learning only (n=2; Table 3). Of the studies using hand-coding, data were coded by the reviewed studies' researchers in 16 studies, while 1 study used crowdsourcing (ie, many Web workers) [30]. All the studies that used machine learning also used initial hand-coding for a subset of the data, except for 2 studies that used topic modeling [31,32]. Topic modeling produces thematically related word clusters from the text [31].

The data collected in the hand-coded studies ranged from a collection period of 1 day to 1 year, while the machine learning studies ranged from 1 month to 5 years of Twitter data. Across the articles that used hand-coding, the number of coders per tweet ranged from 1 to 6. Coding agreement was reported by 20 articles: coding agreement percentage (n=5; 72% to 95.7%), kappa scores (n=13; kappa=.64 to 1.00), Cronbach alpha (n=1; alpha=.61 to 1.00), and both coding agreement percentage and kappa scores (n=1). Kappa values can be interpreted as poor (<.20), fair (.21-.40), moderate (.41-.60), good (.61-.80), and very good (.81-1.00) [51]. Cronbach alpha can be interpreted as unacceptable (<.50), poor (.50-.59), questionable (.60-.69), acceptable (.70-.79), good, (.80-.89), and excellent (>.90) [52]. However, it is important to note that these thresholds are not derived statistically but instead rely on intuitive judgments. The number of tweets coded per study ranged from 171 to 17,098 for hand-coding and from 7362 to 1,669,123 for machine learning. A total of 14 articles included retweets in their total number of tweets. A total of 15 articles reported the number of unique Twitter accounts, which ranged from 2 to 3804 for hand-coding studies and from 23,700 to 166,857 for machine learning studies (see Table 3).

URLs in tweets can provide information that changes the context or meaning of a tweet. Following URLs to their respective webpages can be time-consuming, but it can increase coding accuracy. Machine learning algorithms can analyze the text within URLs but may require human coders to follow them to their respective webpages. A total of 15 articles reported whether they followed URLs (followed: n=10, did not follow: n=5; Table 3). One article provided the most common URLs [26]. The studies tended to show that advertising or commercial tweets were significantly more likely to contain URLs than other types of tweets [25,33,34].

Table 3. Coding methods.

Article	Coding method	No. of coders	No. of tweets coded	Coded retweets	No. of Twitter accounts	Followed URLs	Coding agreement
[8]	Hand-coded by researchers	1: all tweets; 2: subsample 300 tweets	2248: relevance; 2192: content	Yes	NR[a]	No	91%: sentiment; 72%: theme
[25]	Hand-coded by researchers	6: for a subset of 250 tweets; NR for total	17,098: relevance; 10,128: content	Yes, if additional context	NR	Yes	κ=.64 to .70
[26]	Machine learning with initial hand-coding; Python Scikit-Learn	NR	1,669,123	Yes	NR	Yes	NR
[27]	Machine learning and hand-coding; naïve Bayes, k-nearest neighbors, and support vector machines	2: pilot of 1000; 2: random subset of 150; 2: all 7362	7362: relevance; 4215: content	Retweeted posts were only included once	NR	NR	κ>.70 for the random subset of 150
[28]	Hand-coded by researchers	1: all tweets; 2: for 10% subsample	300: complete sample; 300: industry-free sample; 481 of 600: content (duplicates between samples removed)	Yes	148: complete sample; 215: industry-free sample	Yes	κ=.74
[29]	Hand-coded by researchers	2	NR	Yes	Approximately 3400	NR	NR
[30]	Crowdsourcing with initial hand-coding	3	5000: relevance; 4978: content	NR	3804	NR	κ=.66 to .85 among a subset coded by researchers
[31]	Topic modeling with machine learning; MALLET, a command-line implementation of latent Dirichlet allocation (LDA)	NR	319,315: total; 95,738: hookah; 22,513: cigar; 201,064: cigarette	NR	NR	NR	NR
[32]	Topic modeling (LDA) with machine learning	NR	4962	NR	NR	NR	NR
[33]	Machine learning and hand-coding; DiscoverText	2: for a subset of 500 for relevance, 4500 for commercial versus organic, 7500 for cessation	73,672	Yes	23,700	Yes, hand-coded tweets with URLs	κ=.87 to .93
[34]	Hand-coded by researchers	1: all; 2: for subsets of 100 tweets	5000: relevance; 2847: content	NR	NR	Yes	κ=.64 to 1.00
[35]	Hand-coded by researchers	1: all tweets; 3: subsample	133	No	NR	NR	alpha = .61 to 1.00

Article	Coding method	No. of coders	No. of tweets coded	Coded retweets	No. of Twitter accounts	Followed URLs	Coding agreement
[36]	Hand-coded by researchers	3	3935: relevance, foreign language, retweets; 2656 sampled for 288 original tweets for coding	No	346	Yes	κ=.64 to .91
[37]	Hand-coded by researchers; wordcloud R package	NR	171: relevance; 84: content	NR	84	NR	NR
[38]	Hand-coded by researchers	1: all tweets; 2: for 20% of tweets	143,287: identified; 4753: coded for clinical practice guidelines for treating tobacco dependence	NR	153	Yes	>90%
[39]	Hand-coded by researchers	2	684	Yes	306	Yes	NR
[40]	Machine learning and hand-coding; naïve Bayes, LIBLINEAR, Bayesian logistic regression, random forests; keyword comparisons	1: all tweets; 2: subsample of 2000	13,146	NR	2147	No, removed URLs	κ=.87 for subsample
[41]	Machine learning and hand-coding; human detection algorithm; Hedonometrics; key phrasal pattern matching	2: for all tweets from 500 automated accounts and 500 organic accounts as classified by the algorithm; 2: for 4 groups of 500 randomly sampled tweets to gauge accuracy of subcategorical tweet topics	850,000	Yes	131,622: automated accounts; 134717: organic accounts: 188,182: not classified accounts (ie, accounts with <25 tweets)	No, but the algorithm used the count of URLs to distinguish automated accounts from organic accounts; also used keywords in the URLs for the algorithm to determine subcategories of automated accounts	94.6% true-positive rate, 12.9% false-positive rate for the machines on the tweets from the 1000 accounts also coded with human-coding
[42]	Machine learning with initial hand-coding; Python Scikit-Learn; topic modeling with MALLET	2: for a subset of 1000 profiles	224,000 in 2013 sample; 349,401 in 2015 sample	Yes	34,000 in 2013 sample; 100,000 in 2015 sample	No; metadata on the presence of URL links	κ=.88
[43]	Hand-coded by researchers and MySQL pattern matcher	NR	1180	Yes	2: Blu and V2; 537: users retweeting Blu and V2	NR	NR

Article	Coding method	No. of coders	No. of tweets coded	Coded retweets	No. of Twitter accounts	Followed URLs	Coding agreement
[44]	Hand-coded by researchers	1: all tweets; 2: for 20% of tweets (n=358)	2191: relevance; 1790: content	Yes	NR (>21)	NR	κ=.95 for 20% subsample
[45]	Machine learning with initial hand-coding; naïve Bayes classifier, k-nearest neighbors, support vector machines	6: for a subset of 250 tweets; NR for total	17,098: relevance; 10,128: content	Yes, if additional context	NR	NR	κ=.64 to .70
[46]	Hand-coded by researchers	3	1776	No	16	Yes	For 5% of data, 95.7%; κ=.72
[47]	Machine learning with initial hand-coding; naïve Bayes classifier	2: subset of 450 tweets for relevance; 2: subset of 350 tweets for content	245,319: relevance; 193,491: content	NR	166,857	NR; metadata on the presence of URL links	κ=.93
[48]	Hand-coded by researchers	1: all tweets; 2: for 1% of tweets	8645: relevance; 6257: content	Yes	NR	Yes	90% for a 1% sample of tweets
[49]	Hand-coded by researchers	2	900, with 50 tweets per account	Yes	18	NR	84%
[50]	Hand-coded by researchers	2	1519	No	1321	Yes	κ=.84

[a]NR: not reported.

Coded Categories

All the studies developed categories for content. These content areas included one or more of the following: sentiment, theme, location of use, user description, profile photo, or location of user (Table 4).

Sentiment

A total of 9 articles coded for sentiment (Table 4). One article made a distinction between coding for sentiment (ie, emotional tone or affective content: positive, negative, or neutral) and message attitude (ie, pro, con, neutral or do not know) [28]. Two articles coded for sentiment in terms of emotional tone. In 6 articles, sentiment was described in terms of being supportive or against tobacco, tobacco users, or decisions regarding tobacco, which suggests an assessment of message attitude. Furthermore, 1 article assessed valence, but it was not clear whether positive or negative valence suggested an attitude or emotional tone [35].

Topic or Theme

A total of 21 studies coded for topic or theme (Table 4). The most common themes included the following: advertisement, marketing, industry or commercial (n=12 articles); health, safety, harms (n=9); use (n=8); policy, government, regulation, activism (n=7); e-cigarettes for smoking cessation (n=7); flavors (n=7); personal opinion or communication (n=6); risky behaviors or other substances (n=6); cessation (n=5); information (n=5), and craving or need (n=5). One study coded for location of use with 20 categories (eg, school, work) [26]. Finally, 2 studies used topic modeling to explore tobacco content [31,32].

User or Account

A total of 10 studies coded for user description from data found in the user profile, including type of account, age, location, and other characteristics (Table 4). The most common types of user categories coded were personal accounts (n=7 articles), industry accounts (n=5), news (n=5), unclassified (n=5), and bots, automatic, or fake (n=4; ie, automated computer program). None of the articles reported the percentage of tweets that had accounts without user profile information. However, across the articles with a category for tweets with an unclassifiable user description, 0.2%-38% of tweets were unclassified. Across the studies that coded for bots, the percentages of tweets varied drastically from 6.9% to 80.7%.

One study coded Twitter profile photos with 4 categories: gender, age, race, and single person versus multiple people [36]. Visual cues (eg, skin color, background themes, facial features) in the profile pictures were used for coding. Coding for age based on available visual cues in the profile photos tended to be difficult, so coding for age was simplified to 3 broad groups: young, middle age, and older adult. There were few middle-aged and older adults represented in the sample, so a dichotomous variable of young or not young was created [36].

A total of 4 studies coded for the location of the user with one each coding for state (California vs other) [8], country [37], continent [38], and city, state, and country (United States vs other) [39]. Location was identified for 51% to 63% of Twitter profiles. Most accounts that listed a location were from the United States and North America.

Table 4. Coded categories.

Category type	Category	Number of articles and percent of total[a] n (%)	Articles
Relevance	Relevant versus nonrelevant	16 (59)	[8,25,27,30,33,34,36,37,41-45,47-49]
Sentiment		9 (33)	[8,25,27,28,35,41,45,48,50]
	Positive or negative (ie, supportive or against)	6 (22)	[8,25,27,45,48,50]
	Positive or negative (ie, emotional tone)	2 (7)	[28,41]
	Positive or negative valence	1 (4)	[35]
	Neutral or unknown	6 (22)	[8,27,28,45,48,50]
Message attitude	Pro or con	1 (4)	[28]
Type of utterance	Comparison versus attribution versus metonymy	1 (4)	[35]
Topics, themes, or genres		21 (78)	[8,25-28,30,33,34,36-43,45-47,49,50]
	Joke or humorous	3 (11)	[27,28,36]
	Song or music	2 (7)	[30,36]
	Profanity	1 (4)	[36]
	Social relationships	2 (7)	[27,50]
	Sex or romance	1 (4)	[30]
	Image or stereotype	1 (4)	[27]
	Risky behaviors or other substances	6 (22)	[25,27,28,30,36,45]
	Illicit substance use in e-cigarettes	2 (7)	[25,45]
	Preference for another substance	1 (4)	[30]
	Affiliation and preference	1 (4)	[36]
	Flavors	7 (26)	[25,36,39,41-43,45]
	Pleasure	1 (4)	[27]
	Tastes good	1 (4)	[28]
	Craving, desire, and need	5 (19)	[25,27,28,36,45]
	Addiction	1 (4)	[37]
	Type of tobacco product	4 (15)	[27,30,34,38]
	Type of tobacco product brand	1 (4)	[26]
	E-cigarettes' smoke-free aspect	1 (4)	[42]
	Health, safety, harms	9 (33)	[8,25,27,30,33,37,39,45,50]
	Downplayed or refuted harms, harm reduction	2 (7)	[27,42]
	E-cigarettes for smoking cessation	7 (26)	[25,28,33,40-42,45]
	Cessation	5 (19)	[27,30,37,41,46]
	Cessation product	2 (7)	[34,38]
	Socioemotional support tweets regarding quitting smoking	1 (4)	[49]
	Encouraging or engaging tweets regarding quitting smoking	1 (4)	[49]
	Clinical practice guidelines for treating tobacco dependence	1 (4)	[38]
	Demonstration	1 (4)	[36]
	Use	8 (30)	[25,27,28,30,36,37,40,45]
	Use: general	2 (7)	[36,40]
	First-person use or intent	5 (19)	[25,27,28,30,45]
	Second- or third-person experience	4 (15)	[25,27,28,45]

Category type	Category	Number of articles and percent of total[a] n (%)	Articles
	Starting use or smoking initiation	3 (11)	[27,28,37]
	Recent use	1 (4)	[30]
	Underage use	3 (11)	[25,27,45]
	Parental use	2 (7)	[25,45]
	Does not use or does not want to use	1 (4)	[30]
	Secondhand smoke	1 (4)	[46]
	Rejection and prevention	1 (4)	[36]
	Disgust, unattractive, or uncool	2 (7)	[27,30]
	Policy, government, regulation, activism, politics	7 (26)	[8,25,27,28,39,45,46]
	Normalization versus discouragement	1 (4)	[30]
	Getting others started or advocating use	1 (4)	[28]
	Attempt to engage other Twitter users	1 (4)	[28]
	Fear appeals	1 (4)	[47]
	Lies or propaganda	2 (7)	[8,39]
	Advertisement, promotion, marketing, industry, commercial	12 (44)	[8,25-28,30,33,36,38,41,45,50]
	Offering advice	1 (4)	[28]
	Personal opinion or communication	6 (22)	[25,27,28,33,38,45]
	News or update	4 (15)	[25,27,28,45]
	Information	5 (19)	[25,27,28,45,49]
	Science or scientific publication	2 (7)	[37,39]
	Cultural reference	1 (4)	[27]
	Issue salience	1 (4)	[39]
	Commodity	1 (4)	[27]
	Connoisseurship	1 (4)	[27]
	Cheaper than smoking	1 (4)	[28]
	Money	1 (4)	[8]
	Price promotion, discount, coupon	4 (15)	[26,33,34,41]
	Backgrounded	1 (4)	[27]
	Other or undetermined	2 (7)	[36,50]
Domains smoking was compared with for campaign slogans	Personal features; hobby or hype; person or group; social norm; big event; technology and innovation; sex or relation; eating, drinking, and stimulants; school; transport; and campaign	1 (4)	[35]
Links (URLs)	Most common links	1 (4)	[26]
Location of use		1 (4)	[26]
	Class	1 (4)	[26]
	House, room, bed	1 (4)	[26]
	School	1 (4)	[26]
	Public	1 (4)	[26]
	Bathroom	1 (4)	[26]
	Work	1 (4)	[26]
	In front of someone	1 (4)	[26]

Category type	Category	Number of articles and percent of total[a] n (%)	Articles
	Car	1 (4)	[26]
	Restaurant	1 (4)	[26]
	Movie theater	1 (4)	[26]
	Airplanes or airport	1 (4)	[26]
	Store	1 (4)	[26]
	Bars or clubs	1 (4)	[26]
	Dormitory	1 (4)	[26]
	Library	1 (4)	[26]
	Mall	1 (4)	[26]
	Bowling alley	1 (4)	[26]
	Café or coffee shop	1 (4)	[26]
	Hospital	1 (4)	[26]
	Locker room	1 (4)	[26]
Topic modeling		2 (7)	[31,32]
	Hookah topic 1: social locations, leisure time, and positive affect	1 (4)	[31]
	Hookah topic 2: fun, leisure time, and sociability	1 (4)	[31]
	Cigarette topic 1: death and unpleasant smell	1 (4)	[31]
	Cigar topic 1: positive affect and enjoyment	1 (4)	[31]
	Cigar topic 2: luxury alcohol products	1 (4)	[31]
	Tobacco topic 1: tobacco use and substance use	1 (4)	[32]
	Tobacco topic 2: addiction recovery	1 (4)	[32]
	Tobacco topic 3: addiction recovery and tobacco promotion by clubs or bars	1 (4)	[32]
	Tobacco topic 4: tobacco promotion by bars or clubs and marijuana use	1 (4)	[32]
	Tobacco topic 5: antismoking and addiction recovery	1 (4)	[32]
User or account		10 (37)	[8,25,28,29,34,37,41,44,45,49]
	Government	3 (11)	[25,44,45]
	Foundations or nonprofit organizations	4 (15)	[25,44,45,49]
	Public health and health care	1 (4)	[28]
	Researcher or research center	2 (7)	[29,37]
	News	5 (19)	[25,28,37,44,45]
	Reputable news source	2 (7)	[25,45]
	Press, media, or news	3 (11)	[28,37,44]
	Medical news source	1 (4)	[37]
	Personal accounts	7 (26)	[8,25,28,29,37,44,45]
	Personal accounts, everyday people, individuals	6 (22)	[8,25,28,37,44,45]
	Personal accounts with industry ties	1 (4)	[28]
	Person: supporter	1 (4)	[29]
	Person: basic profile (no mention of e-cigarettes)	1 (4)	[29]
	Celebrity, public figures	3 (11)	[25,28,45]

Category type	Category	Number of articles and percent of total[a] n (%)	Articles
	Organic (human)	1 (4)	[41]
	E-cigarette community movement	2 (7)	[25,45]
	Industry	5 (19)	[25,28,29,34,45]
	Industry: retailer or manufacturer	2 (7)	[28,29]
	Retailer or vendor	3 (11)	[25,34,45]
	Tobacco company	2 (7)	[25,45]
	Industry: other (eg, vaping magazine, Web marketer)	1 (4)	[29]
	For-profit organization	1 (4)	[44]
	Entity: general (eg, company, store, advocacy group)	1 (4)	[8]
	Nonperson (eg, musical band)	1 (4)	[29]
	Bots, automatic, fake	4 (15)	[25,28,38,41,45]
	Unclassified or other	5 (19)	[8,29,37,44,49]
Profile photo		1 (4)	[36]
	Single person versus multiple people	1 (4)	[36]
	Gender (male, female, mixed group)	1 (4)	[36]
	Age (babies or children, high school or college, adult)	1 (4)	[36]
	Race (African American, white, Hispanic, Asian, undetermined)	1 (4)	[36]
Location of user		4 (15)	[8,37-39]
	City, state, and country	1 (4)	[39]
	State	1 (4)	[8]
	Country	1 (4)	[37]
	Continent	1 (4)	[38]

[a]Percentages are rounded to the nearest whole percent.

Discussion

Overview

Studies analyzing tobacco-related Twitter data have grown in number in recent years. Although we searched for articles published from 2006 to 2016, articles meeting inclusion criteria for this review were published from 2011 to 2016, with 85% (23/27) of the publications occurring in 2014-2016. Widely used among adolescents and black and Hispanic individuals, Twitter could provide a rich source of tobacco-related data among these groups. One of the benefits of Twitter research is the ability to focus on emerging issues and products that are not yet addressed in surveillance or epidemiological research. Collecting tobacco data in real time via Twitter could be a useful tool for tobacco surveillance, which could help inform tobacco control policies and social media campaigns.

Data Collection Methods

Twitter data can change rapidly as they are being posted daily [37]. If a study only collects data from one point in time, it may not be reflective of data at any other point in time. Twitter studies should consider collecting data at multiple time points or over longer periods of time to decrease the likelihood that results are idiosyncratic to that point in time. However, there may be some research questions that only require sampling one time or sampling directly before and after an event to gauge short-term responses.

The primary sources of data were Gnip and Twitter's Streaming API, which offer different strengths and weaknesses. If a study wants access to all tobacco-related tweets, then Gnip may be more effective. However, if a study is interested in tweets about a specific tobacco campaign or has a constrained budget, then Twitter's Streaming API may be a better data source. Alternatively, if a study is focusing on multiple social media sites (eg, Twitter, Facebook), then using Radian6, Simply Measured, or Sysomos Heartbeat may be appealing because of their cross-platform analysis.

The results of this review suggest that there are some gaps in the types of tobacco products studied by Twitter analyses. E-cigarettes were the product that was addressed by most studies, while none of the studies focused on smokeless tobacco, snus, bidis, or kreteks. The focus on e-cigarettes over other products could be due in part to their recent rise in popularity and recent debates about policies. From 2011 to 2014, e-cigarette use among high school students in the United States significantly increased from 1.5% to 13.4% [53]. Similarly, there was a

significant increase in hookah use from 4.1% to 9.4%. However, there were significant decreases for cigarette (15.8% to 9.2%), snus (2.9% to 1.9%), cigar (11.6% to 8.2%), pipe (4.0% to 1.5%), and bidi (2.0% to 0.9%) use [53]. It could be useful to conduct studies comparing tweets about cigarettes, e-cigarettes, and hookah among adolescents to help understand the changing rates of use and Twitter postings that discuss issues related to use. The rates of e-cigarette, hookah, and cigar use among high school students may be affected by the FDA regulations of these products that went into effect on August 8, 2016. One of the provisions of these regulations is that it will be illegal to sell e-cigarettes, cigars, and hookah tobacco to persons younger than 18 years [54]. Researchers may want to analyze e-cigarette, cigar, and hookah tweets before and after August 8, 2016, to gain real-time insight into adolescents' reactions to the new regulations.

Future Twitter studies could benefit from a standard of reporting data collection methods. Only 4 of the articles in this review reported such rigorous methods of selecting their search keywords as suggested by Kim and colleagues' [11] search filter framework, while 15 articles reported retrieval precision and 4 reported retrieval recall. The terminology around newer products such as e-cigarettes is growing, so it may be difficult to capture all relevant Twitter conversations with one's keywords [25]. Future studies will need to continue to refine and expand search keywords.

Data Coding

The methods of coding were hand-coding, machine learning, or a combination of the two. Machine learning can code larger quantities of data at a quicker rate than hand-coding, but human coders may have greater discretion at coding for the complexities and subtlety of language such as humor, irony, or sarcasm. For example, algorithms developed to detect irony only retrieved 54%-57% of tweets coded as irony by multiple independent human coders [55]. Hand-coding can be subject to bias, but creating coding schemes based on prior literature and working to achieve acceptable levels of interrater reliability can help attenuate individual bias. Hand-coding allows researchers to follow URLs, which can change the meaning of the tweet. Viewing the webpage may provide additional information that may not be discernable from the URL. Studies that require determining subtle differences in context may be better suited to hand-coding a small sample of Twitter data, while studies that rely less on context could code large samples with machine learning.

Coded Categories

The sentiment of tweets could help evaluate whether the responses to pro- and antitobacco efforts are positive or negative as a way of understanding social norms about these products. Clarity and comparability across studies could be improved if a distinction is made between attitude and emotion when coding for sentiment. For example, a tweet could be pro-vaping but have a negative emotional tone or it could be anti-vaping but have a positive tone. Only 1 article clarified the meaning by making an explicit distinction between coding for emotional sentiment and message attitude [28]. This is reflective of coding for stance (in favor of, against, or neutral) versus sentiment (emotional tone) [56].

The three most common themes used for coding were advertisements or marketing; health, safety, harms; and use. Surveillance of these themes could be beneficial to understand whether tobacco advertisements are being circulated on Twitter with the potential of reaching underage individuals, whether the content is making unproven claims about the health and safety of their products, or promoting the use of their products to vulnerable populations such as youth or ethnic minorities.

The most common user account descriptions that the studies coded for were personal accounts, industry, news, unclassified, and bots. Determining the user description and demographic information for the accounts that tweet about tobacco could help determine whether tobacco companies, pro-vaping advocacy groups, or antitobacco efforts are circulating with a potential to reach certain groups (eg, adolescents). Researchers could also monitor how news organizations are presenting tobacco-related information to the public.

It may be difficult to determine the demographics of the person tweeting or of the audience exposed to the tweets, which could be especially problematic when studies want to focus on vulnerable populations (eg, adolescents). A Twitter account could be run by an individual, multiple people (eg, vape shop employees), or bots. To reduce bias, bots should be identified and the tweets from these accounts removed from analysis or identified as tweets originating from automated accounts [57]. Only 4 studies coded for bots with a range from 6.9% to 80.7% of tweets classified as bots. Even if an account is run by an individual, a Twitter profile provides little information. It may or may not include a photograph, profile description, location, website, and birthday. It does not include gender or ethnicity and race, so this information needs to be estimated. Although none of the studies reported the percentage of tweets with accounts missing profile information, 5 studies coded for an unclassified category with 0.2%-38% of tweets being unable to be classified based on the account profile. If profile information is included, it could be used in combination with natural language processing to infer information about the individual from his or her tweets [29]. For example, algorithms have estimated gender with a 75.5% accuracy based on tweets and a 92.0% accuracy based on tweets, screen name, full name, and profile description [58]. A study that combined analysis of text and image processing predicted gender with an accuracy of 85.1% [59]. Additionally, algorithms based on tweets were capable of predicting the exact age of the user within a margin of 4 years, while accuracy for age categories were 93.0% for <20 years, 67.4% for 20-40 years, and 81.6 for >40 years [60].

Recommendations

It is recommended that tobacco Twitter studies adopt methodological standards of reporting and data quality assessment. Important information to consider reporting include data sources (eg, Gnip, Twitter's Streaming API), the date range of tweets collected, the number of tweets coded, whether retweets were coded, whether coders followed URLs, whether images were coded, the categories coded, the decision criteria for each category, the number of unique Twitter accounts, and

the types of Twitter accounts. Studies that use hand-coding should also consider reporting the number of independent coders, the number of tweets coded across coders, and their coding agreement, while it is important for machine learning studies to detail the development and refinement of their algorithms. Providing this information is likely to increase comparability across studies and the ability to replicate results.

Depending on research goals, studies may want to sample Twitter's databases multiple times, especially if they collect from Twitter's Search API or Streaming API, which put limitations on the amount of data that can be collected per sample. Sampling at multiple times is also important for studies that have access to Twitter's Firehose through Gnip, considering that tweets are in real time, which could lead to fluctuations based on real-world events (eg, the FDA's new regulations for e-cigarettes). Following messages over time could also help establish trends in the content of posts. However, some studies may be interested in short-term reactions to an event, which requires different sampling methods than following trends. For example, if a study is interested in short-term reactions to the FDA's new regulations for e-cigarettes, then a single sample before and after the implementation of the new regulations could be sufficient.

When coding for sentiment, researchers could improve clarity by making a distinction between whether they are coding for message attitude or for emotional tone [28]. A clear distinction between the two could improve comparability of sentiment ratings across studies.

If relevant to the research questions, Twitter studies may want to code images and URLs. This added step can be time-consuming, but doing so could change the context or meaning of a tweet. For example, an image could help determine if a tweet should be coded as humor or sarcasm, while following a URL could help determine if the tweet is an advertisement. Failing to code images and URLs could result in missing significant content that could affect coding accuracy and skew results.

Analyzing Twitter user profiles could provide context for tweets. The same pro–e-cigarette tweet could hold very different significance for a study's results if it is from a vape shop versus an adolescent. However, the limitations of analyzing user profiles should be recognized and steps taken to improve accuracy. It should be noted that user profiles may be misleading (eg, a tobacco industry representative posing as an unaffiliated citizen) or profiles of bots. Bots should be identified and potentially removed to reduce bias and improve the quality of data [57]. We cannot assume that the analysis is describing individuals but must instead consider the poster to be an "account" rather than a person.

A limitation of Twitter data is that it does not provide much information on the effects of tweets on behavior. For example, tweeting about tobacco use does not necessarily mean that the person tweeting uses tobacco [40]. Additionally, an individual may tweet about a quit smoking campaign with positive sentiment, but that does not mean that the campaign has influenced his or her smoking behavior. Follow-up studies with the individuals tweeting about the campaign would need to be conducted. None of the studies in the review included follow-up survey studies with individuals from their Twitter sample. Although recruitment of individuals through Twitter may come with its own set of obstacles, this could be an avenue for future Twitter research. Following specific individuals over time could allow for the analysis of changes in message content.

Review Limitations

There are several limitations of this systematic review. First, the results are limited to the databases and search keywords selected, which could have resulted in incomplete retrieval of identified research. Second, this review is limited to its inclusion criteria and the decision rules of the single, independent coder who selected the articles and extracted the data to be included in the review. Individual bias was limited by explicit inclusion, exclusion, and data extraction criteria. However, some studies that were included or excluded for this review may have varied given different inclusion criteria or a different coder. Finally, this review is limited to methodology of categorically coded tobacco Twitter data. Different methodological results and recommendations may have been made if the topic of Twitter data had been different (eg, marijuana) or if the review had focused on different outcomes (eg, popularity of tweets or diffusion of tweets).

Conclusions

Categorically coded Twitter research can be used for certain insights that other survey research does not provide: emerging issues, popular content in real time, changes over time, how tobacco companies and pro-vaping advocacy groups use social media to increase message exposure in the population (eg, youth who might otherwise be protected from tobacco marketing), how tobacco control policies and campaigns can most effectively use social media, arguments by groups that may be incorporated into media message design, and quick reactions to antitobacco media campaigns and regulations. There are several approaches that researchers are taking to this end, each having its own set of strengths and weaknesses. Standards for data collection and coding should be developed to more easily compare and replicate tobacco-related Twitter results. Additional recommendations, dependent on one's research goals, include the following: sample Twitter's databases multiple times, make a distinction between message attitude and emotional tone for sentiment, code images and URLs, analyze user profiles, and identify and remove bots.

Acknowledgments

This work was supported by the National Cancer Institute grant number T32CA009492-29 and the USC (University of Southern California) Tobacco Center of Regulatory Science for Vulnerable Populations grant number 3P50CA180905-04S1.

Authors' Contributions

BAL planned the review, conducted the literature search, synthesized the results, and wrote the manuscript. JBU, TBC, and KC assisted in the conceptualization of the review and edited the manuscript.

Conflicts of Interest

None declared.

References

1. Twitter. 2016 Jun 30. Twitter Usage/Company Facts URL: https://about.twitter.com/company [accessed 2016-11-17] [WebCite Cache ID 6m7ZjJlHW]
2. Twitter. 2016. Truth Initiative Twitter account URL: https://twitter.com/truthinitiative [accessed 2016-11-17] [WebCite Cache ID 6m7a6uQ5G]
3. Twitter. 2016. Blu Twitter account URL: https://twitter.com/blucigs [accessed 2016-11-17] [WebCite Cache ID 6m7aH0aG7]
4. Still Blowing Smoke. 2016. URL: http://stillblowingsmoke.org/ [accessed 2016-11-17] [WebCite Cache ID 6m7aTK5La]
5. Not Blowing Smoke. 2016. URL: http://notblowingsmoke.org/ [accessed 2016-11-17] [WebCite Cache ID 6m7bVdDlq]
6. Twitter. 2016. Still Blowing Smoke account URL: https://twitter.com/stillblwngsmoke [accessed 2016-11-17] [WebCite Cache ID 6m7bf3MPt]
7. Twitter. 2016. Not Blowing Smoke account URL: https://twitter.com/notblwngsmoke [accessed 2016-11-17] [WebCite Cache ID 6m7bpA286]
8. Allem J, Escobedo P, Chu K, Soto DW, Cruz TB, Unger JB. Campaigns and counter campaigns: reactions on Twitter to e-cigarette education. Tob Control 2016 Mar 8:1-4. [doi: 10.1136/tobaccocontrol-2015-052757] [Medline: 26956467]
9. Lenhart A. Pew Research Center. 2015 Apr 19. Teen, Social Media and Technology Overview 2015 URL: http://www.pewinternet.org/2015/04/09/teens-social-media-technology-2015 [accessed 2016-11-18] [WebCite Cache ID 6m7bxa9wU]
10. Duggan M. Pew Research Center. 2015 Aug 19. Mobile Messaging and Social Media - 2015 URL: http://www.pewinternet.org/2015/08/19/mobile-messaging-and-social-media-2015/ [accessed 2016-11-18] [WebCite Cache ID 6m7cE9B8E]
11. Kim Y, Huang J, Emery S. Garbage in, Garbage Out: Data Collection, Quality Assessment and Reporting Standards for Social Media Data Use in Health Research, Infodemiology and Digital Disease Detection. J Med Internet Res 2016 Feb 26;18(2):e41 [FREE Full text] [doi: 10.2196/jmir.4738] [Medline: 26920122]
12. O'Connor RJ. Non-cigarette tobacco products: what have we learnt and where are we headed? Tob Control 2012 Mar;21(2):181-190 [FREE Full text] [doi: 10.1136/tobaccocontrol-2011-050281] [Medline: 22345243]
13. US Food and Drug Administration. 2016 May 05. Recognize tobacco in its many forms URL: http://www.fda.gov/ForConsumers/ConsumerUpdates/ucm392735.htm [accessed 2016-11-17] [WebCite Cache ID 6m7cRBDCk]
14. Twitter. 2016. Public API: GET statuses/user_timeline URL: https://dev.twitter.com/rest/reference/get/statuses/user_timeline [accessed 2016-11-17] [WebCite Cache ID 6m7ccAB9K]
15. Twitter. 2016. Public API: The Search API URL: https://dev.twitter.com/rest/public/search [accessed 2016-11-17] [WebCite Cache ID 6m7dVm3hE]
16. R-project. 2015 Jul 29. twitteR: R Based Twitter Client URL: https://cran.r-project.org/web/packages/twitteR/index.html [accessed 2016-11-16] [WebCite Cache ID 6m7dfGPAu]
17. QSR International. 2015 Jul 07. What is NCapture? URL: http://www.qsrinternational.com/support/faqs/what-is-ncapture [accessed 2016-11-18] [WebCite Cache ID 6m7dqDJge]
18. Social Media Research Foundation. Social Network Importer for NodeXL. 2015 Mar 21. URL: https://socialnetimporter.codeplex.com/ [accessed 2016-11-17] [WebCite Cache ID 6m7dxDOHY]
19. Twitter. 2016. The Streaming APIs URL: https://dev.twitter.com/streaming/overview [accessed 2016-11-17] [WebCite Cache ID 6m7e9SDuI]
20. Twitonomy. 2016. URL: http://www.twitonomy.com/ [accessed 2016-11-17] [WebCite Cache ID 6m7eGH8hl]
21. Gnip Inc. Gnip. 2016. URL: https://www.gnip.com/ [accessed 2016-11-18] [WebCite Cache ID 6m7eNkJO2]
22. Salesforce Marketing Cloud. 2016. Radian6 is now part of Salesforce Marketing Cloud URL: https://www.marketingcloud.com/products/social-media-marketing/radian6/ [accessed 2016-11-18] [WebCite Cache ID 6m7eWXoMH]
23. Simply Measured Inc. Simply Measured. 2016. URL: http://simplymeasured.com/ [accessed 2016-11-18] [WebCite Cache ID 6m7edX0IE]
24. Sysomos Heartbeat. Proactive Social Media Monitoring URL: https://sysomos.com/products/heartbeat [accessed 2016-11-18] [WebCite Cache ID 6m7erpt4U]
25. Cole-Lewis H, Pugatch J, Sanders A. Social listening: a content analysis of e-cigarette discussions on Twitter. J Med Internet Res 2015;17:e243. [doi: 10.2196/jmir.4969]
26. Kim AE, Hopper T, Simpson S, Nonnemaker J, Lieberman AJ, Hansen H, et al. Using twitter data to gain insights into e-cigarette marketing and locations of use: an infoveillance study. J Med Internet Res 2015 Nov;17(11):e251 [FREE Full text] [doi: 10.2196/jmir.4466] [Medline: 26545927]

27. Myslín M, Zhu S, Chapman W, Conway M. Using twitter to examine smoking behavior and perceptions of emerging tobacco products. J Med Internet Res 2013 Aug;15(8):e174 [FREE Full text] [doi: 10.2196/jmir.2534] [Medline: 23989137]
28. van der Tempel J, Noormohamed A, Schwartz R, Norman C, Malas M, Zawertailo L. Vape, quit, tweet? electronic cigarettes and smoking cessation on twitter. Int J Public Health 2016 Mar;61(2):249-256. [doi: 10.1007/s00038-016-0791-2] [Medline: 26841895]
29. Chu K, Unger JB, Allem J, Pattarroyo M, Soto D, Cruz TB, et al. Diffusion of messages from an electronic cigarette brand to potential users through twitter. PLoS One 2015;10(12):e0145387 [FREE Full text] [doi: 10.1371/journal.pone.0145387] [Medline: 26684746]
30. Krauss M, Sowles S, Moreno M. Hookah-related twitter chatter: a content analysis. Prev Chronic Dis 2015;12:e121. [doi: 10.5888/ped12.150140]
31. Jain S, Zhu S, Conway M. Exploring consumer attitudes towards hookah, cigarettes, and cigars using twitter. Tobacco Reg Sci 2015 Oct 01;1(3):198-203. [doi: 10.18001/TRS.1.3.1]
32. Prier KW, Smith MS, Giraud-Carrier C, Hanson CL. Identifying health-related topics on twitter: an exploration of tobacco-related tweets as a test topic. In: Salerno J, editor. Social Computing, Behavioral-Cultural Modeling and Prediction. Berlin, Heidelberg: Springer-Verlag; 2011:18-25.
33. Huang J, Kornfield R, Szczypka G, Emery SL. A cross-sectional examination of marketing of electronic cigarettes on twitter. Tob Control 2014 Jul;23 Suppl 3:iii26-iii30 [FREE Full text] [doi: 10.1136/tobaccocontrol-2014-051551] [Medline: 24935894]
34. Jo CL, Kornfield R, Kim Y, Emery S, Ribisl KM. Price-related promotions for tobacco products on twitter. Tob Control 2016 Jul;25(4):476-479. [doi: 10.1136/tobaccocontrol-2015-052260] [Medline: 26130449]
35. van den Heerik RAM, van Hooijdonk CMJ, Burgers C, Steen GJ. “Smoking Is Sóóó ... Sandals and White Socks”: Co-creation of a Dutch anti-smoking campaign to change social norms. Health Commun 2016 Jun 29:1-8. [doi: 10.1080/10410236.2016.1168000] [Medline: 27355838]
36. Step MM, Bracken CC, Trapl ES, Flocke SA. User and content characteristics of public tweets referencing little cigars. Am J Health Behav 2016 Jan;40(1):38-47. [doi: 10.5993/AJHB.40.1.5] [Medline: 26685812]
37. de Viron S, Suggs L, Brand A, Van Oven H. Communicating genetics and smoking through social media: are we there yet? J Med Internet Res 2013 Sep 09;15(9):e198 [FREE Full text] [doi: 10.2196/jmir.2653] [Medline: 24018012]
38. Prochaska J, Pechmann C, Kim R, Leonhardt J. Twitter = quitter? an analysis of twitter quit smoking social networks. Tobacco Control 2012;21:447-449. [doi: 10.1007/978-3-642-19656-0]
39. Harris JK, Moreland-Russell S, Choucair B, Mansour R, Staub M, Simmons K. Tweeting for and against public health policy: response to the Chicago Department of Public Health's electronic cigarette twitter campaign. J Med Internet Res 2014;16(10):e238 [FREE Full text] [doi: 10.2196/jmir.3622] [Medline: 25320863]
40. Aphinyanaphongs Y, Lulejian A, Brown D, Bonneau R, Krebs P. Text classificationfor automatic detection of e-cigarette use and use for smoking cessation from twitter: a feasibility pilot. Pac Symp Biocomput 2016;21:480-491.
41. Clark EM, Jones CA, Williams JR, Kurti AN, Norotsky MC, Danforth CM, et al. Vaporous marketing: uncovering pervasive electronic cigarette advertisements on twitter. PLoS One 2016 Jul;11(7):e0157304 [FREE Full text] [doi: 10.1371/journal.pone.0157304] [Medline: 27410031]
42. Kavuluru R, Sabbir AK. Toward automated e-cigarette surveillance: Spotting e-cigarette proponents on twitter. J Biomed Inform 2016 Jun;61:19-26. [doi: 10.1016/j.jbi.2016.03.006] [Medline: 26975599]
43. Chu K, Unger J, Cruz T. Electronic cigarettes on twitter - spreading the appeal of flavors. Tob Regul Sci 2015;1:36-41. [doi: 10.18001/TRS.1.1.4]
44. Chung JE. A smoking cessation campaign on twitter: understanding the use of twitter and identifying major players in a health campaign. J Health Commun 2016 May;21(5):517-526. [doi: 10.1080/10810730.2015.1103332] [Medline: 27115179]
45. Cole-Lewis H, Varghese A, Sanders A, Schwarz M, Pugatch J, Augustson E. Assessing electronic cigarette-related tweets for sentiment and content using supervised machine learning. J Med Internet Res 2015 Aug 25;17(8):e208 [FREE Full text] [doi: 10.2196/jmir.4392] [Medline: 26307512]
46. Duke JC, Hansen H, Kim AE, Curry L, Allen J. The use of social media by state tobacco control programs to promote smoking cessation: a cross-sectional study. J Med Internet Res 2014 Jul;16(7):e169 [FREE Full text] [doi: 10.2196/jmir.3430] [Medline: 25014311]
47. Emery S, Szcyzpka G, ABril E. Are you scared yet? evaluating fear appeal messages in tweets about the tips campaign. J Commun 2014;64:278-295. [doi: 10.1111/jcom.12083]
48. McDaniel PA, Patzke H, Malone RE. Twitter users' reaction to a chain pharmacy's decision to end tobacco sales. Tob Induc Dis 2015;13:36 [FREE Full text] [doi: 10.1186/s12971-015-0060-9] [Medline: 26539069]
49. Rocheleau M, Sadasivam RS, Baquis K, Stahl H, Kinney RL, Pagoto SL, et al. An observational study of social and emotional support in smoking cessation Twitter accounts: content analysis of tweets. J Med Internet Res 2015;17(1):e18 [FREE Full text] [doi: 10.2196/jmir.3768] [Medline: 25589009]
50. Unger JB, Escobedo P, Allem J, Soto DW, Chu K, Cruz T. Perceptions of secondhand e-cigarette aerosol among twitter users. Tobacco Reg Sci 2016 Apr 01;2(2):146-152. [doi: 10.18001/TRS.2.2.5]
51. Altman D. Practical statistics for medical research. London: Chapman and Hall; 1991.

52. George D, Mallery P. SPSS for Windows step by step: A simple guide and reference. 11.0 update. 4th edition. Boston, MA: Allyn & Bacon; 2003.
53. Centers for Disease Control and Prevention. 2015 Apr 17. Morbidity and Mortality Weekly Report: Tobacco use among middle and high school students: United States, 2011-2014 URL: http://www.cdc.gov/mmwr/preview/mmwrhtml/mm6414a3.htm [accessed 2016-11-18] [WebCite Cache ID 6m7f7VXzc]
54. US Food and Drug Administration. 2016 Nov 28. The Facts on the FDA's New Tobacco Rule URL: https://www.fda.gov/ForConsumers/ConsumerUpdates/ucm506676.htm [accessed 2017-03-13] [WebCite Cache ID 6owhqygQ1]
55. Reyes A, Rosso P, Veale T. A multidimensional approach for detecting irony in Twitter. Lang Resources & Evaluation 2012 Jul 24;47(1):239-268. [doi: 10.1007/s10579-012-9196-x]
56. Sobhani P, Mohammad S, Kiritchenko S. Detecting stance in tweets and analyzing its interaction with sentiment. 2016 Presented at: Proceedings of the Fifth Joint Conference on Lexical and Computational Semantics; August 2016; Berlin, Germany p. 159-169.
57. Allem J, Ferrara E. The importance of debiasing social media data to better understand e-cigarette-related attitudes and behaviors. J Med Internet Res 2016 Aug 09;18(8):e219 [FREE Full text] [doi: 10.2196/jmir.6185] [Medline: 27507563]
58. Burger JD, Henderson J, Kim G, Zarrella G. Discriminating gender on Twitter. 2011 Presented at: Proceedings of the Conference on Empirical Methods in Natural Language Processing; July 2011; Edinburgh, Scotland, UK.
59. Sakaki S, Miura Y, Ma X, Hattori K, Ohkuma T. Twitter user gender inferences using combined analysis of text and image processing. 2014 Presented at: Proceedings of the 25th International Conference on Computational Linguistics; August 2014; Dublin, Ireland.
60. Nguyen D, Gravel R, Trieschnigg D, Meder T. “How old do you think I am?”: a study of language and age in twitter. 2013 Presented at: Proceedings of the Seventh International AAAI Conference on WeblogsSocial Media; July 2013; Cambridge, MA.

Abbreviations

API: application programming interface
FDA: US Food and Drug Administration

Edited by A Keepanasseril; submitted 30.11.16; peer-reviewed by M Conway, KY Wen; comments to author 22.01.17; revised version received 26.01.17; accepted 23.02.17; published 31.03.17.

Please cite as:
Lienemann BA, Unger JB, Cruz TB, Chu KH
Methods for Coding Tobacco-Related Twitter Data: A Systematic Review
J Med Internet Res 2017;19(3):e91
URL: http://www.jmir.org/2017/3/e91/
doi:10.2196/jmir.7022
PMID:28363883

Original Paper

Online Recruitment: Feasibility, Cost, and Representativeness in a Study of Postpartum Women

Liana S Leach[1], BA, PhD; Peter Butterworth[2,3], BA, MSc, PhD; Carmel Poyser[4], BSc, MSc; Philip J Batterham[4], BSc, MPH, PhD; Louise M Farrer[4], BA, PhD

[1]National Centre for Epidemiology and Population Health, Research School of Population Health, The Australian National University, Canberra, Australia

[2]Centre for Mental Health, Melbourne School of Population and Global Health, University of Melbourne, Melbourne, Australia

[3]Melbourne Institute of Applied Economic and Social Research, University of Melbourne, Melbourne, Australia

[4]Centre for Mental Health Research, Research School of Population Health, The Australian National University, Canberra, Australia

Corresponding Author:
Liana S Leach, BA, PhD
National Centre for Epidemiology and Population Health
Research School of Population Health
The Australian National University
Building 54, Mills Road
Canberra, 2614
Australia
Phone: 61 261259725
Fax: 61 261259725
Email: Liana.Leach@anu.edu.au

Abstract

Background: Online recruitment is feasible, low-cost, and can provide high-quality epidemiological data. However, little is known about the feasibility of recruiting postpartum women online, or sample representativeness.

Objective: The current study investigates the feasibility of recruiting a population of postpartum women online for health research and examines sample representativeness.

Methods: Two samples of postpartum women were compared: those recruited online as participants in a brief survey of new mothers (n=1083) and those recruited face-to-face as part of a nationally representative study (n=579). Sociodemographic, general health, and mental health characteristics were compared between the two samples.

Results: Obtaining a sample of postpartum women online for health research was highly efficient and low-cost. The online sample over-represented those who were younger (aged 25-29 years), were in a de facto relationship, had higher levels of education, spoke only English at home, and were first-time mothers. Members of the online sample were significantly more likely to have poor self-rated health and poor mental health than the nationally representative sample. Health differences remained after adjusting for sociodemographic differences.

Conclusions: Potential exists for feasible and low-cost e-epidemiological research with postpartum populations; however, researchers should consider the potential influence of sample nonrepresentativeness.

(J Med Internet Res 2017;19(3):e61) doi:10.2196/jmir.5745

KEYWORDS

online; Internet; recruitment; feasibility; representativeness; postpartum

Introduction

Internet-based recruitment and data collection for epidemiological research has been shown to offer benefits in terms of feasibility and accessibility (ie, low-cost, time-efficient, access to hard-to-reach populations, broad geographical/global reach) [1,2]. Importantly, evidence is also accumulating that data obtained via Internet-based research methods is reliable and valid [3-5]. Research also suggests that Internet recruitment can provide a sample of young adults broadly representative of the target population of interest [1,2,6,7].

One specific population-group in which Internet-based research and data collection may be highly enabling, but where little research is available, is postpartum women (first 12 months after birth). Previous research investigating postpartum parents has relied largely on recruitment via hospitals and child/maternal health clinics, requiring considerable time and financial investment. However, there may be significant emerging opportunities to recruit this population more efficiently online. Internet usage in this group is likely to be high, given postpartum women are often socially isolated at home, restricted in their mobility, and time-poor [8-10]. Studies successfully recruiting women planning a pregnancy (preconception) [11,12], and small samples of postpartum women [9,10], suggest that online recruitment is feasible. However, little is known about how participants recruited online might differ from postpartum women in the general population.

The current study investigates the feasibility of recruiting a large sample of postpartum women online for health research, and examines sample representativeness. Sociodemographic and health characteristics were compared between an online sample and another postpartum sample that was recruited face-to-face in the context of a nationally representative household survey.

Methods

Participants and Procedure

The Living with a Young Baby Survey

Participants were recruited during January and February 2015 for an anonymous online survey investigating women's general postpartum health and psychological wellbeing. Online recruitment involved two processes: (1) advertisement on a popular Australian pregnancy/infant-focused website *babycentre* [13], and (2) targeted Facebook advertisements. The brief advertisements targeted women who were aged >18 years, who resided in Australia with a young baby: " *Research: Mums Wanted! Researchers are looking for Australian women with a young baby to do a brief survey* ". No incentive was provided for participation. Mental health was not mentioned, in an effort to minimize health selection. Individuals who *clicked* on the advertisement were redirected to the online survey, which took approximately 15 minutes to complete. The Living with a Young Baby Survey (LYBS) was completed by 1083 respondents. The LYBS was approved by The Australian National University Human Research Ethics Committee.

Household Income Labour Dynamics in Australia Study

Comparative data were drawn from waves (time-points) 11 and 13 of the Household Income Labour Dynamics in Australia Study (HILDA), as at these time-points the K-10 Psychological Distress Scale was included in the survey. HILDA is a longitudinal nationally representative household panel survey that has been conducted annually since 2001 [14]. At baseline, the study recruited participants using a multi-stage sampling approach, sampling households within a selection of administrative areas. At baseline, 7682 households were involved, including 13,969 individual household members aged 15-95 years. Completion of the HILDA survey involves both a brief face-to-face personal interview and a paper-pencil self-completed questionnaire. The baseline response rate (66%) and individual-level reinterview rates (96% in wave 13) are comparable or superior to other national household panels around the world [15]. The sample is not static, as new participants enter study households and a significant top-up sample was introduced in wave 11. In addition, weights provided with the dataset enable adjustment for selection and attrition to better reflect the national population [16]. The current HILDA analyses were restricted to postpartum women aged 18-50 years (youngest child <1 year), providing a total of 579 respondents (wave 11 n=288; wave 13 n=291).

Measures

Sociodemographic measures included location (state and remoteness), language spoken at home, age, relationship status, number of children, and level of education completed. *General self-rated health* was measured by asking, "In general, how would you say your own health is?" Possible responses were, "excellent", "very good", "good", "fair", or "poor" [17]. *Psychological distress* was measured using the K-10 Kessler Psychological Distress Scale [18]. The K-10 has ten items representing symptoms of psychological distress, with a total scale score ranging from 10-50. The total score was categorized into four groups: 10-15 (*Low*), 16-21 (*Moderate*), 22-29 (*High*), and 30-50 (*Very High*), consistent with previous research [19,20].

Statistical Analyses

Descriptive statistics for sociodemographic and health variables from both samples are presented. The data reported for the HILDA sample was weighted by the cross-sectional population weights included in waves 11 and 13. Chi-square tests of association identified characteristics that differed between the LYBS and HILDA samples. Univariate logistic regression provided estimates of the magnitude of differences in characteristics between the LYBS and HILDA samples (odds ratios). Multivariate logistic regression (including all variables) tested whether differences in sociodemographic characteristics accounted for differences in general health and mental health between the samples. The weighted HILDA data provides the best possible comparison group available: postpartum women can be identified, self-rated health and the K-10 are reported (comparable measures to the LYBS), and the sample is representative of the Australian population.

Results

Online recruitment from the *babycentre* website took place over 9 days, and 264 surveys were completed. There was no cost associated with advertising on the *babycentre* website. Online recruitment from Facebook took place over 4 days, and 819 surveys were completed. Information provided by the Facebook Ads Manager shows that across the 4 days the advertisement was *clicked on* 2647 times, and the total audience reach was 38,765. Given that 819 surveys were completed, we estimate a conversion rate of *click*-to-completed-survey of 30.94% (819/2647). The total cost for advertising on Facebook was Aus $448.68 (the cost per *click* on the advertisement was Aus $0.17), and cost per survey completion was Aus $0.55.

Multimedia Appendix 1 presents both descriptive information and results from logistic regression models. Compared to mothers in the nationally representative HILDA survey, mothers in the online LYBS were more likely to be located in the Australian Capital Territory (+4.9%), to be located in a remote or rural area (+1.5%), to speak English only at home (+15.4%), to be younger (+7.7%), to be in a de facto relationship (+7.4%), and to be a first-time mother (+16.5%). Mothers in the online survey were less likely to have *not completed* high school (-5.5%). Mothers in the online LYBS had poorer general health and higher psychological distress than those in the HILDA study. Multivariate logistic regression (final column in Multimedia Appendix 1) showed that including all sociodemographic characteristics did little to attenuate the association between poor general health (and poor mental health) and participation in the LYBS online sample (compared to the HILDA sample). Those with fair/poor general health (as opposed to excellent health) remained 2.66 times more likely to participate in the LYBS online sample, compared to the HILDA sample. Those with moderate distress (as opposed to low distress) remained 2.73 time more likely to participate in the LYBS sample; those with high distress were 2.03 times more likely, and those with very high distress were 3.92 times more likely.

Discussion

The online recruitment strategy was highly time-efficient and low-cost. Over a period of thirteen days, 1083 participants were recruited for a total direct cost of Aus $448.68. The recruitment cost for the current online study is significantly lower than has been reported previously in studies conducting online recruitment [1,21]. This result is likely due to the combined strategy of utilizing free website advertising and commercial Facebook advertising, and because the current recruitment approach targeted a very specific population (postpartum women).

There were several significant differences in sociodemographic characteristics between the online LYBS sample and the HILDA sample. One of the greatest differences was that the LYBS recruited more first-time mothers. First-time mothers may be more likely to self-select into research focused on postpartum experiences, given the experience of motherhood is new and highly salient. The higher education levels observed in the LYBS sample likely reflect greater education among social media users in general [22], and in the context of motherhood, those who engage with infant care and development websites. Women in the online LYBS sample had poorer general health and mental health. Previous research has similarly found that online recruitment for mental health surveys attracts a sample with higher levels of psychological distress than is reported by the general population [1]. This finding may be due to the fact that online recruitment is more susceptible to uncontrolled *snowballing* as participants recruit others that they think will be interested in the research. It is also possible that reports from the general population under-represent levels of psychological distress [23,24]. Given that the poorer general health and mental health of postpartum mothers recruited online was not simply due to sociodemographic variation, further investigation is needed to understand who participates in online health research and why, and how the characteristics of this population may affect the results obtained.

The extent to which sample differences associated with recruitment method will impact on the validity of research findings largely depends on the research question being investigated [5,25,26]. The current findings suggest that when the goal is to estimate the prevalence of physical and mental health problems in postpartum women, Internet recruitment will likely provide an overestimate. Alternatively, when the focus is on the relationship between exposure and illness (etiology), Hatch et al state, “representativeness of a source population is arguably not a prerequisite for either internal validity or generalizability” [5]. Nonrepresentation is chiefly a problem when characteristics that differentiate those online from those not online (eg, in this study: primiparity, non-English speaking background, education), distort or moderate the association between exposures and outcomes of interest [25].

This study recruited a large sample of postpartum mothers online, demonstrating both efficiency and low-cost. Online samples may over-represent first-time mothers with higher education and English-speaking backgrounds, as well as those experiencing physical and mental health problems. Unmeasured differences in survey methodology play a selection role (ie, length of survey, anonymity in the LYBS), and the findings may be unique to the Australian context. While the monetary figures clearly demonstrate that online recruitment is low-cost, we did not specifically conduct cost-effective analyses. Researchers adopting Internet methods to investigate postpartum health should consider the implications of sample nonrepresentation in relation to their specific research aims (eg, prevalence, etiology, intervention).

Acknowledgments

This work was supported by Australian Research Council Future Fellowship #FT13101444, National Health and Medical Research Council (NHMRC) Early Career Fellowship #1035803, and NHMRC Career Development Fellowship #1083311. HILDA was initiated and funded by the Australian Department of Social Services (DSS) and is managed by The Melbourne Institute for Applied Economic and Social Research. The findings and views reported in this paper, however, are those of the author and should not be attributed to either DSS or The Melbourne Institute. There have been no involvements that might raise the question of bias in the work reported or in the conclusions, implications, or opinions stated.

Conflicts of Interest

None declared.

Multimedia Appendix 1

[PDF File (Adobe PDF File), 34KB - jmir_v19i3e61_app1.pdf]

References

1. Batterham PJ. Recruitment of mental health survey participants using Internet advertising: content, characteristics and cost effectiveness. Int J Methods Psychiatr Res 2014 Jun;23(2):184-191. [doi: 10.1002/mpr.1421] [Medline: 24615785]
2. Fenner Y, Garland SM, Moore EE, Jayasinghe Y, Fletcher A, Tabrizi SN, et al. Web-based recruiting for health research using a social networking site: an exploratory study. J Med Internet Res 2012 Feb 01;14(1):e20 [FREE Full text] [doi: 10.2196/jmir.1978] [Medline: 22297093]
3. Andreeva V, Galan P, Julia C, Castetbon K, Kesse-Guyot E, Hercberg S. Assessment of response consistency and respective participant profiles in the Internet-based NutriNet-Santé Cohort. Am J Epidemiol 2014 Apr 01;179(7):910-916 [FREE Full text] [doi: 10.1093/aje/kwt431] [Medline: 24521560]
4. Konstan JA, Rosser BR, Ross MW, Stanton J, Edwards WM. The story of subject naught: a cautionary but optimistic tale of Internet survey research. J Comput Mediat Commun 2005;10(2):A-00. [doi: 10.1111/j.1083-6101.2005.tb00248.x]
5. Hatch EE, Hahn KA, Wise LA, Mikkelsen EM, Kumar R, Fox MP, et al. Evaluation of selection bias in an Internet-based study of pregnancy planners. Epidemiology 2016 Jan;27(1):98-104 [FREE Full text] [doi: 10.1097/EDE.0000000000000400] [Medline: 26484423]
6. Harris ML, Loxton D, Wigginton B, Lucke JC. Recruiting online: lessons from a longitudinal survey of contraception and pregnancy intentions of young Australian women. Am J Epidemiol 2015 May 15;181(10):737-746. [doi: 10.1093/aje/kwv006] [Medline: 25883155]
7. Mishra G, Hockey R, Powers J, Loxton D, Tooth L, Rowlands I, et al. Recruitment via the Internet and social networking sites: the 1989-1995 cohort of the Australian Longitudinal Study on Women's Health. J Med Internet Res 2014 Dec 15;16(12):e279 [FREE Full text] [doi: 10.2196/jmir.3788] [Medline: 25514159]
8. Drentea P, Moren-Cross JL. Social capital and social support on the web: the case of an Internet mother site. Sociol Health Illn 2005 Nov;27(7):920-943 [FREE Full text] [doi: 10.1111/j.1467-9566.2005.00464.x] [Medline: 16313523]
9. Maloni JA, Przeworski A, Damato EG. Web recruitment and internet use and preferences reported by women with postpartum depression after pregnancy complications. Arch Psychiatr Nurs 2013 Apr;27(2):90-95. [doi: 10.1016/j.apnu.2012.12.001] [Medline: 23540519]
10. McDaniel B, Coyne SM, Holmes EK. New mothers and media use: associations between blogging, social networking, and maternal well-being. Matern Child Health J 2012 Oct;16(7):1509-1517. [doi: 10.1007/s10995-011-0918-2] [Medline: 22094592]
11. Huybrechts KF, Mikkelsen EM, Christensen T, Riis AH, Hatch EE, Wise LA, et al. A successful implementation of e-epidemiology: the Danish pregnancy planning study 'Snart-Gravid'. Eur J Epidemiol 2010 May;25(5):297-304 [FREE Full text] [doi: 10.1007/s10654-010-9431-y] [Medline: 20148289]
12. Wise L, Rothman KJ, Mikkelsen EM, Stanford JB, Wesselink AK, McKinnon C, et al. Design and conduct of an Internet-based preconception cohort study in North America: Pregnancy Study Online. Paediatr Perinat Epidemiol 2015 Jul;29(4):360-371 [FREE Full text] [doi: 10.1111/ppe.12201] [Medline: 26111445]
13. BabyCenter. 2017. URL: http://www.babycentre.com.au [accessed 2017-02-25] [WebCite Cache ID 6oXbjtqu5]
14. Melbourne Institute of Applied Economic and Social Research. 2002. Household, Income and Labour Dynamics in Australia (HILDA) survey annual report 2002 URL: https://melbourneinstitute.com/downloads/hilda/Annual_Report/A_Report_2002.pdf [accessed 2017-02-25] [WebCite Cache ID 6oXbonLIh]
15. Melbourne Institute of Applied Economic and Social Research. 2011. Household, Income and Labour Dynamics in Australia (HILDA) survey annual report 2011 URL: https://melbourneinstitute.com/downloads/hilda/Annual_Report/A_Report_2011.pdf [accessed 2017-02-25] [WebCite Cache ID 6oXc9Ozal]
16. Watson N. HILDA Project Technical Paper Series No. 2/12. 2012 Dec. Longitudinal and Cross-sectional Weighting Methodology for the HILDA Survey URL: http://www.melbourneinstitute.com/downloads/hilda/bibliography/hilda_technical_papers/htec212.pdf [accessed 2017-02-25] [WebCite Cache ID 6oXcHIJLd]
17. Fayers P, Sprangers MA. Understanding self-rated health. Lancet 2002 Jan 19;359(9302):187-188. [doi: 10.1016/S0140-6736(02)07466-4] [Medline: 11812551]
18. Kessler R, Andrews G, Colpe LJ, Hiripi E, Mroczek DK, Normand SL, et al. Short screening scales to monitor population prevalences and trends in non-specific psychological distress. Psychol Med 2002 Aug;32(6):959-976. [Medline: 12214795]
19. Australian Bureau of Statistics. ABS, Information Paper: Use of the Kessler Psychological Distress Scale in ABS Health Surveys, Australia, 2007-08. 2008. URL: http://www.abs.gov.au/ausstats/abs@.nsf/mf/4817.0.55.001 [accessed 2017-02-25] [WebCite Cache ID 6oXcm3Mtw]

20. Andrews G, Slade T. Interpreting scores on the Kessler Psychological Distress Scale (K10). Aust N Z J Public Health 2001 Dec;25(6):494-497. [Medline: 11824981]
21. Morgan A, Jorm AF, Mackinnon AJ. Internet-based recruitment to a depression prevention intervention: lessons from the Mood Memos study. J Med Internet Res 2013 Feb 12;15(2):e31 [FREE Full text] [doi: 10.2196/jmir.2262] [Medline: 23403043]
22. Perrin A. Pew Research Centre.: Pew Research Centre; 2015. Social Media Usage 2005-2015 URL: http://www.pewinternet.org/2015/10/08/social-networking-usage-2005-2015/ [accessed 2017-02-25] [WebCite Cache ID 6oXdKgM3Z]
23. Slade T, Johnston A, Oakley Browne MA, Andrews G, Whiteford H. 2007 National Survey of Mental Health and Wellbeing: methods and key findings. Aust N Z J Psychiatry 2009 Jul;43(7):594-605. [doi: 10.1080/00048670902970882] [Medline: 19530016]
24. Woolhouse H, Gartland D, Perlen S, Donath S, Brown SJ. Physical health after childbirth and maternal depression in the first 12 months post partum: results of an Australian nulliparous pregnancy cohort study. Midwifery 2014 Mar;30(3):378-384. [doi: 10.1016/j.midw.2013.03.006] [Medline: 23619027]
25. Nohr E, Olsen J. Commentary: epidemiologists have debated representativeness for more than 40 years--has the time come to move on? Int J Epidemiol 2013 Aug;42(4):1016-1017. [doi: 10.1093/ije/dyt102] [Medline: 24062289]
26. Rothman KJ. Six persistent research misconceptions. J Gen Intern Med 2014 Jul;29(7):1060-1064 [FREE Full text] [doi: 10.1007/s11606-013-2755-z] [Medline: 24452418]

Abbreviations

DSS: Department of Social Services
HILDA: Household Income Labour Dynamics in Australia Study
LYBS: Living with a Young Baby Survey
NHMRC: National Health and Medical Research Council

Edited by G Eysenbach; submitted 09.03.16; peer-reviewed by E Hatch, A Barrera, E Nelson, R Nosheny; comments to author 28.06.16; revised version received 18.08.16; accepted 23.11.16; published 08.03.17.

Please cite as:
Leach LS, Butterworth P, Poyser C, Batterham PJ, Farrer LM
Online Recruitment: Feasibility, Cost, and Representativeness in a Study of Postpartum Women
J Med Internet Res 2017;19(3):e61
URL: http://www.jmir.org/2017/3/e61/
doi:10.2196/jmir.5745
PMID:28274906

Original Paper

Rates, Delays, and Completeness of General Practitioners' Responses to a Postal Versus Web-Based Survey: A Randomized Trial

Paul Sebo[1]; Hubert Maisonneuve[1]; Bernard Cerutti[2]; Jean Pascal Fournier[3]; Nicolas Senn[4]; Dagmar M Haller[1]

[1]Primary Care Unit, Faculty of Medicine, University of Geneva, Geneva, Switzerland

[2]Unit of Research and Development in Medical Education, Faculty of Medicine, University of Geneva, Geneva, Switzerland

[3]Department of General Practice, Faculty of Medicine, University of Nantes, Nantes, France

[4]Department of Ambulatory Care and Community Medicine, University of Lausanne, Lausanne, Switzerland

Corresponding Author:
Paul Sebo
Primary Care Unit
Faculty of Medicine
University of Geneva
rue Michel-Servet 1
Geneva, 1211
Switzerland
Phone: 41 223488888
Fax: 41 223486604
Email: paulsebo@hotmail.com

Abstract

Background: Web-based surveys have become a new and popular method for collecting data, but only a few studies have directly compared postal and Web-based surveys among physicians, and none to our knowledge among general practitioners (GPs).

Objective: Our aim is to compare two modes of survey delivery (postal and Web-based) in terms of participation rates, response times, and completeness of questionnaires in a study assessing GPs' preventive practices.

Methods: This randomized study was conducted in Western Switzerland (Geneva and Vaud) and in France (Alsace and Pays de la Loire) in 2015. A random selection of community-based GPs (1000 GPs in Switzerland and 2400 GPs in France) were randomly allocated to receive a questionnaire about preventive care activities either by post (n=700 in Switzerland, n=400 in France) or by email (n=300 in Switzerland, n=2000 in France). Reminder messages were sent once in the postal group and twice in the Web-based group. Any GPs practicing only complementary and alternative medicine were excluded from the study.

Results: Among the 3400 contacted GPs, 764 (22.47%, 95% CI 21.07%-23.87%) returned the questionnaire. Compared to the postal group, the participation rate in the Web-based group was more than four times lower (246/2300, 10.70% vs 518/1100, 47.09%, *P*<.001), but median response time was much shorter (1 day vs 1-3 weeks, *P*<.001) and the number of GPs having fully completed the questionnaire was almost twice as high (157/246, 63.8% vs 179/518, 34.6%, *P*<.001).

Conclusions: Web-based surveys offer many advantages such as reduced response time, higher completeness of data, and large cost savings, but our findings suggest that postal surveys can be still considered for GP research. The use of mixed-mode approaches is probably a good strategy to increase GPs' participation in surveys while reducing costs.

(J Med Internet Res 2017;19(3):e83) doi:10.2196/jmir.6308

KEYWORDS

participation rate; response time; completeness; survey methods; primary care

Introduction

Three main methods of collecting general data were usually used in the past in clinical and epidemiological research (face-to-face interviews, telephone interviews, and postal surveys). They were considered to be more or less equivalent in terms of validity of the data obtained, although postal surveys had the advantage of promoting more truthful responses to sensitive questions, costing less, and requiring fewer staff [1].

More recently, Web-based surveys became a new and popular method for collecting data because they are simple to use, inexpensive (no costs for printing, postage and data entry), less time-consuming (immediate survey delivery, real-time data tracking, and no data entry) and correct (high data quality because a structured format minimizes entry of erroneous or unacceptable data and automatic data transfer minimizes data entry errors) [2-6]. In addition, they offer other advantages, such as flexibility in display design (questions can be revised or removed, new questions can be added according to preliminary results) and almost no limit to the number of respondents [2,4,6,7]. Web-based surveys were developed to study various conditions [8-13] or assess the efficacy of Internet-based programs, for example, in preventing smoking relapse [14]. Several authors compared Web-based and postal surveys in terms of validity and/or reliability, and showed little or no difference between the two methods of data collection [8,10-13].

However, Web-based surveys can give rise to specific concerns about response rates, ethical issues (ie, whether researchers truly can promise anonymity and confidentiality, what constitutes informed consent), and selection bias (in relation to age, socioeconomical and education-related bias in access to the Internet) [2]. In particular, response rates seem clearly lower compared to postal-based surveys both in studies surveying nondoctors (either patients or general population) [5,15,16] and doctors [3,7,17,18]. Only a few studies have directly compared postal and Web-based surveys among physicians, usually using mixed-mode designs (making it more difficult to interpret); none to our knowledge studied this question among general practitioners (GPs). Using these mixed-mode designs, Beebe et al [3] (n=326 physicians) showed a statistically significant difference in response rates between two groups after one reminder (post 57% vs Web 47%), but the difference observed was not significant after switching groups for the second reminder, and McMahon et al [7] (n=181 pediatricians) found that response rates after one reminder were 41% by post and 26% in the Web-based group.

In general, response rates to surveys conducted among doctors are lower than surveys among nondoctors [3,6,19]. In addition, GPs are known to be more difficult to recruit than other doctors [20] and more likely to drop out [3,20,21]. Therefore, the results of studies conducted among physicians other than GPs will not necessarily be similar to studies performed in primary care settings.

The aim of our study was to determine through a randomized design whether Web-based surveys are feasible in primary care by assessing GPs' participation rates, response times, and completeness of data using two modes of questionnaire delivery: postal and Web-based.

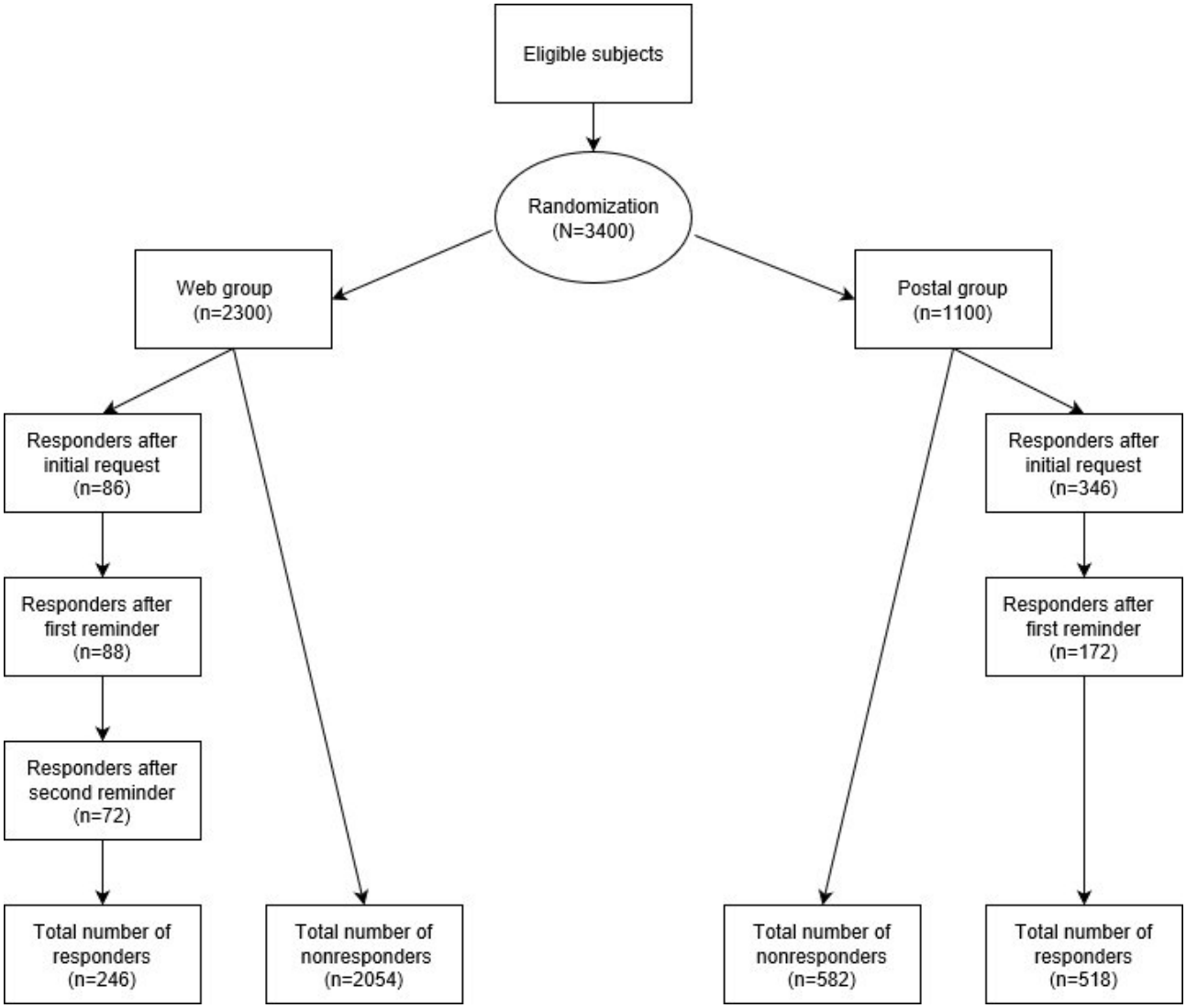
Methods

Study Site, Population, and Sample Size

This randomized trial was conducted in Western Switzerland (canton of Geneva and Vaud) and France (Alsace and Pays de la Loire) in 2015, as part of a study of GPs' preventive practices. Based on previous studies, the expected participation rate was 50% in the postal-based group and 20% in the Web-based group [7,19,22]. Assuming that we wanted to detect a difference in response rate of 10% or more between the two groups, with a power of 80%, and a type I error rate of 5%, a minimum of 408 questionnaires had to be completed in each group (postal and Web-based), which led to a sample size of 816 for the postal group and 2040 for the Web-based group. Taking into account anticipated incomplete questionnaires, both samples were increased by approximately 250, and rounded to 1100 and 2300, respectively.

For this purpose, a random sample of 600 community-based GPs practicing in the canton of Geneva was selected from a sampling frame consisting of all the GPs who were members of the physicians' professional organization and had a valid and available email address. They were allocated to the postal (n=300) or the Web-based (n=300) group at random, using simple (unrestricted) randomization based on computer-generated random numbers, and invited to participate by post or by email. In addition, 400 GPs were randomly selected in the canton of Vaud. Because a list of email addresses was unavailable (the professional organization of this canton does not make the list available for research purposes), all GPs from the canton of Vaud were allocated to the postal group. Therefore, in Switzerland, 700 GPs were included in the postal group and 300 in the Web-based group. The same procedure—recruitment of community-based GPs by post (n=400) and by email (n=2000)—was carried out in France (Alsace and Pays de la Loire). Reminder messages (once for the postal group and twice for the Web-based group) were sent at one-month intervals. No monetary incentives were offered to the participating GPs. All community-based GPs practicing were eligible for the study, except those practicing only complementary and alternative medicine. There were no other exclusion criteria. This recruitment process is summarized in Figure 1.

Figure 1. Flowchart of the study.



Data Collection

In Switzerland, a research assistant contacted each randomly selected GP by email or by post, according to group allocation. In France, this task was done by the Union Régionale des Professionnels de Santé Alsace and Pays de la Loire. They informed the GPs about the aim of our study and the practical procedures for completing the questionnaire. The postal letters included a cover letter and a stamped return envelope. GPs were asked to send the completed questionnaires back to the research assistant. Participants in the Web-based group received the same cover letter in a Web-based format. GPs were asked to connect through a hyperlink and complete the online questionnaire. The paper questionnaire was designed first, closely following recently published recommendations for optimal survey content and layout [23]. The Web-based questionnaire was then created to be as similar as possible to the paper version, including regarding text formatting. Note that completion of all questions was not required before submission of the Web-based questionnaire.

The questionnaire included questions about sociodemographic characteristics (age, gender, location of the practice, certification, number of working days per week, number of working years in the current practice), as well as about current preventive practices, scored on a five-point Likert scale ranging from “never performed” to “always performed.” There were 37 items in the questionnaire. The analysis of these data will be presented in a separate paper.

The questionnaire was piloted with seven GPs in a primary care clinic (Centre Médical des Trois Chêne, Geneva, Switzerland) to identify any source of difficulties or misunderstanding that colleagues may face when responding to the questions, but only minor changes in the layout were suggested during this pretest phase. All collected data remained confidential. Only the research assistants knew the name and participation code of the GPs taking part in our study; they were not known at any time by the researchers who analyzed the study data. Tacit consent was presumed from the doctors if they completed a questionnaire. We did not collect data about nonresponding GPs. For Switzerland, a waiver from obtaining informed consent was granted by the Research Ethics Committee of Geneva (approval by the Ethics Committee is unnecessary in Switzerland when only physicians are surveyed about their practice), whereas for France the research protocol was approved by the Research Ethics Committee of Nantes (ref: 2015-09-06).

Statistical Analysis

Comparisons of different categorical variables within contingency tables were made with chi-square tests. Continuous data were summarized by means and standard deviations, and comparisons were made with *t* tests. When the data were clearly asymmetric, medians and interquartile ranges were used, and Wilcoxon rank sum tests for comparisons. The association between the completeness of questionnaires and various variables (gender, age group, country, type of survey, number of half-days worked per week, number of working years) were investigated with generalized linear models with binomial link functions. All significant variables were included in a multivariate model, and the final multivariate model was chosen with a backward and forward stepwise procedure based on the Akaike information criterion [24].

All analyses were run on R 2.15.3 (the R Foundation for Statistical Computing) and TIBCO Spotfire S+ 8.1 for Windows (TIBCO Software Inc).

Results

Characteristics and Representability of the Respondents

Among the 3400 GPs who were contacted at random (2400 in France and 1000 in Switzerland), 764 (22.47%) responded to the survey, 336 (14.00%) in France and 428 (42.80%) in Switzerland. Figure 1 shows the flowchart of the study and Table 1 presents the sociodemographic characteristics of the responders according to the group (postal vs Web-based) they were allocated to. There were more men (62.5%, 318/509 vs 49.8%, 110/221, *P*=.002) and GPs older than 55 years (51.6%, 266/516 vs 39.6%, 88/222, *P*=.005) in the postal group compared to the Web-based group. The GPs' profiles were similar in the two countries (men: 59.1%, 240/406 in Switzerland vs 58.0%, 188/324 in France, *P*=.82; age ≥55 years: 48.9%, 199/407 vs 46.8%, 155/331, *P*<.001).

Table 1. General practitioners' characteristics (N=764).

Characteristics	Web group	Postal group	*P*	Total
Gender, n/N (%)[a]			.002	
Male	110/221 (49.8)	318/509 (62.5)		428/730 (58.6)
Female	111/221 (50.2)	191/509 (37.5)		302/730 (41.4)
Age group (years), n/N (%)[a]			.005	
<35	16/222 (7.2)	13/516 (2.5)		29/738 (3.9)
35-44	53/222 (23.9)	104/516 (20.2)		157/738 (21.3)
45-54	65/222 (29.3)	133/516 (25.8)		198/738 (26.8)
55-64	75/222 (33.8)	207/516 (40.1)		282/738 (38.2)
>64	13/222 (5.9)	59/516 (11.4)		72/738 (9.8)
Mean number of half-days worked per week, mean (SD)	8.5 (2.1)	8.6 (2.3)	.80	8.6 (2.2)
Number of working years in the current practice, mean (SD)	16.4 (11.0)	18.7 (11.0)	.01	18.0 (11.1)

[a] n=number with factor considered; N=number of data available. Denominators do not all add to 764 because of missing values.

For Switzerland, the postal sample appeared to be slightly more representative of all community-based GPs (median age of 56 years and 78% men in 2015 [25]) regarding the median age (median 53.7, IQR 45.7-60.9, for the postal group, and 45.9, IQR 42.9-60.9, for the Web-based group) and gender distribution (men: 215/353, 60.9%, for the postal group and 25/53, 47%, for the Web-based group).

For France (data from Pays de Loire only; median age of 51 years and 57% men in 2013 [26]) the results were similar. The postal sample appeared to be slightly more representative of all community-based GPs regarding the median age (50.9, IQR 41.5-58.7, for the postal group and 47.1, IQR 38.5-56.2, for the Web-based group) and gender distribution (men: 58/97, 60%, for the postal group and 53/114, 46.5%, for the Web-based group).

Differences in Response Rates Between Groups

Table 2 compares GPs' participation rates in the two groups. Overall, participation rates were more than four times higher in the postal compared to the Web-based group (47.09%, 518/1100 vs 10.70%, 246/2300, *P*<.001), although the difference was less pronounced in Switzerland (50.7%, 355/700, vs 24.3%, 73/300, *P*<.001) compared to France (40.8%, 163/400 vs 8.65%, 173/2000, *P*<.001). In the Web-based group, the rates of GPs who participated after the initial request, the first, or the second reminder were relatively similar, whereas in the postal group, only half as many GPs completed the questionnaire after the reminder compared to the initial request.

Table 2. General practitioners' participation rates in Web and postal groups stratified by country and by initial request or reminders.

Characteristics	Web group		Postal group		*P*	Total	
	n/N (%)[a]	95% CI	n/N (%)[a]	95% CI		n/N (%)[a]	95% CI
Switzerland	73 /300 (24.3)	19.5%-29.2%	355/700 (50.7)	47.0%-54.4%	<.001	428/1000 (42.8)	39.7%-45.9%
Initial request	21 (7.0)	4.1%-9.9%	245 (35.0)	31.5%-38.5%		266 (26.6)	23.9%-29.3%
Reminder #1	22 (7.3)	4.4%-10.3%	110 (15.7)	13.0%-18.4%		132 (13.2)	11.1%-15.3%
Reminder #2	30 (10.0)	6.6%-13.4%	NA	NA		30 (3.0)	1.9%-4.1%
France	173/2000 (8.7)	7.4%-9.9%	163/400 (40.8)	35.9%-45.6%	<.001	336/2400 (14.0)	12.6%-15.4%
Initial request	65 (3.3)	2.5%-4.0%	101 (25.3)	21.0%-29.5%		166 (6.9)	5.9%-7.9%
Reminder #1	66 (3.5)	2.5%-4.1%	62 (15.5)	12.0%-19.0%		128 (5.3)	4.4%-6.2%
Reminder #2	42 (2.1)	1.5%-2.7%	NA	NA		42 (1.8)	1.2%-2.3%
Total	246/2300 (10.7)	9.4%-12.0%	518/1100 (47.1)	44.4%-50.0%	<.001	764/3400 (22.5)	21.1%-23.9%
Initial request	86 (3.7)	3.0%-4.5%	346 (31.5)	28.7%-34.2%		432 (12.7)	11.6%-13.8%
Reminder #1	88 (3.8)	3.0%-4.6%	172 (15.6)	13.5%-17.8%		260 (7.6)	6.8%-8.5%
Reminder #2	72 (3.1)	2.4%-3.8%	NA	NA		72 (2.1)	1.6%-2.6%

[a] n=number of GPs agreeing to participate; N=number of GPs contacted in the Web group (Geneva: n=300; Vaud: n=0; Alsace: n=1000; Pays de la Loire: n=1000) and in the postal group (Geneva: n=300; Vaud: n=400; Alsace: n=200; Pays de la Loire: n=200).

Differences in Response Time Between Groups

In Switzerland, the median response time for paper surveys was approximately 3 weeks for initial request and reminder (Table 3), whereas in France, the response time was 1 week for initial request and 3 weeks for the reminder. In contrast, in both countries, GPs allocated to the Web-based group completed the questionnaire about 1 day after they had received the initial request, the first, or the second reminder.

Table 3. General practitioners' responses times in the two groups stratified by country and by initial request or reminders.

Characteristics	Web group (days), median (IQR)	Postal group (days), median (IQR)	*P*	Total (days), median (IQR)
Switzerland	1 (1-3)	21 (16-26)	<.001	19 (14-24)
Initial request	2 (1-8)	21 (16-27)	<.001	21 (14-24)
Reminder #1	0 (0-1)	19 (15-25)	<.001	18 (14-25)
Reminder #2	1 (1-3)	NA	NA	1 (1-3)
France	1 (0-2)	10 (6-24)	.78	4 (0-10)
Initial request	1 (0-2)	6 (6-10)	<.001	5 (1-6)
Reminder #1	0 (0-3)	25 (20-28)	<.001	6 (0-24)
Reminder #2	1 (0-2)	NA	NA	1 (0-2)
Total	1 (0-3)	20 (14-25)	<.001	14 (2-23)
Initial request	1 (0-3)	20 (13-23)	<.001	16 (6-23)
Reminder #1	0 (0-3)	21 (17-28)	<.001	15 (1-25)
Reminder #2	1 (1-2)	NA	NA	1 (1-2)

Differences in Questionnaire Completion Between Groups

The number of GPs who fully completed the Web-based questionnaire was nearly 1.5 times higher in Switzerland and more than twice as high in France compared to the completion of postal questionnaires (Table 4). In multivariate analyses, the proportion of fully completed questionnaires was higher for GPs working more (>8 half-days per week), those being less experienced (≤18 years in the current practice), and those having completed the online version of the questionnaire, whereas GP gender and age group, and location of the practice (Switzerland or France) were not associated with full completion of the questionnaire (Table 5).

Table 4. Number of general practitioners in the Web and postal groups who fully completed the questionnaire, without missing data, stratified by country.

Characteristics	Web group		Postal group		*P*	Total	
	n/N (%)[a]	95% CI	n/N (%)[a]	95% CI		n/N (%)[a]	95% CI
Switzerland	37/73 (50.7)	39.2%-62.15%	125/355 (35.2)	30.2%-40.2%	<.001	162/428 (37.9)	33.3%-42.4%
France	120/173 (69.4)	62.5%-76.2%	54/163 (31.1)	25.9%-40.3%	<.001	174/336 (51.8)	46.4%-57.1%
Total	157/246 (63.8)	57.8%-69.8%	179/518 (34.6)	30.5%-38.7%	<.001	336/764 (44.0)	40.5%-47.5%

[a] n=number of GPs having agreed to participate; N=number of GPs contacted.

Table 5. Proportion of fully completed questionnaires by sociodemographic characteristics of the responders and type of survey.

Characteristics	n/N (%)	OR (95% CI)	*P*	AOR (95% CI)[a]	*P*
Gender			.07		
Male	185/428 (43.2)	1.00			
Female	151/302 (50.0)	1.31 (0.98-1.77)			
Age group (years)			<.001		
<35	16/29 (55.2)	1.00			
35-44	78/157 (49.7)	0.80 (0.31-2.02)			
45-54	107/198 (54.0)	0.96 (0.38-2.41)			
55-64	115/282 (40.8)	0.56 (0.23-1.39)			
>64	20/72 (27.8)	0.31 (0.11-0.90)			
Number of half-days worked per week			.005		.007
≤ 8	127/315 (40.3)	1.00		1.00	
>8	209/412 (50.7)	1.52 (1.13 2.05)		1.55 (1.12-2.16)	
Number of working years in the current practice			.009		.01
≤18	192/374 (51.3)	1.00		1.00	
>18	144/346 (41.6)	0.68 (0.50-0.91)		0.66 (0.48-0.91)	
Location of the practice			<.001		
France	174/336 (51.8)	1.00			
Switzerland	162/428 (37.9)	0.57 (0.42-0.76)			
Type of survey			<.001		<.001
Postal	179/518 (34.6)	1.00		1.00	
Web	157/246 (63.8)	3.34 (2.43-4.59)		4.42 (3.10-6.30)	

[a] Adjusted for all variables listed in the table.

Discussion

Main Findings

Participation rates were more than four-fold higher when GPs were sent a questionnaire by post than by email. But Web-based questionnaires were completed in a timelier and more complete manner compared to postal questionnaires. The findings were similar in two French-speaking countries.

Comparison With Previous Studies

Response Rates

Previous studies compared participation rates between postal and Web-based surveys among nonphysicians (either general population or patients), and their findings point in the same direction. In a study involving Danish women referred for mammography (N=376), Kongsved et al [15] showed that the response rates were much higher for the postal (73%) compared to the Web-based survey (18%); the questionnaire consisted of 17 pages with a total of 119 items. Another study by Bergeson et al [5] assessed patient experiences with care in Minnesota (N=1392) and found a response rate of 33% by post (vs 14% for a Web-based survey). Finally, Sinclair et al [16] carried out a community survey of greywater use (N=1621) and showed a 10.5% and 7.5% response rate for a personalized and for a generic postal survey, vs 4.7% and 2.2% for a personalized and for a generic postal invitation with Web survey. The low participation rates in this study are probably explained by the

study population (general population) and the absence of reminders.

Only a few studies addressed this issue among physicians, usually using mixed-mode designs (making it more difficult to interpret), and none to our knowledge among GPs. These studies provide findings similar to our own results. Beebe et al [3] recruited 326 physicians from 12 divisions within the Mayo Clinic Department of Medicine and showed a statistically significant difference in response rate between the two groups after one reminder (post 57% vs Web 47%), but the difference observed was not significant after switching groups for the second reminder (mail/Web group, that is mail invitation, mail reminder #1, Web reminder #2: 71%; Web/mail group: 63%, *P*=.07). According to the authors, the relatively high response rates in this study could be explained by the fact that Mayo Clinic physicians are particularly interested in research. In a survey that recruited 181 pediatricians in Georgia (USA), McMahon et al [7] found that response rates after one reminder were 41% by post and 26% in the Web-based group. Lower participation rates (26% in the postal vs 11% in the Web-based group) were noted by Hardigan and al [17] among practicing dentists (N=1232) in the Florida Tobacco Control Survey. Finally, in a survey on attitudes toward the screening and treatment of hepatitis C (N=398), Kroth and al [18] showed that, despite a total of five rounds of online solicitation, 24% of the responses were in paper form (the paper version was mailed only to the nonrespondents to the online solicitation).

We found that sending reminders improved response rates in both groups. However, in the Web-based group, the percentage of GPs agreeing to participate after the initial request and the reminders was relatively similar, whereas in the postal group, approximately half as many GPs completed the questionnaire following a reminder (vs the initial request). This finding means that it is probably useful in Web-based surveys to send two or even more reminders, which can be done without extra costs, bearing in mind that too many reminders may be considered as possible harassment of potential respondents [20]. In contrast, the cost-effectiveness of reminders can be unfavorable in postal surveys because costs are constant whereas participation rates tend to decrease for reminders. In addition to increasing the number of reminders, the following strategies have been shown to favor higher response rates in Web-based surveys: attaching a copy of the questionnaire to the email invitation, recruiting physicians with prior experience with Web surveys, using mixed-mode designs (Web/mail or mail/Web), and allowing participants to respond in the desired format (Web or mail) using personalized invitations and financial incentives [3,6,20,27,28].

An important limitation with Web-based surveys is related to the fact that the physicians' email addresses are neither always available nor regularly updated, which happens less frequently for postal addresses because the physicians have to use them in their current practice for contact with the authorities, other health care providers, and their patients.

Response Time

We found that the median response time was significantly longer for postal compared to Web-based surveys; therefore, if a short response time is required, a Web-based survey could be more appropriate than a postal survey. Our results are only in part explained by the fact that postal surveys require several days to be sent and returned, whereas Web-based surveys and responses are delivered immediately. Indeed, the much longer delay for postal responses suggests a longer delay both in opening the invitation letter and in sending back responses. Our results are consistent with previous research; in Beebe et al's study [3], the median response time was approximatively two days shorter in the Web/postal compared to the postal/Web group, but the results are somewhat difficult to analyze because the two groups were mixed. In McMahon et al's study [7], 23% of the Web-based questionnaires were returned on the same day versus none in the postal survey. Finally, Akl et al [29] randomized 119 residents in a university-based internal medicine residency program and showed a shorter mean response time (by 3.8 days) in a Web-based survey compared to a postal survey.

Completeness of Questionnaires

We showed that the number of GPs who fully completed the questionnaire was higher in the Web-based group compared to the postal group. Although both questionnaire formats were designed to be similar, this suggests that completion of a Web-based questionnaire is less subject to attention errors than a paper questionnaire. Our data are supported by Kongsved et al's study [15] in which 98% filled in a complete questionnaire in the Web group versus 35% in the mail group, and by McMahon et al's study [7] in which 2.1% of the questions were not answered in the mail group versus only 0.4% in the Web group. Interestingly, in the multivariate analysis, the proportion of fully completed questionnaires was higher for GPs working more days in the practice and for those who were less experienced, who may be more interested in research because they were involved in clinical research in their recent residency program. To our knowledge, this study is the first to describe questionnaire completeness by sociodemographic characteristics of the respondents. Because completeness of data are as important as response rates, researchers should probably take these results into consideration when conducting questionnaire studies among GPs.

Profile of Responders

The GPs in the postal group were a little more representative of the whole community than the Web-based group. The finding that older GPs were less prone to participate in the Web survey is well-known [2]. Interestingly, despite the conventional tendency to view the Web as rather male dominated, we also showed that women were more likely than men to participate in the Web survey; this is probably related to the demographic shift in GP populations with a higher proportion of female GPs in the younger age groups (<35 years: 72% women vs 28% men; >64 years: 15% women vs 85% men). This is also in line with recent studies that tend to show that a growing number of women take part in online research and that, with regard to gender, the so-called "digital divide" tends to disappear [4]. These findings also highlight the fact that in primary care a Web-based survey is more likely to introduce selection bias than a postal survey.

Limitations

First, only GPs with an available email address (with the exception of GPs practicing in the canton of Vaud) and those practicing in Western Switzerland and two regions in France were invited to participate; this sample could not be representative of all GPs practicing in French-speaking parts of Europe. In addition, email addresses are not regularly updated (which is less the case for postal addresses because the doctors have to use them regularly in their current practice), which could lead to a decrease in participation rates because doctors with incorrect email address may not have received the request to take part in the study. Second, the sampling procedure led to having groups that were considerably different in size. In addition, group allocation cannot be considered to be completely at random because all GPs from the canton of Vaud were allocated to the postal group due to unavailability of email addresses; this might have introduced selection bias. However, the multivariate analysis suggested this bias to be minimal because patterns of responses depended more on the format of the survey and the experience in years and level of activity of the GPs than on the location of activity. Third, only 764 GPs took part in the survey when 816 were expected. The targeted sample size could not be reached in the Web-based group, although the objective was reached for the postal group. Therefore, the study was slightly underpowered. Fourth, although the theme of our study could theoretically have an influence on participation rates, we did not think that this was the case because our results were comparable with many previous studies on various topics and populations. Fifth, the Web- and the paper-based questionnaires were similar regarding text formatting. However, we cannot entirely exclude some degree of measurement error related to minor differences between the two versions of the questionnaire. For example, responders to the four-paged paper version had to turn pages and continue on the back of the sheet, which was not the case for responders of the Web version. However, analysis of the pattern of missing responses showed that there were not significantly more missing responses for questions on the back page of the paper questionnaire compared to the Web version. Finally, we applied recommended strategies to improve response rates in the development of the paper questionnaire. The Web-based questionnaire was then created to be very similar to the paper version, but the extent to which the design of the Web-based version was equally optimal to favor response is unknown.

In conclusion, Web-based surveys lead to reduced response times, higher completeness of data, and cost savings, but postal surveys can still be considered for studies involving GPs to limit low response rates and selection bias. Mixed-mode approaches (postal and Web-based surveys) are probably a good strategy to increase GPs' participation in surveys while reducing costs. Researchers can use these mixed-mode designs in two ways: they can allow respondents to respond to the form that is the most appropriate for them or they can use a two-step strategy, including an initial postal survey followed by a Web-based reminder to nonrespondents or vice versa.

Acknowledgments

We would like to warmly thank all the doctors who participated in this study; Antoine Abhamon, our research assistant; the Unions Régionales des Professionnels de Santé (URPS) Alsace and Pays de la Loire, who were responsible for the organizational aspects of the project; and Drs Jean-Baptiste Caillard, Thomas Herault, and Fabien Rougerie for their support and assistance throughout the study.

This project was supported by institutional funding from the Faculty of Medicine, University of Geneva, and by a funding from the URPS Alsace and Pays de la Loire and from the Collège Regional des Généralistes Enseignants Alsace.

Conflicts of Interest

None declared.

References

1. Armstrong B, White E, Saracci R. Principles of Exposure Measurement in Epidemiology. Oxford: Oxford University Press; 1992.
2. Rhodes SD, Bowie DA, Hergenrather KC. Collecting behavioural data using the world wide web: considerations for researchers. J Epidemiol Community Health 2003 Jan;57(1):68-73 [FREE Full text] [Medline: 12490652]
3. Beebe TJ, Locke GR, Barnes SA, Davern ME, Anderson KJ. Mixing web and mail methods in a survey of physicians. Health Serv Res 2007 Jun;42(3 Pt 1):1219-1234 [FREE Full text] [doi: 10.1111/j.1475-6773.2006.00652.x] [Medline: 17489911]
4. Schleyer TK, Forrest JL. Methods for the design and administration of web-based surveys. J Am Med Inform Assoc 2000;7(4):416-425 [FREE Full text] [Medline: 10887169]
5. Bergeson SC, Gray J, Ehrmantraut LA, Laibson T, Hays RD. Comparing Web-based with mail survey administration of the consumer assessment of healthcare providers and systems (CAHPS(®)) clinician and group survey. Prim Health Care 2013 Apr 15;3:pii [FREE Full text] [doi: 10.4172/2167-1079.1000132] [Medline: 24078901]
6. Dykema J, Stevenson J, Day B, Sellers SL, Bonham VL. Effects of incentives and prenotification on response rates and costs in a national web survey of physicians. Eval Health Prof 2011 Dec;34(4):434-447 [FREE Full text] [doi: 10.1177/0163278711406113] [Medline: 21613242]

7. McMahon SR, Iwamoto M, Massoudi MS, Yusuf HR, Stevenson JM, David F, et al. Comparison of e-mail, fax, and postal surveys of pediatricians. Pediatrics 2003 Apr;111(4 Pt 1):e299-e303. [Medline: 12671142]
8. Basnov M, Kongsved SM, Bech P, Hjollund NH. Reliability of short form-36 in an Internet- and a pen-and-paper version. Inform Health Soc Care 2009 Jan;34(1):53-58. [doi: 10.1080/17538150902779527] [Medline: 19306199]
9. Brown JJ, Weisler RH. Posttraumatic stress disorder screening practices: a 2010 internet assessment of customary care. Prim Care Companion CNS Disord 2011;13(5):pii [FREE Full text] [doi: 10.4088/PCC.11m01171] [Medline: 22295271]
10. Clayer M, Davis A. Can the Toronto Extremity Salvage Score produce reliable results when used online? Clin Orthop Relat Res 2011 Jun;469(6):1750-1756 [FREE Full text] [doi: 10.1007/s11999-010-1715-8] [Medline: 21128031]
11. De Vera MA, Ratzlaff C, Doerfling P, Kopec J. Reliability and validity of an internet-based questionnaire measuring lifetime physical activity. Am J Epidemiol 2010 Nov 15;172(10):1190-1198 [FREE Full text] [doi: 10.1093/aje/kwq273] [Medline: 20876666]
12. Touvier M, Méjean C, Kesse-Guyot E, Pollet C, Malon A, Castetbon K, et al. Comparison between web-based and paper versions of a self-administered anthropometric questionnaire. Eur J Epidemiol 2010 May;25(5):287-296. [doi: 10.1007/s10654-010-9433-9] [Medline: 20191377]
13. Verkuil B, Brosschot JF. The online version of the Dutch Penn State Worry Questionnaire: factor structure, predictive validity and reliability. J Anxiety Disord 2012 Dec;26(8):844-848. [doi: 10.1016/j.janxdis.2012.08.002] [Medline: 23023163]
14. Elfeddali I, Bolman C, Candel MJ, Wiers RW, De Vries H. Preventing smoking relapse via Web-based computer-tailored feedback: a randomized controlled trial. J Med Internet Res 2012;14(4):e109 [FREE Full text] [doi: 10.2196/jmir.2057] [Medline: 22903145]
15. Kongsved SM, Basnov M, Holm-Christensen K, Hjollund NH. Response rate and completeness of questionnaires: a randomized study of Internet versus paper-and-pencil versions. J Med Internet Res 2007;9(3):e25 [FREE Full text] [doi: 10.2196/jmir.9.3.e25] [Medline: 17942387]
16. Sinclair M, O'Toole J, Malawaraarachchi M, Leder K. Comparison of response rates and cost-effectiveness for a community-based survey: postal, internet and telephone modes with generic or personalised recruitment approaches. BMC Med Res Methodol 2012;12:132 [FREE Full text] [doi: 10.1186/1471-2288-12-132] [Medline: 22938205]
17. Hardigan PC, Succar CT, Fleisher JM. An analysis of response rate and economic costs between mail and web-based surveys among practicing dentists: a randomized trial. J Community Health 2012 Apr;37(2):383-394. [doi: 10.1007/s10900-011-9455-6] [Medline: 21858591]
18. Kroth PJ, McPherson L, Leverence R, Pace W, Daniels E, Rhyne RL, et al. Combining web-based and mail surveys improves response rates: a PBRN study from PRIME Net. Ann Fam Med 2009;7(3):245-248 [FREE Full text] [doi: 10.1370/afm.944] [Medline: 19433842]
19. Asch SM, Kerr EA, Keesey J, Adams JL, Setodji CM, Malik S, et al. Who is at greatest risk for receiving poor-quality health care? N Engl J Med 2006 Mar 16;354(11):1147-1156. [doi: 10.1056/NEJMsa044464] [Medline: 16540615]
20. Cunningham CT, Quan H, Hemmelgarn B, Noseworthy T, Beck CA, Dixon E, et al. Exploring physician specialist response rates to web-based surveys. BMC Med Res Methodol 2015;15:32 [FREE Full text] [doi: 10.1186/s12874-015-0016-z] [Medline: 25888346]
21. Cull WL, O'Connor KG, Sharp S, Tang SS. Response rates and response bias for 50 surveys of pediatricians. Health Serv Res 2005 Feb;40(1):213-226 [FREE Full text] [doi: 10.1111/j.1475-6773.2005.00350.x] [Medline: 15663710]
22. Cummings SM, Savitz LA, Konrad TR. Reported response rates to mailed physician questionnaires. Health Serv Res 2001 Feb;35(6):1347-1355 [FREE Full text] [Medline: 11221823]
23. Edwards PJ, Roberts I, Clarke MJ, Diguiseppi C, Wentz R, Kwan I, et al. Methods to increase response to postal and electronic questionnaires. Cochrane Database Syst Rev 2009(3):MR000008. [doi: 10.1002/14651858.MR000008.pub4] [Medline: 19588449]
24. Burnham K, Anderson D. Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach. New York: Springer; 2002.
25. Cohidon C, Cornuz J, Senn N. Primary care in Switzerland: evolution of physicians' profile and activities in twenty years (1993-2012). BMC Fam Pract 2015;16:107 [FREE Full text] [doi: 10.1186/s12875-015-0321-y] [Medline: 26292762]
26. Le Breton-Lerouvillois G. La Démographie Médicale en Région Pays-de-la-Loire: Situation en 2013. Paris: Conseil National de l'Ordre des Médecins; 2013. URL: https://www.conseil-national.medecin.fr/sites/default/files/pays_de_la_loire_2013_0.pdf [accessed 2016-12-08] [WebCite Cache ID 6mbGZvDhV]
27. McLean SA, Feldman JA. The impact of changes in HCFA documentation requirements on academic emergency medicine: results of a physician survey. Acad Emerg Med 2001 Sep;8(9):880-885 [FREE Full text] [Medline: 11535480]
28. Potts HW, Wyatt JC. Survey of doctors' experience of patients using the Internet. J Med Internet Res 2002;4(1):e5 [FREE Full text] [doi: 10.2196/jmir.4.1.e5] [Medline: 11956037]
29. Akl EA, Maroun N, Klocke RA, Montori V, Schünemann HJ. Electronic mail was not better than postal mail for surveying residents and faculty. J Clin Epidemiol 2005 Apr;58(4):425-429. [doi: 10.1016/j.jclinepi.2004.10.006] [Medline: 15862729]

Abbreviations

GP: general practitioner

Edited by G Eysenbach; submitted 02.07.16; peer-reviewed by J Ziegenfuss, D Ruijter; comments to author 17.11.16; revised version received 15.12.16; accepted 22.12.16; published 22.03.17.

Please cite as:
Sebo P, Maisonneuve H, Cerutti B, Fournier JP, Senn N, Haller DM
Rates, Delays, and Completeness of General Practitioners' Responses to a Postal Versus Web-Based Survey: A Randomized Trial
J Med Internet Res 2017;19(3):e83
URL: http://www.jmir.org/2017/3/e83/
doi:10.2196/jmir.6308
PMID:28330830

Original Paper

Behavioral Indicators on a Mobile Sensing Platform Predict Clinically Validated Psychiatric Symptoms of Mood and Anxiety Disorders

Skyler Place[1], PhD; Danielle Blanch-Hartigan[2], PhD, MPH; Channah Rubin[1], BA; Cristina Gorrostieta[1], PhD; Caroline Mead[1], BA; John Kane[1], PhD; Brian P Marx[3,4], PhD; Joshua Feast[1], MBA; Thilo Deckersbach[5], PhD; Alex "Sandy" Pentland[6], PhD; Andrew Nierenberg[5], MD; Ali Azarbayejani[1], PhD

[1]Cogito Corporation, Boston, MA, United States

[2]Department of Natural and Applied Sciences, Bentley University, Waltham, MA, United States

[3]VA Boston Healthcare System, National Center for Posttraumatic Stress Disorder, Boston, MA, United States

[4]Boston University School of Medicine, Boston, MA, United States

[5]Massachusetts General Hospital, Harvard Medical School, Bipolar Clinic and Research Program, Department of Psychiatry, Boston, MA, United States

[6]Massachusetts Institute of Technology, Media Lab, Cambridge, MA, United States

Corresponding Author:
Danielle Blanch-Hartigan, PhD, MPH
Department of Natural and Applied Sciences
Bentley University
175 Forest Street
106 Jennison
Waltham, MA, 02452
United States
Phone: 1 7818912066
Fax: 1 7818912066
Email: danielleblanch@gmail.com

Abstract

Background: There is a critical need for real-time tracking of behavioral indicators of mental disorders. Mobile sensing platforms that objectively and noninvasively collect, store, and analyze behavioral indicators have not yet been clinically validated or scalable.

Objective: The aim of our study was to report on models of clinical symptoms for post-traumatic stress disorder (PTSD) and depression derived from a scalable mobile sensing platform.

Methods: A total of 73 participants (67% [49/73] male, 48% [35/73] non-Hispanic white, 33% [24/73] veteran status) who reported at least one symptom of PTSD or depression completed a 12-week field trial. Behavioral indicators were collected through the noninvasive mobile sensing platform on participants' mobile phones. Clinical symptoms were measured through validated clinical interviews with a licensed clinical social worker. A combination hypothesis and data-driven approach was used to derive key features for modeling symptoms, including the sum of outgoing calls, count of unique numbers texted, absolute distance traveled, dynamic variation of the voice, speaking rate, and voice quality. Participants also reported ease of use and data sharing concerns.

Results: Behavioral indicators predicted clinically assessed symptoms of depression and PTSD (cross-validated area under the curve [AUC] for depressed mood=.74, fatigue=.56, interest in activities=.75, and social connectedness=.83). Participants reported comfort sharing individual data with physicians (Mean 3.08, SD 1.22), mental health providers (Mean 3.25, SD 1.39), and medical researchers (Mean 3.03, SD 1.36).

Conclusions: Behavioral indicators passively collected through a mobile sensing platform predicted symptoms of depression and PTSD. The use of mobile sensing platforms can provide clinically validated behavioral indicators in real time; however, further validation of these models and this platform in large clinical samples is needed.

(J Med Internet Res 2017;19(3):e75) doi:10.2196/jmir.6678

KEYWORDS

mHealth; post-traumatic stress disorders; depression; behavioral symptoms

Introduction

Exploring behavioral patterns has tremendous potential to aid clinicians and patients in the real-time recognition and treatment of symptoms and symptom clusters across a variety of disease states. Early recognition of subsyndromal mood and anxiety symptoms is crucial to reduce the pernicious impact of chronic psychological distress and loss of function [1-3]. Yet many affected individuals are unable to receive timely and adequate mental health resources [4]. A key objective of the National Institute of Mental Health's strategic plan is to identify clinically useful behavioral indicators to determine effective intervention strategies and deliver those interventions at the appropriate time in the illness trajectory [5].

Barriers exist to objectively identifying behaviors in real time. Mobile sensing platforms allow for not only the collection of these behavioral indicators, but also can provide the complex architecture for securely storing, analyzing, and providing feedback. More specifically, mobile sensing platforms can enable the identification and tracking of behaviors (eg, mood, fatigue, social connectedness, physical isolation) from digital trace data passively collected from sensors embedded on mobile devices (eg, call logs, global positioning system [GPS] meta-data, phone activity) [6]. However, this approach still needs to be clinically validated and shown to be scalable; the application of mobile sensing platforms in clinical care is currently limited. A National Institutes of Health (NIH) mHealth evidence workshop cautioned that, despite the promise of these new technologies in clinical care, theoretically based empirical evidence is needed [7]. In this paper, we present a field trial of the implementation of a mobile sensing platform to provide clinically validated behavioral indicators of symptoms of depression and PTSD.

Digital trace data has been used to track and predict behavioral outcomes (eg, GPS data to predict friendship networks amongst graduate students [8] and location and proximity data to predict changes in an individual's physical health [9]). Mobile sensing platforms using mobile phone trace data have shown some success in predicting self-reported depression [10,11]; however, they have focused on diagnosis and not on tracking specific symptoms.

In addition to digital trace data, the analysis of cues from the voice as a behavioral indicator is another tremendous potential of mobile sensing platforms. Vocal cues can provide valuable insights into physical and mental states, not just through lexical content but also through prosody, voice quality, and overall tone of voice. Analysis of vocal cues can be useful for inferring emotional states [12,13]; for example, various aspects of vocal prosody have been linked to depression severity [14-16]. Vocal cue analysis has recently been supported in a clinical use-case for detecting and telemonitoring symptom progression in Parkinson's disease [17-19]. However, the potential of audio processing has been largely understudied for clinical use cases outside of Parkinson's disease. Human experts can identify changes in speech properties of depressed individuals [20] and recent efforts have been made to quantify these vocal features [21].

There is potential to complement or replace traditional methods of assessing mental health and well-being that rely upon an individual's self-report (ie, surveys, questionnaires, diagnostic interviews) with passively collected indicators derived from behavioral indicators and vocal cues [22,23]. Clinicians would not have to rely solely on subjective retrospective self-reports of symptoms during clinical visits that are subject to recall and other response biases or symptom tracking by self-report that has low levels of sustained adherence.

Identifying and tracking behavioral indicators also has clinical relevance beyond diagnosis. These behavioral indicators, even if validated against a specific clinical symptom, can provide information to clinicians across a number of chronic or comorbid conditions. Patients can have complex profiles of symptoms and underlying behavioral patterns that can be present in a number of different conditions. The Research Domain Criteria (RDoC) framework emphasizes the clinical value for considering underlying behaviors and thinking across categories of mental health disorders [24]. Having access to objectively measured behavioral indicators can help subtype mood and anxiety disorders and recognize patterns in individual patients.

The goal of this project was to test the ability of a scalable mobile sensing platform to collect, store, and analyze objective behavioral indicators underlying mood and anxiety disorder symptoms. Specifically, we hypothesized that models of behavioral indicators of PTSD and depression symptoms could be derived from mobile-based digital trace data and auditory signals.

Methods

Participants

Participants were recruited through Web-based advertisements and local veteran service organizations. Inclusion criteria included: age 18 years or older, English speaker, current subscriber to a cellular and data plan, and if they reported at least one symptom of PTSD or depression (measured by Primary Care PTSD Screen, PC-PTSD [25] and the Patient Health Questionnaire, PHQ-2 [26]). Participants were excluded if they had plans to permanently leave the Boston area during the study period, if they shared their cell phone with another person, or if they were active duty military personnel. The protocol was approved by an accredited institutional review board, and all participants were consented. Additional participant and safety details are included in the Multimedia Appendix 1.

Procedure

Each participant had an initial study visit, 12 weeks of data gathering, and a follow-up study visit. At the initial visit, participants completed a baseline questionnaire and replaced their existing phone with the study mobile phone, equipped with

the research app. At the time of the study, the app was limited to a small number of devices so participants transferred their cellular plan and all personal cell phone data to the study Android device. The mobile app has since been made compatible across devices.

The participant used the phone as usual during the study period. In addition, participants were instructed to leave an audio diary entry in the app at least weekly. The audio recordings were designed to be short "voicemail" style entries about how participants were feeling or how their days were going. At the end of the 12-week period, participants returned to the study site and completed a semistructured clinical interview with a trained clinician. In addition, participants completed a close-out survey, were debriefed, and the mobile app and all study data were removed from the phone. Participants were paid US $15/hour for study visits, a 1-time US $50 incentive for completing at least 75% of all assessments, and given the option of keeping the study phone (valued at US $300).

Mobile Sensing Platform Architecture

Digital trace data (Table 1) were collected on intermittent fixed schedules from predefined, configured probes built into the phone's operating system. The mobile software data gathering and analysis platform was developed by Cogito with funding from the Defense Advance Research Project Agency (DARPA). Details of the platform architecture are in the Multimedia Appendix 1.

Table 1. Categories of digital trace data.

Category	Description
Activity	The physical handling of the phone. The phone hardware includes an accelerometer and gyroscope. The gyroscope was used to determine angle, velocity, direction, and acceleration of the phone. The accelerometer provided data on rate of rotation on the X, Y, and Z axes.
Social	How the user is interacting with others through the phone. Collected the time and deidentified descriptor (see privacy and security section) of all outgoing and ingoing phone calls and SMS (short message service, SMS) text messages.
Location	Where the phone is physically located. Data are combined by the phone operating system from global positioning service (GPS), mobile phone tower triangulation, and WiFi network locations. These data consist of time stamp, longitude, and latitude readings. WiFi network names, WiFi access, usernames, or passwords were not collected.
Device interaction	When the phone is being used. Data are timestamps of when the phone screen is turned on or activated and when it is turned off. We did not record app usage, keystrokes, or any other measure of active use of the device.
Device information	Data describing the physical device. This included variables on phone make, model, battery status (% full), and phone operating system version. This data is used for quality assurance testing.
Vocal cues	Digital recordings of audio-diaries were processed to extract measurements related to speaking, rate, prosody, intonation, and voice quality. These measurements were computed on short-term overlapping frames and then aggregated (using descriptive statistics) over the entire audio diary entry. The lexical content of the recordings was not analyzed.

Data Privacy and Security

Participants' data were protected by a NIH Certificate of Confidentiality, and utilized best practice technical approaches to protect identity. This included both the identity of the participant and the identity of individuals in contact, and communication log data. Encrypted, hashed, deidentified descriptors were used to label all data. Details on the security methodology and algorithms are in the Multimedia Appendix 1.

Measures

Main Outcomes

Clinical symptoms were measured through validated clinical interviews with a trained clinician using the depression and PTSD modules of the Structured Clinical Interview for Mental Disorders (SCID) [27]. We focused our models on 4 behaviorally relevant symptoms (Depression A1—depressed mood most of the day; Depression A2—diminished interest or pleasure in all or most activities; Depression A6—fatigue or loss of energy; and PTSD C2—avoids activities, places, people).

Digital Trace Data

Digital trace data collected on the mobile sensing platform fell into 6 categories (Table 1). Details on the schedule and quantity of digital trace data gathered are in the Multimedia Appendix 1.

Additional Measures

Participants reported age, race, current annual income, employment, level of education, marital status, and veteran status. At the end of the study participants also reported ease of use, interest in future use, and their perception of data sharing and privacy concerns for both personal and anonymized health data. All questions were 1 (very unlikely) to 5 (very likely) Likert questions. Survey instruments are in Multimedia Appendix 1.

Analysis

The primary analysis goal was to develop models utilizing digital trace data and audio recordings to predict the presence of clinician-assessed psychiatric symptoms. A secondary goal was to have a model whose inputs reflected types and quantities of data that could be realistically gathered in real world clinical environments with minimal participant burden. A third goal to facilitate clinical utilization and acceptability by clinicians was to develop models with a hypothesis-driven approach from data sources which were clinically interpretable.

Three symptoms of depression and PTSD with strong behavioral components were chosen a priori as target goals for model

development from digital trace data (fatigue, interest in activities, and social connectedness). One symptom of depression (depressed mood) was chosen a priori as the target goal for model development from audio data. For each symptom, a binary target variable was defined to characterize the presence or absence of the symptom. To best model clinically relevant differences in symptomatology, we removed participants with only subthreshold symptoms or with insufficient information to determine symptom presence or absence.

Participant audio and mobile features were derived using the most recent week of data (1 week of data prior to the SCID symptom assessment). An initial modeling approach predicted symptoms based on the entire 12-week data collection period; however, high discrimination of the classes was achieved using features temporally closer to the target assessment.

For the 3 models trained from digital trace data, we used the following methodology. Through data exploration and a priori hypotheses on how mobile phone data would capture relevant behaviors, we created a set of features using descriptive statistics over a week of digital trace data. These features included means, counts, and standard deviations of social and location digital trace data (number of texts received in a week, number of minutes spent on outbound calls in the last week). Activity data (see Table 1) was excluded after extensive data exploration failed to yield conceptually meaningful features. Device interaction data (see Table 1) was excluded as data exploration determined that events were being captured for push notifications, not just user-device interactions; therefore, the interaction data did not accurately represent user behavior. The set of features were further reduced by cluster analysis to enhance interpretability on the resulting models and improve prediction performance. The set of 14 features were reduced to prevent over-fitting and interfeature correlation. Reduction method consisted of identifying clusters of features such that within group correlation is high with respect to between group correlation, which highlighted 3 clusters that conceptually matched a digital trace data source (location, social calls, social texts). From each grouping, a single feature was chosen, that conceptually matched a digital trace data source (location, social calls, social texts). Combinations of features were utilized to create 10 candidate models per symptom. Candidate models consisted of logistic models describing the probability of having the corresponding symptom. To select the best model from the list of candidates, we utilized 10-fold cross validation. The model with the highest cross-validated area under the ROC curve (AUC) was chosen (Table 2).

Table 2. Model characteristics and performance.

Symptom target	Input features	Cross-validated area under the curve (AUC)
Depressed mood most of the day	MeanPitchVar+MeanVocalEffort+MeanVocalEffort:MeanPitchVar	.74
Diminished interest or pleasure in all or most activities	sms.address.count+travel.distance.sum	.56
Fatigue or loss of energy	call.out.sum+sms.address.count	.75
Avoid activities, places, people	call.out.sum+sms.address.count+(call.out.sum)(sms.address.count)	.83

One model was trained using audio data. The audio was sampled at 8 kHz with 16-bit precision. M4A compression was applied to audio captured by the device and the format was subsequently converted to waveform audio file format (WAV) before computing audio features. Candidate audio features to develop this model were defined by an existing library of features providing temporal information about voice quality, prosody, and intonation. Audio features were aggregated over time as an average and standard deviation per recording. A penalized logistic regression approach, least absolute shrinkage and selection operator (LASSO) [28], was used to reduce the initial set of features using the depressed mood symptom target as the dependent variable. The penalization parameter was determined by 10-fold cross validation. Five candidate features remained following this reduction strategy. Candidate models were created in the same fashion as models created from digital trace data, and the model with the highest cross-validated AUC following 10-fold cross-validation was selected.

Details on the set of reduced features for each model, representative examples from the published literature which include audio feature algorithms, and the full list of candidate models are in the Multimedia Appendix 1.

Results

Participant Characteristics

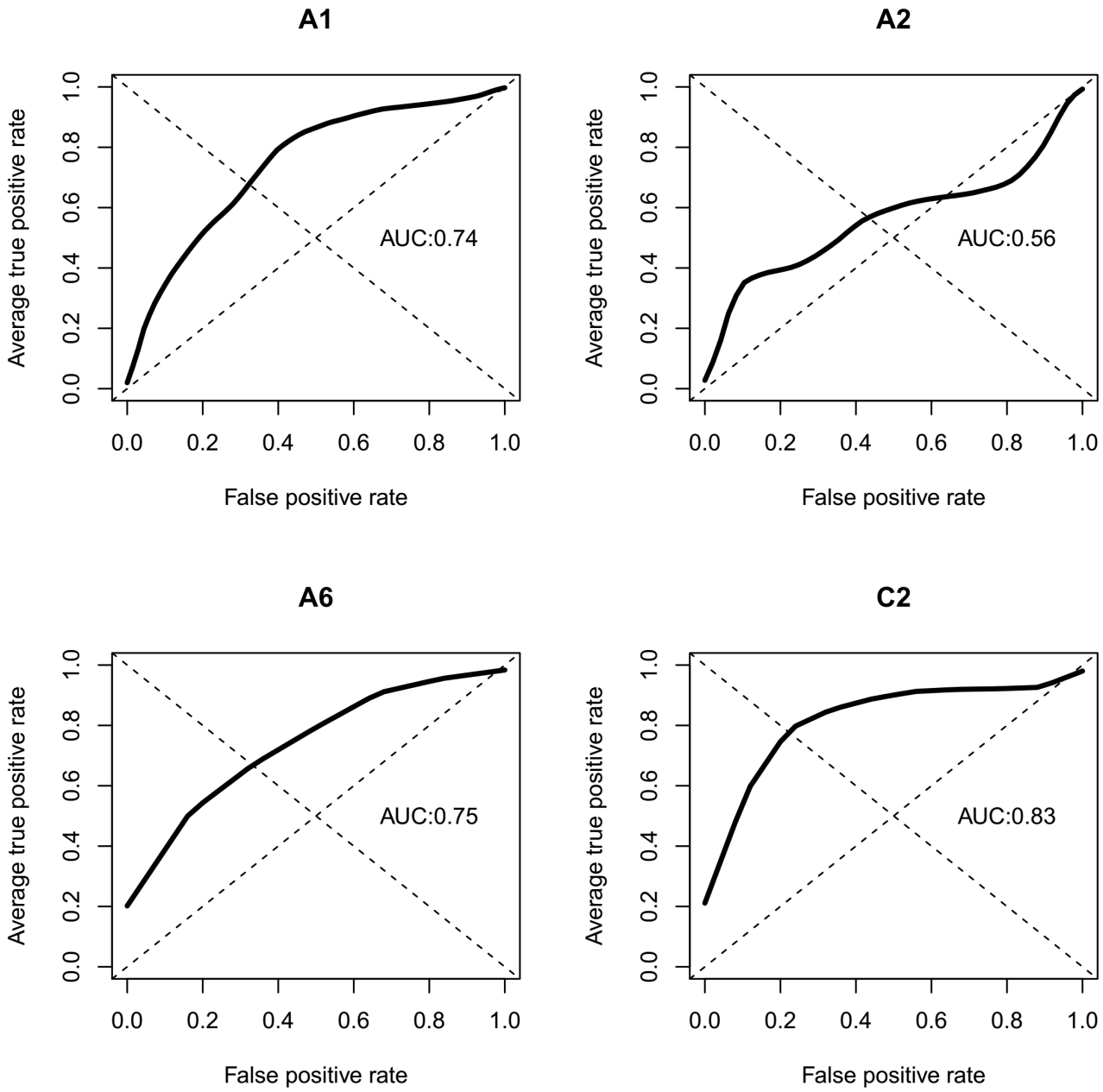
Ninety-five participants consented to the study. One participant called to withdraw and 21 were withdrawn due to technical issues (most commonly a lost phone). Seventy-three participants (67% male, 48% non-Hispanic white, 33% veterans) completed all 12-weeks of data collection and a follow-up clinical interview (see Multimedia Appendix 1 for detailed participant characteristics). There was no difference between participants who completed the study and those that did not in terms of demographic characteristics. All participants reported a lifetime history of at least one traumatic event during clinical interview. The 12-week dataset included 1217 participant audio recordings and 51,080,131 digital trace data points.

Models and Model Fit

The resulting models for each of the 4 dependent variable modeling targets are in Table 2. See Figure 1 for receiver operating characteristic (ROC) curves. Detailed model performance, including positive and negative predictive values and accuracy at varying thresholds, are included in the Multimedia Appendix 1. At a .5 threshold, models predicted

clinically assessed symptoms of depression and PTSD (cross-validated AUCs for depressed mood .74, fatigue .56, interest in activities .75, and social connectedness .83). No information on prior diagnoses, current comorbidities, lifestyle, or demographic characteristics was included in the models to achieve these AUCs.

Figure 1. Receiver operating characteristic curves (ROC) and area under the curve (AUC).



Acceptability and Feasibility

Adherence to study protocol was extremely high with 96% of participants (N=70) completing at least one audio diary per week. Participants found the mobile app easy to use (Mean 4.05, SD 1.14) and were interested in using the app in the future (Mean 3.59, SD 1.27). Overall participants reported moderate comfort with sharing individual data (Mean 2.48, SD 0.91) and anonymized data (Mean 3.00, SD 1.23). There were no significant differences by age, gender, or veteran status in comfort with sharing. For individualized data, participants were most comfortable with sharing their personal data with primary care physicians (Mean 3.08, SD 1.22), mental health providers (Mean 3.25, SD 1.39), and medical researchers (Mean 3.03, SD 1.36). Participants reported the least comfort with sharing individualized data with insurance providers (Mean 1.63, SD 1.06) and friends (Mean 1.95, SD 1.12).

Discussion

Principal Findings

In this study, we examined the use of a mobile sensing platform to passively collect and analyze digital trace and voice data as behavioral indicators of clinically-validated symptoms of

depression and PTSD. The models were predictive of clinician-assessed symptoms of depressed mood, fatigue, interest in activities, and social connectedness. This study highlights the power of mobile sensing platforms to gather and compute behavioral indicators of symptoms of mood and anxiety disorders.

To have clinical relevance, modeling strategies for behavioral indicators should be parsimonious (simple), objective (unbiased), and actionable (useful in real-world clinical settings). The modeling strategy in the present analysis was parsimonious, distilling millions of raw data points into a small number of key features. The models were objective, utilizing passively collected digital trace data, and extracted vocal features of tonality and speaking style. Adding self-reported predictors to the models reported here did not significantly increase model performance. There are established biases in self-reporting of mental states [29]. Biases may be intentional or unintentional, as the disorders themselves can change the way patients perceive their own actions. Finally, the models were actionable. Predictive models that have been developed for other disorders, including suicidality, rely on social determinants and clinical characteristics that have to be extracted from electronic medical records [30]. Clinicians may want to know what is happening with patients, but asking patients may not always yield accurate results [31,32].

In addition, these behavioral indicators described by the models are clinically meaningful. The models are based off of clinician-assessed symptoms and not self-report instruments. These models include behavioral indicators that can capture clinically relevant levels of impairment. For example, the models are not concerned with whether a patient walked 9000 steps versus 10,000 steps, but if the patient has not left his house in the past week. The models use macro level behavioral features as inputs, such as distance traveled over a week, to predict symptoms that both patients and clinicians can understand and potentially intervene upon. The present models do not assume a causal link between the behavioral indicators and clinical symptoms; nevertheless, the models may be of interest to clinicians and patients regardless of their causal pathways, as a means to track and alert to changes in symptoms in real time. These models are not designed to replace clinical decision-making or diagnose a particular mental health condition, but rather to assess a number of behavioral indicators underlying mood and anxiety disorders. These models are a resource to provide clinically relevant behavioral information to augment clinical care.

This study also demonstrates the feasibility and scalability of mobile sensing platforms for capturing, sharing, and analyzing behavioral data. Participants reported comfort with sharing even personal data with clinicians and medical researchers. They were willing to use the app in the future and reported that it did not change their behavior. Future research should continue to explore feasibility and acceptability of mobile sensing platforms in different populations and how this technology influences patients, clinicians, and their interactions.

Clinical Relevance

The behavioral indicator algorithms created in this study, when embedded within a high performance, highly scalable, cloud computing architecture has distinct advantages in clinical care. The models described in this study utilize only 1 week of data to make predictions, and when implemented in high-performance code can automatically provide new values on patient behavior daily. Continuous monitoring can allow patients, clinicians, or researchers to view changes and symptom trajectories over time. Passive data collection does not burden patients and is free from self-report subject to biases. Current mHealth approaches for mental health assessment, including mobile phone apps and ecological momentary assessment methods, may be an improvement over paper and pencil measures, but still require direct patient input [22,33] and many have not been clinically validated [34]. The use of mobile sensing platforms could allow clinicians and researchers to track episode onset, symptom progression, and relapse across populations with less patient burden [35]. Patients, caregivers, or clinicians could be alerted to clinically meaningful changes in behavioral indicators between clinical interactions.

Currently, there is no real-time, continuous, and objective solution to collect, analyze, and track mental health symptom-related behavioral indicators. The mobile sensing approach validated in this study streamlined the process by which clinically-relevant indicators are gathered and transmitted. Each year, approximately 6.9% of all US adults experience major depressive disorder [36] and PTSD prevalence is estimated at 6.8% over a lifetime [37]. The rates of mood and anxiety disorders are even higher among veterans. Although estimates vary by study and service era, current prevalence of depression in veterans of Iraq and Afghanistan conflicts is estimated at 13.8% [38] and lifetime prevalence of PTSD is estimated at 30.9% for male veterans and 26.9% for female veterans [39]. This innovative mobile sensing platform offers a scalable approach to care for millions with mood and anxiety disorders.

This study sample is larger than in previous studies of mobile sensing platforms and weighted to prevalence estimates; however, it is a convenience sample from the community and, subsequently, results may lack generalizability. Although the study demonstrates proof of concept, additional validation of these models is needed.

The use of this innovative approach could empower patient health self-management. Recognizing warning signs and changes in behavioral patterns can be difficult for patients with mood and anxiety disorders, particularly those without easy access to care or those experiencing symptoms near episode onset. Disability and behavior are closely linked and mental disorders represent a leading cause of disability. Developing approaches to identify functional disability that do not rely on self-report is an important extension to this work. Patients who are more aware of their own behavioral patterns, particularly between or before clinical appointments, maybe are better able to manage their own health. A mobile sensing platform could be capable of providing real-time feedback to patients and clinicians, enabling increased help seeking behavior and access to needed

care. Patients and providers can use behavioral data to recognize more objectively and definitively when they need to seek additional services. These additional novel metrics may enable clinicians to provide better informed clinical care. The efficacy of such a platform in improving clinical care remains to be tested.

Acknowledgments

This material is based upon work supported by the Defense Advance Research Project Agency (DARPA) and Space and Naval Warfare Systems Center Pacific under Contract No. N66001-11-C-4094. The views, opinions, or findings contained in this material are those of the authors and should not be interpreted as representing the official views or policies of the US Department of Defense or the US Government. The authors thank DARPA Prime, BBN-Raytheon, and in particular our contact for the DARPA grant, Michael Crystal (BBN-Raytheon, no compensation received from Cogito Corporation), for their support throughout this project. Additional thanks to Bonnie Zimmer, MSW LICSW (paid consultant to Cogito Corporation, clinical social worker in private practice) and Ann Stewart, LICSW (The Red Sox Foundation and Massachusetts General Hospital, no compensation received from Cogito Corporation) for clinical support and guidance.

Conflicts of Interest

Skyler Place, Channah Rubin, Cristina Gorrostieta, Caroline Mead, John Kane, Joshua Feast, and Ali Azarbayejani are paid employees of Cogito Corporation. Sandy Pentland is a cofounder of Cogito Corporation. Danielle Blanch-Hartigan is a paid consultant to Cogito Corporation. Brian Marx was formally a paid consultant to Cogito Corporation. Andrew Nierenberg and Thilo Deckersbach are subcontractors and coinvestigators on a series of government and industry funded projects to Cogito Corporation. Cristina Gorrostieta (Cogito Corporation) conducted all the data analyses. She had full access to all of the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis.

Multimedia Appendix 1

Supplemental material: eMethods: Participants - Mobile sensing platform architecture - Data security & privacy - Digital trace data - Survey instrument; eTables: Features Tables - Audio Feature Definitions - Candidate Models - Participant Characteristics - Detailed model performance.

[PDF File (Adobe PDF File), 1MB - jmir_v19i3e75_app1.pdf]

References

1. Zajecka JM. Residual symptoms and relapse: mood, cognitive symptoms, and sleep disturbances. J Clin Psychiatry 2013;74(Suppl 2):9-13 [FREE Full text] [doi: 10.4088/JCP.12084su1c.02] [Medline: 24191972]
2. Judd LL, Akiskal HS, Maser JD, Zeller PJ, Endicott J, Coryell W, et al. A prospective 12-year study of subsyndromal and syndromal depressive symptoms in unipolar major depressive disorders. Arch Gen Psychiatry 1998 Aug;55(8):694-700. [Medline: 9707379]
3. Judd LL, Paulus MJ, Schettler PJ, Akiskal HS, Endicott J, Leon AC, et al. Does incomplete recovery from first lifetime major depressive episode herald a chronic course of illness? Am J Psychiatry 2000 Sep;157(9):1501-1504. [doi: 10.1176/appi.ajp.157.9.1501] [Medline: 10964869]
4. Gunn WB, Blount A. Primary care mental health: a new frontier for psychology. J Clin Psychol 2009 Mar;65(3):235-252. [doi: 10.1002/jclp.20499] [Medline: 19156779]
5. NIMH.NIH. 2015. National Institute of Mental Health Strategic Plan for Research URL: https://www.nimh.nih.gov/about/strategic-planning-reports/nimh_strategicplanforresearch_508compliant_corrected_final_149979.pdf [accessed 2017-02-24] [WebCite Cache ID 6oWLMwu0o]
6. Harari GM, Lane ND, Wang R, Crosier BS, Campbell AT, Gosling SD. Using smartphones to collect behavioral data in psychological science: opportunities, practical considerations, and challenges. Perspect Psychol Sci 2016 Nov;11(6):838-854. [doi: 10.1177/1745691616650285] [Medline: 27899727]
7. Kumar S, Nilsen WJ, Abernethy A, Atienza A, Patrick K, Pavel M, et al. Mobile health technology evaluation: the mHealth evidence workshop. Am J Prev Med 2013 Aug;45(2):228-236 [FREE Full text] [doi: 10.1016/j.amepre.2013.03.017] [Medline: 23867031]
8. Eagle N, Pentland AS, Lazer D. Inferring friendship network structure by using mobile phone data. Proc Natl Acad Sci U S A 2009 Sep 8;106(36):15274-15278 [FREE Full text] [doi: 10.1073/pnas.0900282106] [Medline: 19706491]
9. Madan A, Cebrian M, Lazer D, Pentland A. Social sensing for epidemiological behavior change. 2010 Presented at: 12th ACM international conference on Ubiquitous computing; 2010; Copenhagen, Denmark p. 291-300. [doi: 10.1145/1864349.1864394]

10. Saeb S, Zhang M, Karr CJ, Schueller SM, Corden ME, Kording KP, et al. Mobile phone sensor correlates of depressive symptom severity in daily-life behavior: an exploratory study. J Med Internet Res 2015;17(7):e175 [FREE Full text] [doi: 10.2196/jmir.4273] [Medline: 26180009]
11. Ben-Zeev D, Scherer EA, Wang R, Xie H, Campbell AT. Next-generation psychiatric assessment: using smartphone sensors to monitor behavior and mental health. Psychiatr Rehabil J 2015 Sep;38(3):218-226 [FREE Full text] [doi: 10.1037/prj0000130] [Medline: 25844912]
12. Tawari A, Trivedi MM. Speech emotion analysis: exploring the role of context. IEEE Trans Multimedia 2010 Oct;12(6):502-509. [doi: 10.1109/TMM.2010.2058095]
13. Scherer S, Schwenker F, Palm G. Classifier fusion for emotion recognition from speech. In: Advanced intelligent environments. US: Springer; 2009:95-117.
14. France DJ, Shiavi RG, Silverman S, Silverman M, Wilkes DM. Acoustical properties of speech as indicators of depression and suicidal risk. IEEE Trans Biomed Eng 2000 Jul;47(7):829-837. [doi: 10.1109/10.846676] [Medline: 10916253]
15. Yang Y, Fairbairn C, Cohn JF. Detecting depression severity from vocal prosody. IEEE Trans Affect Comput 2013;4(2):142-150 [FREE Full text] [doi: 10.1109/T-AFFC.2012.38] [Medline: 26985326]
16. Schwenker F, Scherer S, Magdi Y, Palm G. The GMM-SVM supervector approach for the recognition of the emotional status from speech. In: Artificial Neural Networks - ICANN 2009: 19th International Conference, Limassol, Cyprus, September 14-17, 2009, Proceedings, Part I. Berlin Heidelberg: Springer; 2009:894-903.
17. Tsanas A, Little MA, McSharry PE, Ramig LO. Accurate telemonitoring of Parkinson's disease progression by noninvasive speech tests. IEEE Trans Biomed Eng 2010 Apr;57(4):884-893. [doi: 10.1109/TBME.2009.2036000] [Medline: 19932995]
18. Tsanas A, Little MA, McSharry PE, Spielman J, Ramig LO. Novel speech signal processing algorithms for high-accuracy classification of Parkinson's disease. IEEE Trans Biomed Eng 2012 May;59(5):1264-1271. [doi: 10.1109/TBME.2012.2183367] [Medline: 22249592]
19. Arora S, Venkataraman V, Zhan A, Donohue S, Biglan KM, Dorsey ER, et al. Detecting and monitoring the symptoms of Parkinson's disease using smartphones: a pilot study. Parkinsonism Relat Disord 2015 Jun;21(6):650-653. [doi: 10.1016/j.parkreldis.2015.02.026] [Medline: 25819808]
20. Darby JK, Hollien H. [Vocal and speech patterns of depressive patients]. Folia Phoniatr (Basel) 1977;29(4):279-291. [Medline: 604242]
21. Mundt JC, Vogel AP, Feltner DE, Lenderking WR. Vocal acoustic biomarkers of depression severity and treatment response. Biol Psychiatry 2012 Oct 1;72(7):580-587 [FREE Full text] [doi: 10.1016/j.biopsych.2012.03.015] [Medline: 22541039]
22. Onnela J, Rauch SL. Harnessing smartphone-based digital phenotyping to enhance behavioral and mental health. Neuropsychopharmacology. Neuropsychopharmacology 2016 Jun;41(7):1691-1696. [doi: 10.1038/npp.2016.7] [Medline: 26818126]
23. Torous J, Kiang MV, Lorme J, Onnela J. New tools for new research in psychiatry: a scalable and customizable platform to empower data driven smartphone research. JMIR Ment Health 2016 May 05;3(2):e16 [FREE Full text] [doi: 10.2196/mental.5165] [Medline: 27150677]
24. Insel TR. The NIMH research domain criteria (RDoC) project: precision medicine for psychiatry. Am J Psychiatry 2014 Apr;171(4):395-397. [doi: 10.1176/appi.ajp.2014.14020138] [Medline: 24687194]
25. Prins A, Ouimette P, Kimerling R, Camerond RP, Hugelshofer DS, Shaw-Hegwer J, et al. The primary care PTSD screen (PC-PTSD): development and operating characteristics. Primary Care Psychiatry 2003;9(1):9-14. [doi: 10.1185/135525703125002360]
26. Kroenke K, Spitzer RL, Williams JB. The Patient Health Questionnaire-2: validity of a two-item depression screener. Med Care 2003 Nov;41(11):1284-1292. [doi: 10.1097/01.MLR.0000093487.78664.3C] [Medline: 14583691]
27. First M, Spitzer R, Gibbon M, Williams J. User's guide for the structured clinical interview for DSM-IV personality disorders (SCID-II). Washington, DC: American Psychiatric Association; 1996:1-80.
28. Tibshirani R. Regression shrinkage and selection via the lasso. J R Stat Soc Series B Stat Methodol 1996;58(1):267-288 [FREE Full text]
29. Prusoff BA, Klerman GL, Paykel ES. Concordance between clinical assessments and patients' self-report in depression. Arch Gen Psychiatry 1972 Jun;26(6):546-552. [Medline: 5027118]
30. McCarthy JF, Bossarte RM, Katz IR, Thompson C, Kemp J, Hannemann CM, et al. Predictive modeling and concentration of the risk of suicide: implications for preventive interventions in the US Department of Veterans Affairs. Am J Public Health 2015 Sep;105(9):1935-1942. [doi: 10.2105/AJPH.2015.302737] [Medline: 26066914]
31. Van de Mortel TF. Faking it: social desirability response bias in self-report research. Aust J Adv Nurs 2008;25(4):40.
32. Linaker OM, Moe A. The COOP/WONCA charts in an acute psychiatric ward. Validity and reliability of patients' self-report of functioning. Nord J Psychiatry 2005;59(2):121-126. [doi: 10.1080/08039480510022918] [Medline: 16195109]
33. Moore RC, Depp CA, Wetherell JL, Lenze EJ. Ecological momentary assessment versus standard assessment instruments for measuring mindfulness, depressed mood, and anxiety among older adults. J Psychiatr Res 2016 Apr;75:116-123. [doi: 10.1016/j.jpsychires.2016.01.011] [Medline: 26851494]
34. Torous J, Levin ME, Ahern DK, Oser ML. Cognitive behavioral mobile applications: clinical studies, marketplace overview, and research agenda. Cogn Behav Pract 2016 Jun:-. [doi: 10.1016/j.cbpra.2016.05.007]

35. Torous J, Staples P, Onnela J. Realizing the potential of mobile mental health: new methods for new data in psychiatry. Curr Psychiatry Rep 2015 Aug;17(8):602 [FREE Full text] [doi: 10.1007/s11920-015-0602-0] [Medline: 26073363]
36. Substance Abuse and Mental Health Services Administration (SAMHSA). SAMHSA. Rockville, MD: Substance Abuse and Mental Health Services Administration; 2013. Results from the 2012 national survey on drug use and health: Summary of national findings URL: https://www.samhsa.gov/data/sites/default/files/NSDUHresults2012/NSDUHresults2012.pdf [accessed 2017-02-24] [WebCite Cache ID 6oWNzjlLb]
37. Kessler RC, Berglund P, Demler O, Jin R, Merikangas KR, Walters EE. Lifetime prevalence and age-of-onset distributions of DSM-IV disorders in the National Comorbidity Survey Replication. Arch Gen Psychiatry 2005 Jun;62(6):593-602. [doi: 10.1001/archpsyc.62.6.593] [Medline: 15939837]
38. Tanielian TL, Jaycox L. Invisible wounds of war psychological and cognitive injuries, their consequences, and services to assist recovery. Santa Monica, CA: RAND; 2008.
39. Kulka R, Schlenger W, Fairbank J, Hough R, Jordan B, Marmar C. Trauma and the Vietnam war generation: Report of findings from the National Vietnam Veterans Readjustment Study. Philadelphia, PA, US: Brunner/Mazel; 1990.

Abbreviations

AUC: area under the curve
GPS: global positioning system
LASSO: least absolute shrinkage and selection operator
PTSD: post-traumatic stress disorder
RDoC: Research Domain Criteria
ROC: receiver operating characteristic
SCID: Structured Clinical Interview for Mental Disorders

Edited by G Eysenbach; submitted 22.09.16; peer-reviewed by P Indic, M Little, JP Onnela; comments to author 24.11.16; revised version received 28.12.16; accepted 11.02.17; published 16.03.17.

Please cite as:
Place S, Blanch-Hartigan D, Rubin C, Gorrostieta C, Mead C, Kane J, Marx BP, Feast J, Deckersbach T, Pentland A", Nierenberg A, Azarbayejani A
Behavioral Indicators on a Mobile Sensing Platform Predict Clinically Validated Psychiatric Symptoms of Mood and Anxiety Disorders
J Med Internet Res 2017;19(3):e75
URL: http://www.jmir.org/2017/3/e75/
doi:10.2196/jmir.6678
PMID:28302595

Original Paper

Virtual Versus In-Person Focus Groups: Comparison of Costs, Recruitment, and Participant Logistics

Douglas J Rupert[1], MPH; Jon A Poehlman[1], PhD; Jennifer J Hayes[1], BA; Sarah E Ray[2], MA; Rebecca R Moultrie[1], AA

[1]Center for Communication Science, RTI International, Research Triangle Park, NC, United States

[2]Center for Communication Science, RTI International, Atlanta, GA, United States

Corresponding Author:
Douglas J Rupert, MPH
Center for Communication Science
RTI International
3040 Cornwallis Road
Research Triangle Park, NC, 27709
United States
Phone: 1 919 541 6495
Fax: 1 919 541 7384
Email: drupert@rti.org

Abstract

Background: Virtual focus groups—such as online chat and video groups—are increasingly promoted as qualitative research tools. Theoretically, virtual groups offer several advantages, including lower cost, faster recruitment, greater geographic diversity, enrollment of hard-to-reach populations, and reduced participant burden. However, no study has compared virtual and in-person focus groups on these metrics.

Objective: To rigorously compare virtual and in-person focus groups on cost, recruitment, and participant logistics. We examined 3 focus group modes and instituted experimental controls to ensure a fair comparison.

Methods: We conducted 6 1-hour focus groups in August 2014 using in-person (n=2), live chat (n=2), and video (n=2) modes with individuals who had type 2 diabetes (n=48 enrolled, n=39 completed). In planning groups, we solicited bids from 6 virtual platform vendors and 4 recruitment firms. We then selected 1 platform or facility per mode and a single recruitment firm across all modes. To minimize bias, the recruitment firm employed different recruiters by mode who were blinded to recruitment efforts for other modes. We tracked enrollment during a 2-week period. A single moderator conducted all groups using the same guide, which addressed the use of technology to communicate with health care providers. We conducted the groups at the same times of day on Monday to Wednesday during a single week. At the end of each group, participants completed a short survey.

Results: Virtual focus groups offered minimal cost savings compared with in-person groups (US $2000 per chat group vs US $2576 per in-person group vs US $2,750 per video group). Although virtual groups did not incur travel costs, they often had higher management fees and miscellaneous expenses (eg, participant webcams). Recruitment timing did not differ by mode, but show rates were higher for in-person groups (94% [15/16] in-person vs 81% [13/16] video vs 69% [11/16] chat). Virtual group participants were more geographically diverse (but with significant clustering around major metropolitan areas) and more likely to be non-white, less educated, and less healthy. Internet usage was higher among virtual group participants, yet virtual groups still reached light Internet users. In terms of burden, chat groups were easiest to join and required the least preparation (chat = 13 minutes, video = 40 minutes, in-person = 78 minutes). Virtual group participants joined using laptop or desktop computers, and most virtual participants (82% [9/11] chat vs 62% [8/13] video) reported having no other people in their immediate vicinity.

Conclusions: Virtual focus groups offer potential advantages for participant diversity and reaching less healthy populations. However, virtual groups do not appear to cost less or recruit participants faster than in-person groups. Further research on virtual group data quality and group dynamics is needed to fully understand their advantages and limitations.

(J Med Internet Res 2017;19(3):e80) doi:10.2196/jmir.6980

KEYWORDS

focus groups; virtual systems; online systems; videoconferencing; qualitative research; communication; mobile apps; diabetes mellitus

Introduction

Overview

Qualitative research is a critical component of public health interventions and evaluations, providing in-depth information that can be difficult to obtain through surveys and other quantitative methods [1]. Focus groups, in particular, are a valuable tool for identifying and dissecting the knowledge, attitudes, and perceptions that influence individuals' behavior as well as the barriers and facilitators to behavioral change [2,3]. Market research firms are increasingly offering and promoting the use of virtual focus groups to collect qualitative data. Virtual groups are generally defined as qualitative research sessions in which multiple individuals congregate remotely to discuss a specific topic [4,5]. Virtual focus groups may be conducted via phone, chat, or video platforms and may be held either synchronously during a 1- to 2-hour period or asynchronously over multiple days [5].

Presumed Benefits of Virtual Focus Groups

Hypothetically, virtual focus groups offer multiple advantages over traditional, in-person focus groups [5-7]. First, virtual groups theoretically should be less expensive, eliminating travel costs for research staff as well as other incidental expenses (eg, snacks, parking fees). Second, the turnaround time for virtual groups should be faster because they eliminate travel between research sites and, depending on the platform, offer instantaneous transcripts. Third, virtual sessions should facilitate greater geographic diversity both within and across focus groups by enrolling participants from a greater number of locations.

Fourth, virtual groups should reach populations that are often excluded from or under-represented in traditional focus groups, including rural residents, individuals with less than a high school education, individuals of lower socioeconomic status, and individuals with health and mobility impairments. Likewise, researchers can theoretically use virtual platforms to convene individuals from rare populations (eg, low prevalence health conditions) where in-person gatherings would be impossible. Finally, virtual focus groups should reduce the travel and logistical burdens on participants, resulting in higher show rates and faster recruitment.

Evidence on Virtual Focus Groups

Despite the promise of this methodology, the evidence base for virtual focus groups is extremely thin. Few studies have examined whether these hypothesized benefits materialize in practice, and even fewer studies have rigorously compared traditional and virtual focus groups on the aforementioned metrics [6,8]. In terms of cost, no study has directly compared the expenses of virtual and in-person focus groups, although several articles report anecdotal cases of unquantified cost savings [5,7-10]. Likewise, only 1 study has directly examined recruitment differences between traditional and chat focus groups, finding that virtual group participants were slightly younger than in-person participants [11]. Finally, no study has compared the participant logistics of traditional and virtual focus groups, such as travel and preparation time, technology requirements, interference from nearby individuals, and barriers to participation.

In addition to the lack of direct comparisons between traditional and virtual groups, the few studies that have examined virtual focus groups have several methodological limitations. First, the modes compared in the literature are very limited. Most of the chat-based groups have been asynchronous rather than real time [7,8,12-14], and only 1 study has examined video-based groups [15], which are an increasingly common offering from market research firms. Second, almost all virtual group participants either have been drawn from panels of known Internet users [14,16] or represent specialized populations, such as college students and active military personnel [6-9,13,15,17], all of which restrict the generalizability of study findings.

Third, many studies employ weak or subjective measurements to assess virtual focus group characteristics, including expert rankings, research team commentaries, or self-reported participant preferences [4,7,9,14,16,18]. Almost no study has employed rigorous or objective measures to assess virtual group performance [15]. Finally, most studies that compare focus group modes lack strong experimental controls, with studies frequently employing different moderators, durations, sample sizes, and eligibility criteria by mode [8,16] or failing to employ any type of comparison group [7,9,13]. Ultimately, these limitations dilute the evidence base on virtual focus groups, making it difficult to assess their true benefits and limitations.

The purpose of this study was to address these gaps in evidence by rigorously comparing virtual and traditional focus groups on cost, recruitment, and participant logistics. We examined 3 focus group modes—in-person, live chat, and video—and instituted strong experimental controls to ensure a fair comparison. Specifically, we sought to answer the following 3 research questions:

1. Cost: How do costs—both projected and actual—differ by focus group mode?

2. Recruitment: How do recruitment timing and participant characteristics differ by focus group mode?

3. Participant Logistics: How do participant logistics—such as difficulty attending and preparation time—differ by focus group mode?

Methods

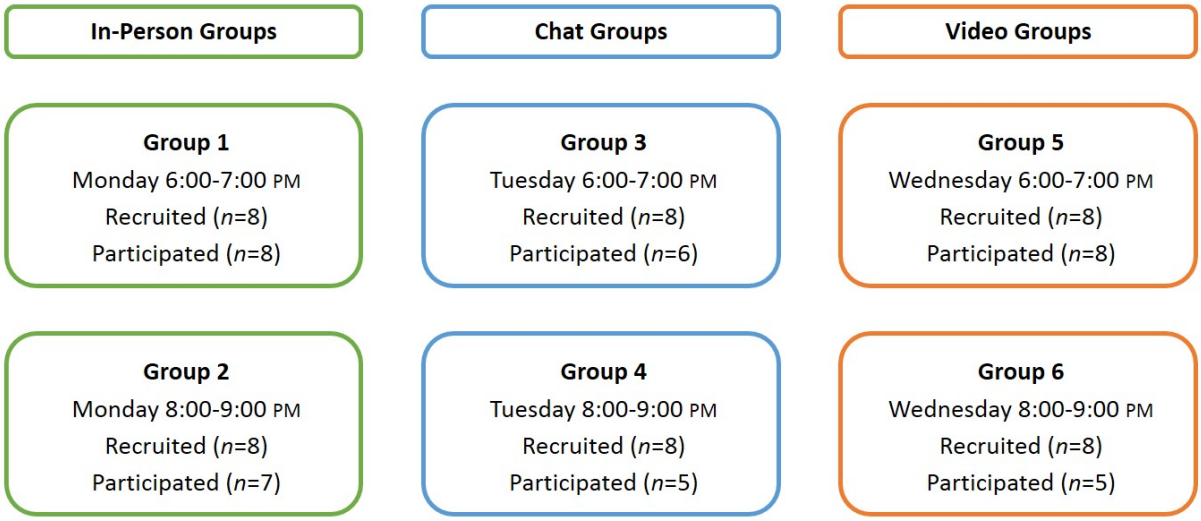
Study Design

We planned and conducted a series of 6 focus groups using 3 different modes—in-person (n=2), live chat (n=2), and video (n=2)—with individuals who had type 2 diabetes (Figure 1). The topic of all 6 groups was using technology to communicate with health care providers, with groups discussing actual and

intended use of email, patient portals, and wearable devices to share health information with clinicians. We selected this topic because it was applicable across all demographic groups (eg, sex, age, race, education, income), and we structured the moderator guide so that participants could engage in group discussion regardless of their actual technology use.

We conducted the groups in August 2014 at the same times of day (6:00-7:00 PM and 8:00-9:00 PM Eastern Daylight Time) from Monday to Wednesday during a single week. We held the in-person groups in Atlanta, GA. Each focus group comprised a 1-hour moderated discussion, and we recruited 8 people per group with no over-recruitment. We ultimately enrolled 48 individuals in the study, of which 39 individuals participated in the groups.

Figure 1. Study design and sample sizes.



Study Population, Eligibility Criteria, and Participant Quotas

The study population comprised individuals diagnosed with type 2 diabetes. We selected this population because public health professionals often conduct formative research with a specific illness population and because type 2 diabetes is broadly distributed across demographic groups (ie, study population does not introduce demographic bias).

Within this population, we established eligibility criteria to ensure participants could actively engage in the discussion and were not frequent research participants (Table 1). These criteria mirror the eligibility criteria typically adopted in other formative health research studies [19-21]. For the virtual focus groups, we also required individuals to have sufficient Internet access to participate in an online discussion. Chat group participants needed to have at least a dial-up Internet connection, and video group participants needed to have a high-speed Internet connection.

Table 1. Participant eligibility criteria.

Eligibility criteria	Rationale
Age 18 years or older	Ensures participants are adults who can consent to enroll in the study Ensures participants make own health care decisions
Diagnosed with type 2 diabetes	Ensures participants are members of the same illness population
Not employed in health care or research fields	Reduces likelihood of response bias
No focus group participation in last 6 months	Reduces likelihood of response bias
English as primary language	Ensures participants can adequately engage in group discussions
Have a regular health care provider	Ensures focus group topic is relevant to participants
Visited health care provider within the last year	Ensures focus group topic is relevant to participants

In addition to these eligibility criteria, we also set several participant quotas for the recruitment firms (Table 2). We established these quotas both to ensure greater demographic diversity within each focus group and to assess recruitment of key demographic groups by mode (for example, we wanted to examine recruitment of less educated individuals for virtual groups versus in-person groups).

Table 2. Participant enrollment quotas.

Category	Enrollment quota per group
Age	Minimum of 2 individuals aged 35 years or younger
	Minimum of 2 individuals aged 55 years or older
Sex	Minimum of 3 males
	Minimum of 3 females
Education	Minimum of 2 individuals with a high school education or less
	Minimum of 2 individuals with some college education or more

Platform and Vendor Selection

First, we solicited bids from multiple vendors to measure average recruitment and hosting costs by focus group mode. Specifically, we collected itemized bids from 4 recruitment firms to recruit participants for all 3 modes, and we collected itemized bids from these same 4 firms to host the in-person focus groups. We also collected itemized bids from 6 Web-based vendors (3 live chat, 3 video) for hosting the virtual groups. We used these bids to calculate projected costs by mode.

We ultimately selected a single recruitment firm to host the in-person focus groups and recruit participants for all 6 groups across all modes. We selected a single firm to ensure consistent recruitment practices across modes and eliminate the possibility that recruitment strategies would confound mode-versus-mode comparisons. We also selected one online vendor to host the chat focus groups and another online vendor to host the video focus groups.

When selecting vendors, we considered cost competitiveness, past experience, and—in the case of virtual focus groups—platform functionality. Specifically, we required that the virtual group platform offer interactive capabilities, such as electronic consent forms, Web-based exit questionnaires, handout sharing and markup, and polling. Although we did not use all of these features in this study, these capabilities are essential for many scientific and public health research studies and were important to consider. Given that the study discussed personal health information, we also required platforms that could guarantee data privacy and that would protect participants' identities at a level sufficient to meet or exceed the institutional review board guidelines. Because no publicly available platform (eg, Skype, Facetime) offered these features, we ultimately selected proprietary virtual group platforms to ensure that study results were generalizable to a wide range of scientific research studies.

Recruitment and Enrollment

We selected a single recruitment firm with national reach to recruit participants for all 3 modes. The firm was blinded to the study's purpose and was not aware that we would be comparing recruitment by mode. To ensure experimental control, the recruitment firm assigned separate recruiters to each focus group mode and provided them with identical eligibility requirements (with the exception of Internet speed). To minimize bias (eg, mode preference) and learning curves (eg, recruitment skill improves with time), the recruiters were blinded to recruitment efforts for the other modes. Recruiters had 2 weeks to complete enrollment. During the enrollment period, recruiters provided daily spreadsheets showcasing their progress.

For all modes, recruiters identified potential participants using contact databases and advertisements. The recruiters contacted potentially eligible individuals by telephone, screened them for eligibility using a 25-item questionnaire, and scheduled eligible individuals for focus groups at preselected dates and times.

Data Collection

We developed a semistructured moderator guide containing questions and probes on the topics of email communication with health care providers, patient portals, sensors and wearable devices (eg, Fitbit), automatic sharing of personal biomedical information (eg, blood glucose level) with health care providers, and privacy concerns or preferences related to technology. A single moderator conducted all 5 focus groups using this same guide.

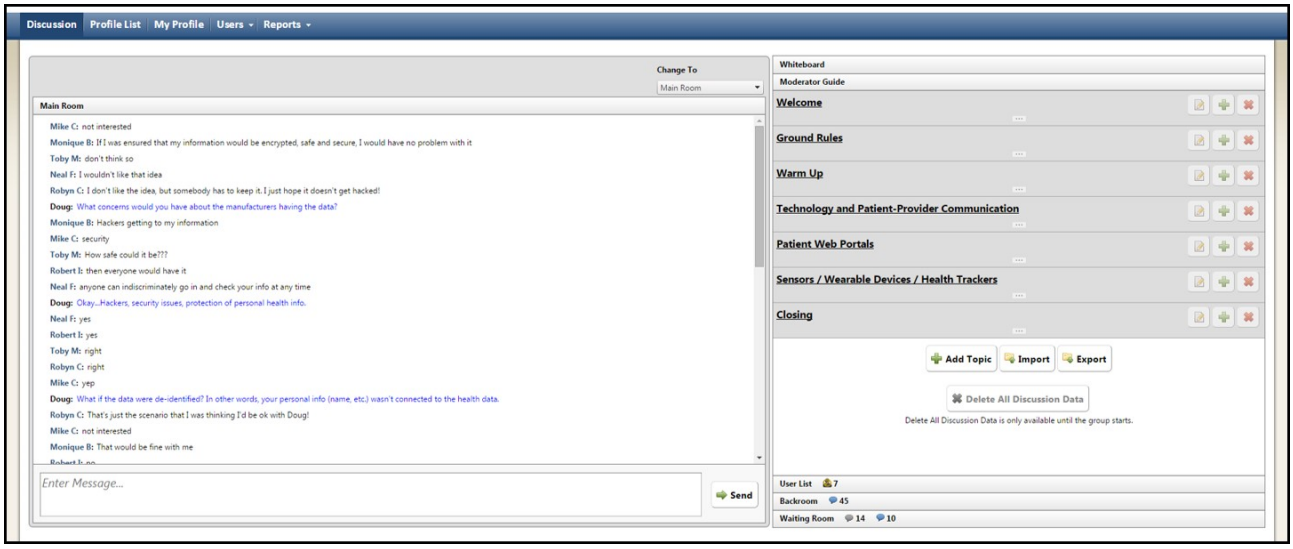
In-Person Groups

For the in-person groups, we conducted 1-hour focus groups at a market research facility in Atlanta. We administered written informed consent to participants upon arrival, and a trained moderator conducted each group by asking questions, probing for details, and leading participants in verbal discussion. At the end of each session, participants completed a 15-item hardcopy exit questionnaire. We also audio and video recorded the sessions and produced verbatim transcripts. Participants received a US $75 incentive after each session.

Chat Groups

For the chat groups, we conducted 1-hour focus groups on a real-time live chat platform hosted by an online research vendor (see Figure 2 for example). We had emailed consent forms and confirmation letters with login credentials to participants several days in advance. Upon login, participants acknowledged an electronic consent form. A trained moderator posted questions and probes from the guide in a chat box, and participants typed responses visible to the entire group. At the end of each session, participants completed an 18-item Web-based exit questionnaire. The platform produced verbatim transcripts of the typed responses from the open group discussion. Participants were mailed a US $75 incentive after each session.

Figure 2. Chat focus group screenshot example.



Video Groups

For the video groups, we conducted 1-hour focus groups on a real-time video platform hosted by an online research vendor (see Figure 3 for example). We had emailed consent forms and confirmation letters with login credentials to participants several days in advance. Upon login, participants acknowledged an electronic consent form. Participants joined the sessions using Web cameras connected to their computer or mobile devices. (If participants did not have a Web camera, we supplied them with one.) The moderator and participants were able to see and hear other individuals' video feeds on screen, and the moderator conducted each group by asking questions, probing for details, and leading participants in verbal discussion. At the end of each session, participants completed an 18-item Web-based exit questionnaire. We also audio and video recorded the sessions and produced verbatim transcripts. Participants were mailed a US $75 incentive after each session.

During all groups, a note taker documented major themes in the discussion. The note taker also completed an observer worksheet that documented participant logistics, such as late arrivals, no-shows, cancellations, and early departures.

Figure 3. Video focus group screenshot example.

[1] Does not depict actual study participants.

Measurements and Data Analysis

Cost Measures

We measured projected costs by averaging the itemized cost estimates on the vendor bids we solicited. Specifically, we measured recruitment costs (eg, recruitment, proposed incentives) by averaging costs for each mode across the 4 recruitment firm bids. We used an identical approach for measuring in-person facility costs (eg, facility rental, video recording, transcripts) and virtual platform costs (eg, platform rental, management fees, recording, transcripts). We measured projected travel expenses by securing an estimate through our institutional travel vendor for roundtrip airfare, 1 night of lodging, and 1 day of meals.

We measured actual costs by tracking invoiced expenses for the 6 focus groups. We calculated actual costs on 3 levels: costs per participant, costs per group with actual participation (ie, actual show rates), and costs per group assuming full participation (ie, 8 participants per group).

Recruitment Measures

We measured recruitment in several ways. First, we examined enrollment timing by using the daily recruitment updates to assess how quickly participants were enrolled in each mode. Second, we examined show rates for each focus group mode. Third, we examined participant demographics by mode, including age, sex, race, education, income, employment status, geographic location, urban-rural classification, public transportation use, Internet use, health care utilization, body mass index (BMI), and personal health ratings. We captured this demographic information on the recruitment screener. Most measures were self-reported; however, we calculated BMI using self-reported height and weight, and we identified participants' urban-rural classifications by comparing their home zone improvement plan (ZIP) codes against the National Center for Health Statistics' urban-rural classification scheme for counties [22].

Participant Logistic Measures

We measured participant logistics using several items from the exit questionnaire. Specifically, we assessed willingness to participate in future groups ("If I were invited to join another [in-person or online] focus group, I would do it."); perceived difficulty attending ("How easy or difficult was it to join today's focus group?"); and preparation and travel time ("How much time did you spend preparing for and traveling to today's focus group?"). The first 2 items had 6-point response scales, and the last item was open ended (measured in minutes).

For virtual group participants, we also assessed participant location during groups ("Where were you during today's focus

group?"); device used by participant ("What type of device did you use to join today's focus group?"); and individuals in the vicinity ("How many people—besides yourself—were near you during the focus group?"). We calculated percentages and means by mode for the participant logistic measures.

Results

Costs (Projected and Actual)

The differences in projected costs by focus group mode were minimal, with in-person groups projected to cost US $3000 per session compared with US $2515 per session for chat groups and US $3028 per session for video groups (Table 3 and Figure 4). Although video recording and travel costs were notably higher for in-person groups, the platform costs and miscellaneous costs were considerably higher for chat and video groups. Recruitment and recommended incentive costs were nearly identical across modes.

Several competing factors accounted for the minimal differences in price. On one hand, recording costs were minimal for video groups and nonexistent for chat groups, and transcription costs also were nonexistent for chat sessions. Likewise, neither chat nor video groups budgeted for research team travel. On the other hand, platform rental for chat and video groups was slightly more expensive than space rental for in-person focus groups, and virtual groups also projected multiple miscellaneous expenses, such as management fees (US $250 per project), incentive mailing fees (US $5-10 per participant), and Web cameras (US $75 per participant, as needed).

Table 3. Projected costs by focus group mode.

Expense category (in US $)[a]	Focus group mode		
	In-person	Chat	Video
Facility or platform	394	467	522
Recruitment	984	1088	1104
Incentives	720	784	768
Video recording	106	N/A[b]	12
Transcription	156	No cost	203
Miscellaneous	40	176	419
Travel	600	No cost	No cost
Total expenses	3000	2515	3028

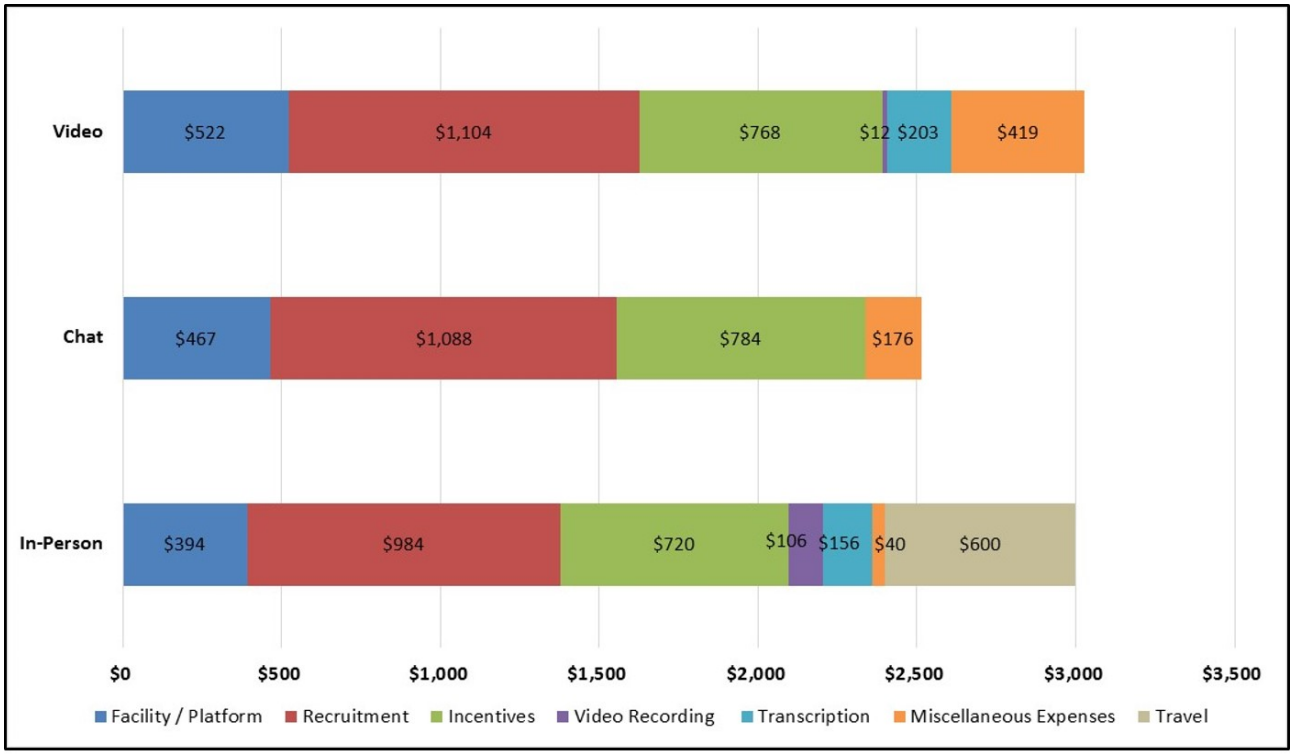
[a]Depicts projected cost per 1-hour focus group, assuming full participation (n=8).

[b]N/A: not applicable.

The differences in actual costs by focus group mode were very similar to the differences in projected costs, with chat groups (US $2000) being less expensive than in-person (US $2666) and video (US $2675) groups (Table 4 and Figure 5). Although in-person groups incurred travel costs and higher recording and transcription fees, virtual groups still had higher platform costs and more miscellaneous expenses (eg, incentive mailing fees, participant Web cameras). Of particular note, 14 of the 16 individuals enrolled in the video groups did not have Web cameras, and the research team needed to purchase cameras for these participants (We standardized actual incentive amounts across the study to ensure a fair comparison on recruitment metrics. Consequently, incentive costs did not differ by mode).

Actual costs for the focus groups in all 3 modes were slightly lower than projected for several reasons. First, the facility or platform costs were lower than anticipated, especially for video groups. Second, we offered slightly lower participant incentives than recommended by the recruitment firms (US $75 vs US $90-98); we selected this incentive based on the amount typically approved for federal government research. Finally, actual travel costs were only two-thirds of projected travel expenses.

Figure 4. Projected costs by focus group mode.



[1] Depicts projected cost per one-hour focus group, assuming full participation (n=8).

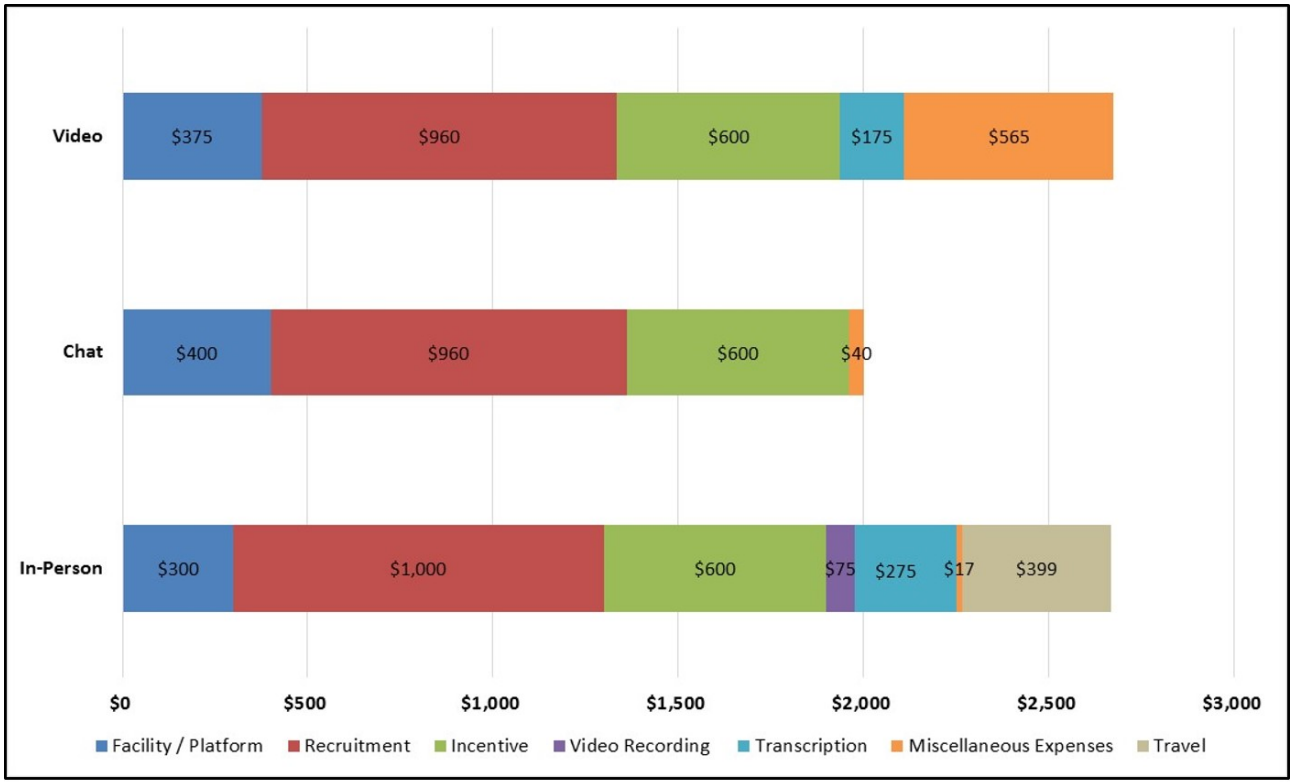
Table 4. Actual costs by focus group mode.

Expense category (in US $)[a]	Focus group mode		
	In-person	Chat	Video
Facility or platform	300	400	375
Recruitment	1000	960	960
Incentives	600	600	600
Video recording	75	N/A[b]	No cost
Transcription	275	No cost	175
Miscellaneous	17	40	565
Travel	399	No cost	No cost
Total expenses	2666	2000	2675

[a]Depicts actual cost per 1-hour focus group, assuming full participation (n=8). In reality, incentive costs were slightly lower, given that not all enrolled participants attended the groups.

[b]N/A: not applicable.

Figure 5. Actual costs by focus group mode.



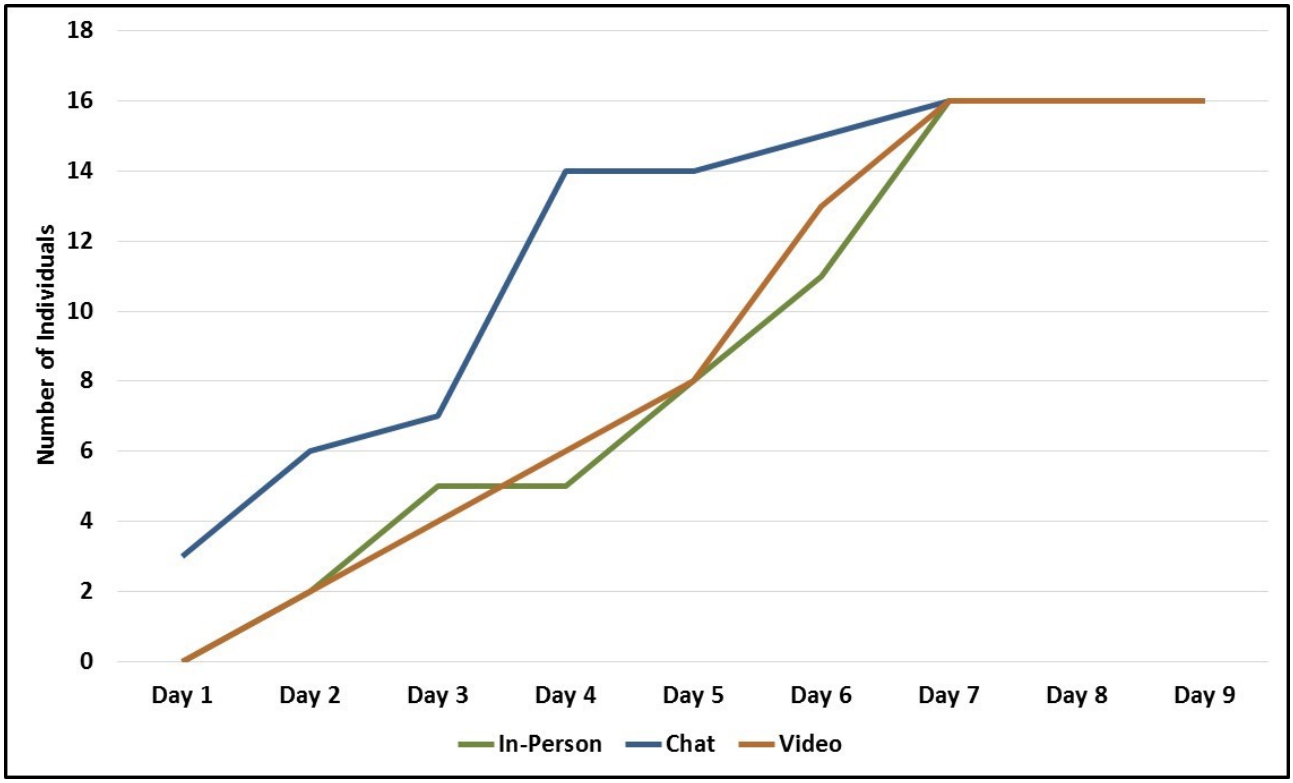
[1] Depicts actual cost per one-hour focus group, assuming full participation (n=8). In reality, incentive costs were slightly lower, given that not all enrolled participants attended the groups.

Recruitment and Participant Characteristics

Enrollment Timing

Recruiters for all 3 focus group modes were able to complete participant enrollment in 7 days, with enrollment finishing on the same day for all 3 modes (Figure 6). However, chat groups initially enrolled participants more quickly, with in-person and video groups lagging behind until the final day. Conversely, the in-person and video group enrollment rates were nearly identical throughout the recruitment period.

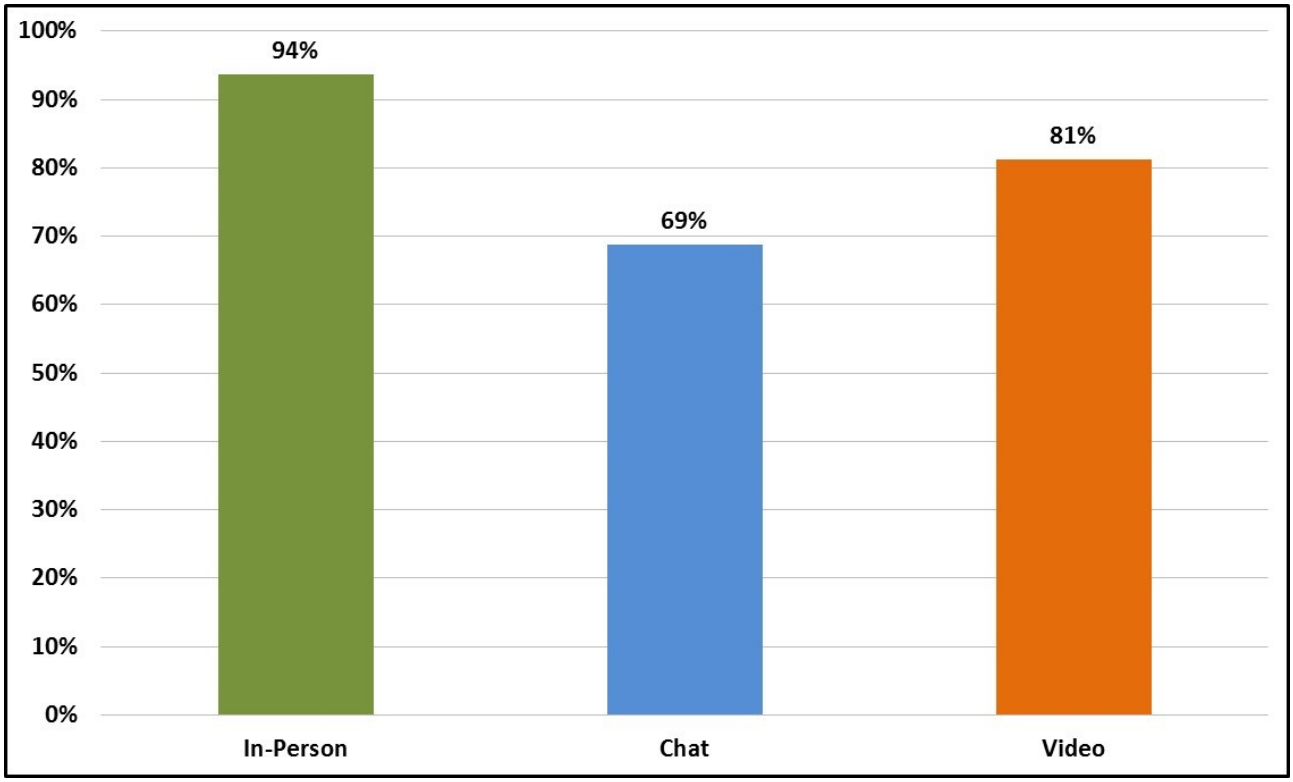
Figure 6. Participant enrollment timing by focus group mode.



Show Rates

Although all 6 focus groups were fully enrolled (n=8 per group), show rates were noticeably different by group mode (Figure 7). In-person groups had the highest show rates (94%, 15/16), followed by video groups (81%, 13/16) and chat groups (69%, 11/16). Of the enrolled individuals who did not participate, those in the in-person and video groups all cancelled by alerting recruiters a few hours in advance that they would not be able to attend. By contrast, most of the enrolled individuals who did not participate in the chat groups simply did not show up for the sessions or were excluded because they arrived more than 20 minutes late.

Figure 7. Show rates by focus group mode.



Geographic Diversity

As anticipated, chat and video group enrollees represented a greater number of US states and a slightly broader range of geographic areas in terms of urban-rural classification (Figure 8). Specifically, individuals enrolled in the in-person groups all came from the same state (Georgia) and resided in central and fringe metro areas. In contrast, chat and video groups enrolled individuals from 11 states and 10 states, respectively, and enrolled at least a few individuals who resided in more outlying areas (eg, medium metro, micropolitan). Nevertheless, chat and video groups still recruited heavily from central and fringe metro areas and no focus group mode enrolled individuals from rural areas.

We also examined enrolled individuals' geographic proximity to the recruitment firm's satellite offices to determine how often virtual group recruiters relied on satellite office participant databases (rather than searching nationally for potential participants). Surprisingly, more than half of the individuals enrolled in the chat (62%, 10/16) and video (81%, 13/16) groups lived within 50 miles of one of the recruitment firm's 14 satellite offices (Figure 9). Thus, the geographic diversity of the virtual group enrollees was closely tied to the recruitment firm's physical locations rather than dispersed across the country.

Figure 8. Number of participants per urban-rural classification by focus group mode.

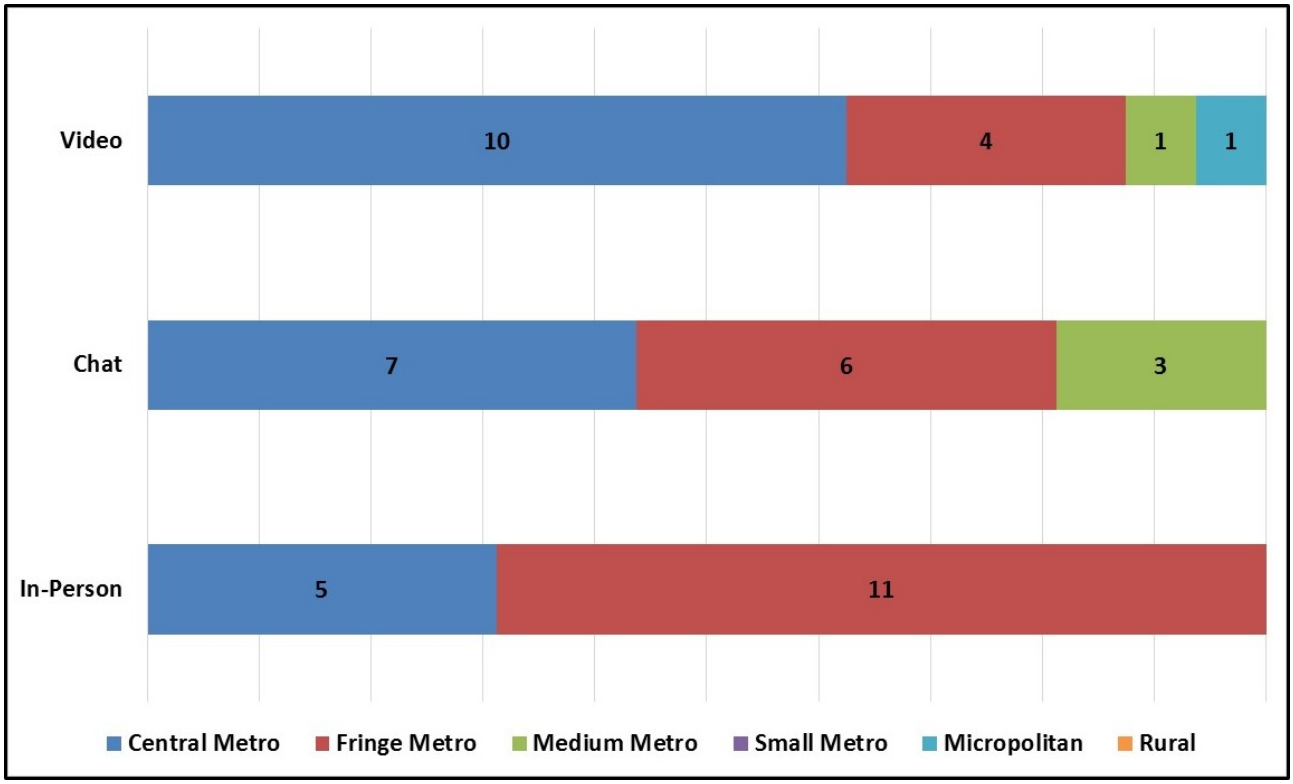
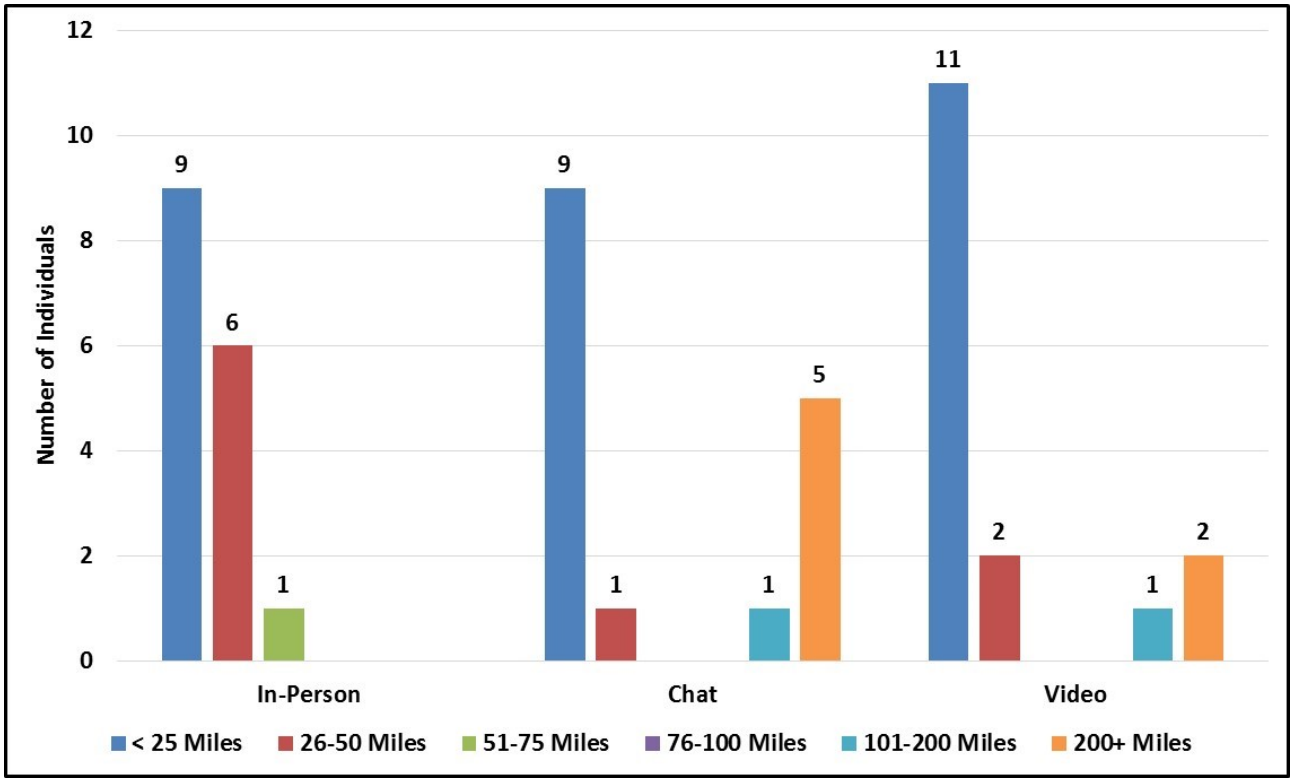


Figure 9. Participant proximity to recruitment firm offices by focus group mode.



Participant Demographics

We examined age, race or ethnicity, education, and income by focus group mode. On average, individuals enrolled in the in-person groups (mean 52.3 years) were slightly older than individuals enrolled in the chat (mean 48.8 years) and video (mean 45.2 years) groups. Moreover, no video group enrollees were older than age 60 years, whereas almost 40% (6/16) of the in-person and chat group enrollees were 61 years or older.

In terms of race, the chat and video group enrollees were more diverse than the in-person group enrollees. The chat and video groups also enrolled individuals from a broader range of racial categories, representing 4-5 different races (white, black, Hispanic, Asian, and other) whereas the in-person groups enrolled only white and black individuals.

Chat and video groups enrolled individuals with a broader range of educational backgrounds than those enrolled in the in-person groups. Although chat groups enrolled the highest number of individuals with post-graduate degrees, the chat and video groups also enrolled a sizeable number of individuals with less than a high school education—25% (4/16) and 38% (6/16), respectively. In contrast, the in-person groups were unable to recruit any individuals without a high school education, despite considerable pressure to do so from the research team.

On average, chat group enrollees had a noticeably higher household income (mean US $71,438) than in-person (mean US $51,750) and video (mean US $49,500) group enrollees. In-person groups also enrolled a greater number of low-income individuals (household income less than US $30,000) than the other 2 two modes. This is surprising, given that in-person groups were unable to recruit individuals with less than a high school education, and it might suggest that in-person groups were more likely to recruit educated individuals who worked low-income jobs or were retired.

Internet Use

Given that virtual focus groups require online access, we examined differences in daily Internet use by focus group mode. Not surprisingly, chat and video group enrollees reported higher daily Internet use at both work and home than in-person group enrollees (Figures 10 and 11). However, chat and video groups still enrolled a sizeable number of light Internet users, especially individuals who use the Internet infrequently at work.

Figure 10. Participant's Internet use at work by focus group mode.

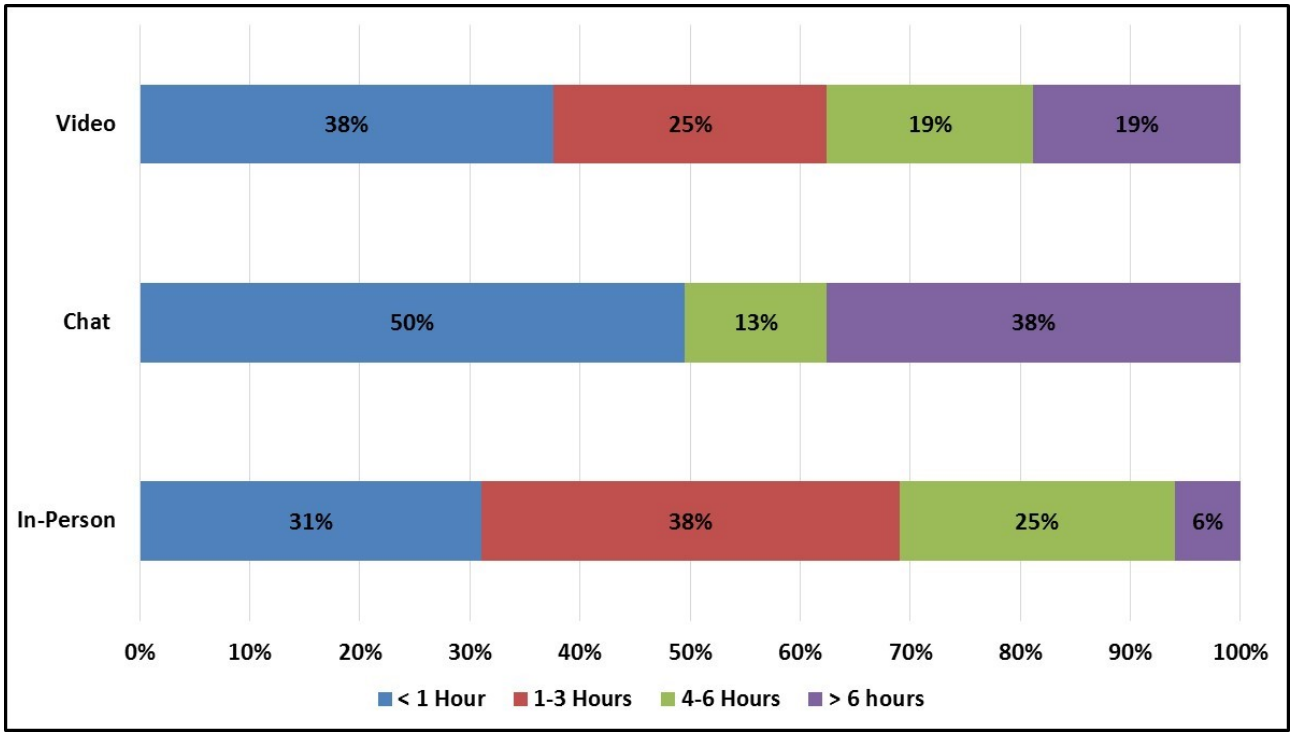
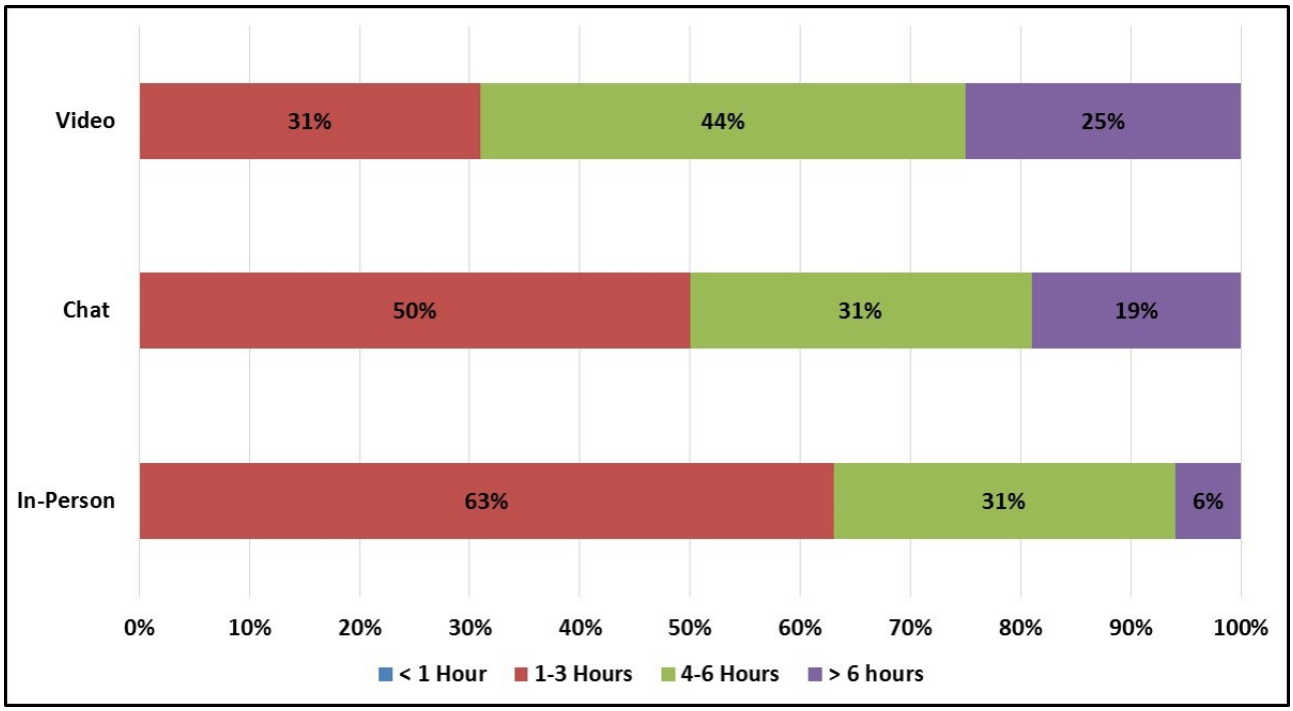


Figure 11. Participant's Internet use at home by focus group mode.



Health Status and Utilization

Chat and video group enrollees reported being in poorer health than in-person group enrollees on several dimensions, including BMI (Figure 12) and health interference in daily activities (Figures 13 and 14). Chat group enrollees, in particular, reported very high BMI values, with all but 1 categorized as obese. One video group enrollee also reported a BMI value of 65.2, more than double the threshold for obesity. Likewise, chat and video group enrollees were more likely to report that their health interfered with their relationships and, among video group enrollees, their jobs.

In terms of health care utilization, fewer video group enrollees reported visiting a health care provider in the last month compared with enrollees in the other modes. Specifically, only 31% (5/16) of video group enrollees reported visiting a health care provider in the previous month compared with 63% (10/16) of chat enrollees and 75% (12/16) of in-person enrollees. On one hand, this finding may suggest better health among video group enrollees because they use health services less frequently.

On the other hand, this finding may indicate that video group enrollees encounter more barriers to accessing health services (eg, no insurance, long distance from facility).

Figure 12. Participant body mass index (BMI) values by focus group mode.

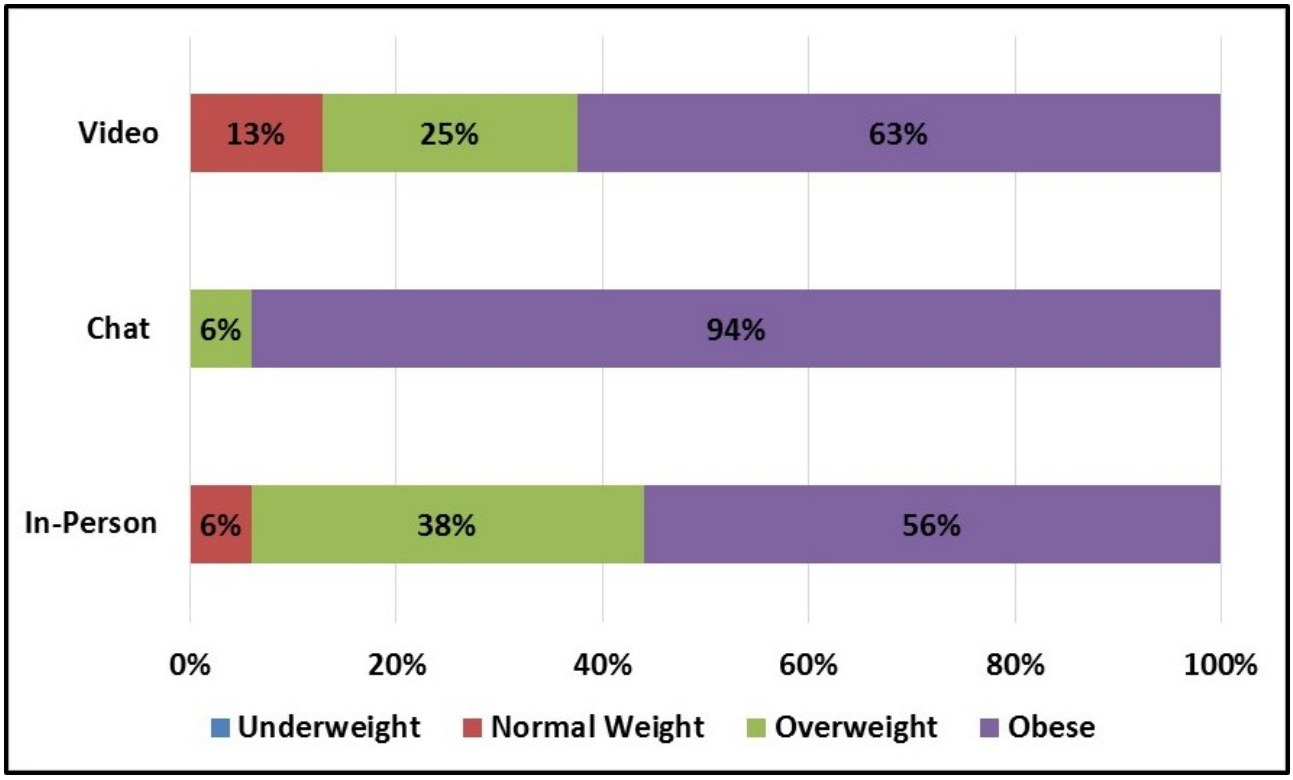


Figure 13. Frequency of participant health interference in relationships by focus group mode.

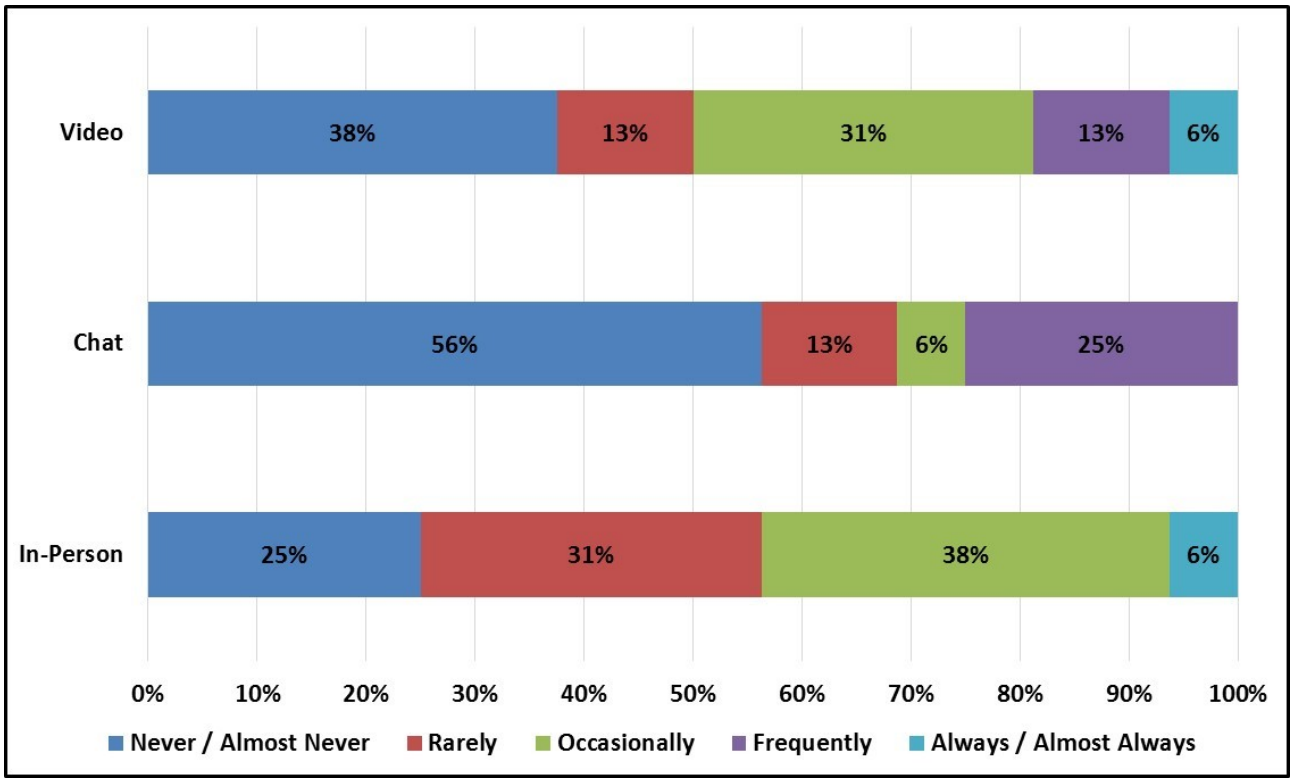
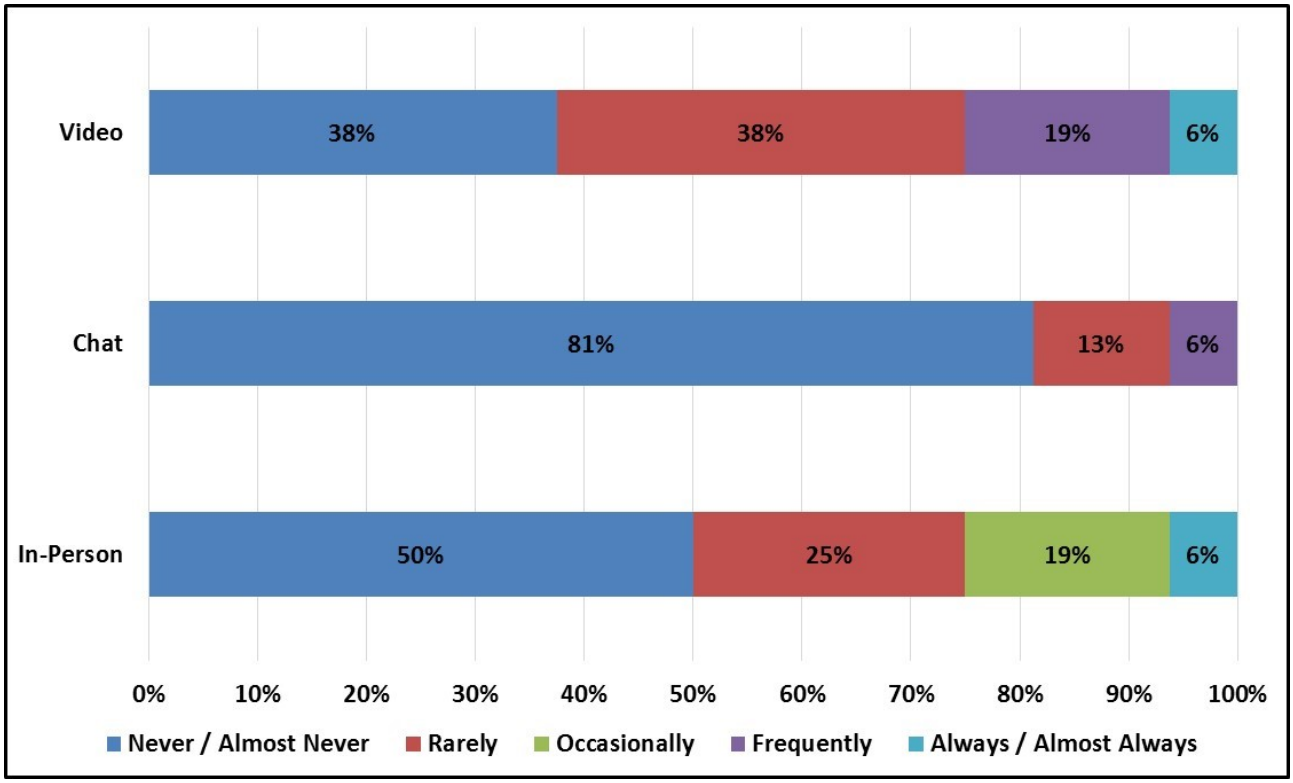


Figure 14. Frequency of participant health interference with job by focus group mode.

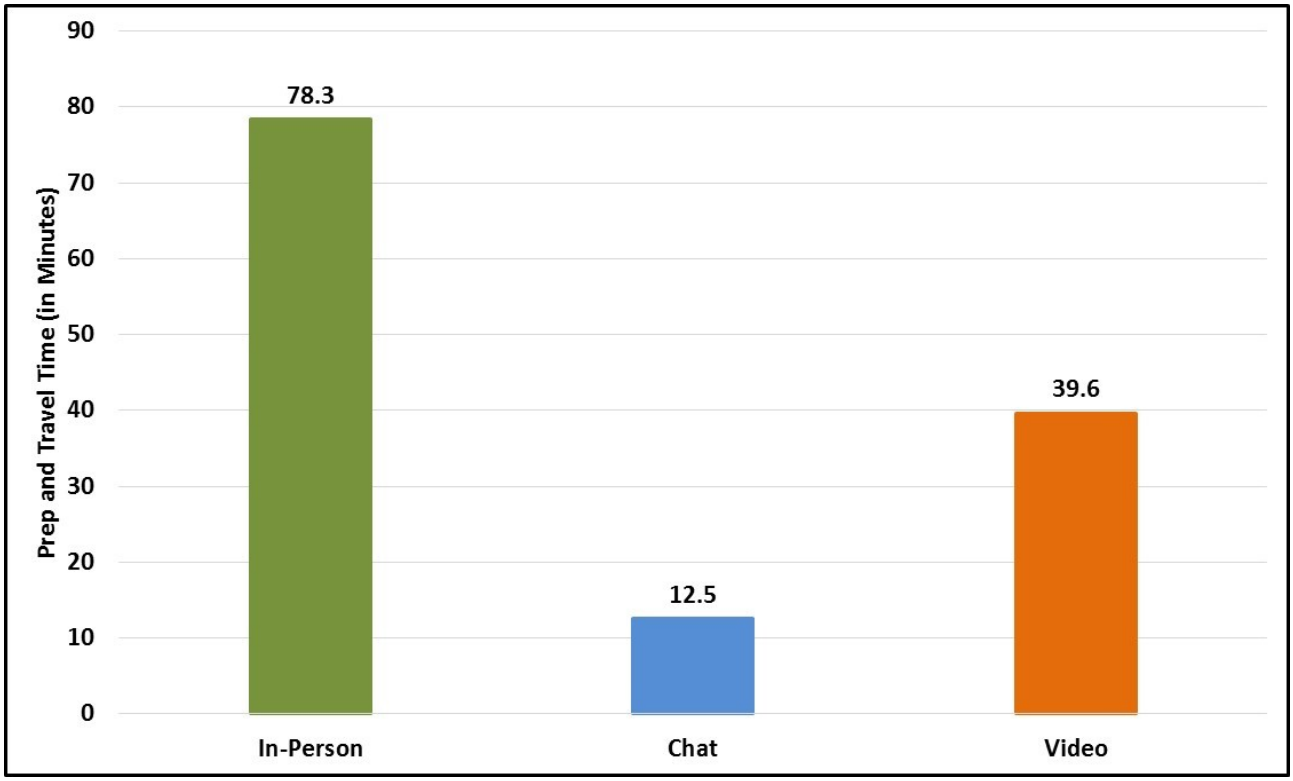


Participant Logistics

Overall, individuals who joined in-person focus groups reported a slightly higher participation burden than individuals who joined chat and video groups. Specifically, participants rated in-person groups as slightly more difficult to join (mean rating 2.4 out of 6.0) than chat (mean rating 1.0) and video (mean rating 1.3) groups. Participants also spent noticeably more time preparing for and traveling to in-person groups than they did preparing for virtual groups (Figure 15). Nevertheless, this additional burden was not reflected in participants' intentions, and participants generally expressed a high willingness to participate in future focus groups using the same mode (in-person = mean rating 5.9 out of 6.0; chat = mean rating 6.0; video = mean rating 5.5).

In terms of virtual focus group logistics, almost all chat and video group participants reported joining the sessions from their homes (91% [10/11] and 100% [13/13], respectively), and all virtual group participants logged onto the platforms using a laptop or desktop computer. No participant joined a group using a tablet, smartphone, or other mobile device. Most participants (71%, 17/24) also joined the virtual groups with no other individuals in their immediate vicinity. However, a handful of participants reporting having 1 or 2 individuals nearby, and 2 chat group participants reported having 5 or more people near them during the sessions.

Figure 15. Average participant preparation and travel time by focus group mode.



Discussion

Principal Findings

There are a number of hypothetical benefits to conducting virtual focus groups, including reduced expenses, faster turnaround time, increased participant diversity, recruitment of hard-to-reach and low-prevalence populations, and reduced participant burden. However, no study has rigorously compared virtual and in-person focus groups to determine whether these benefits materialize in practice. This study presented the most thorough research to date on virtual focus group costs, recruitment, and logistics, and the findings provided a concrete evidence base for understanding the advantages and limitations of virtual qualitative research. We address each of the proposed benefits of virtual focus groups in the following sections.

Claim #1: Virtual Focus Groups Are Less Expensive

We found that both projected and actual cost differences by focus group mode were minimal, with chat groups costing only about US $500- US $600 less and video groups costing approximately the same as in-person groups. Although virtual groups do eliminate the need for travel, they typically incur other costs (eg, management fees, Web camera purchases) that offset this potential savings. Moreover, despite the allegedly decreased burden of participating in a virtual group, recruitment firms often recommended monetary incentives that were higher for virtual groups than for in-person groups.

Show rates are another cost factor that should be considered. We experienced a higher number of cancellations and no-shows among virtual group participants, especially in the chat groups (One possible explanation is that the lack of visual interaction in chat groups leads to a decreased sense of accountability among enrolled individuals). Because cancellations and no-shows still incur recruitment costs and because researchers may need to enroll additional individuals to offset those who withdraw, this reduced participation has cost implications.

Claim #2: Virtual Focus Groups Provide Faster Data

Factoring in travel time, chat and video groups clearly require less time to conduct, and they enable researchers to conduct groups in multiple geographic locations within a single day. That said, when excluding travel, we found that research team preparation time for virtual groups was roughly equivalent to preparation time for in-person groups. On one hand, virtual groups might require additional preparation activities that were unnecessary during in-person groups, such as uploading questions into the chat platform, programming electronic consent forms and exit questionnaires, sending confirmation letters, mailing incentives, and addressing unforeseen technology issues (eg, Web camera not working properly).

On the other hand, virtual groups might streamline or eliminate activities that require more time during in-person groups, such as transcripts (instantly available from chat groups), bathroom breaks (no need for participants to ask), and directing lost participants (no facility for participants to locate). Other activities—such as participant enrollment—did not appear to differ by focus group mode.

Claim #3: Virtual Focus Groups Increase Participant Diversity

Virtual groups theoretically increase participant diversity by accessing a broader range of individuals, particularly in terms of geography. Our study findings support this claim. Participants in the chat and video groups represented a wider range of geographic areas in terms of both states and urban-rural

classification. However, we also found that most virtual group participants lived in close proximity to 1 of the recruitment firm's satellite offices, suggesting that virtual participant diversity was likely tied to the recruitment firm's geographic locations. In addition, people residing in rural areas still were not represented in virtual groups. These findings suggest that researchers might need to push recruiters to enroll individuals outside of their established recruitment pools in order to maximize the geographic reach offered by virtual groups.

We also found that virtual groups enrolled more diverse participants in terms of race or ethnicity, education, and household income. In particular, virtual groups were more effective at enrolling individuals with less than a high school education, individuals of Asian and Hispanic background, and individuals with higher household incomes—all groups that could be difficult to reach through focus groups. Conversely, in-person and chat groups were equally adept at enrolling individuals of different ages, but video groups were unable to enroll any individuals older than 60 years.

One hypothetical limitation was that virtual focus groups might enroll primarily individuals who were frequent Internet users and who were adept at online technology. Although daily Internet use was generally higher among virtual group participants than among in-person group participants, virtual groups still enrolled a sizeable number of light Internet users who spent less than 1 hour per day on the Internet at work. This might suggest that virtual groups were able to enroll individuals who worked in blue collar, retail, and service fields and spent little time working at a computer.

Claim #4: Virtual Focus Groups Capture Hard-to-reach Populations

Given that we intentionally recruited individuals with diabetes for this study, we did not assess whether virtual focus groups enrolled individuals with low-prevalence health conditions. Nevertheless, we did find that virtual groups enrolled less healthy individuals than in-person groups, as evidenced by the number of participants with extremely high BMIs (extremely obese) and the number of individuals who reported that their health interfered with their relationships. As mentioned previously, virtual focus groups also enrolled more individuals from hard-to-reach demographic groups, such as those with a high school education or less, those with high family incomes, and those of a race other than black or white.

Claim #5: Virtual Focus Groups Reduce Participant Burden

We found that chat and video groups slightly reduced the logistical burden on participants compared with in-person focus groups. On average, virtual group participants spent noticeably less time preparing for the sessions, and they rated the groups as easier to join. However, virtual groups did have higher cancellation and no-show rates, and there was no meaningful difference by mode in participants' willingness to join another focus group using the same methodology. These findings suggest that, although virtual focus groups reduced participant burden, this feature did not necessarily translate into more active or engaged participants.

Limitations

This study offered several advantages over previous research on virtual focus groups, including a direct comparison of 3 modes, objective measurements, and strong experimental controls. Nevertheless, this study does have limitations that need to be considered when interpreting the findings. First, the study had a modest sample size, in terms of both groups and participants, which increases the likelihood that a few data points might have skewed the findings. Second, we conducted the study with a single illness population (type 2 diabetes), which might restrict the generalizability of the findings. Although we selected this population because of its demographic diversity, it is possible that replicating this study with other populations would lead to different results.

Third, we used a single recruitment firm to identify and enroll participants, which means that some results might be influenced by the firm's procedures and infrastructure. We selected a single firm to standardize recruitment across modes, and we blinded recruiters to the other modes to ensure impartiality. Although we believe these experimental controls outweigh the potential for firm bias, it is a possible limitation. Finally, we encountered a video group scheduling error (ie, second video group scheduled at 7:00-8:00 PM Eastern Daylight Time instead of 8:00-9:00 PM Eastern Daylight Time) that inferred with the study's experimental time control. All participants were aware of the group starting time. Therefore, it is unlikely that this error affected show rates; however, the earlier timeslot might have affected participant availability.

Future Research

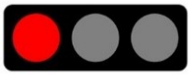
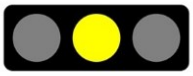
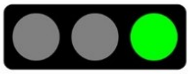
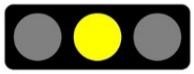
This study is an important and rigorous step in establishing the evidence base on virtual focus groups. However, we recommend additional research in this area to broaden and deepen our understanding of this methodology. First, future research should replicate and expand this study with more focus groups, more participants, and more timeslots to see whether the results change. Second, future studies should examine additional populations beyond type 2 diabetes, including hard-to-reach groups (eg, health care providers, individuals with physical disabilities) and sensitive illness populations (eg, sexually transmitted illnesses, clinical depression). Finally, this study addressed only the cost, recruitment, and logistical aspects of virtual focus groups, and future studies need to compare virtual and in-person groups in terms of data quality (eg, group dynamics, data breath, data saturation). This final piece is particularly important, as the recruitment and logistical advantages of virtual focus groups may carry less weight if these groups produce inferior data.

Conclusions

Virtual focus groups are an appealing and reasonable option for health and medical researchers who seek faster data, increased participant diversity, and inclusion of hard-to-reach populations in qualitative research (Figure 16). Although virtual groups are unlikely to save money, they do appear to reach a broader range of individuals in terms of geography, race or ethnicity, education, and income, and they seem to offer an advantage when it comes to reaching less healthy individuals and those

with mobility impairments. Virtual focus groups also reduce the burden on participants by eliminating travel and minimizing preparation time; however, these advantages do not translate into more active participation, and virtual groups are likely to have higher cancellation and no-show rates. Future research on virtual focus groups should examine data quality to determine whether the methodology's recruitment and logistical advantages lead to useful data.

Figure 16. Summary of hypothetical advantages and supporting evidence for virtual focus groups.

Hypothetical advantage	Evidence rating	Evidence summary
Virtual focus groups are less expensive		Chat groups only slightly less expensive. Video groups same cost as in-person groups.
Virtual focus groups provide faster data		Travel time eliminated. However, researcher prep time same or longer as in-person groups.
Virtual focus groups increase participant diversity		Offer greater geographic, racial, educational, and income diversity. Still reach light Internet users. Watch for geographic clustering near recruitment firm offices.
Virtual focus groups capture hard-to-reach populations		Able to reach less healthy populations (e.g., obese). More evidence needed on low-prevalence health conditions.
Virtual focus groups reduce participant burden		Easier for participants to join, less prep time needed. However, higher cancellation and no-show rates.

Acknowledgments

We would like to thank Craig Lefebvre, Lauren McCormack, and Amy Roussel (RTI International) for their feedback on the study design, research questions, and measurements. This research was funded by an internal professional development award from RTI International. Award funding was not tied to study results, interpretation, or implications. The findings and conclusions in this article are those of the authors and do not necessarily reflect the opinions of RTI International as an organization.

Authors' Contributions

All authors meet the requirements for authorship and submission. DR and JP conceptualized the study, oversaw all study activities, and co-led the writing of the article. DR also led all focus groups and data collection. JH summarized relevant literature and wrote sections of the article. SR collected data and wrote sections of the article. RM coordinated data collection, scheduled all focus groups, and wrote sections of the article.

Conflicts of Interest

None declared.

References

1. National Cancer Institute. Making Health Communication Programs Work: A Planner's Guide, Pink Book. Washington, DC: US Department of Health and Human Services; Jan 01, 2004.
2. Gibbs G, Flick U, Kvale S, Angrosino MV, Barbour RS, Banks M, et al. Analyzing Qualitative Data. In: The Sage Qualitative Research Kit. London: SAGE; 2007.
3. Morgan D, Krueger RA, King JM. The Focus Group Guidebook. In: Focus group kit. Thousand Oaks, California: SAGE Publications; 1998.
4. Clapper D, Massey AP. Electronic focus groups: a framework for exploration. Information & Management 1996 Jan;30(1):43-50. [doi: 10.1016/0378-7206(95)00041-0]
5. Sweet C. Designing and conducting virtual focus groups. Qualitative Mrkt Res 2001 Sep;4(3):130-135. [doi: 10.1108/13522750110393035]
6. Underhill C, Olmsted MG. An experimental comparison of computer-mediated and face-to-face focus groups. Soc Sci Comput Rev 2003 Nov 01;21(4):506-512. [doi: 10.1177/0894439303256541]

7. Tates K, Zwaanswijk M, Otten R, van DS, Hoogerbrugge PM, Kamps WA, et al. Online focus groups as a tool to collect data in hard-to-include populations: examples from paediatric oncology. BMC Med Res Methodol 2009 Mar 03;9:15 [FREE Full text] [doi: 10.1186/1471-2288-9-15] [Medline: 19257883]
8. Nicholas DB, Lach L, King G, Scott M, Boydell K, Sawatzky BJ, et al. Contrasting internet and face-to-face focus groups for children with chronic health conditions: outcomes and participant experiences. Int J Qual Methods 2010 Mar;9(1):105-121. [doi: 10.1177/160940691000900102]
9. Fox F, Morris M, Rumsey N. Doing synchronous online focus groups with young people: methodological reflections. Qual Health Res 2007;17(4):539-547.
10. Kenny AJ. Interaction in cyberspace: an online focus group. J Adv Nurs 2005 Feb;49(4):414-422. [doi: 10.1111/j.1365-2648.2004.03305.x] [Medline: 15701156]
11. Brubaker L, Richter HE, Barber MD, Hsu Y, Rahn DD, Menefee S, et al. Pelvic floor disorders clinical trials: participant recruitment and retention. Int Urogynecol J 2013 Jan;24(1):73-79 [FREE Full text] [doi: 10.1007/s00192-012-1824-x] [Medline: 22669424]
12. Banfield MA, Barney LJ, Griffiths KM, Christensen HM. Australian mental health consumers' priorities for research: qualitative findings from the SCOPE for Research project. Health Expect 2012:e1.
13. Monolescu D, Schifter C. Online focus group: a tool to evaluate online students' course experience. Internet High Educ 2000;2(2-3):171-176.
14. Popenko NA, Devcic Z, Karimi K, Wong BJ. The virtual focus group: a modern methodology for facial attractiveness rating. Plast Reconstr Surg 2012;130(3):455e-461e. [doi: 10.1097/PRS.0b013e31825dcb48]
15. Abrams K, Wang Z, Song Y, Galindo-Gonzalez S. Data richness trade-offs between face-to-face, online audiovisual, and online text-only focus groups. Soc Sci Comput Rev 2015;33(1):80-96.
16. Brüggen E, Willems P. A critical comparison of offline focus groups, online focus groups and e-Delphi. Int J Market Res 2009;51(3):363. [doi: 10.2501/S1470785309200608]
17. Gravlee C, Bernard H, Maxwell C, Jacobsohn A. Mode effects in free-list elicitation: comparing oral, written, and web-based data collection. Soc Sci Comput Rev 2012 Sep 10;31(1):119-132. [doi: 10.1177/0894439312455312]
18. Chase L, Alvarez J. Internet research: the role of the focus group. Libr Inf Sci Res 2000;22(4):357-369.
19. Rupert DJ, Moultrie RR, Read JG, Amoozegar JB, Bornkessel AS, O'Donoghue AC, et al. Perceived healthcare provider reactions to patient and caregiver use of online health communities. Patient Educ Couns 2014 Sep;96(3):320-326. [doi: 10.1016/j.pec.2014.05.015] [Medline: 24923652]
20. Rupert DJ, Poehlman JA, Damon SA, Williams PN. Risk and protective behaviours for residential carbon monoxide poisoning. Inj Prev 2013 Apr;19(2):119-123 [FREE Full text] [doi: 10.1136/injuryprev-2012-040339] [Medline: 22653781]
21. Damon SA, Poehlman JA, Rupert DJ, Williams PN. Storm-related carbon monoxide poisoning: an investigation of target audience knowledge and risk behaviors. Soc Mark Q 2013;19(3):188-199. [doi: 10.1177/1524500413493426]
22. Ingram DD, Franco SJ. CDC. Atlanta: National Center for Health Statistics; 2014. 2013 NCHS urban-rural classification scheme for counties URL: https://www.cdc.gov/nchs/data/series/sr_02/sr02_166.pdf [accessed 2017-03-06] [WebCite Cache ID 6ollvRLCJ]

Abbreviations

BMI: body mass index

Edited by G Eysenbach; submitted 10.11.16; peer-reviewed by D Morgan, JD Robinson; comments to author 25.01.17; revised version received 09.02.17; accepted 10.02.17; published 22.03.17.

Please cite as:
Rupert DJ, Poehlman JA, Hayes JJ, Ray SE, Moultrie RR
Virtual Versus In-Person Focus Groups: Comparison of Costs, Recruitment, and Participant Logistics
J Med Internet Res 2017;19(3):e80
URL: http://www.jmir.org/2017/3/e80/
doi:10.2196/jmir.6980
PMID:28330832

Original Paper

Use of Videos Improves Informed Consent Comprehension in Web-Based Surveys Among Internet-Using Men Who Have Sex With Men: A Randomized Controlled Trial

Eric William Hall[1], MPH; Travis H Sanchez[1], DVM, MPH; Aryeh D Stein[2], MPH, PhD; Rob Stephenson[3], MSc, PhD; Maria Zlotorzynska[1], MPH, PhD; Robert Craig Sineath[1], MPH; Patrick S Sullivan[1], DVM, PhD

[1]Department of Epidemiology, Rollins School of Public Health, Emory University, Atlanta, GA, United States

[2]Hubert Department of Global Health, Rollins School of Public Health, Emory University, Atlanta, GA, United States

[3]Department of Health Behavior and Biological Sciences, School of Nursing and The Center for Sexuality and Health Disparities, University of Michigan, Ann Arbor, GA, United States

Corresponding Author:
Eric William Hall, MPH
Department of Epidemiology
Rollins School of Public Health
Emory University
GCR 432
1518 Clifton Road
Atlanta, GA, 30322
United States
Phone: 1 404 727 8965
Fax: 1 404 727 8737
Email: eric.w.hall@emory.edu

Abstract

Background: Web-based surveys are increasingly used to capture data essential for human immunodeficiency virus (HIV) prevention research. However, there are challenges in ensuring the informed consent of Web-based research participants.

Objective: The aim of our study was to develop and assess the efficacy of alternative methods of administering informed consent in Web-based HIV research with men who have sex with men (MSM).

Methods: From July to September 2014, paid advertisements on Facebook were used to recruit adult MSM living in the United States for a Web-based survey about risk and preventive behaviors. Participants were randomized to one of the 4 methods of delivering informed consent: a professionally produced video, a study staff-produced video, a frequently asked questions (FAQs) text page, and a standard informed consent text page. Following the behavior survey, participants answered 15 questions about comprehension of consent information. Correct responses to each question were given a score of 1, for a total possible scale score of 15. General linear regression and post-hoc Tukey comparisons were used to assess difference (P<.001) in mean consent comprehension scores. A mediation analysis was used to examine the relationship between time spent on consent page and consent comprehension.

Results: Of the 665 MSM participants who completed the comprehension questions, 24.2% (161/665) received the standard consent, 27.1% (180/665) received the FAQ consent, 26.8% (178/665) received the professional consent video, and 22.0% (146/665) received the staff video. The overall average consent comprehension score was 6.28 (SD=2.89). The average consent comprehension score differed significantly across consent type (P<.001), age (P=.04), race or ethnicity (P<.001), and highest level of education (P=.001). Compared with those who received the standard consent, comprehension was significantly higher for participants who received the professional video consent (score increase=1.79; 95% CI 1.02-2.55) and participants who received the staff video consent (score increase=1.79; 95% CI 0.99-2.59). There was no significant difference in comprehension for those who received the FAQ consent. Participants spent more time on the 2 video consents (staff video median time=117 seconds; professional video median time=115 seconds) than the FAQ (median=21 seconds) and standard consents (median=37 seconds). Mediation analysis showed that though time spent on the consent page was partially responsible for some of the differences in comprehension, the direct effects of the professional video (score increase=0.93; 95% CI 0.39-1.48) and the staff-produced video (score increase=0.99; 95% CI 0.42-1.56) were still significant.

Conclusions: Video-based consent methods improve consent comprehension of MSM participating in a Web-based HIV behavioral survey. This effect may be partially mediated through increased time spent reviewing the consent material; however, the video consent may still be superior to standard consent in improving participant comprehension of key study facts.

Trail Registration: Clinicaltrials.gov NCT02139566; https://clinicaltrials.gov/ct2/show/NCT02139566 (Archived by WebCite at http://www.webcitation.org/6oRnL261N).

(J Med Internet Res 2017;19(3):e64) doi:10.2196/jmir.6710

KEYWORDS

informed consent; surveys and questionnaires; HIV

Introduction

Men who have sex with men (MSM) continue to be the group most impacted by human immunodeficiency virus (HIV) in the United States. In 2014, 70% of all new infections in the United States occurred among MSM [1]. In many regions around the world, HIV incidence rates among MSM have been increasing [2-4]. This increase has been theorized, in part, to result from an increase of MSM using the Internet to facilitate sexual relationships through partner selection websites [4-7]. There is evidence that men who arrange Web-based sexual encounters may have increased odds of having unprotected anal intercourse [5,6,8], which presents an increased risk for HIV transmission. As a result, Web-based HIV research and prevention opportunities can be particularly valuable.

Increasing Internet accessibility also can improve the delivery of health services among all MSM. Internet usage is highest among young Americans: 96% of 18-29 year olds used the Internet in 2015 [9]. This high coverage provides an opportunity to deliver Web-based HIV prevention and treatment information on a large scale [10,11]. However, HIV prevention research is first needed to establish the efficacy of prevention programs.

The Internet is increasingly being used to recruit underserved MSM and engage them in HIV prevention research [7]. Compared with face-to-face interviews, Web-based research can collect data from a large number of people in a short period of time [12]. Web-based studies can anonymously include MSM who might be stigmatized if their sexual interests were known publicly [13]. Furthermore, Web-based research can provide access to men who might be at an elevated risk of HIV infection. Compared with participants recruited at physical venues, MSM recruited through the Web are more likely to self-report sexually transmitted infections (STIs), anal intercourse, and unprotected anal intercourse [14-18].

However, there are challenges to using the Internet for HIV prevention research [19]. The data can be subject to selection bias because of demographic differences in Internet use and access [9]. Willingness to click on an advertisement, provide consent, and begin a Web-based survey differs by age, race, education, and urbanicity of residence [20]. Furthermore, there are ethical challenges that are unique to Web-based research [11,21]. When consent is administered on the Web, it is difficult to confirm the age, competency, and comprehension of a potential participant [22]. This is particularly concerning because many of the topics that are typically covered in HIV prevention research (sexual behavior, drug use, commercial sex work, and so on) contain confidential and sensitive information.

Most Web-based consent processes involve the respondent reading a document and then indicating that they agree to participate by clicking a button. This is similar to the lengthy terms-of-service documents that many Internet and mobile phone services require, and participants may develop a habit of accepting documents without fully reviewing them. An ethical review of a Web-based HIV prevention study called for innovative ways to provide informed consent to participants [23]. Better alignment between a participant's typical Web-based experience and the informed consent process may improve their consumption of Web-based consent information.

Many Web-based social media interactions now involve the use of photos and videos [24]. Internet users may also be more acclimated to list-based or brief summaries when they do choose to consume textual information. Previous Web-based research with MSM has also shown that brief consent summaries may improve consent information comprehension over a standard consent process, but this research was done before Web-based interactions became more photo- or video-focused and therefore we did not examine this alternative [19,25]. The objective of our study was to assess if comprehension of key informed consent facts improved through the use of these new techniques versus a standard text-based informed consent form.

Methods

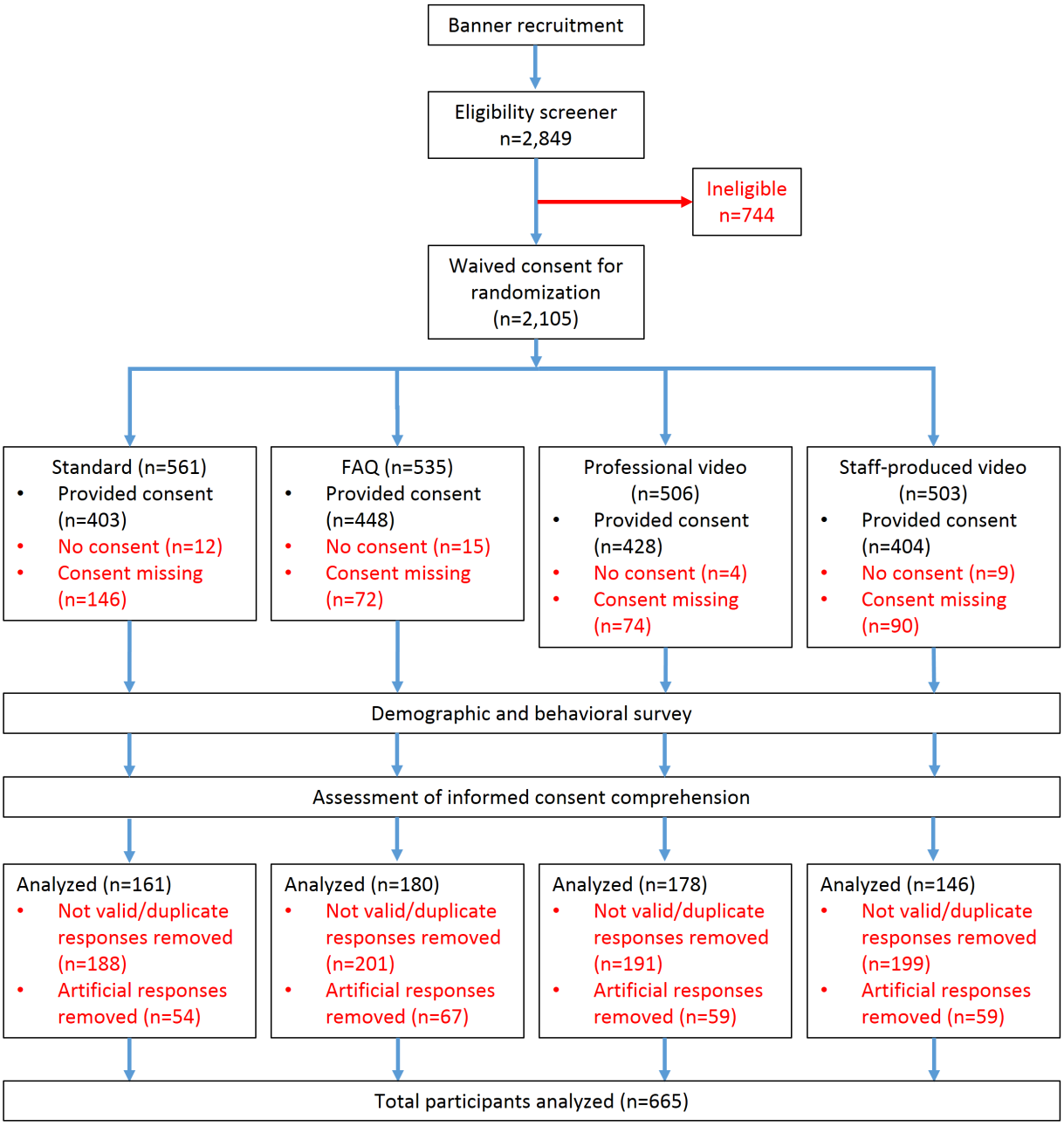
Study Design

A randomized trial (Clinicaltrials.gov NCT02139566) was conducted to assess alternative methods of administering informed consent in a Web-based HIV behavioral research study with young adult MSM residing in the United States. All survey data were collected through a Health Insurance Portability and Accountability Act (HIPAA)–compliant Web-based survey platform (SurveyGizmo, Boulder, CO). Participants were recruited from Facebook from July 28 to September 8, 2014. Banner advertisements were presented to men who indicated they were interested in media and services targeted toward gay men or interested in men on their Facebook profiles. Respondents who clicked on the banners were taken to an eligibility screener. To be eligible, respondents had to report that they lived in the United States, were 18 to 34 years old, were male at birth, and have had sex with another man in the past 12 months. Eligible respondents were randomized by the survey into one of the 4 consent groups: standard informed consent, frequently asked questions (FAQs) informed consent,

a video informed consent produced by a professional company (Tom Coggia, iconAtomic), or a video informed consent produced by study staff (Figure 1). The 3 alternative consent methods were developed based on the results of 5 online focus groups which engaged MSM in discussions about how they considered and interacted with Web-based informed consent documents.

The standard informed consent consisted of a consent document presented in a scrolling window. The FAQ informed consent consisted of 19 questions (e.g., "What is the purpose of the study?","Who is conducting the study?") that revealed short paragraphs of information when the participant clicked on them (Multimedia Appendix 1). The 2 video consents were each about 3.5 minutes long and covered the same topics as the standard and FAQ consents, but used a different script written by the researchers and were presented verbally with bullets of key points highlighted on the screen (Multimedia Appendices 2 and 3). The level of reading comprehension was consistent between the video scripts, the standard consent, and the FAQ consent. All respondents also had the option to open or download the full standard informed consent document as a portable document format (PDF). This study was reviewed and approved by Emory University Institutional Review Board that approved a waiver of written informed consent. After providing consent, participants were directed to a 15-minute Web-based survey on demographic characteristics, sexual history, HIV and STI testing history, and use of HIV prevention services. At the end of the survey, we asked 15 questions about essential elements covered in the informed consent process (Multimedia Appendix 4). Participants did not know there would be questions about the informed consent process. Participants who completed the survey and answered the comprehension questions were given a US $20 Amazon gift card sent by email.

Figure 1. Consort diagram and study scheme for participant enrollment of men who have sex with men in a randomized trial of informed consent methods, United States, 2014.



Measures

The primary analytic outcome was total consent comprehension score, which was calculated by tabulating the number of consent comprehension questions that each participant got correct. Similar to previously used methods, 1 point was given for each correct question for a minimum total score of 0 and maximum of 15 [26]. The 15 consent comprehension questions were selected from a larger pool of 29 possible questions with multiple questions from each of the essential elements of informed consent [27]. Before the main study, consent comprehension questions were pilot-tested among MSM recruited in the same manner as the main study (banner advertisements on Facebook) and were presented with a standard consent form. The pilot study recruited 132 MSM who completed a short survey and the 29 comprehension questions. The data were analyzed for frequencies of correct responses to the 29 pilot consent comprehension questions. At least one question from each of the essential elements was retained. If there were multiple questions from a category, questions with very low or very high correct response rates were removed in order to allow for variability in responses to assess difference in comprehension during the main study.

Time spent on consent was measured by amount of time the respondent stayed on the Web page with the consent information. In order to progress to the next page, respondents had to click on a button indicating they agree (or do not agree)

to participate in the study. Demographic variables (race or ethnicity, highest level of education, and age) were summarized as frequencies and percentages in each consent group. Race or ethnicity was coded into 6 categories: Asian, black, Hispanic, multiracial, American Indian or Alaska Native, and white.

Analyses

All analyses were done in SAS version 9.4 (SAS Institute, Cary, NC). As recommended for Web-based participant recruitment, a deduplication protocol was determined before data collection to remove duplicate or artificial survey attempts [28]. While implementing the study, we noted a pattern of frequent responses from the same Internet protocol (IP) address, very short survey completion times, and unusual email addresses—all indicators of potential fraudulent responses from artificial hacking or bot programs possibly aimed at getting the monetary incentive. We immediately put measures in place to curtail this activity, including Completely Automated Public Turing test to tell Computers and Humans Apart (CAPTCHA) codes and verification of email addresses submitted for incentives [29]. For purposes of analyses, the following types of responses were excluded: a single IP address submitted 5 or more responses, total completed survey time was less than 5 minutes, or invalid email address as determined by staff member reviewing all submitted emails. All analysis was intent-to-treat and all participants were analyzed in the original randomization groups.

Pearson chi-square tests were used to assess demographic differences by consent randomization group. Time spent on consent and consent comprehension score were analyzed as continuous variables and summarized using medians and interquartile ranges (IQRs) or means and standard deviations, where appropriate. Residuals for both continuous variables were gauged for normality by inspecting histograms, boxplots, and quantile-quantile probability plots. As the time spent on consent was not normally distributed, Kruskal-Wallis tests were used to assess differences in time spent on consent by demographic variables. Time spent on consent was log transformed to assess the correlation with consent comprehension score using Pearson correlation coefficient.

Factors associated with consent comprehension score were assessed using general linear regression. Analyses were conducted for consent group, age, race, and highest level of education. Least-squares means and 95% CIs were reported. When overall significant differences were found, pairwise comparisons were considered using the Tukey-Kramer method [30]. A mediation analysis was conducted to determine if time spent on consent was a mediator in the relationship between consent group and consent comprehension score. Direct effects were calculated to determine the change in consent comprehension score when time on consent was held constant. Indirect effects were calculated to determine the change in consent comprehension score when consent group was held constant and time on consent differed by the amount it differed between consent groups. We used methods for mediation analysis with a multicategorical independent variable as described by Hayes et al [31]. Asymmetric bootstrap procedure with 10,000 resamples was used to calculate bias-corrected CIs around the relative indirect effect estimates. All statistical tests were assessed at alpha=.05.

Results

Of the 2849 survey responses, there were 665 (23.34%, 665/2849) eligible participants who consented to be part of the study and complete the survey and comprehension questions (Figure 1). There were 161 (24.2%, 161/665) participants in the standard consent group, 180 (27.1%, 180/665) in the FAQ group, 178 (26.8%, 178/665) in the professional video group, and 146 (22.0%, 146/665) in the staff video group (Table 1). When subsetting the data to only include survey responses submitted after the artificial hacking or bot was curtailed, the percentage of valid, eligible surveys didn't substantially differ (data not presented).

Table 1. Demographic characteristics of 665 men who have sex with men who participated in a randomized trial of informed consent methods, by informed consent group randomization, United States, 2014.

Demographic characteristics			Consent group				
		n (%)	Standard n (%)	FAQ[a] n (%)	Professionally produced video n (%)	Staff-produced video n (%)	*P* value[b]
All			161 (24.2)	180 (27.1)	178 (26.8)	146 (22.0)	
Race or ethnicity							.13
	Asian or Pacific Islander	24 (3.6)	3 (1.9)	11 (6.2)	4 (2.3)	6 (4.1)	
	Black	22 (3.3)	5 (3.1)	5 (2.8)	10 (5.7)	2 (1.4)	
	Hispanic	147 (22.1)	31 (19.4)	45 (25.4)	39 (22.3)	32 (22.1)	
	Multiracial	26 (3.9)	6 (3.8)	8 (4.5)	8 (4.6)	4 (2.8)	
	American Indian or Alaska Native	8 (1.2)	1 (0.6)	5 (2.8)	0 (0)	2 (1.4)	
	White	430 (64.7)	114 (71.3)	103 (58.2)	114 (65.1)	99 (68.3)	
	Missing	8 (1.2)					
Highest level of schooling							.73
	≤ High school	88 (13.3)	18 (11.2)	25 (14.0)	29 (16.4)	16 (11.0)	
	Some college	288 (43.4)	75 (46.6)	73 (40.8)	80 (45.2)	60 (41.1)	
	Bachelor's degree	183 (27.6)	46 (28.6)	52 (29.1)	42 (23.7)	43 (29.5)	
	Graduate or professional	104 (15.7)	22 (13.7)	29 (16.2)	26 (14.7)	27 (18.5)	
	Missing	2					
Age, in years							.34
	18-22	164 (24.7)	46 (28.6)	45 (25.0)	47 (26.4)	26 (17.8)	
	23-27	236 (35.5)	57 (35.4)	58 (32.2)	65 (36.5)	56 (38.4)	
	28-34	265 (39.8)	58 (36.0)	77 (42.8)	66 (37.1)	64 (43.8)	
Age in years, median (IQR[c])		26 (23-29)	25 (22-28)	26 (23-30)	26 (22-29)	26 (24-30)	.10

[a]FAQ: frequently asked questions.

[b]Calculated by Pearson chi-square test for categorical variables and by Kruskal-Wallis test for age as a continuous variable. All tests were assessed with alpha=.05.

[c]IQR: Interquartile range.

Of the participants that reported a race or ethnicity, most identified as white (64.7%, 430/657) or Hispanic (22.1%, 147/657). The median age was 26 years (IQR: 23-39). There were 88 (13.3%, 88/663) participants who reported that they did not receive any schooling past high school. There was not a statistically significant difference in participant race, highest level of education, or age across the 4 randomized consent groups.

Among all participants, the median time spent on the informed consent was 60.5 seconds (Table 2). The time spent on the consent page significantly differed by type of consent (*P*<.001), with participants in the professional video (median=115.0 seconds, IQR: 37.0-208.0) and staff video (median=117.0 seconds, IQR: 39.0-212.0) consent groups spending more time than participants in the standard (median=37.0 seconds, IQR: 14.5-88.0) and FAQ (median=20.5 seconds, IQR: 11.0-82.0) consent groups. About 36.5% (65/178) of participants in the professional video group and 19.9% (29/146) of participants in the staff video group stayed on the consent page for an amount of time that was longer than the duration of the video (professional video: 200 seconds; staff video: 218 seconds, Figure 2). The amount of time spent on the informed consent page differed by race (*P*=.02) and education (*P*=.01). Time on consent was significantly correlated with consent comprehension score in all consent groups (all *P*<.001; Figure 2).

Table 2. Time spent on informed consent and consent comprehension scores among men who have sex with men in a randomized trial of informed consent methods, United States, 2014.

Characteristic	Time spent on consent[a]				Consent comprehension score					
								Post-hoc Tukey comparisons[f]		
	n	Median	IQR[b]	*P* value[c]	n	Mean (SD)[d]	ANOVA[e] *P* value	Mean difference	95% CI	*P* value
All	660	60.5	17.0-162.0		665	6.3 (2.9)				
Consent group				<.001			<.001			
Standard	160	37.0	14.5-88.0		161	5.5 (2.7)		Reference[h]		
FAQ[g]	178	20.5	11.0-82.0		180	5.2 (2.7)		−0.30	−1.06 to 0.47	.75
Professional video	177	115.0	37.0-208.0		178	7.3 (2.8)		1.79	1.02 to 2.55	<.001
Staff video	145	117.0	39.0-212.0		146	7.3 (2.8)		1.79	0.99 to 2.59	<.001
Race or ethnicity				.02			<.001			
Asian	24	71.0	18.0-212.0		24	4.9 (2.8)		−1.77	−3.47 to −0.07	.04
Black	22	95.5	30.0-200.0		22	7.1 (3.0)		0.44	−1.33 to 2.22	.98
Hispanic	147	25.0	13.0-146.0		147	5.5 (2.4)		−1.19	−1.97 to −0.42	<.001
Multiracial	26	67.0	24.0-201.0		26	6.4 (2.4)		−0.22	−1.86 to 1.41	.99
American Indian or Alaska Native	8	27.5	17.0-64.5		8	3.1 (1.6)		−3.52	−6.41 to −0.63	.01
White	425	68.0	20.0-166.0		430	6.7 (3.0)		Reference		
Highest level of education				.008			.001			
≤ High school	88	39.0	14.0-103.5		88	5.5 (2.0)		Reference		
Some college	287	53.0	15.0-154.0		288	6.2 (3.0)		0.63	−0.26 to 1.53	.27
Bachelor's degree	180	65.0	18.0-173.0		183	6.4 (3.0)		0.85	−0.10 to 1.81	.10
Graduate or professional	103	93.0	37.0-205.0		104	7.1 (3.0)		1.61	0.55 to 2.68	.01
Age, in years				.14			.04			
18-22	164	52.5	16.0-157.0		164	6.1 (2.5)		Reference		
23-27	234	71.0	21.0-168.0		236	6.7 (3.0)		0.57	−0.12 to 1.26	.13
28-34	262	53.5	15.0-165.0		265	6.1 (3.0)		−0.03	−0.70 to 0.64	.99

[a]In seconds.

[b]IQR: interquartile range.

[c]Calculated using Kruskal-Wallis tests with alpha=.05.

[d]SD: standard deviation.

[e]ANOVA: analysis of variance.

[f]The null value for mean difference is 0.0.

[g]FAQ: frequently asked questions.

[h]Reference: Reference category for all mean difference comparisons within each characteristic.

Participants in each consent group recorded the following mean comprehension scores: professional video, 7.3 (SD 2.8, range: 1-13); staff-produced video, 7.3 (SD 2.8, range: 1-13); FAQ, 5.2 (SD 2.7, range: 0-13); and standard consent, 5.5 (SD 2.7, range: 1-13; Figure 3). The average consent comprehension score differed significantly by consent group (*P*<.001), age (*P*=.04), race or ethnicity (*P*<.001), and highest level of education (*P*=.001; Table 2). On average, participants in the professional video group scored 1.79 (95% CI 1.02-2.55) points higher than participants in the standard consent group. Similarity, the average score of participants who were in the staff-produced video group was 1.79 (95% CI 0.99-2.59) points higher than the standard consent group. There was not a significant difference in the average scores of participants in the FAQ and standard consent groups (*P*=.75). The number of correct responses significantly differed by consent group for 11 of the 15 consent comprehension questions (Multimedia Appendix 5).

From the mediation analysis, the direct effects of both the professional video (0.93, 95% CI 0.39-1.48) and the staff-produced video (0.99, 95% CI 0.42-1.56) were statistically significant when time spent on consent was held constant. Both video consent methods (relative to standard consent) also indirectly influence consent comprehension score through time spent on consent (professional video 95% CI 0.57-1.16, staff-produced video 95% CI 0.55-1.10; Table 3). This indicates that time on consent functions as a partial mediator of the effect of consent group on consent comprehension score.

Table 3. Results of a mediation analysis of time spent on consent in the relationship between consent type and consent comprehension score among 665 men who have sex with men in a randomized trial of informed consent methods, United States, 2014.

Consent group	Total effect			Direct effect			Relative indirect effect through time spent on consent		
	Coefficient[a]	SE[e]	95% CI	Coefficient[b]	SE	95% CI	Coefficient[c]	Boot SE	BC[d] bootstrap 95% CI
FAQ[f]	−0.32	0.30	−0.91 to 0.26	−0.27	0.27	−0.80 to 0.25	−0.05	0.11	−0.27 to 0.16
Professional video	1.78	0.30	1.19 to 2.36	.93	0.28	0.39 to 1.48	.85	0.15	0.57 to 1.16
Staff video	1.80	0.31	1.19 to 2.42	.99	0.29	0.42 to 1.56	.81	0.14	0.55 to 1.10

[a]Total change in average consent comprehension score.

[b]Change in consent comprehension score when time spent on consent is held constant.

[c]Change in consent comprehension score when consent group is held constant and time spent on consent changes by the amount it differs by consent group.

[d]BC: bias-corrected; number of bootstrap resamples=10,000.

[e]SE: standard error.

[f]FAQ: frequently asked questions.

Figure 2. Correlation of time spent on informed consent and consent comprehension score, by type of informed consent, among 665 MSM in a randomized trial of informed consent methods, United States, 2014.

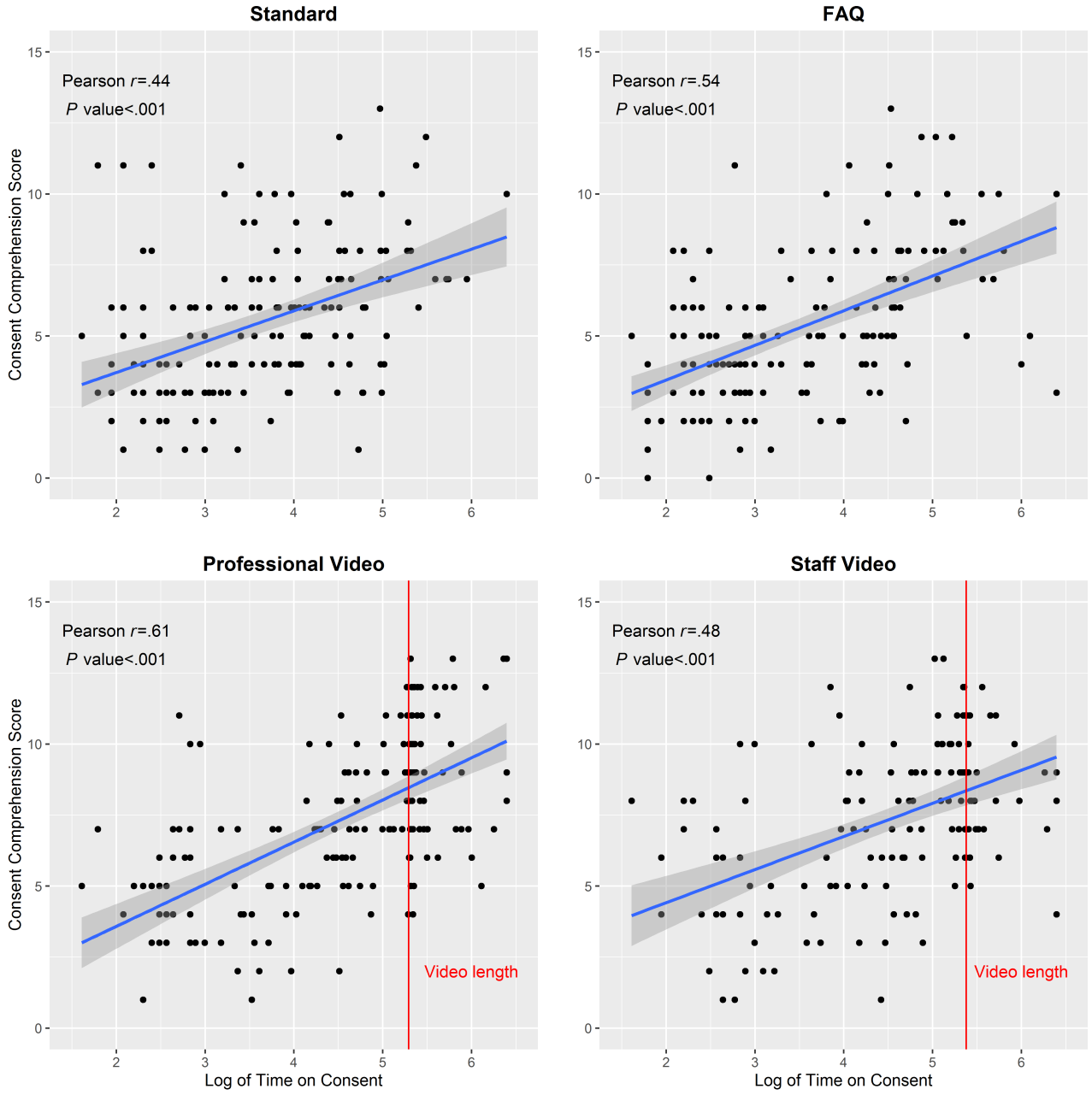
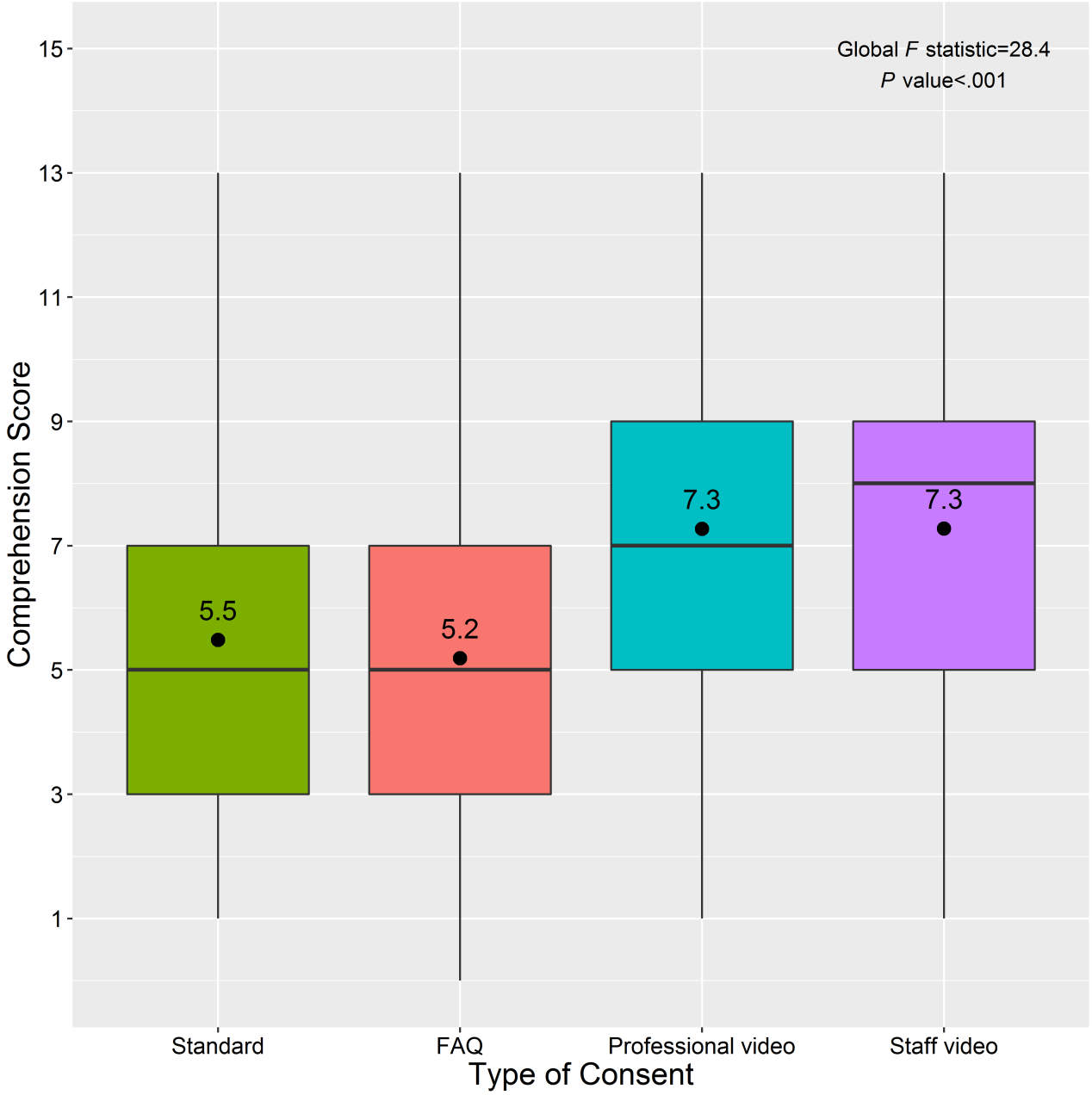


Figure 3. Boxplot of consent comprehension scores, by type of informed consent, among 665 men who have sex with men in a randomized trial of informed consent methods, United States, 2014.



Discussion

Principal Findings

The results of our study highlight the need to improve consent comprehension in Web-based HIV research among MSM. On average, participants were missing one-half (video consent group) to two-thirds (standard consent group) of the informed consent information deemed essential for understanding their participation in it. In particular, questions that focused on personal health information, contacts for questions about the study, or HIPAA authorization were answered correctly by less than half of the participants in every group (Multimedia Appendix 5).

Our study also indicates that video-based methods of administering informed consent can lead to increased consent comprehension (compared with traditional consent forms) in Web-based research. Similar improvements in consent comprehension have been seen when video-based consent is used in clinical and surgical settings [32]. The 4 questions that did not see a difference in consent comprehension focused on benefits of participating (“What benefit can you or others reasonably expect from this research?”), voluntary participation (“True or false: Participation in this study is voluntary”), questions about the study (“Who can you contact if you have questions, concerns or complaints about the study?”), and revoking HIPAA authorization (“What will happen if you revoke your authorization?”). The latter 2 questions were incorrectly answered by the majority of participants in each consent group. Researchers may need to provide these questions early in the consent process or find other ways to highlight this information.

Although the median amount of time spent on consent was much higher in the 2 video groups, many participants still passed through the consent page quickly. About half of the participants in the FAQ group spent less than 20.5 seconds on the consent page, indicating many people read very quickly or did not read every question. Whereas the videos may have held the attention of participants longer, half of the participants in the video group spent less than 2 minutes on the consent page, even though the videos were 3.5 minutes long. It is not clear why time spent on consent varied by race or ethnicity.

Comprehension test scores were almost identical between the 2 video groups, indicating a professionally produced video does not provide additional understanding of informed consent over a staff-produced video. The FAQ group did not demonstrate higher consent comprehension compared with the standard group. This could be because the FAQ consent page still required participants to click on each question and read several lines of text, thus making it less accessible to participants who have lower levels of literacy or susceptible to respondents who click on each question but don't read each one. In addition, the FAQ consent page did not prohibit users from "clicking through" to the survey, even if they did not open any of the FAQ topics.

Time spent on the consent page was significantly correlated with consent comprehension across all groups. Although the correlation coefficient was largest among participants in the FAQ group, comprehension improved as time spent on the consent page (and, presumably, engagement with the consent materials) increased in all groups. The results from the mediation analysis indicate that time on consent does not entirely explain the association between consent group and consent comprehension, which means participants may understand the content in the video better than other methods. Our survey design did not prevent participants from clicking through to the next page before the entire video had finished, and a large number of participants in each group remained on the video page for a period of time that was shorter than the duration of the video. To ensure that all participants are presented the necessary consent information, researchers could design a survey that prevents the user from advancing to the next page until after the video has finished.

Limitations

There are several limitations to consider when interpreting these results. First, with all Web-based research, decisions made during the deduplication process can lead to biased results [28,33]. As a result of possible fraudulent activity, we had to exclude a large proportion of surveys from this analysis to be most conservative. Though other Web-based research studies that offer incentives have reported similar levels of potentially duplicate or fraudulent surveys, this was uncharacteristically high for many other Web-based research studies with MSM that use a similar deduplication process [33-35]. It is possible our process excluded legitimate responses or failed to exclude some fraudulent responses which may bias findings.

Second, we tested comprehension of consent materials using a survey only among respondents who provided their consent to participate in the main study. We don't know how many legitimate respondents clicked on the advertisement, encountered the consent materials, and then declined to participate. Although the data would allow us to determine the number of survey responses that landed on the consent page and the number that agreed to participate by each consent randomization group, those data include artificial and duplicate responses. It is possible that one of the consent methods was more likely to discourage participation in the study and if this behavior was also associated with consent, comprehension could lead to biased results. Similarly, all participants had the opportunity to download a PDF version of the written consent and we were not able to track which participants did that. It is possible, though unlikely, that participants in some groups were more likely to download the PDF version. Future implementation of alternative consent methods will also include this option in order to ensure that guidelines for the documentation of informed consent are met [27].

Recruitment for this study targeted social media users and this convenience sample may not be representative of the general community of MSM. However, this type of recruitment is common in research among MSM and allows researchers to collect behavior data for a large number of MSM in a short period of time [36]. Finally, our sample is limited to MSM who report living within the United States, which limits the generalizability of these results to other groups or cultures affected by the HIV epidemic. Although some Web-based HIV risk and prevention research has been conducted in low or middle income countries [37,38], there are additional ethical concerns that must be considered when designing informed consent procedures in these settings [39,40].

Despite these limitations, this research advances the development of innovative methods of administering informed consent in Web-based HIV prevention research among MSM. Although our study tested consent comprehension for a comparative analysis, researchers might also include a Web-based quiz of consent information and require correct answers before proceeding to the actual research [22]. Web-based HIV prevention researchers can also adapt the "teaching then testing" methods used in drug or alcohol abuse research [41].

Future Research

Future research should continue to improve on these innovative consent procedures and ensure they are optimized for use on mobile phone apps. About 64% of American adults now own a mobile phone that can access the Internet [42] and there is indication that MSM continue to adopt mobile phone technology faster than the general population [43] and are increasingly seeking sexual partners on mobile phone apps [7]. Social media mobile sites and mobile phone apps will continue to be a source of recruitment for HIV prevention research. In the current environment, using videos to administer informed consent are likely to be more effective than traditional, text-based approaches. People are currently used to interacting and receiving information through Web-based videos and matching research processes to these user experiences can improve how we conduct key activities. However, the continued use of studies similar to this one will help determine best practices as they arise and develop in this changing environment.

Acknowledgments

This work was funded by the Eunice Kennedy Shriver National Institute of Child Health and Human Development (R21HD074502) and the Emory University Center for AIDS Research (P30AI050409).

Conflicts of Interest

None declared.

Multimedia Appendix 1

Screenshot of frequently asked questions (FAQ) consent form used in a randomized trial of informed consent methods, United States, 2014.

[PNG File, 84KB - jmir_v19i3e64_app1.png]

Multimedia Appendix 2

Staff-produced video used in a randomized trial of informed consent methods, United States, 2014.

[MP4 File (MP4 Video), 52MB - jmir_v19i3e64_app2.mp4]

Multimedia Appendix 3

Professional video used in a randomized trial of informed consent methods, United States, 2014.

[MP4 File (MP4 Video), 40MB - jmir_v19i3e64_app3.mp4]

Multimedia Appendix 4

Consent comprehension questions used in a randomized trial of informed consent methods, United States, 2014.

[PDF File (Adobe PDF File), 306KB - jmir_v19i3e64_app4.pdf]

Multimedia Appendix 5

Number of correct responses to each consent comprehension question, by type of informed consent, in a randomized trial of informed consent methods, United States, 2014.

[PDF File (Adobe PDF File), 274KB - jmir_v19i3e64_app5.pdf]

References

1. CDC. HIV Surveillance Report URL: http://www.cdc.gov/hiv/library/reports/surveillance/ [accessed 2016-09-22] [WebCite Cache ID 6kimQTkmm]
2. Beyrer C, Sullivan P, Sanchez J, Baral SD, Collins C, Wirtz AL, et al. The increase in global HIV epidemics in MSM. AIDS 2013 Nov 13;27(17):2665-2678. [doi: 10.1097/01.aids.0000432449.30239.fe] [Medline: 23842129]
3. Beyrer C, Baral SD, van Griensven F, Goodreau SM, Chariyalertsak S, Wirtz AL, et al. Global epidemiology of HIV infection in men who have sex with men. Lancet 2012 Jul 28;380(9839):367-377 [FREE Full text] [doi: 10.1016/S0140-6736(12)60821-6] [Medline: 22819660]
4. Sullivan PS, Hamouda O, Delpech V, Geduld JE, Prejean J, Semaille C, Annecy MSM Epidemiology Study Group. Reemergence of the HIV epidemic among men who have sex with men in North America, Western Europe, and Australia, 1996-2005. Ann Epidemiol 2009 Jun;19(6):423-431. [doi: 10.1016/j.annepidem.2009.03.004] [Medline: 19460672]
5. Liau A, Millett G, Marks G. Meta-analytic examination of online sex-seeking and sexual risk behavior among men who have sex with men. Sex Transm Dis 2006 Sep;33(9):576-584. [doi: 10.1097/01.olq.0000204710.35332.c5] [Medline: 16540884]
6. Garofalo R, Herrick A, Mustanski BS, Donenberg GR. Tip of the Iceberg: young men who have sex with men, the Internet, and HIV risk. Am J Public Health 2007 Jun;97(6):1113-1117. [doi: 10.2105/AJPH.2005.075630] [Medline: 17463378]
7. Grov C, Breslow AS, Newcomb ME, Rosenberger JG, Bauermeister JA. Gay and bisexual men's use of the Internet: research from the 1990s through 2013. J Sex Res 2014;51(4):390-409 [FREE Full text] [doi: 10.1080/00224499.2013.871626] [Medline: 24754360]

8. Lewnard JA, Berrang-Ford L. Internet-based partner selection and risk for unprotected anal intercourse in sexual encounters among men who have sex with men: a meta-analysis of observational studies. Sex Transm Infect 2014 Jun;90(4):290-296. [doi: 10.1136/sextrans-2013-051332] [Medline: 24518249]
9. Perrin A, Duggan M. Pew Research Center. Americans' Internet Access: 2000-2015 URL: http://www.pewinternet.org/2015/06/26/americans-internet-access-2000-2015/ [accessed 2016-09-22] [WebCite Cache ID 6kimkD5Sj]
10. Sullivan PS, Carballo-Diéguez A, Coates T, Goodreau SM, McGowan I, Sanders EJ, et al. Successes and challenges of HIV prevention in men who have sex with men. Lancet 2012 Jul 28;380(9839):388-399 [FREE Full text] [doi: 10.1016/S0140-6736(12)60955-6] [Medline: 22819659]
11. Chiasson MA, Parsons JT, Tesoriero JM, Carballo-Dieguez A, Hirshfield S, Remien RH. HIV behavioral research online. J Urban Health 2006 Jan;83(1):73-85 [FREE Full text] [doi: 10.1007/s11524-005-9008-3] [Medline: 16736356]
12. Rhodes SD, Bowie DA, Hergenrather KC. Collecting behavioural data using the world wide web: considerations for researchers. J Epidemiol Community Health 2003 Jan;57(1):68-73 [FREE Full text] [Medline: 12490652]
13. Ross MW. Typing, doing, and being: sexuality and the internet. J Sex Res 2005 Nov;42(4):342-352. [doi: 10.1080/00224490509552290] [Medline: 19827239]
14. Sanchez T, Smith A, Denson D, Dinenno E, Lansky A. Internet-based methods may reach higher-risk men who have sex with men not reached through venue-based sampling. Open AIDS J 2012;6:83-89 [FREE Full text] [doi: 10.2174/1874613601206010083] [Medline: 23049657]
15. Hernandez-Romieu AC, Sullivan PS, Sanchez TH, Kelley CF, Peterson JL, Del RC, et al. The comparability of men who have sex with men recruited from venue-time-space sampling and facebook: a cohort study. JMIR Res Protoc 2014;3(3):e37 [FREE Full text] [doi: 10.2196/resprot.3342] [Medline: 25048694]
16. Shi L, Wei C, McFarland W, Yan H, Li J, Raymond HF. Comparing samples of men who have sex with men recruited online and in venues, Jiangsu Province, China, 2013. LGBT Health 2016 Jun;3(3):238-242 [FREE Full text] [doi: 10.1089/lgbt.2014.0129] [Medline: 26651841]
17. Tsui HY, Lau JT. Comparison of risk behaviors and socio-cultural profile of men who have sex with men survey respondents recruited via venues and the internet. BMC Public Health 2010;10:232 [FREE Full text] [doi: 10.1186/1471-2458-10-232] [Medline: 20444297]
18. Evans AR, Wiggins RD, Mercer CH, Bolding GJ, Elford J. Men who have sex with men in Great Britain: comparison of a self-selected internet sample with a national probability sample. Sex Transm Infect 2007 Jun;83(3):200-5; discussion 205 [FREE Full text] [doi: 10.1136/sti.2006.023283] [Medline: 17135330]
19. Pequegnat W, Rosser BR, Bowen AM, Bull SS, DiClemente RJ, Bockting WO, et al. Conducting Internet-based HIV/STD prevention survey research: considerations in design and evaluation. AIDS Behav 2007 Jul;11(4):505-521. [doi: 10.1007/s10461-006-9172-9] [Medline: 17053853]
20. Sullivan PS, Khosropour CM, Luisi N, Amsden M, Coggia T, Wingood GM, et al. Bias in online recruitment and retention of racial and ethnic minority men who have sex with men. J Med Internet Res 2011;13(2):e38 [FREE Full text] [doi: 10.2196/jmir.1797] [Medline: 21571632]
21. Sullivan PS, Grey JA, Simon RB. Emerging technologies for HIV prevention for MSM: what we have learned, and ways forward. J Acquir Immune Defic Syndr 2013 Jun 1;63 Suppl 1:S102-S107 [FREE Full text] [doi: 10.1097/QAI.0b013e3182949e85] [Medline: 23673879]
22. Curtis BL. Social networking and online recruiting for HIV research: ethical challenges. J Empir Res Hum Res Ethics 2014 Feb;9(1):58-70 [FREE Full text] [doi: 10.1525/jer.2014.9.1.58] [Medline: 24572084]
23. Bull SS, Breslin LT, Wright EE, Black SR, Levine D, Santelli JS. Case study: an ethics case study of HIV prevention research on Facebook: the Just/Us study. J Pediatr Psychol 2011;36(10):1082-1092 [FREE Full text] [doi: 10.1093/jpepsy/jsq126] [Medline: 21292724]
24. Purcell K. Pew Research Center. 2013. Online Video 2013 URL: http://www.pewinternet.org/2013/10/10/online-video-2013/ [accessed 2016-09-22] [WebCite Cache ID 6kimtL6go]
25. Rosser BR, Gurak L, Horvath KJ, Oakes JM, Konstan J, Danilenko GP. The challenges of ensuring participant consent in internet-based sex studies: a case study of the Men's INTernet Sex (MINTS-I and II) studies. J Comput Mediat Commun 2009 Apr;14(3):602-626 [FREE Full text] [doi: 10.1111/j.1083-6101.2009.01455.x] [Medline: 24204104]
26. Arora A, Rajagopalan S, Shafiq N, Pandhi P, Bhalla A, Dhibar DP, et al. Development of tool for the assessment of comprehension of informed consent form in healthy volunteers participating in first-in-human studies. Contemp Clin Trials 2011 Nov;32(6):814-817. [doi: 10.1016/j.cct.2011.05.012] [Medline: 21664491]
27. US Department of Health Human Services. HHS. 2009. Code of Federal Regulations URL: https://www.hhs.gov/ohrp/sites/default/files/ohrp/policy/ohrpregulations.pdf [accessed 2017-02-21] [WebCite Cache ID 6oRrUeX4O]
28. Bauermeister J, Pingel E, Zimmerman M, Couper M, Carballo-Diéguez A, Strecher VJ. Data quality in web-based HIV/AIDS research: handling invalid and suspicious data. Field methods 2012 Aug 1;24(3):272-291 [FREE Full text] [doi: 10.1177/1525822X12443097] [Medline: 23180978]
29. Von AL, Blum M, Hopper N, Langford J. CAPTCHA: Using hard AI problems for security. In: Advances in Cryptology — EUROCRYPT 2003. EUROCRYPT 2003. Lecture Notes in Computer Science, vol 2656. Berlin, Heidelberg: Springer; 2003:294-311.

30. Tukey JW. Comparing individual means in the analysis of variance. Biometrics 1949 Jun;5(2):99-114. [Medline: 18151955]
31. Hayes AF, Preacher KJ. Statistical mediation analysis with a multicategorical independent variable. Br J Math Stat Psychol 2014 Nov;67(3):451-470. [doi: 10.1111/bmsp.12028] [Medline: 24188158]
32. Schenker Y, Fernandez A, Sudore R, Schillinger D. Interventions to improve patient comprehension in informed consent for medical and surgical procedures: a systematic review. Med Decis Making 2011;31(1):151-173. [doi: 10.1177/0272989X10364247] [Medline: 20357225]
33. Grey JA, Konstan J, Iantaffi A, Wilkerson JM, Galos D, Rosser BR. An updated protocol to detect invalid entries in an online survey of men who have sex with men (MSM): how do valid and invalid submissions compare? AIDS Behav 2015 Oct;19(10):1928-1937. [doi: 10.1007/s10461-015-1033-y] [Medline: 25805443]
34. Sanchez TH, Sineath RC, Kahle EM, Tregear SJ, Sullivan PS. The annual American men's internet survey of behaviors of men who have sex with men in the United States: protocol and key indicators report 2013. JMIR Public Health Surveill 2015;1(1):e3 [FREE Full text] [doi: 10.2196/publichealth.4314] [Medline: 27227126]
35. Sanchez T, Zlotorzynska M, Sineath C, Kahle E, Sullivan P. The annual American men's internet survey of behaviors of men who have sex with men in the United States: 2014 key indicators report. JMIR Public Health Surveill 2016;2(1):e23 [FREE Full text] [doi: 10.2196/publichealth.5476] [Medline: 27244770]
36. Sanchez T, Smith A, Denson D, Dinenno E, Lansky A. Developing a web-based HIV behavioral surveillance pilot project among men who have sex with men. Open AIDS J 2012;6:224-231 [FREE Full text] [doi: 10.2174/1874613601206010224] [Medline: 23091580]
37. Stephenson R, Chard A, Finneran C, Sullivan P. Willingness to use couples voluntary counseling and testing services among men who have sex with men in seven countries. AIDS Care 2014 Feb;26(2):191-198 [FREE Full text] [doi: 10.1080/09540121.2013.808731] [Medline: 23786340]
38. Chard AN, Finneran C, Sullivan PS, Stephenson R. Experiences of homophobia among gay and bisexual men: results from a cross-sectional study in seven countries. Cult Health Sex 2015;17(10):1174-1189 [FREE Full text] [doi: 10.1080/13691058.2015.1042917] [Medline: 26096688]
39. Chiu CJ, Menacho L, Fisher C, Young SD. Ethics issues in social media-based HIV prevention in low- and middle-income countries. Camb Q Healthc Ethics 2015 Jul;24(3):303-310. [doi: 10.1017/S0963180114000620] [Medline: 26059956]
40. Krogstad DJ, Diop S, Diallo A, Mzayek F, Keating J, Koita OA, et al. Informed consent in international research: the rationale for different approaches. Am J Trop Med Hyg 2010 Oct;83(4):743-747 [FREE Full text] [doi: 10.4269/ajtmh.2010.10-0014] [Medline: 20889858]
41. Aldridge J, Charles V. Researching the intoxicated: informed consent implications for alcohol and drug research. Drug Alcohol Depend 2008 Mar 1;93(3):191-196. [doi: 10.1016/j.drugalcdep.2007.09.001] [Medline: 17945437]
42. Smith A. Pew Research Center. US smartphone use in 2015 URL: http://www.pewinternet.org/2015/04/01/us-smartphone-use-in-2015/ [accessed 2016-09-22] [WebCite Cache ID 6kinCDjOc]
43. Grov C, Ventuneac A, Rendina HJ, Jimenez RH, Parsons JT. Perceived importance of five different health issues for gay and bisexual men: implications for new directions in health education and prevention. Am J Mens Health 2013 Jul;7(4):274-284 [FREE Full text] [doi: 10.1177/1557988312463419] [Medline: 23093075]

Abbreviations

CAPTCHA: Completely Automated Public Turing test to tell Computers and Humans Apart
FAQ: frequently asked questions
HIV: human immunodeficiency virus
HIPAA: Health Insurance Portability and Accountability Act
IP: internet protocol
IQR: interquartile range
MSM: men who have sex with men
PDF: portable document format
STI: sexually transmitted infection

Edited by G Eysenbach; submitted 28.09.16; peer-reviewed by B Curtis, S Badawy, P Frew; comments to author 01.12.16; revised version received 07.12.16; accepted 08.02.17; published 06.03.17.

Please cite as:
Hall EW, Sanchez TH, Stein AD, Stephenson R, Zlotorzynska M, Sineath RC, Sullivan PS
Use of Videos Improves Informed Consent Comprehension in Web-Based Surveys Among Internet-Using Men Who Have Sex With Men: A Randomized Controlled Trial
J Med Internet Res 2017;19(3):e64
URL: http://www.jmir.org/2017/3/e64/
doi:10.2196/jmir.6710
PMID:28264794

Original Paper

Website Use and Effects of Online Information About Tobacco Additives Among the Dutch General Population: A Randomized Controlled Trial

Dominique A Reinwand[1,2], PhD; Rik Crutzen[2], PhD; Anne S Kienhuis[3], PhD; Reinskje Talhout[3], PhD; Hein de Vries[2], PhD

[1]Rehabilitative Gerontology, Faculty of Human Science, University of Cologne, Cologne, Germany

[2]CAPHRI, Department of Health Promotion, Maastricht University, Maastricht, Netherlands

[3]Center for Health Protection, National Institute for Public Health and the Environment, Bilthoven, Netherlands

Corresponding Author:
Dominique A Reinwand, PhD
Rehabilitative Gerontology
Faculty of Human Science
University of Cologne
Herbert Lewin Str 2
Cologne, 50931
Germany
Phone: 49 2214701529
Fax: 49 2214703793
Email: d.reinwand@uni-koeln.de

Abstract

Background: As a legal obligation, the Dutch government publishes online information about tobacco additives to make sure that it is publicly available. Little is known about the influence this website ("tabakinfo") has on visitors and how the website is evaluated by them.

Objective: This study assesses how visitors use the website and its effect on their knowledge, risk perception, attitude, and smoking behavior. The study will also assess how the website is evaluated by visitors using a sample of the Dutch general population, including smokers and nonsmokers.

Methods: A randomized controlled trial was conducted, recruiting participants from an online panel. At baseline, participants (N=672) were asked to fill out an online questionnaire about tobacco additives. Next, participants were randomly allocated to either one of two experimental groups and invited to visit the website providing information about tobacco additives (either with or without a database containing product-specific information) or to a control group that had no access to the website. After 3 months, follow-up measurements took place.

Results: At follow-up (n=492), no statistically significant differences were found for knowledge, risk perception, attitude, or smoking behavior between the intervention and control groups. Website visits were positively related to younger participants (B=–0.07, 95% CI –0.12 to –0.01; t_{11}=–2.43, *P*=.02) and having a low risk perception toward tobacco additives (B=–0.32, 95% CI –0.63 to –0.02; t_{11}=–2.07, *P*=.04). In comparison, having a lower education (B=–0.67, 95% CI –1.14 to –0.17; t_{11}=–2.65, *P*=.01) was a significant predictor for making less use of the website. Furthermore, the website was evaluated less positively by smokers compared to nonsmokers (t_{324}=–3.55, *P*<.001), and males compared to females (t_{324}=–2.21, *P*=.02).

Conclusions: The website did not change perceptions of tobacco additives or smoking behavior. Further research is necessary to find out how online information can be used to effectively communication about the risks of tobacco additives.

Trial Registration: Nederlands Trial Register NTR4620; http://www.trialregister.nl/trialreg/admin/rctview.asp?TC=4620 (Archived by WebCite at http://www.webcitation.org/6oW7w4Gnj)

(J Med Internet Res 2017;19(3):e60) doi:10.2196/jmir.6785

KEYWORDS

tobacco additives; information dissemination; website use; website evaluation; RCT

Introduction

Cigarettes and other manufactured tobacco products contain numerous additives [1]. A total of 673 different tobacco additives are used during the production of cigarettes [2]. Indeed, one cigarette contains, on average, 68 different additives [3]. Additives such as sugar or vanillin may be perceived as being harmless, but they can develop into harmful carcinogenic substances (eg, formaldehyde) during the combustion process [1,4,5]. Some tobacco additives are thought to cause a higher bioavailability of nicotine, which increases nicotine addiction [6]. From the perspective of those in the tobacco industry, one of the main reasons to use tobacco additives is to improve the taste and to make the smoke milder and consumable [6].

Since 2003, tobacco producers and importers in the Netherlands are legally obliged to list all tobacco additives used in their products and to provide this information to the government [7]. Information about the amount of additives used—including their function and known impact on health—must be declared [8]. Since 2012, the Dutch National Institute for Public Health and the Environment (RIVM) has published information about tobacco additives on a dedicated website ("tabakinfo"). The aim of this website is to provide neutral and objective information to the general population. In contrast to the Netherlands, most other European countries that collect information about tobacco additives do not publish this information. In the United Kingdom, for example, this information is not published because it is feared that people might misunderstand the information [3].

Only two studies—one in Australia [9] and one in the United States [10]—have investigated the influence of the public dissemination of information about components in tobacco or smoke. Results from focus group interviews in Australia demonstrated that information about tobacco additives is desired by the general public. In this case, despite feeling that information about tobacco additives might have an influence on smoking behavior, they were not interested in looking up this information. Moreover, some information appeared to be too detailed or confusing. Additionally, information related to tobacco additives led to some misconceptions. For example, some people thought that cigarettes with fewer additives were less dangerous than cigarettes containing more additives [9].

A cross-sectional quantitative survey in the United States looked at the impact of information about smoke components on smokers and nonsmokers by assessing awareness, worries, and smoking discouragement [10]. The study participants were aware of six of the 20 mentioned components. Reading information related to tobacco additives was associated with increased levels of worry about the harmfulness of substances that were added to tobacco products, such as tobacco additives, compared to substances that naturally occur in cigarette smoke. However, smoke components differ from tobacco additives. Tobacco additives are intentionally added during the manufacturing process of the tobacco product to improve taste or product quality, whereas smoke components are substances to which smokers are exposed to during use of the product. Still, the US study is informative with regard to how people perceive information on tobacco additives.

It is conceivable that information about tobacco additives and smoke components may not be interesting to everyone. Smokers are known to be less interested in information about the risks of smoking [11] in comparison to nonsmokers. It might also be that people with different educational levels understand information about tobacco additives and smoke components differently, or might even misunderstand them. Previous studies that underpin this assumption are rarely about tobacco additives, but have been conducted for potential reduced exposure products (PREPs) (eg, light cigarettes). Some studies demonstrated that smokers believed that light products were less harmful compared to regular cigarettes [12,13]. Although the information given about PREPs in these studies did not contain any statements about positive health outcomes, smokers thought that these products were healthier compared to regular cigarettes [12]. Furthermore, those with a higher education level were more aware of those products. However, no differences were found with regard to risk perceptions of these products when comparing different educational levels [14].

Another study about reactions to reduced risk tobacco advertisements found that people with a lower educational level misinterpreted the advertisement and thought those products were completely free of health risks [15]. Moreover, those with a higher educational level used PREPs more frequently than those with a lower educational level [16]. These results show that information can have a misleading impact. There is a risk that smokers and nonsmokers, as well as people with different educational levels, can misunderstand information provided online about tobacco additives and smoke components. Given that information on tobacco additives is mandatory nowadays, it is important to find out whether this information may result in changing knowledge, risk perceptions, and attitudes about tobacco additives. Indeed, these determinants are important in the intention-forming process in health-related behaviors [17-19]. Currently, there is a lack of insight into these determinants in the Netherlands.

The current website has the aim to provide information to the general public, not to change attitudes and behavior, a goal also strictly guided by policies from the Dutch government on these matters. Yet, because the website could influence attitudes and behavior, these outcomes were also the subject of our evaluation, as well as assessing potential negative side effects of the new website on these outcomes. Therefore, the first aim of our study is to describe the effects of a randomized controlled trial (RCT) in terms of knowledge, risk perception, attitude change about tobacco additives, and smoking behavior. This will be in terms of the amount of daily smoked cigarettes. The second aim is to assess the usage of the website. Finally, we will describe evaluation of the website by the visitors.

Methods

The Tabakinfo Website

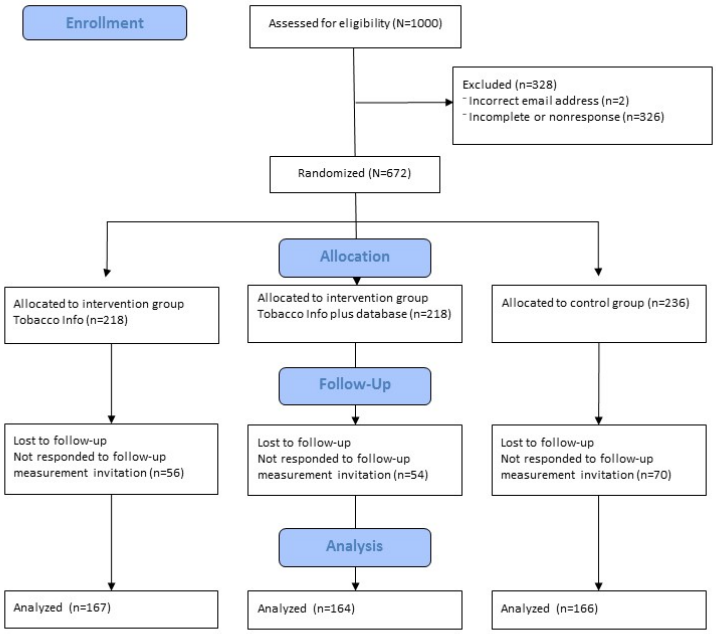
For this study, we made a copy of the original website and deactivated hyperlinks to information that was not about tobacco

additives, which were also available on the original website. By doing this, we wanted to reduce attention to other information and be able to measure the usage behavior of participants. On the tabakinfo website, participants could find information about legislation on tobacco additives, why tobacco additives are used, and in which products they are added. Furthermore, 15 specific tobacco additives (eg, sugar or vanillin) were listed on fact sheets. These provided more information about the function of these additives and their potential harms. The website also provided a database where visitors could search for specific products by brand and get an overview about composition and additives.

Design

An RCT with three study groups (tobacco info, tobacco info plus database, and control group) and two measurement times (baseline and follow-up) was conducted (Figure 1). Participants in all three groups had to fill in the baseline questionnaire. Afterwards, the tobacco info group was asked to visit the website and to read information about tobacco additives. The tobacco info plus database group was asked to visit the website and also the database. They looked up information about tobacco additives for a cigarette brand and type. This information was not available on the website for participants who were allocated to the tobacco info group. This manipulation was done in order to investigate the possible effect of visiting the database. The follow-up measurement took place after 3 months. During the 3 month, participants had the opportunity to visit the original website. The control group only had to complete the baseline and follow-up questionnaires but did not have access to the website. Ethical approval of the Regional Medical Ethics committee in the Netherlands was not necessary because participants in this study were not subjected to procedures or required to follow certain rules of behavior (the criteria for ethical approval) [20].

Figure 1. Flowchart of study design.



Procedure and Participants

Participants older than 18 years were recruited via a Dutch independent Internet panel that operates in line with ISO standards [21]. All members of this panel expressed their willingness to participate in scientific research studies. In total, 1000 panel members were invited to participate in this study, of which 672 panel members did (67.21% response rate). Data collection took place in July (baseline) and October (3-month follow-up) 2014. After receiving an email invitation, participants had 1 week to fill in the questionnaire. One email reminder was sent to participants who had not responded after 5 days. To be able to make a comparison between smoker and nonsmoker perceptions of the website, we aimed to have 50% of the participants in the study sample as smokers and 50% as nonsmokers. Panel members received an email with a link to the online questionnaire. At the end of the questionnaire, a unique link to the study website was included for participants from one of the experimental groups. The email for participants allocated to the control group did not include this. The use of a unique link for each participant from the experimental groups provided the opportunity to track behavior of the participants on the website (ie, what parts of the website the participants visited) using the software ComScore by Sitestat [22]. The invitation for the follow-up measurement was only sent to participants who completed the baseline questionnaire. Participants received incentives in the form of points they can collect and exchange for vouchers.

Questionnaire

The questionnaire was administered in Dutch. Personal characteristics were assessed only at baseline and included gender (1=male, 2=female), age (1=18-19, 2=20-24, 3=25-29, 4=30-34, 5=35-39, 6=40-44, 7=45-49, 8=50-54, 9=55-59,

10=60-64, 11=65 years or older), educational status (1=low: no education, primary or lower vocational school, 2=middle: secondary vocational school or high school, 3=high: higher education or university), and income (1=low: <€12,500; 2=middle: €26,000-€39,000; 3=high: €39,000 or more). As described by Statistics Netherlands, distribution of the sample was representative of the Dutch population with regard to age, gender, income, and level of education [23].

At baseline, participants were asked with one item if they already knew the website: "Are you familiar with the tabakinfo website?" (1=yes, 2=no).

Smoking behavior, knowledge, risk perception, and attitude toward tobacco additives were measured at baseline and at 3-month follow-up. In line with World Health Organization guidelines for measurements of smoking status, all participants were asked whether they smoked at least once a month (1=yes, 0=no) [24]. If yes, participants were categorized as smokers. They were then asked to indicate through multiple choice questions how many they smoked a day, the brand they smoked, and the type of cigarettes they smoked (eg, cigarillos, cigars, rolling tobacco, or pipe).

Omega was used to assess the internal structure of the scales [25]; this value represents a less biased alternative to Cronbach alpha [26]. Similar to Cronbach alpha, omega values can range from 0 to 1, where higher values indicate a more coherent internal structure. In other words, the proportion of variance due to a general factor (ie, omega) provides important information about the extent to which a scale score estimates a latent variable common to all items [27]. Knowledge about tobacco additives was assessed with 10 right or wrong statements (omega=0.82, 95% CI 0.80-0.84). To dichotomize knowledge, we recoded "I don't know" answers as answering the question wrong because they did not know the right answer. For example, "Tobacco with additives consists of less carcinogenic substances compared to tobacco without additives" (wrong).

To measure risk perception toward tobacco additives, six items were assessed; three items covered cognitive aspects and three items covered affective aspects of risk perception. These questions were based on earlier questionnaires assessing risk perception and were adjusted for tobacco additives [28]. Answers were given on a five-point Likert scale (5=totally agree, 1=totally disagree; omega=0.81, 95% CI 0.77-0.84). "If I smoke tobacco with additives I have a high chance of getting cancer," is an example item of a cognitive risk perception item.

Attitude toward tobacco additives was assessed by 14 items on a five-point Likert scale (5=totally agree, 1=totally disagree); there were eight items about the pros with regard to tobacco additives and six items about the cons of tobacco additives (pro: omega=0.83, 95% CI 0.79-0.86; con: omega=0.79, 95% CI 0.76-0.82). These questions were based on an earlier questionnaire assessing attitude and were adjusted for tobacco additives [29,30]. As an example, for a pro attitude item, participants were asked, "If I would smoke tobacco with additives, I feel good."

Items with regard to the evaluation of the questionnaire was included in all studies that were conducted in the panel used in the study at hand. On a visual analog five-point Likert scale, participants rated whether they evaluated the questionnaire as interesting (1) or uninteresting (5), pleasurable (1) or not (5), too long (5) or too short (1), and difficult (1) or easy (5) to answer. We also documented how many minutes it took to complete the questionnaire.

The follow-up measurement contained items about the website evaluation, which were only answered by participants from the two experimental groups. Participants' perceptions about the website were assessed using 10 concepts about diverse aspects of the website, such as content, layout, language, or navigation [31]. These different constructs are useful in evaluating different aspects of visitor experiences on the website. Positive experiences are associated with detailed use [32]. Completeness and layout were assessed with two items, and the other concepts with three items each: efficiency (eg, I easily find information I am looking for on this website; omega=0.92, 95% CI 0.90-0.93), effectiveness (eg, The website provides useful information; omega=0.92, 95% CI 0.91-0.94), enjoyment (eg, I found my visit on this website enjoyable; omega=0.96, 95% CI 0.95-0.96), active trust (eg, I would act on the information presented on this website if needed; omega=0.89, 95% CI 0.86-0.91) [31,33], relevance (eg, The information on the website was new to me; omega=0.78, 95% CI 0.71-0.80), understanding (eg, I found many words on the website difficult to understand; omega=0.83, 95% CI 0.82-0.88), completeness (eg, The website contains enough information; r=.79), layout (eg, I found that the layout of the website looks good; r=.29), recommendation to others (eg, I would recommend the website to others; omega=0.83, 95% CI 0.84-0.89), and intention to revisit (eg, I would revisit the website again; omega=0.84, 95% CI 0.81-0.87). All items could be answered on a five-point Likert scale (1=totally disagree, 3=neither disagree nor agree, 5=totally agree).

Website use was assessed based on tracking the unique links that participants received. Using these unique links, it was possible to see which pages each participant visited.

Statistical Analyses

Data was analyzed using the software SPSS version 24 and R was used to calculate omegas as scale reliability for knowledge, risk perception, attitude pro, attitude con, and website evaluation constructs. Logistic regression analysis was used to evaluate whether dropout was higher among specific subgroups. Frequencies and descriptive statistics were used to describe the study sample at baseline.

Linear regression using the enter method was used to analyze differences in knowledge, risk perception, attitude, and the amount of cigarettes smoked per day between control and study groups after visiting the website. All analysis was done for these five outcome variables, with age, gender, level of education and income, and smoking status as independent variables. Further, interaction between of study group*age, study group*smoking status, and study group*education were used in order to determine any differences among the subgroups using the website. If one of these interaction terms were significant at a

P value of ≤.05, analyses on the relevant subgroups were undertaken. All analyses were corrected for age, gender, educational level, income, and smoking status. When analyzing the follow-up effects of knowledge, risk perception, attitude pro, attitude con, and smoking, the corresponding baseline measurement was included on each of the analysis.

Linear regression analyses using the enter method were conducted to determine the predictors of website use. For this purpose, a continuous dependent variable was calculated which summed up the numbers of visited pages within the website. This included specific information about tobacco additives. Independent variables included age, gender, education, income, knowledge, attitude, risk perception, smoking status, amount of smoked cigarettes, duration of questionnaire, and questionnaire evaluation. In total, 19 participants reported at baseline that they knew the website (tobacco info group: n=9; tobacco info plus database group: n=7; control group: n=3) and were excluded from analyses.

Finally, we analyzed how the website was evaluated by the different subgroups. For this purpose, we conducted independent sample *t* tests to compare males with females, and smokers with nonsmokers, using ANOVA Tukey post hoc tests we compared educational and income levels and used the means of each of the 10 evaluation concepts (efficiency, effectiveness, active trust, relevance, enjoyment, understanding, completeness, layout, recommendation to others, and intention to revisit).

Results

Sample Characteristics and Attrition

The study sample (Table 1) consisted of 672 participants (tobacco info plus database group: n=218; tobacco info group: n=218; control group: n=236) at baseline (male: 54.8%, 368/672). In terms of age and gender, the distribution within the study sample was in line with distribution in the Netherlands, as was distribution of income (high: 31.7%, 213/672; middle: 45.2%, 304/672; low: 23.1%, 155/672) and education (high: 24.6%, 165/672; middle: 43.2%, 290/672; low: 32.3%, 217/672). Furthermore, the sample consisted of 53.9% (632/672) smokers. Table 2 presents perceptions about tobacco additives at baseline and follow-up for each of the study groups.

Table 1. Participant characteristics (N=672).

Variables	n (%)
Age (years)	
18-19	23 (3.4)
20-24	58 (8.9)
25-29	61 (9.1)
30-34	50 (7.4)
35-39	47 (7.0)
40-44	60 (8.9)
45-49	69 (10.3)
50-54	66 (9.8)
55-59	61 (9.1)
60-64	50 (7.4)
≥65	127 (18.9)
Gender	
Male	368 (54.8)
Female	304 (45.2)
Smoking	
Smoker	362 (53.9)
Nonsmoker	310 (46.1)
Education	
Low	217 (32.3)
Middle	290 (43.2)
High	165 (24.6)
Income	
Low	155 (23.1)
Middle	304 (45.2)
High	213 (31.7)

Table 2. Baseline and 3-month follow-up perceptions about tobacco additives.

Variables	Tobacco info group, mean (SD)	Tobacco info plus database group, mean (SD)	Control group, mean (SD)
Knowledge (score 0-1)			
Baseline	0.42 (0.28)	0.44 (0.28)	0.42 (0.29)
Follow-up	0.56 (0.28)	0.54 (0.28)	0.53 (0.26)
Risk perception (score 1-5)			
Baseline	3.45 (0.62)	3.42 (0.61)	3.41 (0.61)
Follow-up	3.47 (0.61)	3.49 (0.60)	3.45 (0.66)
Attitude pro (score 1-5)			
Baseline	2.64 (0.52)	2.54 (0.62)	2.60 (0.64)
Follow-up	2.53 (0.58)	2.53 (0.60)	2.43 (0.63)
Attitude con (score 1-5)			
Baseline	3.40 (0.58)	3.50 (0.64)	3.43 (0.67)
Follow-up	3.48 (0.66)	3.49 (0.61)	3.55 (0 73)

The loss to follow-up after 3 months was 26.8% (180/672). Dropout was significantly higher for participants with a lower income (OR 0.42, 95% CI 0.19-0.90, *P*=.03), higher within the experimental group tobacco info plus database (OR 2.17, 95% CI 1.12-4.22, *P*=.02), and among participants who indicated at baseline that they did not like filling out the questionnaire (OR 0.48, 95% CI 0.30-0.77, *P*=.002). With regard to all other variables, including age, gender, smoking behavior, educational level, evaluation of the questionnaire (clear: mean 1.9, SD 1.1; interesting: mean 2.1, SD 1.2; length: mean 3.0, SD 0.6; pleasurable: mean 2.1, SD 1.1), for knowledge, attitude, and risk perception, no significant differences in dropout were found.

Determinants and Smoking Status

Table 2 indicates the mean values of tobacco additives perceptions at baseline and follow-up. ANOVA and Tukey post hoc tests did not reveal differences between the three study groups. As can been seen in Tables 3 and 4, after the 3-month follow-up, when compared to the control group, neither one of the measured concepts with regard to tobacco additives (knowledge: R^2=.39; risk perception: R^2=.30; attitude pro: R^2=.34; attitude con: R^2=.38) or the amount of cigarettes smoked per day (R^2=.62) changed within the experimental groups. Furthermore, smokers had significantly lower risk perception and a less negative attitude toward tobacco additives than nonsmokers. None of the interaction terms were significant, indicating that there were no subgroup effects for smoking status, people with different educational levels, and age groups.

Table 3. Linear regression results for knowledge, risk perception, and attitude pro after follow-up measurement among study groups.

Variables	Knowledge		Risk perception		Attitude pro	
	B (95% CI)[a]	*P*	B (95% CI)[a]	*P*	B (95% CI)[a]	*P*
Tobacco info group	0.08 (–0.14, 0.30)	.46	0.22 (–0.31, 0.75)	.42	–0.11 (–0.61, 0.39)	.67
Tobacco info + database group	0.16 (–0.07, 0.38)	.17	0.52 (–0.02, 1.07)	.06	–0.27 (–0.78, 0.24)	.30
Age	0.00 (–0.01, 0.01)	.77	0.02 (–0.01, 0.05)	.24	0.01 (–0.02, 0.03)	.70
Gender	–0.01 (–0.06, 0.03)	.52	–0.03 (–0.14, 0.07)	.52	–0.04 (–0.13, 0.06)	.48
Education low	–0.03 (–0.12, 0.07)	.61	–0.17 (–0.41, 0.06)	.14	0.05 (–0.17, 0.27)	.68
Education middle	–0.05 (–0.14, 0.03)	.21	–0.10 (–0.30, 0.10)	.34	–0.06 (–0.25, 013)	.54
Income low	0.01 (–0.04, 0.07)	.65	–0.10 (–0.23, 0.04)	.16	0.05 (–0.08, 0.17)	.48
Income middle	0.00 (–0.05, 0.04)	.93	0.10 (–0.01, 0.21)	.08	0.02 (–0.09, 0.12)	.76
Smoking status	0.05 (–0.02, 0.12)	.17	0.19 (0.02, 0.36)	.02	–0.11 (–0.27, 0.05)	.19
Tobacco info group*age	0.00 (–0.02, 0.02)	.96	–0.02 (–0.06, 0.02)	.27	0.01 (–0.03, 0.05)	.60
Tobacco info + database group*age	–0.01 (–0.03, 0.01)	.26	–0.02 (–0.06, 0.02)	.23	0.00 (–0.04, 0.04)	.90
Tobacco info group*education low	0.01 (–0.12, 0.14)	.88	0.06 (–0.26, 0.38)	.71	–0.05 (–0.35, 0.26)	.77
Tobacco info group*education middle	0.05 (–0.07, 0.17)	.43	0.09 (–0.19, 0.38)	.52	0.00 (–0.27, 0.28)	>.99
Tobacco info + database group*education low	–0.02 (–0.16, 0.11)	.76	0.03 (–0.30, 0.35)	.87	0.25 (–0.06, 0.56)	.11
Tobacco info + database group*education middle	0.01 (–0.11, 0.14)	.82	–0.14 (–0.44, 0.16)	.35	0.12 (–0.17, 0.40)	.42
Tobacco info group*smoking	–0.04 (–0.14, 0.06)	.39	–0.09 (–0.32, 0.15)	.48	0.06 (–0.17, 0.29)	.60
Tobacco info + database group*smoking	–0.04 (–0.14, 0.06)	.42	–0.17 (–0.41, 0.07)	.16	0.16 (–0.06, 0.39)	.16
Baseline assessment[b]	0.60 (0.52, 0.67)	<.001	0.47 (0.39, 0.55)	<.001	0.55 (0.47, 0.63)	<.001

[a] Unstandardized B.

[b] Baseline assessment for the corresponding outcome.

Table 4. Linear regression results for outcome for attitude con and cigarettes per day after follow-up measurement among study groups.

Variables	Attitude con		Cigarettes per day	
	B (95% CI)[a]	*P*	B (95% CI)[a]	*P*
Tobacco info group	0.12 (–0.42, 0.65)	.66	–0.22 (–0.86, 0.42)	.50
Tobacco info + database group	0.40 (–0.14, 0.95)	.15	–0.23 (–0.86, 0.40)	.48
Age	0.00 (–0.03, 0.03)	.90	0.02 (–0.03, 0.07)	.38
Gender	0.06 (–0.05, 0.16)	.27	0.00 (–0.16, 0.15)	.96
Education low	0.06 (–0.18, 0.29)	.64	0.09 (–0.23, 0.41)	.59
Education middle	–0.13 (–0.34, 0.07)	.21	–0.02 (–0.32, 0.29)	.92
Income low	–0.02 (–0.16, 0.11)	.75	0.04 (–0.15, 0.23)	.66
Income middle	0.06 (–0.06, 0.17)	.32	0.13 (–0.03, 0.30)	.11
Smoking status	0.27 (0.10, 0.45)	<.001	—	
Tobacco info group*age	0.00 (–0.04, 0.04)	.85	0.00 (–0.07, 0.06)	.91
Tobacco info + database group*age	–0.01 (–0.05, 0.03)	.69	–0.01 (–0.07, 0.06)	.87
Tobacco info group*education low	–0.12 (–0.44, 0.20)	.48	0.0.9 (–0.36, 0.55)	.68
Tobacco info group*education middle	0.23 (–0.06, 0.53)	.11	0.32 (–0.13, 0.77)	.16
Tobacco info + database group*education low	–0.34 (–0.67,–0.01)	.40	0.21 (–0.25, 0.68)	.37
Tobacco info + database group*education middle	–0.13 (–0.44, 0.17)	.39	0.19 (–0.28, 0.66)	.42
Tobacco info group*smoking	–0.17 (–0.41, 0.07)	.17	—	
Tobacco info + database group*smoking	–0.15 (–0.39, 0.09)	.23	—	
Baseline assessment[B]	0.54 (0.46, 0.63)	<.001	0.75 (0.67, 0.84)	<.001

[a] Unstandardized B.

[b] Baseline assessment for the corresponding outcome.

Website Use

All participants from the tobacco info group visited the website. From the tobacco info plus database group, 159 of 163 participants (97.6%) visited the website; specific information about tobacco additives, another part of the website, was looked up by 129 of 346 (37.3%) participants from both groups. Only a small minority of participants browsed the website to gain more information about one of the 14 specific tobacco additives, such as vanilla (23/346, 6.7%) or sugar (25/346, 7.2%). Of those participants from the tobacco info plus database group who visited the website (n=159), 33.3% (53/159) visited the subwebsite that provided the database.

The model in Table 5 accounted for 78% of the total variance in predicting website usage. It shows that being younger (B=–0.07, t_{11}=–2.43, *P*=.03) and having a low risk perception toward tobacco additives (B=–0.32, t_{11}=–2.07, *P*=.04) were significant predictors for website usage, but having a lower educational level (B=–0.67, t_{11}=–2.65, *P*=.01) was a significant predictor for using the website less.

Table 5. Linear regression analysis of predictors of website use (n=337).

Variables	Website use	
	B (95% CI)[a]	*P*
Smoking	–0.08 (–0.43, 0.27)	.66
Age	–0.07 (–0.12,–0.01)	.02
Gender	–0.15 (–0.48, 0.18)	.37
Education low	–0.66 (–1.14,–0.17)	.05
Education middle	–0.32 (–0.75, 0,11)	.14
Income low	–0.03 (–0.51, 0.44)	.89
Income middle	–0.10 (–0.49, 0.29)	.61
Knowledge	0.43 (–0.21, 1.07)	.19
Risk perception	–0.32 (–0.63,–0.02)	.04
Attitude pro	–0.05 (–0.39, 0.28)	.75
Attitude con	0.10 (–0.25, 0.46)	.57

[a] Unstandardized B.

Website Evaluation

The website was evaluated through use of the following 10 concepts: efficiency, effectiveness, active trust, relevance, enjoyment, understanding, completeness, layout, recommendation to others, and intention to revisit. Table 6 shows the mean scores of these scales for the study sample. As can be seen, participants tended to evaluate all these concepts negatively because their answers ranged between disagree and neutral.

Table 6. Website evaluation.

Evaluation variables	Tobacco info group, mean (SD) (n=158)	Tobacco info plus database group, mean (SD) (n=157)
Enjoyment	2.73 (0.70)	2.38 (0.65)
Layout	2.58 (0.58)	2.67 (0.63)
Intention to revisit	2.74 (0.77)	2.88 (0.86)
Active trust	2.60 (0.70)	2.61 (0.69)
Recommending to others	2.64 (0.70)	2.66 (0.73)
Relevance	2.38 (0 65)	2.48 (0.59)
Completeness	2.32 (0.58)	2.41 (0.69)
Efficiency	2.28 (0.62)	2.38 (0.65)
Effectiveness	2.25 (0.61)	2.33 (0.64)
Understanding	2.22 (0.66)	2.34 (0.65)

[a] Scales: 1=totally disagree, 3=neither disagree nor agree, 5=totally agree.

For all these concepts, we did not find significant differences between participants with different educational levels, income groups, or different ages. Females (mean 2.70, SD 0.73) were found to put a significantly higher degree of active trust in the website compared to males (mean 2.53, SD 0.65; mean difference=–0.17, 95% CI –0.33 to –0.17; t_{324}=–2.21, P=.03). Furthermore, smokers (mean 2.66, SD 0.79) compared to nonsmokers (mean 2.90, SD 0.80) had a significantly lower intention to revisit the website again (mean difference=–0.32, 95% CI –0.49 to –0.14; t_{324}=–3.55, P<.001).

Discussion

Main Findings

This study showed that participants at baseline did not have a high level of knowledge about tobacco additives, which is in line with a recent study from the United States [34]. Furthermore, our participants had no strong positive or negative risk perception or attitude regarding tobacco additives.

We demonstrated that, at 3-month follow-up, visiting the website resulted in no changes between the control group and the experimental groups. This could be explained by several factors. Firstly, the participants from our study visited the website only once. This single exposure to new information might not be

enough to expect any changes in smoking behavior, nor determinants such as knowledge, risk perception, or attitude [35]. Furthermore, participants showed low levels of engagement with the website in terms of visiting subwebsites and specific information. Also, participants did not evaluate the website positively and had low intentions to revisit the website, which could have limited the impact of the website. In addition, it might also be possible that participants from the control group became curious about the topic of tobacco additives and searched for more information about this topic (eg, using Google). Those participants might have visited the original website and received the same information about tobacco additives as the experimental groups or found other information about tobacco additives on the Web.

Secondly, with regard to predictors of website use, we found that personal characteristics, such as an older age, higher educational level, and having a high risk perception toward tobacco additives is associated with more extensive website use. This included visiting more subwebsites (ie, looking up specific information about tobacco additives). This is in line with previous findings that people with a higher socioeconomic status are more interested in health-related topics and search the Internet more frequently for health information [36,37]. Furthermore, a content analysis of information provided online on tobacco indicated that most websites require high grades of reading levels [38]. These requirements might explain why participants with a lower educational level visited less information on the evaluated website. Besides, participants with a low risk perception toward tobacco additives had visited the website in more detail, supporting earlier findings indicating that high risk perception is associated with the avoidance of seeking information [39,40].

In general, the overall results show that only one-third of the participants from the experimental group visited the database, a minority visited the website in detail, and that the evaluation of the website was not positive. It is important to mention that smokers had a lower intention to revisit the website again and these people are the target markets of this website. Visitors should benefit from the website, for example, in terms of increasing their knowledge. There may be several reasons why people do not visit a website in-depth: the layout, design, structure, and function of a website are essential elements that determine how a website will be perceived and used [41]. The evaluation of the website indicates important concepts of the website should be improved, such as understanding, relevance, trust, or enjoyment. Improving the website might motivate visitors to stay longer on the website, which in turn increases the chance that visitors receive information about tobacco additives. Furthermore, the website might be improved by changing the navigation into a clearer display, thereby avoiding hyperlinks within the text. Instead, a navigation path next to the main text may be provided. Given the fact that participants from this study did not visit pages with in-depth information about specific tobacco additives, it might be worth changing the content of the information into more general and less-specific information. This is because the majority of participants only looked at general information. More in-depth information about tobacco additives might be offered on the website and could also be labeled as “in-depth information” to help visitors to distinguish between broad and deep information. It might also be helpful to guide visitors through the website using options such as “what information are you looking for?” These recommendations must be examined in further studies.

Limitations

There are some limitations to this study. Our participants were recruited via a panel and received a reward for participating. The information given on the website might have a greater impact on people who visited the website proactively. Yet, inclusion of visitors only could have jeopardized the generalizability of our findings to the overall Dutch population.

Our model of predicted website use had a very low explained variance (8%). This indicates that there must be other variables that we did not measured and are related to website use. It is conceivable that there are topic-related variables about tobacco additives that we did not assess, such as interest, or variables associated with Internet usage, including health literacy [42].

Furthermore, we only assessed which hyperlinks participants used; other measures of website engagement, such as time people spent online, might be of interest.

Another limitation of our study is that we cannot preclude the possibility that participants from the control group visited the original website between the baseline and the follow-up measurement. This is because the website could be found on the Web just as other websites about tobacco additives on the Web (eg, by means of a Google search). Finally, hyperlinks to topics other than tobacco additives on the website were deactivated, possibly resulting in bias to the observed website usage.

Conclusion

With the tabakinfo website and disclosure of information about tobacco additives, the Dutch government fulfills their requirements to inform the general population [43]. Visiting the website did not influence knowledge, risk perception, attitudes toward tobacco additives, and smoking behavior. If website-based health-related information is to have an impact on concepts such as knowledge, risk perception, or attitude, it will be necessary to adapt the website more to the needs of the visitors. That could be achieved, for example, by making the information easily accessible on the home page, thus avoiding long browsing activities. This is because most of the participants only visited the first subwebsite. Furthermore, it may be desirable to make the website more attractive to those who are less educated and to smokers. It might also be also possible that the website was not encouraging participants to look up information in detail, but this notion needs to be evaluated through further studies. Additionally, as the needs of visitors may differ, tailoring the information to their needs is recommended. Indeed, further research is needed to gain a deeper insight into these needs and requirements.

Acknowledgments

This work and the website were supported by the Dutch Ministry of Health, Welfare and Sport.

Authors' Contributions

All authors contributed to this paper. DR conducted the overall analyses and drafted the paper. HdV and RC contributed to the design of the study. All authors contributed to the interpretation of the data and to the writing of the paper. All authors revised the manuscript critically for important intellectual content and read and approved the final manuscript.

Conflicts of Interest

AK and RT work at the RIVM, which is responsible for the content of the website. The study design, questionnaire development, participant recruitment, and data analysis were performed independently of the RIVM. HdV is the scientific director of Vision2Health, a company that licenses evidence-based innovative computer-tailored health communication tools. DR and RC declare that they have no competing interests.

Multimedia Appendix 1

Questionnaire about tobacco additives.

[PDF File (Adobe PDF File), 26KB - jmir_v19i3e60_app1.pdf]

References

1. Schmoldt A, Benthe HF, Haberland G. Digitoxin metabolism by rat liver microsomes. Biochem Pharmacol 1975 Sep 1;24(17):1639-1641. [Medline: 10]
2. Schenk E, van de Nobelen S, Pennings J, Kienhuis A, Talhout R. Additieven in Nederlandse tabaksproducten. Trendanalyse gegevens 2010-2014. Bilthoven: RIVM; 2015. URL: http://www.rivm.nl/dsresource?objectid=64870e2a-0180-401b-8b3a-ae381c260bad&type=org&disposition=inline [WebCite Cache ID 6oAeN5pyv]
3. Talhout R. Rapportage en regelgeving van tabaksingredienten. Een vergelijking van Nederland met andere landen. Bilhoven: RIVM; 2012. URL: http://www.rivm.nl/bibliotheek/rapporten/340610004.pdf [WebCite Cache ID 6oAej3fxP]
4. Swenberg JA, Moeller BC, Lu K, Rager JE, Fry RC, Starr TB. Formaldehyde carcinogenicity research: 30 years and counting for mode of action, epidemiology, and cancer risk assessment. Toxicol Pathol 2013 Feb;41(2):181-189 [FREE Full text] [doi: 10.1177/0192623312466459] [Medline: 23160431]
5. Songur A, Ozen O, Sarsilmaz M. The toxic effects of formaldehyde on the nervous system. In: Reviews of Environmental Contamination and Toxicology. New York: Springer; 2010:105-118.
6. Rabinoff M, Caskey N, Rissling A, Park C. Pharmacological and chemical effects of cigarette additives. Am J Public Health 2007 Nov;97(11):1981-1991. [doi: 10.2105/AJPH.2005.078014] [Medline: 17666709]
7. RIVM. 2015. EMTOC URL: http://www.rivm.nl/en/Topics/T/Tobacco/EMTOC [accessed 2016-10-11] [WebCite Cache ID 6lArXirOf]
8. Kienhuis A, Schenk E, Talhout R. Publieksinformatie over additieven in tabaksproducten, the right to know. Bilthoven: RIVM; 2013. URL: http://www.rivm.nl/dsresource?objectid=d6ebd9de-3044-41c8-8275-4a863162af73&type=org&disposition=inline [WebCite Cache ID 6oAfFreOn]
9. Public Health Value of Disclosed Cigarette Ingredients and Emissions Data. North Sydney, Australia: Ipsos-Eureka Social Research Institute; 2009. URL: http://www.tobaccolabels.ca/wp/wp-content/uploads/2013/12/Australia-2009-Public-Health-Value-of-Disclosed-Cigarette-Ingredients...-Govt-Report1.pdf [WebCite Cache ID 6oAfUTZXm]
10. Hall MG, Ribisl KM, Brewer NT. Smokers' and nonsmokers' beliefs about harmful tobacco constituents: implications for FDA communication efforts. Nicotine Tob Res 2014 Mar;16(3):343-350 [FREE Full text] [doi: 10.1093/ntr/ntt158] [Medline: 24151139]
11. Weinstein ND, Nicolich M. Correct and incorrect interpretations of correlations between risk perceptions and risk behaviors. Health Psychol 1993 May;12(3):235-245. [Medline: 8500454]
12. Hamilton W, Norton GD, Ouellette T, Rhodes W, Kling R, Connolly G. Smokers' responses to advertisements for regular and light cigarettes and potential reduced-exposure tobacco products. Nicotine & Tobacco Res 2004 Dec;6(6):353-362. [doi: 10.1080/14622200412331320752]
13. Cummings KM, Hyland A, Giovino GA, Hastrup JL, Bauer JE, Bansal MA. Are smokers adequately informed about the health risks of smoking and medicinal nicotine? Nicotine Tob Res 2004 Dec;6 Suppl 3:S333-S340. [Medline: 15799596]

14. O'Connor RJ, Hyland A, Giovino GA, Fong GT, Cummings KM. Smoker awareness of and beliefs about supposedly less-harmful tobacco products. Am J Prev Med 2005 Aug;29(2):85-90. [doi: 10.1016/j.amepre.2005.04.013] [Medline: 16005803]
15. Shiffman S, Pillitteri JL, Burton SL, Di Marino ME. Smoker and ex-smoker reactions to cigarettes claiming reduced risk. Tob Control 2004 Mar;13(1):78-84 [FREE Full text] [Medline: 14985602]
16. Parascandola M, Augustson E, Rose A. Characteristics of current and recent former smokers associated with the use of new potential reduced-exposure tobacco products. Nicotine Tob Res 2009 Dec;11(12):1431-1438 [FREE Full text] [doi: 10.1093/ntr/ntp157] [Medline: 19915081]
17. Fishbein M, Ajzen I. Predicting and Changing Behavior: The Reasoned Action Approach. New York: Routledge; 2011.
18. Grube JW, Morgan M, McGree ST. Attitudes and normative beliefs as predictors of smoking intentions and behaviours: a test of three models. Br J Soc Psychol 1986 Jun;25 (Pt 2):81-93. [Medline: 3719241]
19. Tyas SL, Pederson LL. Psychosocial factors related to adolescent smoking: a critical review of the literature. Tob Control 1998;7(4):409-420 [FREE Full text] [Medline: 10093176]
20. Centrale Commissie Mensgebonden Onderzoek. 2015. Uw onderzoek: WMO-plichtig of niet URL: http://www.ccmo.nl/nl/uw-onderzoek-wmo-plichtig-of-niet [accessed 2016-10-11] [WebCite Cache ID 6lArkh8Ad]
21. Flycatcher Internet Research. 2015. URL: http://www.flycatcher.eu/eng/home/ [accessed 2016-10-11] [WebCite Cache ID 6lArqMaG0]
22. ComScore. 2016. URL: http://sitestat.com/privacy/index.php?lang=en_US [accessed 2016-11-21] [WebCite Cache ID 6mBa84Mxg]
23. Centraal Bureau voor de Statistiek. 2016. URL: https://www.cbs.nl/ [accessed 2016-12-14] [WebCite Cache ID 6mkedyU7W]
24. Guidelines for Controlling and Monitoring the Tobacco Epidemic. Geneva: World Health Organization; 1998.
25. Peters GJ. The alpha and the omega of scale reliability and validity. Eur Health Psychol 2014;16(2):56-69.
26. Dunn TJ, Baguley T, Brunsden V. From alpha to omega: a practical solution to the pervasive problem of internal consistency estimation. Br J Psychol 2014 Aug;105(3):399-412. [doi: 10.1111/bjop.12046] [Medline: 24844115]
27. Revelle W, Zinbarg RE. Coefficients alpha, geta, omega, and the glb: comments on Sijtsma. Psychometrika 2008 Dec 11;74(1):145-154. [doi: 10.1007/s11336-008-9102-z]
28. Janssen E, van OL, Lechner L, Candel M, de Vries H. Thinking versus feeling: differentiating between cognitive and affective components of perceived cancer risk. Psychol Health 2012;27(7):767-783. [doi: 10.1080/08870446.2011.580846] [Medline: 21767108]
29. Schulz DN, Kremers SP, van Osch LA, Schneider F, van Adrichem MJ, de Vries H. Testing a Dutch web-based tailored lifestyle programme among adults: a study protocol. BMC Public Health 2011;11:108 [FREE Full text] [doi: 10.1186/1471-2458-11-108] [Medline: 21324181]
30. Schulz DN, Kremers SP, Vandelanotte C, van Adrichem MJ, Schneider F, Candel MJ, et al. Effects of a web-based tailored multiple-lifestyle intervention for adults: a two-year randomized controlled trial comparing sequential and simultaneous delivery modes. J Med Internet Res 2014;16(1):e26 [FREE Full text] [doi: 10.2196/jmir.3094] [Medline: 24472854]
31. Crutzen R, Cyr D, de Vries NK. Bringing loyalty to e-Health: theory validation using three internet-delivered interventions. J Med Internet Res 2011;13(3):e73 [FREE Full text] [doi: 10.2196/jmir.1837] [Medline: 21946128]
32. Crutzen R, Cyr D, de Vries NK. The role of user control in adherence to and knowledge gained from a website: randomized comparison between a tunneled version and a freedom-of-choice version. J Med Internet Res 2012;14(2):e45 [FREE Full text] [doi: 10.2196/jmir.1922] [Medline: 22532074]
33. Crutzen R, Beekers N, van Eenbergen M, Becker M, Jongen L, van Osch L. E-loyalty towards a cancer information website: applying a theoretical framework. Psychooncology 2014 Jun;23(6):685-691. [doi: 10.1002/pon.3471] [Medline: 24408565]
34. Boynton MH, Agans RP, Bowling JM, Brewer NT, Sutfin EL, Goldstein AO, et al. Understanding how perceptions of tobacco constituents and the FDA relate to effective and credible tobacco risk messaging: A national phone survey of U.S. adults, 2014-2015. BMC Public Health 2016;16:516 [FREE Full text] [doi: 10.1186/s12889-016-3151-5] [Medline: 27333921]
35. Snyder L, Hamilton M. A meta-analysis of US health campaign effects on behaviormphasize enforcement, exposure, and new information, and beware the secular trend. In: Public Health Communications: Evidence for Behavior Change. Hillsdale, NJ: Lawrence Earlbaum; 2002:357-384.
36. van Deursen AJ, van Dijk JA. The digital divide shifts to differences in usage. New Media Soc 2013 Jun 07;16(3):507-526. [doi: 10.1177/1461444813487959]
37. Mesch G, Mano R, Tsamir J. Minority status and health information search: a test of the social diversification hypothesis. Soc Sci Med 2012 Sep;75(5):854-858. [doi: 10.1016/j.socscimed.2012.03.024] [Medline: 22633160]
38. Margolis KA, Bernat JK, Keely OE, Delahanty JC. Online information about harmful tobacco constituents: a content analysis. Nicotine Tob Res 2016 Aug 30:220. [doi: 10.1093/ntr/ntw220] [Medline: 27613931]
39. Witte K. Fear control and danger control: a test of the extended parallel process model (EPPM). Commun Monogr 2009 Jun 02;61(2):113-134. [doi: 10.1080/03637759409376328]
40. Hammond D. Health warning messages on tobacco products: a review. Tob Control 2011 Sep;20(5):327-337. [doi: 10.1136/tc.2010.037630] [Medline: 21606180]

41. Garrett J. The Elements of User Experience: User-Centered Design for the Web and Beyond. Berkeley: Macmillan Computer Pub; 2011.
42. Diviani N, van den Putte B, Giani S, van Weert JC. Low health literacy and evaluation of online health information: a systematic review of the literature. J Med Internet Res 2015;17(5):e112 [FREE Full text] [doi: 10.2196/jmir.4018] [Medline: 25953147]
43. Partial Guidelines for Implementation of Articles 9 and 10 of the WHO Framework Convention on Tobacco Control. 2012. URL: http://www.who.int/fctc/guidelines/Guideliness_Articles_9_10_rev_240613.pdf?ua=1 [accessed 2017-02-10] [WebCite Cache ID 6oAgi8p3Z]

Abbreviations

PREP: potential reduced exposure product
RCT: randomized controlled trial

Edited by G Eysenbach; submitted 11.10.16; peer-reviewed by G Nagelhoud, E Neter, A Lazard, M Lorenz; comments to author 17.11.16; revised version received 19.12.16; accepted 06.01.17; published 14.03.17.

Please cite as:
Reinwand DA, Crutzen R, Kienhuis AS, Talhout R, de Vries H
Website Use and Effects of Online Information About Tobacco Additives Among the Dutch General Population: A Randomized Controlled Trial
J Med Internet Res 2017;19(3):e60
URL: http://www.jmir.org/2017/3/e60/
doi:10.2196/jmir.6785
PMID:28292739

Original Paper

Patient Portals as a Tool for Health Care Engagement: A Mixed-Method Study of Older Adults With Varying Levels of Health Literacy and Prior Patient Portal Use

Taya Irizarry[1], RN, MSN; Jocelyn Shoemake[2], MPH; Marci Lee Nilsen[1], RN, PhD; Sara Czaja[3], PhD; Scott Beach[3], PhD; Annette DeVito Dabbs[1], RN, FAAN, PhD

[1]School of Nursing, University of Pittsburgh, Pittsburgh, PA, United States

[2]Department of Psychology, Ohio State University, Columbus, OH, United States

[3]University Center for Social and Urban Research, University of Pittsburgh, Pittsburgh, PA, United States

Corresponding Author:
Taya Irizarry, RN, MSN
School of Nursing
University of Pittsburgh
336 Victoria Building
3500 Victoria Street
Pittsburgh, PA, 15261
United States
Phone: 1 412 624 4722
Fax: 1 412 383 7227
Email: tai19@pitt.edu

Abstract

Background: Growing evidence that patient engagement improves health outcomes and reduces health care costs has fueled health providers' focus on patient portals as the primary access point for personal health information and patient-provider communication. Whereas much attention has been given to identifying characteristics of older adults who do and do not adopt patient portals and necessary adaptions to portal design, little is known about their attitudes and perceptions regarding patient portal use as a tool for engagement in their health care within the context of health literacy, experience navigating Web-based health information, and previous patient portal use.

Objective: The specific aims of this study were to explore attitudes toward portal adoption and its perceived usefulness as a tool for health care engagement among adults (65 years and older) who have varying levels of health literacy and degrees of prior patient portal use.

Methods: A phone survey of 100 community dwelling adults gathered sociodemographic, health, and technology related information. Older adults were purposefully selected for 4 follow-up focus groups based on survey responses to health literacy and previous patient portal use. A mixed-method approach was used to integrate phone survey data with thematic analysis of 4 focus groups. Due to variability in attitudes between focus group participants, an individual case analysis was performed and thematic patterns were used as the basis for subgroup formation.

Results: Differences in health literacy, comfort navigating health information on the Web, and previous portal experience explained some but not all differences related to the 7 themes that emerged in the focus groups analysis. Individual cases who shared attitudes were arranged into 5 subgroups from least to most able and willing to engage in health care via a patient portal. The subgroups' overall portal adoption attitudes were: (1) Don't want to feel pushed into anything, (2) Will only adopt if required, (3) Somebody needs to help me, (4) See general convenience of the portal for simple tasks and medical history, but prefer human contact for questions, and (5) Appreciates current features and excited about new possibilities .

Conclusions: Most of the older adults are interested in using a patient portal regardless of health literacy level, previous patient portal adoption, or experience navigating health information on the Web. Research targeting informal caregivers of older adults who are unable or unwilling to engage with information technology in health care on their own is warranted. Health care organizations should consider tailored strategies to meet the needs of older adults (and their informal caregivers) and explore alternative workflows that integrate patient portal information into phone conversations and face-to-face contact with health care providers.

(J Med Internet Res 2017;19(3):e99) doi:10.2196/jmir.7099

KEYWORDS

patient portals; patient participation; telemedicine; aged; health literacy; access to information; patient preference

Introduction

Growing evidence that patient engagement improves health outcomes and reduces health care costs [1] coupled with government reforms to promote efficiency, quality, and safety [2] has fueled health providers' focus on patient portals as the primary access point for personal health information and patient-provider communication [3]. Additionally, health systems have added convenience features such as prescriptions refills and appointment scheduling. Consequently, consumer adoption of patient portals is becoming increasingly critical for receipt of quality health care including interactions with health providers outside of clinical visits and quick access to one's personal health information.

Although the intention of patient portals is to promote patient engagement, numerous large-scale survey studies have demonstrated that older adults are less likely to adopt portals even though they utilize the greatest proportion of health care resources [4-6]. Low adoption rates are most pronounced among older adults who have less access to and experience with technology, less education, and who demonstrate low health literacy and numeracy skills [7,8] **.** These barriers have been aptly described as the "gray digital divide" [9].

Qualitative studies examining the known barriers that contribute to the gray digital divide discuss a strong need for supplementary support to assist vulnerable patients with portal navigation, particularly those with limited health literacy [10,11]. In addition, user-satisfaction, usability, and task analysis studies [12-15] have focused on older adults' experiences navigating portal functionality and evaluating their performance of specific tasks. Results from these studies indicate that previous computer experience and adequate health literacy and numeracy are strong contributing factors to ones' ability to successfully perform health management tasks using a patient portal.

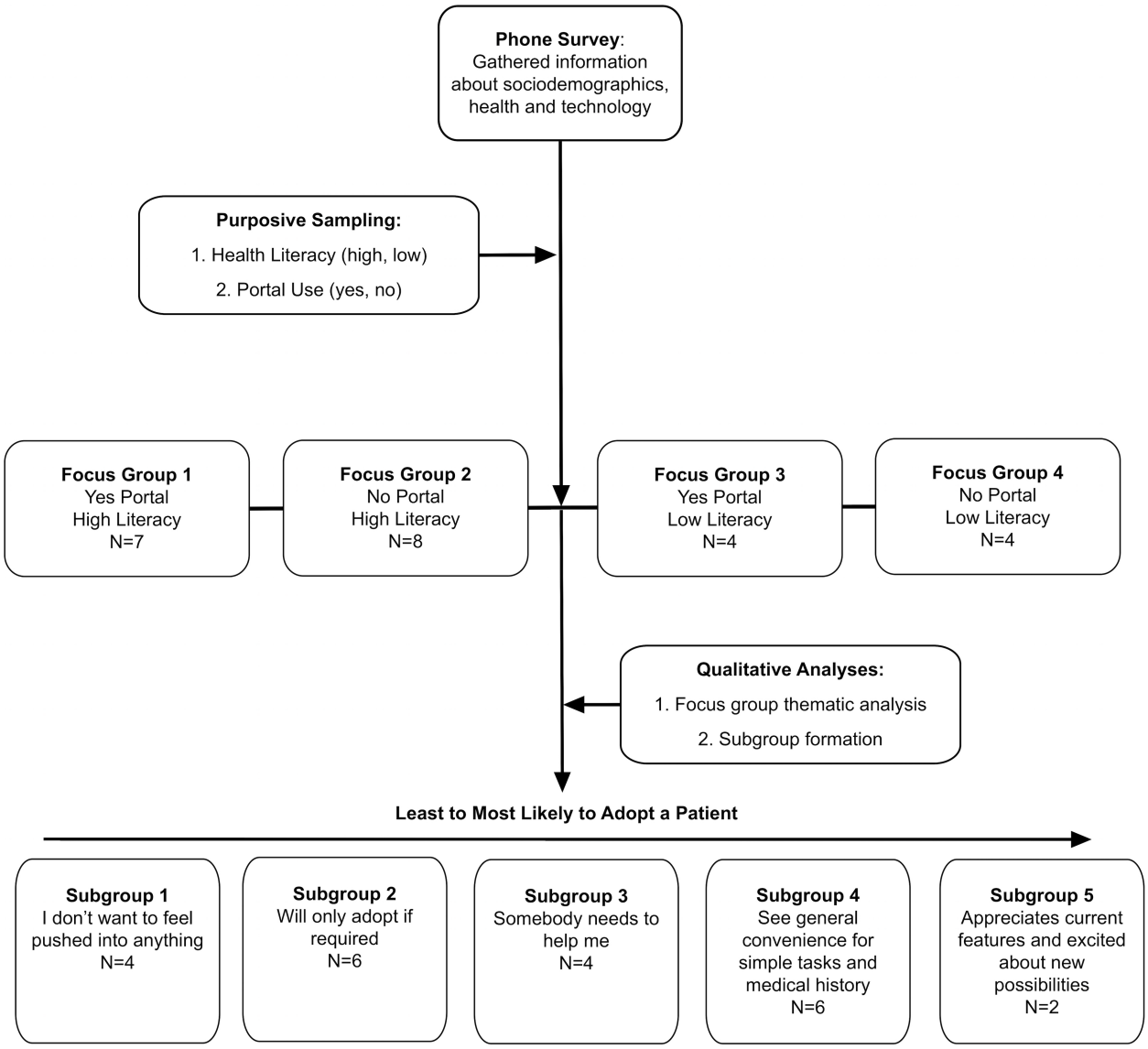
Whereas much attention has been given to identifying the characteristics of older adults who do and do not adopt patient portals and necessary adaptions to portal functionality, little is known about their attitudes and perceptions regarding portal use as a tool for engagement in their health care within the context of health literacy level, experience navigating Web-based health information, and previous patient portal use. A better understanding of the relationships connecting these concepts could help health care systems align their organizational practices and system design (ie, people, process, and technology) [16] to better meet the unique needs of the older adult populations they serve [8,10]. Therefore, the specific aims of this study were to explore attitudes toward portal adoption and its perceived usefulness as a tool for health care engagement among older adults with varying levels of health literacy and degrees of prior patient portal use.

Methods

Design

A mixed-method study design was chosen for the purpose of complementarity [17], meaning findings from quantitative and qualitative methods were integrated in a complementary fashion to produce a more complete understanding of the phenomena of interest. Integration occurred at multiple stages. First, a phone survey was used to assess sociodemographic and other health- and technology-related characteristics in a sample of 100 community dwelling older adults. Second, subsamples of older adults were purposefully selected to participate in follow-up focus groups based on responses to health literacy score and patient portal use. Third, the quantitative data from the phone survey and thematic analysis results of the focus group were integrated to form a rich description of older adults' experiences with navigating health information on the Web and attitudes toward the patient portal's usefulness as a tool for engagement in their health care. Finally, distinct patterns among individuals across all 4 focus groups were identified, resulting in the formation of subgroups ranging from least to mostly likely to adopt a patient portal. Refer to Figure 1 for an overview of the study design. Ethics approval for this research was obtained through the University of Pittsburgh Institutional Review Board (REN16060086/PRO15050313).

Figure 1. Overview of mixed-method study design.



Sample

A convenience sample of English-speaking community-dwelling older adults (defined as living in an independent residence, ie, not in an assisted living facility or nursing home) who were at least 65 years of age and cognitively equipped to answer a full battery of questions were recruited from the institutional review board (IRB)–approved research registry maintained by University of Pittsburgh's University Center for Social and Urban Research (UCSUR). The registry (n=1000) was assembled from a variety of population-based surveys conducted by UCSUR in the Pittsburgh area; thus, it is generally representative of the regional population.

A final sample of 100 registry participants was contacted by phone by experienced interviewers and a 45-minute survey was completed. Each participant received US $10 for participation. Seventy-five of the phone survey participants agreed to be contacted for possible participation in a follow-up focus group. Those who agreed were then stratified by level of health literacy and portal use based on responses to two survey items: (1) whether or not the participant had experience using a patient portal (yes, no), and (2) level of health literacy (high, low). Purposive sampling was used to include representation of participants in terms of age, race, and gender in each of the 4 focus groups. The first 10 participants who met criteria for one of the 4 groups and were willing and able to commit to a common meeting date and time were invited to participate.

Phone Survey Data Collection

The selection of survey measures was based on evidence-based criteria related to the gray digital divide. Instruments were selected from the core battery for demographics, health and disability from the Quality of Life Technology (NSF 0540865), and the core battery from the Center for Research and Education in Aging and Technology Enhancement (CREATE) which includes sociobehavioral factors proposed by the Institute of Medicine for inclusion in the electronic health record [18].

Sociodemographics

Demographic variables included age (years), gender, race (white, African American, other), marital status (single or divorced or widowed, married or living with partner), education (high school degree or general educational development [GED], associate

or vocational training, bachelor's, master's degree), and income (less than or US $14,999, US $15,000-US $39,000, US $40,000-US $69,000, US $70,000 or more).

Health Status and Disease Burden

Health characteristics included the number of comorbidities endorsed from a list of 12 potential conditions with the option of "other" and responses to 2 questions about perceived health status: (1) "In general, would you say your health is" (poor or fair, good, very good, or excellent), and (2) "How often are your daily activities limited due to your physical health" (never or seldom, sometimes, often, or always).

Health Literacy

The brief health literacy screen (BHLS), a subjective health literacy screening measure [19], includes 3 Likert-scale questions addressing one's confidence completing medical forms, reading hospital materials, and understanding written health information. On the basis of previous research [20,21], the response cutoff that optimized sensitivity and specificity for low health literacy was ≤ 3 (ie, "somewhat" or "some of the time"), meaning only participants who scored a 4 or 5 on all 3 questions were considered to have adequate health literacy (although we refer to the dichotomy as high and low literacy for simplicity in the study results). These questions have been previously shown to correlate with the Short Test of Functional Health Literacy in Adults (S-TOFHLA) and Rapid Estimate of Adult Literacy in Medicine as criterion standards [20-22].

Level of Engagement in Health and Health Care

The short form Patient Activation Measure (PAM-13) created by Hibbard and colleagues, a reliable and valid tool that is widely used in both research and clinical settings [23-25], was used to measure engagement. The PAM-13 includes 13 items with 4 response options ranging from 1(disagree strongly) to 4 (agree strongly). The raw score is then converted to an overall score ranging from 0-100 using the table provided in PAM licensing materials [25]. Scores are categorized into one of four levels of activation ranging from least to most active: level one (47.0 or lower), patients believe taking an active role in their health is important but are unprepared for this role; level two (47.1 to 55.1), patients have some knowledge but still struggle to manage their medical conditions; level three (55.2 to 67.0), patients begin to take action in terms of self-management but do not have the skills to support or sustain their behavior; and level four (67.1 or above), patients have adopted self-management behaviors and work on maintaining them in stressful life situations.

Technology Use and Attitudes

Two items of the Technology and Computer/Web Experience Questionnaire [18] were included in the survey. First, use of technology was measured by whether participants ever searched on the Web for health-related information: "During the past three months, did you or someone who helps you look online for information about any of the following topics: health condition or treatment, medications, health service availability, health professionals, and or health care facilities?" (yes or no). Second, they were asked to rate their attitudes about technology on a scale from 1 (not at all) to 10 (completely) for each of the following descriptors: "To what extent do you believe that technology makes life:" (1) easy and convenient, (2) complicated, (3) gives people control over their daily lives, (4) makes people dependent, (5) comfortable, (6) stressful, (7) brings people together, (8) makes people isolated, (9) increases personal safety and security, and (10) reduces privacy. Negative descriptors were reverse scored and a final score was an average of the total sum, thus a higher score means a more positive attitude.

Portal Use

Previous use of a patient portal was determined by a yes or no response to the question, "Have you ever used a patient portal on your own?" Answers were dichotomized as no or yes (all or some of the time). Further clarification was gained by asking, "Does someone help you access the portal or access the portal on your behalf?" If yes, some or all the time .

Focus Group Data Collection

Health Literacy

Since previous research had identified health literacy and numeracy as contributing factors for patient engagement and patient portal use [14,26], in addition to screening for health literacy using BHLS, we administered the newest vital sign (NVS) to focus group participants. The NVS is a 6-item objective health literacy measure of both reading comprehension and numeracy abilities [27]. Each participant was taken to a private space before the focus group discussions and given a copy of the nutrition label and the NVS questions. Scoring of the NVS is based upon number of correct answers with 0-1=high likelihood of limited health literacy, 2-3=possibility of limited health literacy, and 4-6=adequate health literacy skills. The NVS time to completion was also recorded; any participant taking longer than 7 minutes to complete the screening questions is considered to have low health literacy. The NVS has been previously shown to correlate with the Short Test of Functional Health Literacy in Adults (S-TOFHLA) and Rapid Estimate of Adult Literacy in Medicine short form (REALM-SF) as criterion standards [27,28].

Attitudes Toward Technology and Portal Use

Experienced moderators facilitated the focus groups using a topical guide geared toward understanding participants' attitudes toward technology, experiences of accessing health-related information on the Web, and perceived ability to use and benefit from patient portal functions. In addition to the open discussion, an interactive patient portal demonstration, developed by Czaja and colleagues at CREATE [12,29], University of Miami, was used to stimulate conversation about common features of patient portals and to seek participants' reactions.

Each of the 4 focus group sessions lasted approximately 1 hour and participants were provided light refreshments and US $40 for participation. The focus group discussions were audio recorded, transcribed verbatim by trained transcriptionists, and reviewed for accuracy by the focus group facilitators before being uploaded in to Atlas.ti version 7.5 (Scientific Software Development GmbH), the software program chosen to organize data for coding and visualization. The transcripts were

supplemented with session notes and linked to participant survey data.

Phone Survey Analysis

Descriptive statistical analyses of all variables of interest (demographics, health characteristics, engagement, health literacy, technology use and attitudes, and portal use) were organized by level of study participation: (1) participants who only completed the phone survey (n=26), (2) those contacted for possible follow-up focus group participation (n=51), and (3) focus group participants (n=23). In order to determine differences between the 3 groups, chi-square test of association was used for dichotomous variables (gender, race, marital status, technology use for health-related activities, health literacy, and portal use). Post hoc testing using standardized adjusted residuals was used to determine the source of any significant result using a threshold of ±2 [30]. Kruskal-Wallis H was used for ordinal and continuous variables (age, education, income, all health characteristics, engagement and technology attitudes). Subsequent pairwise comparisons were performed using Dunn [31] procedure with a Bonferroni correction for multiple comparisons and adjusted *P* values were presented. SPSS version 24.0 software (SPSS Inc) was used for all analyses.

Focus Groups Analysis

The thematic analysis [32] of qualitative transcript data of the focus groups was initiated by a lead coder who used open coding to describe the views of participants regarding the following topics: (1) experience with technology for health-related information and (2) impressions about the patient portal demonstration and its potential usefulness in promoting personal engagement in health care. A second coder then reviewed the initial codes and added new codes when she felt existing codes were needed. Both coders met to reach consensus on final codes. A third coder joined the team to collapse codes into themes through a process of consensus. Descriptive statistical analyses of demographics, health characteristics, engagement, health literacy, technology use and attitudes, and portal use were performed to in order to identify quantitative differences and similarities between focus groups.

Case Analysis

The stratification of focus groups by patient portal use and health literacy explained some but not all differences in attitudes toward accessing Web-based health information, portal adoption, and perceptions of usefulness of patient portal functionality. Due to the variability in attitudes between individuals within the focus groups, an individual case analysis was performed by all 3 coders in order to identify potential patterns. The case analysis began by linking passages from the transcript to individuals, which made it possible to connect the codes from individuals to the themes identified in the focus groups. Individuals' thematic patterns were then displayed in a matrix and subgroups were identified based upon the similarities and differences between them (Table 5). In addition, characteristics including demographics, health characteristics, health literacy, and technology use and attitudes were calculated for each subgroup in order to identify possible quantitative similarities and differences.

Results

Quantitative Results by Level of Study Participation

Survey results by level of study participation and group difference statistics are displayed in Table 1. Significant group differences were found for race (*P*=.03), searching on the Web for health information (*P*=.01), education (*P*=.01), income (*P*=.001), health status (*P*=.003), and engagement (*P*=.001). No statistically significant group differences were found for any of the other variables (age, gender, marital status, limited due to physical health, health literacy, or attitude toward technology). The post hoc analysis revealed statistically significant differences between the phone survey group, the follow-up call, and focus group attendees, but not between the follow-up call group and focus group attendees (see Tables 2 and 3). Post hoc testing indicated more African American or other focus group attendees than would be expected by chance (adjusted residual 2.6). This finding supports the success of the purposive sampling technique, which was meant to encourage the most racially diverse participant representation possible. Post hoc testing also indicated fewer survey participants searched on the Web for health information than would be expected by chance (adjusted residual −2.6). The purposive sampling method did not control for technology-related variables, therefore, this finding suggests that people who are more familiar with accessing Web-based health information may be more willing to participate in research related to technology in health care.

Table 1. Between group differences by level of study participation: phone survey, follow-up call, focus group.

Participant characteristics		Phone survey (n=26)	Follow-up call (n=51)	Focus group (n=23)	Group differences
Sociodemographics					
Age in years (n=100), Mean (range)		76.58 (65-93)	74.69 (65-97)	72.61 (65-82)	X^2_2=2.6, *P*=.27
Gender (n=100), n (%)	Female	12 (46.2)	30 (58.8)	13 (52.2)	X^2_2=0.8, *P*=.79
Race (n=99), n (%)					X^2_2= 6.9, *P*=.03
	White	21 (80.8)	44 (86.3)	13 (56.5)	
	African American	2 (7.7)	7 (13.7)	6 (26.1)	
	Other	3 (11.5)	0 (0)	3 (13.0)	
	Refused[a]			1 (4.3)	
Marital status (n=100), n (%)	Married or living with partner	8 (30.8)	28 (54.9)	12 (52.2)	X^2_2=4.2, *P*=.12
Education (n=99), n (%)					X^2_2=8.9, *P*=.01
	Less than high school degree	2 (7.7)	2 (3.9)	0 (0)	
	High school degree or general educational development	12 (46.2)	12 (23.5)	3 (13.0)	
	Associate or vocational training	6(23.1)	14 (27.5)	9 (39.1)	
	Bachelor's degree	3 (11.5)	12 (23.5)	8 (34.8)	
	Master's degree	2 (7.7)	11 (21.6)	3 (13.0)	
Income in US $ (n=96), n (%)					X^2_2=13.0, *P*=.001
	<$14,999	7 (26.9)	6 (11.8)	3 (13.0)	
	$15,000-$39,999	14 (53.8)	13 (25.5)	4 (17.4)	
	$40,000-$69,999	5 (19.2)	17 (33.3)	8 (34.8)	
	>$70,000	0 (0)	13 (25.5)	4 (17.4)	
	Refused[a]	2 (50)	2 (3.9)		
Health characteristics					
Current health status (n=100), n (%)					X^2_2=11.3, *P*=.003
	Poor to fair	16 (61.5)	10 (19.6)	5 (21.7)	
	Good	6 (23.1)	26 (51.0)	9 (39.1)	
	Very good to excellent	2 (7.69)	15 (29.4)	9 (39.1)	
Daily activities limited due to physical health (n=100), n (%)					X^2_2=4.4, *P*=.11
	Never	5 (19.2)	14 (27.5)	12 (52.2)	
	Seldom	5 (19.2)	12 (23.5)	13 (13.0)	
	Sometimes	14 (53.8)	19 (37.3)	6 (26.1)	
	Often to always	2 (7.7)	6 (11.8)	2 (8.7)	
Number of comorbidities (n=100), mean (range)		2 (0-5)	1.94 (0-6)	1.56 (0-5)	X^2_2=2.4, *P*=.08
Health literacy (n=100), n (%)					

Participant characteristics		Phone survey (n=26)	Follow-up call (n=51)	Focus group (n=23)	Group differences
	Brief health literacy screen (Hi)	9 (34.6)	28 (54.9)	15 (65.2)	X^2_2=4.9, *P*=.08
Engagement					
Patient Activation Measure (PAM, n=100; 0-100) mean (range)		59.49 (41.70-80.0)	71.71 (37.30-100.0)	77.05 (52.90-100.0)	X^2_2 15.1, *P*=.001
	PAM level 1, n (%)	3 (11.5)	1 (2.0)	0 (0)	
	PAM level 2, n (%)	4 (15.4)	4 (7.8)	1 (4.3)	
	PAM level 3, n (%)	13 (50)	19 (37.3)	6 (26.1)	
	PAM level 4, n (%)	6 (23.1)	27 (52.9)	16 (69.6)	
Technology attitude (n=99)					
	Searched on the Web for health-related information (yes), n (%)	9 (34.6)	35 (68.6)	12 (52.2)	X^2_2=8.9, *P*=.01
	Technology attitudes (score 0-10), mean (range)	5.72 (2.7-9.6)	6.33 (3.4-8.8)	6.26 (4-8.5)	
Portal use (n=100), n (%)					
	Ever use a patient portal on your own (yes)	0 (0)	25 (49.0)	11 (47.8)	X^2_2=19.8, *P*=<.001
	Someone helps you use portal[b]				
	Yes, all of the time	0 (0)	3 (5.9)	0 (0)	
	Yes, sometimes	6 (23.1)	4 (7.8)	6 (26.1)	
	Someone accesses the portal on your behalf[b]				
	Yes, all of the time	2 (7.7)	4 (7.8)	0 (0)	
	Yes, sometimes	5 (19.2)	3 (5.9)	4 (17.4)	

[a]Participants chose not to supply information.

[b]Not analyzed due to sample size.

Table 2. Post hoc analysis of dichotomous variables.

Dichotomous Demographic Variables[a]		Phone survey only	Follow-up call	Focus group attendee
Race				
	White	21 (0.3)	44 (1.9)	13 (−2.6)
	African American or Other	5 (−0.3)	9 (−1.9)	7 (2.6)
Searched on the Web for health information				
	Yes	9 (−2.6)	35 (2.7)	12 (−0.5)
	No	17 (2.6)	15 (−2.7)	11 (0.5)

[a]Adjusted residuals appear in parenthesis next to observed frequencies.

Table 3. Post hoc analysis of ordinal variables.

Ordinal Demographic Variables	Adjusted significance, *P* value	Group[a] (mean rank)
Education	.03	PS (35.84), AFC (53.74)
	.02	PS (35.84), FGA (57.33)
	>.99	AFC, FGA
Income	.001	PS (31.54), AFC (54.03)
	.02	PS (31.54), FGA (52.50)
	>.99	AFC, FGA
Current health status	.007	PS (35.15), AFC (55.09)
	.01	PS (35.15), FGA (57.67)
	>.99	AFC, FGA
Engagement (patient activation measure)	.007	PS (32.69), AFC (53.87)
	.001	PS (32.69), FGA (63.15)
	.59	AFC, FGA

[a]PS: phone survey, AFC: follow-up call, FGA: focus group attendee.

Quantitative Results by Focus Group

Of the 40 volunteers who committed to one of the 4 focus groups, only 23 attended the sessions. There was no follow-up with the participants who didn't make it to the sessions, so the reason for not attending in unknown. The yes portal, low literacy group and no portal, low literacy group had 4 participants, the yes portal, high literacy group had 7, and the no portal, high literacy group had 8. Survey results per focus group are displayed in Table 4. The no portal, low literacy group had the widest age range (65-82 years) and included 2 participants in their early eighties. Three out of 4 of the focus groups had an equal proportion of males and females. The no portal, high literacy group had fewer males (n=3) and was the most racially diverse with 3 African Americans and 3 others who did not identify themselves as white, but did not specify. The majority of participants in all focus groups had education beyond high school or GED, with half having a college degree. Both yes portal groups had more white and married participants, as well as more participants with higher average income.

The majority of participants reported their perceived health status was "good," to "excellent," and only 2 in the yes portal high literacy group and 2 in the no portal low literacy group reported "poor or fair" health. All the groups had at least one participant report his or her physical condition limits daily activities "sometimes." The no portal low literacy group had 1 participant who reported "often" being physically limited. This same participant reported having 5 comorbidities, where the range across all other groups was 0-3.

The NVS scores for the high literacy groups (yes portal and no portal) ranged from 2-6 and 0-6, respectively. The NVS scores for the low literacy groups (yes portal and no portal) ranged from 5-6 and 0-5, respectively. The BHLS and NVS scores were not always in alignment, meaning fewer participants were found to have high literacy using the NVS than with the BHLS screening instrument. No notable differences were found in PAM scores, searching on the Web for health-related information, or attitudes toward technology with the exception that for the technology items the average scores for the no portal low literacy group were slightly lower.

Table 4. Characteristics of the sample by focus group.

Patient characteristics		Group 1 Yes portal, high literacy (n=7)	Group 2 No portal, high literacy (n=8)	Group 3 Yes portal, low literacy (n=4)	Group 4 No portal, low literacy (n=4)
Sociodemographics					
Age in years, mean (range)		73.29 (66-80)	73.88 (66-80)	69.00 (66-73)	72.50 (65-82)
Gender, n (%)					
	Male	4 (57.1)	3 (37.5)	2 (50)	2 (50)
Race, n (%)					
	White	5 (71.4)	2 (25)	3 (75)	3 (75)
	African American	2 (28.6)	3 (37.5%)	0 (0)	1 (25)
	Other	0 (0)	3 (37.5)	1 (25)	0 (0)
	Refused[a]		1(12.5)		
Marital status, n (%)					
	Single or divorced or widowed	2 (28.6)	5 (87.5)	0 (0)	2 (50)
	Married or living with partner	5 (71.4)	1 (12.5)	4 (100)	2 (50)
Education, n (%)					
	High school degree or general educational development	1 (14.3)	1 (12.5)	0 (0)	1 (25)
	Associate or vocational training	2 (28.6)	4 (50)	0 (0)	1 (25)
	Bachelor's degree	2 (28.6)	3 (37.5)	2 (50)	1 (25)
	Master's degree	2 (28.6)	0 (0)	2 (50)	1 (25)
Income in US $, n (%)					
	<$14,999	1 (14.3)	2 (25)	0 (0)	0 (0)
	$15,000-$39,999	1 (14.3)	2 (25)	1 (25)	0 (0)
	$40,000-$69,999	1 (14.3)	3 (37.5)	1 (25)	3 (75)
	>$70,000	2 (28.6)	0 (0)	2 (50)	0 (0)
	Refused[a]		1 (12.5)		1 (25)
Health characteristics					
Current health status, n (%)					
	Poor to fair	2 (28.6)	1 (12.5)	0 (0)	2 (50)
	Good	3 (42.9)	3 (37.5)	3 (75)	0 (0)
	Very good to excellent	2 (28.6)	4 (50)	1 (25)	2 (50)
Daily activities limited due to physical health, n (%)					
	Never	4 (57.1)	5 (62.5)	1 (25)	2 (50)
	Seldom	1 (14.3)	1 (12.5)	0 (0)	1(25)
	Sometimes	1 (14.3)	2 (25)	3 (75)	0 (0)
	Often to always	1 (14.3)	0 (0)	0 (0)	1(25)
Number of comorbidities mode (range)		1 (1-2)	1 (0-5)	1 (1-2)	2 (0-2)
Health literacy					

Patient characteristics		Group 1 Yes portal, high literacy (n=7)	Group 2 No portal, high literacy (n=8)	Group 3 Yes portal, low literacy (n=4)	Group 4 No portal, low literacy (n=4)
Newest vital sign (possible score 0-6) Mean (range)[b]		4.71 (2-6)	3.2 (0-6)	5.5 (5-6)	2.75 (0-5)
Newest vital sign time to complete (minutes, seconds) Mean (range)[b]		3.28 (3.44-5.38)	5.43 (2.7-16.33)	3.46 (2.53-7.0)	5.12 (2.41-7.23)
Engagement					
PAM[c] (possible score 0-100) Mean (range)		77.80 (56.40-100.00)	82.74 (52.90-100.00)	66.40 (56.40-77.50)	76.21 (56.40-100.00)
	PAM[c] level 1, n (%)	0 (0)	0 (0)	0 (0)	0 (0)
	PAM level 2, n (%)	0 (0)	1 (12.5)	0 (0)	0 (0)
	PAM level 3, n (%)	2 (28.6)	1 (12.5)	2 (50)	1 (25)
	PAM level 4, n (%)	5 (71.4)	6 (75)	2 (50)	3 (75)
Technology attitude					
	Searched on the Web for health-related information (yes)	4 (57)	4 (50)	2 (50)	2 (50)
	Technology attitudes (possible score 0-10), mean (range)	6.60 (5-8.5)	6.45 (5-8)	6.68 (5-8)	4.88 (4-6)
Portal Use					
	Someone helps you use portal some or all of the time (yes), n (%)	3 (42.9)	1 (12.5)	1 (25)	1 (25)
	Someone accesses the portal on your behalf some or all of the time (yes), n (%)	1 (14.3)	1 (12.5)	1 (25)	1 (25)

[a]Participants chose not to supply information.

[b]The newest vital sign (NVS) measured during the focus group sessions only.

[c]PAM: patient activation measure.

Qualitative Results by Focus Group

Seven major themes were identified and arranged in order from least to most positive experience and attitudes toward technology use for health care and portal adoption. The 7 themes included: (1) limited or poor relationship with technology, (2) fears and frustrations with technology and portal, (3) prefers phone over secure messaging for communication (outside of clinical visit), (4) willing to adopt the portal with support, (5) good relationship with technology, (6) Internet as source of health information, and (7) portal is helpful. The following detailed explanation of each of the 7 themes includes all the original codes that informed the theme and is framed by the focus group stratification criteria (previous portal use and health literacy level).

Limited or Poor Relationship With Technology

Participants in every focus group expressed at least some degree of negativity toward technology use in health care as well as in everyday encounters. Difference in reasons why varied by literacy level. The most common negative sentiment among participants in the high health literacy groups was difficulty troubleshooting without having access to live technical support and feeling pressured to adopt new communication methods (eg, instant messaging, video calls) that seemed unnecessary. In contrast, many participants in the no portal, low health literacy group mentioned having had little experience using computers and did not have the Internet access in their homes. Many no portal, low literacy group participants noted that they had no computer training as part of their job and retired before computers were a regular part of the working environment. Many felt afraid of making a mistake or felt stigmatized by their lack of knowledge and therefore either avoided using computers all together or relied heavily on family members to help them.

Fear and Frustration With Technology and Portal

Fears about personal health data security risks were shared in every focus group. Those participants in the high health literacy groups were more articulate in describing specific instances in

which they felt uncomfortable, whereas those in the low health literacy groups spoke more generally about security fears. Despite the level of specific risk described, some participants cited the risk as being too great, whereas others felt it was an everyday risk we all must "get used to." In all instances, participants felt that speaking over the phone with a trusted health care professional was the safest and most secure way to share health information.

Participants with more experience navigating the health care system and seeing several medical specialists shared experiences where they felt burdened by the task of circulating the most current health information even though all health care providers had access to the same electronic health record. These people also described instances where medical history in the patient portal was outdated or incorrect which led to a sense of frustration, as it was not clear what action should be taken to correct it. A few participants from the high literacy group who had experience using the patient portal secure message function recounted instances in which they either did not get a response, or they were not entirely certain that they understood the response. In either case, they experienced anxiety and frustration. They also expressed discomfort with the idea that other staff aside from the primary health care provider could respond to personal secure messages via the patient portal because they did not personally know them.

Prefers Phone Over Secure Messaging for Communication (Outside of Clinical Visit)

Those who experienced anxiety and frustration using the patient portal's secure message feature were convinced that phone was the only way to communicate, ask questions, and clarify information. The no portal high literacy group voiced the most opposition to secure messaging in health care encounters and the most strongly in support of direct clinician communication by phone. When probed further, there were instances where the participant felt comfortable speaking with a nurse, nurse practitioner, or physician assistant (as opposed to an office manager) as long as they worked closely with his or her provider and felt they had enough clinical expertise. Those with no portal experience in both the high and low literacy groups were more positive about the potential for secure messaging and the idea of having access to their provider outside the clinical visit.

Willing to Adopt Portal With Support

Participants in both the low health literacy groups expressed an interest in portal training. Those with less experience with computers had less confidence in their ability to learn and expressed an interest in having someone access the system on their behalf. The high literacy portal users expressed fear and anxiety around doing something wrong and had recommendations for a "task specific" training in which users would learn the purpose of each function (eg, information searches, secure messaging, appointment, and medication refill requests) and gain hands-on experience with navigation, use, and troubleshooting.

Happy to Engage With Technology

Participants who were eager to learn new things, had exposure to computer use at work, or sought out resources to learn computer skills, expressed appreciation for the conveniences of technology. They were eager to share what they had learned in terms of accessing computers outside of the home, where to take computer classes, and ways to troubleshoot. The level of enthusiasm and interest did not parallel levels of health literacy or portal use. However, participants in the high literacy no portal group were the most philosophical in their debate about technology degrading the quality of human relationships, and were the least willing to discuss potential benefits to their personal lives. High literacy portal users who had negative experiences with information sharing among providers or negative experiences with the portal expressed both positive and negative feelings about their relationship with technology in the health care setting, yet generally felt the positives outweighed the negatives.

Internet as a Good Source of Health Information

The degree to which participants felt the information they received from their providers was enough for them was variable across all focus groups. Some participants from the high literacy portal user group felt the portal lacked individualized information and that it was more helpful to search on the Web. In contrast, some low literacy groups (portal and nonportal users) believed the health information presented in the portal was the more trusted and reliable, and therefore it made them feel more confident about the information in comparison to looking up information on the Internet.

Sees Portal as Beneficial

Convenience features of the portal were perceived as useful by participants regardless of whether they had used a particular function before and whether or not they could personally see themselves using it. Overall, participants in good health thought it was helpful to have all their personal medical information, including clinician contact numbers, all in one place for easy reference, but felt that other more complex functions would be helpful for people who were managing a serious illness or chronic condition or were acting as a caregiver. In most cases, they considered their current non-technological methods (eg, paper copy of their post-visit summary) to be satisfactory. Participants who had experience with serious illness and chronic disease management were vocal about the convenience of having timely access to lab results and having them stored electronically, but felt that the kind of communication they needed to have with their providers was too involved for secure messaging.

Qualitative Case Results

Individual cases who shared attitudes toward engaging in one's health care via technology and portal adoption were arranged into 5 subgroups from least to most able and willing to engage in health care via a patient portal. All individual participants were placed in a subgroup except for 1 from the no portal, high literacy group who could not be categorized due to minimal verbal participation, and therefore scarcity of codes attributed to her. Table 5 presents the matrix of individuals in relationship to the themes and demonstrates the prevalence of themes in each focus group. Table 6 presents a description of the 5 subgroups, including the number of individuals from each of

the 4 original focus groups, and summarizes the general attitudes toward adoption of technology for health care engagement and adoption of the portal.

Table 5. Endorsement of themes by individual participants within focus group.

Focus group[a]	Individual participant	Limited or poor relation-ship with technology	Fears and frustrations with technology (may extend to portal)	Prefers phone as primary mode of communication	Portal willing with support	Good relationship without technology	Internet as source of health information and education	Portal is helpful
No portal low literacy	Carol	X	X	X				X[b]
	Henry	X	X	X				X[c]
	Sheila			X		X	X	X
	Willy	X	X	X				X[c]
No portal high literacy	Elsie	X	X	X	X			X[c]
	Mick		X	X	X			X[c]
	Brian			X				X[c]
	Gerald	X		X				
	May			X				X[c]
	Francis	X						
	Jane		X		X			X[c]
	Mary							X[c]
Yes portal low literacy	John		X	X		X	X	X
	Lynn					X		X
	Rick	X	X	X				X[c]
	Terry		X	X				X[c]
Yes portal high literacy	Rob	X	X	X				X[c]
	Sue		X	X		X	X	X
	Gary		X	X				X[c]
	Tim	X		X	X			
	Anne							X
	Ray		X	X	X	X	X	X
	Lily	X						X[c]

[a]Themes in columns left to right from most negative to most positive.

[b]Considered the portal to be generally helpful but not for their personal use.

[c]Considered the portal to be helpful only for viewing lab results.

Table 6. Attitudes of subgroups arranged according to the least (subgroup 1) to most likely (subgroup 5) to adopt health technologies and patient portal.

Subgroup	Original focus group				Attitudes toward adoption of technology for health care engagement	Attitudes toward adoption of the portal
	YPHL[a]	YPLL[b]	NPHL[c]	NPLL[d]		
Subgroup 1 n=4	2			2	Don't think the benefits are worth the hassle or risk	I don't want to feel pushed into anything
Subgroup 2 n=6		1	5		Satisfied as things are	Will only adopt if required
Subgroup 3 n=4	2		1	1	Technology is the way of the future, but too difficult to learn new things	Somebody needs to help me
Subgroup 4 n=6	2	2	1	1	Comfortable with technology, but prefer to talk to a person for personal health-related issues	Sees general convenience of the portal for simple tasks and medical history
Subgroup 5 n=2	1	1			Thrilled with technology for information and communication with no reservations.	Appreciates current features and excited about new possibilities

[a]YPHL=yes portal, high literacy.

[b]YPLL=yes portal, low literacy.

[c]NPHL=no portal, high literacy.

[d]NPLL=no portal, low literacy.

Quantitative Case Results

The smallest of the 5 subgroups (n=2) comprised women who expressed complete comfort with technology and were previous portal users. Although neither of them had a college degree, they both scored a 6 out 6 on the NVS and completed the literacy measure in less than 3 minutes (far less time than of any other subgroup where at least half of the group had a college degree). They were both married, nearly the same age (70 and 71 years), reported an income over US $70,000 a year, and expressed having good, but not excellent health. All other subgroups were more diverse in terms of demographics, NVS scores, and attitudes toward technology. Subgroup 3 expressed interest in using technology yet acknowledged they needed assistance to do so. This group had the widest range in age, NVS scores and completion time, and the highest number of comorbidities.

Discussion

Principal Findings

This study used a mixed-method approach to gain greater insight into the experiences and attitudes toward the patient portal as an engagement tool among older adults with varying levels of health literacy and patient portal experience. Numerous topics revealed in the focus groups were consistent with current literature on the adoption of patient portals by older adults including fears about security issues, interest in convenience features (eg, appointment scheduling, prescription refills), and lack of access to a computer or Internet as a major barrier to portal adoption [33]. The unique contributions of this study are the findings regarding attitudes of older adults toward adoption of a patient portal for engagement in health care, specifically in the context of health literacy and previous experience navigating Web-based health information.

Focus group participants were assembled based on health literacy level (assessed using the BHLS) and previous experience using a patient portal. This stratification explained some but not all differences in attitudes toward portal adoption and perceptions of usefulness of patient portal functionality. Neither the health literacy scores from the BHLS nor the NVS was directly correlated with previous patient portal adoption or perceived usefulness except in the case of subgroup 5 which comprised 2 women who scored 6 out of 6 on the NVS, expressed complete comfort with technology, and were regular portal users. The qualitative findings revealed that health literacy was a contributing factor to confidence accessing and evaluating health care information on the Web (this evidence is corroborated by Diviani [34]). However, health literacy was not directly related to one's motivation to engage in health care via a patient portal as evidenced by the case analysis and subsequent subgroupings.

Qualitative findings also revealed that previous experience accessing and evaluating health care information on the Web was not entirely correlated with prior patient portal adoption or perceptions about its potential usefulness. In fact, some members of the no portal focus groups (both high and low health literacy) who did not have experience navigating health information on the Web were motivated to explore their health care institution's patient portal following the patient portal simulation, that included examples of how the patient portal functionality could be helpful within a personally relevant context. This finding is supported by Melenhorst, Rogers, and Bouwhuis [35] who identified the "decisive role of perceived benefits" as the main motivator of adoption among older adult users.

Presently, most patient portals are introduced via an email with a time-sensitive link and sign-on instructions, or a postcard given in a clinical visit or sent through the mail. These methods may be appropriate for younger generations who are more comfortable with navigating the Web-based environment

including sharing personal information and social interaction. However, this research demonstrates that older adults require an initial introduction that highlights contextually relevant benefits and addresses their particular needs and concerns. Both high and low health literacy groups felt that specific task-based training was an important, yet lacking, resource that would help build confidence and understanding of when, why, and how to navigate the features included in patient portals.

Additionally, deliberate outreach and tailored training of informal caregiver proxy users is recommended in the case of older adults who recognize and appreciate the potential benefits of the patient portal as a tool for engagement, yet lack computer access or perceive themselves as unfit to manage their own health information. The potential role of informal caregivers as a key factor to improve access and use of patient portals by older adults who are unable to engage on their own is recognized in the literature [35-37], yet very little research exists regarding the experience of acting as a proxy-user, effective strategies for encouraging portal adoption among proxy-users, and what design features could be enhanced to encourage engagement with both the older adult of concern and his or her health care provider [38].

Whereas a patient portal adoption campaign tailored to older adults could convince and encourage greater numbers of older adults to use patient portals, many participants in this as well as other studies [7,39] express fear that they won't always understand the personal health information available to them and felt that the secure message function was a poor substitute for direct clinician-patient interaction necessary to clarify things. This sentiment remained strong across all focus groups as evidenced by the case analysis matrix (Table 5) and was especially true for older adults managing complex conditions. Frequent portal users also described instances where medical history in the patient portal was outdated or incorrect which led to a sense of frustration and concern, as it was not clear what action to take to correct it. A possible solution for these issues, concerns, and frustrations may be restructuring care team workflows. Examples include making it possible for patients and informal caregivers to call and talk to a clinic representative who can answer simple questions, verify and update information presented in the patient portal (which would then update the electronic health record), and triage more serious issues when appropriate. In addition, further integration of patient portal use during face-to-face encounters with health care providers and phone interactions with clinic representatives could transform the patient portal into an information resource for all parties and may incentivize older adults and caregivers to use it as a tool for health care engagement outside of clinical visits. Refer to Table 7 for a brief description of findings, implications, and recommendations.

Reliability, Validity, Trustworthiness, and Rigor

Quantitative and qualitative procedures were performed according to the assumptions of each paradigm. Survey methods, including sampling, use of validated measures, the collection of data using standard formatting by trained phone interviewers, and standard analysis techniques ensured reliability and validity of quantitative analysis [40]. Trustworthiness and rigor of qualitative focus group phase of the study included purposive sampling to ensure as much heterogeneity with regard to participant characteristics as possible [41]. The focus groups were led by an experienced facilitator using a semistructured guide and field observer. All focus group sessions were audiorecorded, transcribed verbatim, and open coding was initially performed to provide thick description of exchanges during the focus groups [42]. Three researchers performed thematic analyses independently and interpretation of the findings and meaning making was achieved by consensus [43]. Quantitative and qualitative data sources were integrated using matrices [44] and merged according to recommended methods [40].

Table 7. Summary of findings, implications, and recommendations.

Findings	Implications	Recommendations
Health literacy: A contributing factor to confidence accessing and evaluating health care information on the Web. Yet, not directly related to one's motivation to engage in health care via a patient portal.	Health literacy is not a primary barrier to patient portal adoption, but may impact confidence in navigating its features.	Offer specific task-based training to build confidence and understanding of when, why, and how to navigate the features included in patient portals.
Perceptions of portal usefulness: Not entirely correlated with prior patient portal adoption or previous experience accessing and evaluating health care information on the Web.	Older adults are motivated to adopt a portal when the initial introduction highlights contextually relevant benefits and addresses their particular needs and concerns.	Create a patient portal adoption campaign tailored to the needs and concerns of older adults.
Some willing adopters are unable: Some older adults appreciate the potential benefits of the patient portal as a tool for engagement, yet lack computer access or perceive themselves as unfit to manage their own health information.	The potential role of informal caregivers as a key factor to improve access and use of patient portals by older adults who are unable to engage on their own is supported here and recognized in the literature.	Design deliberate outreach and tailored training of informal caregiver proxy users.
Portal as source of information, but not a stand-alone solution: Many feared they wouldn't always understand portal information and felt secure messaging was a poor substitute for direct clinician-patient interaction necessary to clarify things.	Most of the older adults believe the portal is convenient for simple tasks and medical history, but is not sufficient as a stand-alone engagement tool.	Explore alternative workflows that integrate portal use into face-to-face clinical encounters and offer access to personnel with the skills to review and respond to questions over the phone or triage more serious issues if appropriate.
Errors in portal information are a source of concern: Frequent users found outdated or incorrect medical history in the portal and were unsure what to do about it.	The usefulness of the portal is diminished when the information is not accurate and promotes dissatisfaction when no clear avenue of correction is available.	Explore alternative workflows that offer access to personnel with ability to change, up-date, and validate missing or inaccurate portal information.

Limitations

This study used a convenience sample of older adults from a Pittsburgh regional research registry. Whereas a strong effort was made to achieve the most representative sample of those in jeopardy of experiencing the gray digital divide, the comparison between phone survey, follow-up, and focus group attendees revealed statistically significant differences between survey participates and focus group participants. Most notably, focus group participants were more educated, with higher income, better health status, higher engagement scores, and more experience with searching for health information on the Web than survey participants. Alternative recruitment methods designed to target older adults that are least likely to be interested in focus group participation, such as semistructured interviews over the phone or home visits, are warranted. However, the sample characteristics of the focus group participants are representative of the average older American adult population, and therefore helpful for understanding general attitudes toward technology in health care and portal adoption.

In regards to the focus group sample, characteristics of interest were generally comparable across groups; however, both low literacy groups were nearly half the size of the high literacy groups despite having had the same number of participant invitations accepted. A more equal representation of low literacy participants may have added further clarification of the differences in attitudes and preferences as compared with the high literacy groups. Also, 3 of the 4 groups were 75% white. Having more input from a wider range of racial backgrounds may have added more detail regarding differences in adoption attitudes toward the patient portal related to race. Additionally, this study was conducted in English, which limited our findings to fluent English speakers. Whereas most patient portals are currently in English only, continued work in other languages is necessary to meet the needs of our increasingly diverse communities. Descriptive statistics of the focus groups and subgroups were explored and reported in order to provide a rich multidimensional description; however, the small sample size of the focus groups and subgroups did not allow for statistical inferences between participant characteristics and the attitudes toward health care engagement via a patient portal.

Finally, the question, "Does someone help you access the portal or access the portal on your behalf?" (If yes, some or all of the time) was asked in the survey phase; however due to the minimal number of positive answers within any of the groups, it was not possible to explore the concept in a mixed-method approach (although the data is reported). Instead, the role of the caregiver proxy was discussed in detail qualitatively. Further research focused on addressing older adult informal caregivers' attitudes and experiences as proxy users is needed.

Conclusions

The study used quantitative and qualitative methods in a complementary fashion to produce a more complete understanding of older adults' attitudes toward using a patient portal as a health care engagement tool within the context of health literacy, previous experience with patient portal, and accessing Web-based information. Only a minority of older adults believe that the security risks or trouble learning something new is not worth it; most of the older adults are interested in using a patient portal regardless of health literacy level, previous patient portal adoption, or experience navigating health information on the Web. Health care organizations should consider the following strategies to align people, process, and technology in order to meet the needs of the older adults they serve: (1) create a patient portal adoption campaign tailored to

the needs of older adult so that the benefits are communicated in a contextually relevant way, (2) offer task-specific training so they feel they have the support they need to confidently use the functionality, (3) specifically target informal caregiver proxy users as part of the adoption campaign and training, and (4) explore alternative workflows that give patients access to personnel with the skills to review and respond to questions over the phone about personal health information within the portal, as well as change, up-date, validate missing or inaccurate information, and triage more serious issues when appropriate. Such organizational strategies would transform the patient portal from a repository of information with a secure message function to a tool designed to support engagement, information sharing, and enhanced communication between care teams, patients, and informal caregivers.

Acknowledgments

Funding for this study was granted by the Aging Institute of University of Pittsburgh Medical Center (UPMC), Pittsburgh. No review or approval of this manuscript for publication was required.

Conflicts of Interest

None declared.

References

1. Forbat L, Cayless S, Knighting K, Cornwell J, Kearney N. Engaging patients in health care: an empirical study of the role of engagement on attitudes and action. Patient Educ Couns 2009 Jan;74(1):84-90. [doi: 10.1016/j.pec.2008.07.055] [Medline: 18790594]
2. Ferman JH. Healthcare reform: Medicare & Medicaid Innovation Center. CMS provision will test new payment and service delivery models. Healthc Exec 2010;25(4):58, 60. [Medline: 20649172]
3. GPO. Medicare and Medicaid programs; electronic health record incentive program URL: https://www.gpo.gov/fdsys/pkg/FR-2010-07-28/pdf/2010-17207.pdf [accessed 2016-11-26] [WebCite Cache ID 6mJPgLifR]
4. Smith SG, O'Conor R, Aitken W, Curtis LM, Wolf MS, Goel MS. Disparities in registration and use of an online patient portal among older adults: findings from the LitCog cohort. J Am Med Inform Assoc 2015 Apr 25;22(4):888-895. [doi: 10.1093/jamia/ocv025] [Medline: 25914099]
5. Sarkar U, Karter AJ, Liu JY, Adler NE, Nguyen R, López A, et al. Social disparities in internet patient portal use in diabetes: evidence that the digital divide extends beyond access. J Am Med Inform Assoc 2011 May 01;18(3):318-321 [FREE Full text] [doi: 10.1136/jamia.2010.006015] [Medline: 21262921]
6. Graetz I, Gordon N, Fung V, Hamity C, Reed ME. The digital divide and patient portals: internet access explained differences in patient portal use for secure messaging by age, race, and income. Med Care 2016 Aug;54(8):772-779. [doi: 10.1097/MLR.0000000000000560] [Medline: 27314262]
7. Gordon NP, Hornbrook MC. Differences in access to and preferences for using patient portals and other ehealth technologies based on race, ethnicity, and age: a database and survey study of seniors in a large health plan. J Med Internet Res 2016 Mar 04;18(3):e50 [FREE Full text] [doi: 10.2196/jmir.5105] [Medline: 26944212]
8. Nahm E, Sagherian K, Zhu S. Use of patient portals in older adults: a comparison of three samples. Stud Health Technol Inform 2016;225:354-358. [Medline: 27332221]
9. Morris A. E-literacy and the grey digital divide: a review with recommendations. JIL 2007;1(3):13-28 [FREE Full text] [doi: 10.11645/1.3.14]
10. Turner AM, Osterhage K, Hartzler A, Joe J, Lin L, Kanagat N, et al. Use of patient portals for personal health information management: the older adult perspective. AMIA Annu Symp Proc 2015;2015:1234-1241 [FREE Full text] [Medline: 26958263]
11. Latulipe C, Gatto A, Nguyen HT, Miller DP, Quandt SA, Bertoni AG, et al. Design considerations for patient portal adoption by low-income, older adults. Proc SIGCHI Conf Hum Factor Comput Syst 2015 Apr;2015:3859-3868 [FREE Full text] [doi: 10.1145/2702123.2702392] [Medline: 27077140]
12. Zarcadoolas C, Vaughon WL, Czaja SJ, Levy J, Rockoff ML. Consumers' perceptions of patient-accessible electronic medical records. J Med Internet Res 2013;15(8):e168 [FREE Full text] [doi: 10.2196/jmir.2507] [Medline: 23978618]
13. Czaja SJ, Zarcadoolas C, Vaughon WL, Lee CC, Rockoff ML, Levy J. The usability of electronic personal health record systems for an underserved adult population. Hum Factors 2015 May;57(3):491-506. [doi: 10.1177/0018720814549238] [Medline: 25875437]
14. Taha J, Sharit J, Czaja SJ. The impact of numeracy ability and technology skills on older adults' performance of health management tasks using a patient portal. J Appl Gerontol 2014 Jun;33(4):416-436. [doi: 10.1177/0733464812447283] [Medline: 24781964]
15. Tieu L, Schillinger D, Sarkar U, Hoskote M, Hahn KJ, Ratanawongsa N, et al. Online patient websites for electronic health record access among vulnerable populations: portals to nowhere? J Am Med Inform Assoc 2016 Jul 8:- Epub ahead of print. [doi: 10.1093/jamia/ocw098] [Medline: 27402138]

16. Leavitts H. Applied organisation change in industry: structural, technical and human approaches. In: New Perspectives in Organisational Research. New York: Wiley; 1964:55-71.
17. Greene JC, Caracelli VJ. JSTOR. 1989. Toward a conceptual framework for mixed-method evaluation designs: Educational evaluation and policy analysis URL: http://www.jstor.org/stable/1163620 [accessed 2017-03-15] [WebCite Cache ID 6ozNxesBa]
18. Czaja SJ, Charness N, Fisk AD, Hertzog C, Nair SN, Rogers WA, et al. Factors predicting the use of technology: findings from the center for research and education on aging and technology enhancement (CREATE). Psychol Aging 2006 Jun;21(2):333-352 [FREE Full text] [doi: 10.1037/0882-7974.21.2.333] [Medline: 16768579]
19. Chew LD, Bradley KA, Boyko EJ. Brief questions to identify patients with inadequate health literacy. Fam Med 2004 Sep;36(8):588-594 [FREE Full text] [Medline: 15343421]
20. Chew LD, Griffin JM, Partin MR, Noorbaloochi S, Grill JP, Snyder A, et al. Validation of screening questions for limited health literacy in a large VA outpatient population. J Gen Intern Med 2008 May;23(5):561-566 [FREE Full text] [doi: 10.1007/s11606-008-0520-5] [Medline: 18335281]
21. Wallace LS, Rogers ES, Roskos SE, Holiday DB, Weiss BD. Brief report: screening items to identify patients with limited health literacy skills. J Gen Intern Med 2006 Aug;21(8):874-877. [doi: 10.1111/j.1525-1497.2006.00532.x] [Medline: 16881950]
22. Powers BJ, Trinh JV, Bosworth HB. Can this patient read and understand written health information? JAMA 2010 Jul 7;304(1):76-84. [doi: 10.1001/jama.2010.896] [Medline: 20606152]
23. Greene J, Hibbard JH. Why does patient activation matter? an examination of the relationships between patient activation and health-related outcomes. J Gen Intern Med 2012 May;27(5):520-526 [FREE Full text] [doi: 10.1007/s11606-011-1931-2] [Medline: 22127797]
24. Fowles JB, Terry P, Xi M, Hibbard J, Bloom CT, Harvey L. Measuring self-management of patients' and employees' health: further validation of the Patient Activation Measure (PAM) based on its relation to employee characteristics. Patient Educ Couns 2009 Oct;77(1):116-122. [doi: 10.1016/j.pec.2009.02.018] [Medline: 19356881]
25. Hibbard JH, Stockard J, Mahoney ER, Tusler M. Development of the patient activation measure (PAM): conceptualizing and measuring activation in patients and consumers. Health Serv Res 2004 Aug;39(4 Pt 1):1005-1026 [FREE Full text] [doi: 10.1111/j.1475-6773.2004.00269.x] [Medline: 15230939]
26. Frosch DL, Elwyn G. Don't blame patients, engage them: transforming health systems to address health literacy. J Health Commun 2014;19 Suppl 2:10-14. [doi: 10.1080/10810730.2014.950548] [Medline: 25315579]
27. Weiss BD, Mays MZ, Martz W, Castro KM, DeWalt DA, Pignone MP, et al. Quick assessment of literacy in primary care: the newest vital sign. Ann Fam Med 2005;3(6):514-522 [FREE Full text] [doi: 10.1370/afm.405] [Medline: 16338915]
28. Kirk JK, Grzywacz JG, Arcury TA, Ip EH, Nguyen HT, Bell RA, et al. Performance of health literacy tests among older adults with diabetes. J Gen Intern Med 2012 May;27(5):534-540 [FREE Full text] [doi: 10.1007/s11606-011-1927-y] [Medline: 22095571]
29. Mitzner TL, Boron JB, Fausset CB, Adams AE, Charness N, Czaja SJ, et al. Older adults talk technology: technology usage and attitudes. Comput Human Behav 2010 Nov 1;26(6):1710-1721 [FREE Full text] [doi: 10.1016/j.chb.2010.06.020] [Medline: 20967133]
30. Agresti A. An introduction to categorical data analysis. Hoboken, NJ: Wiley-Interscience; 2007.
31. Dunn OJ. Multiple comparisons using rank sums. Technometrics 1964 Aug;6(3):241. [doi: 10.2307/1266041]
32. Boyatzis RE. Transforming qualitative information: thematic analysis and code development. Thousand Oaks, CA: Sage Publications; 1998.
33. Irizarry T, DeVito DA, Curran CR. Patient portals and patient engagement: a state of the science review. J Med Internet Res 2015 Jun 23;17(6):e148 [FREE Full text] [doi: 10.2196/jmir.4255] [Medline: 26104044]
34. Diviani N, van den Putte B, Giani S, van Weert JC. Low health literacy and evaluation of online health information: a systematic review of the literature. J Med Internet Res 2015 May 07;17(5):e112 [FREE Full text] [doi: 10.2196/jmir.4018] [Medline: 25953147]
35. Melenhorst A, Rogers WA, Bouwhuis DG. Older adults' motivated choice for technological innovation: evidence for benefit-driven selectivity. Psychol Aging 2006 Mar;21(1):190-195. [doi: 10.1037/0882-7974.21.1.190] [Medline: 16594804]
36. Mayberry LS, Kripalani S, Rothman RL, Osborn CY. Bridging the digital divide in diabetes: family support and implications for health literacy. Diabetes Technol Ther 2011 Oct;13(10):1005-1012 [FREE Full text] [doi: 10.1089/dia.2011.0055] [Medline: 21718098]
37. Kim E, Stolyar A, Lober WB, Herbaugh AL, Shinstrom SE, Zierler BK, et al. Challenges to using an electronic personal health record by a low-income elderly population. J Med Internet Res 2009;11(4):e44 [FREE Full text] [doi: 10.2196/jmir.1256] [Medline: 19861298]
38. Wolff JL, Berger A, Clarke D, Green JA, Stametz R, Yule C, et al. Patients, care partners, and shared access to the patient portal: online practices at an integrated health system. J Am Med Inform Assoc 2016 Nov;23(6):1150-1158. [doi: 10.1093/jamia/ocw025] [Medline: 27026614]

39. Lyles CR, Sarkar U, Schillinger D, Ralston JD, Allen JY, Nguyen R, et al. Refilling medications through an online patient portal: consistent improvements in adherence across racial/ethnic groups. J Am Med Inform Assoc 2016 Apr;23(e1):e28-e33. [doi: 10.1093/jamia/ocv126] [Medline: 26335983]
40. Craig CL, Marshall AL, Sjöström M, Bauman AE, Booth ML, Ainsworth BE, et al. International physical activity questionnaire: 12-country reliability and validity. Med Sci Sports Exerc 2003 Aug;35(8):1381-1395. [doi: 10.1249/01.MSS.0000078924.61453.FB] [Medline: 12900694]
41. Daly J, Willis K, Small R, Green J, Welch N, Kealy M, et al. A hierarchy of evidence for assessing qualitative health research. J Clin Epidemiol 2007 Jan;60(1):43-49. [doi: 10.1016/j.jclinepi.2006.03.014] [Medline: 17161753]
42. Jenks C. Thick description: Toward an interpretive theory of culture, in Culture: Critical Concepts in Sociology. In: Culture critical concepts in sociology. London: Routledge; 2003.
43. Graneheim UH, Lundman B. Qualitative content analysis in nursing research: concepts, procedures and measures to achieve trustworthiness. Nurse Educ Today 2004 Feb;24(2):105-112. [doi: 10.1016/j.nedt.2003.10.001] [Medline: 14769454]
44. Miles M, Huberman A. Qualitative data analysis: an expanded sourcebook. Thousand Oaks, CA: Sage Publications; 1994.

Abbreviations

BHLS: brief health literacy screen
GED: general educational development
NVS: newest vital sign
PAM-13: Patient Activation Measure
REALM-SF: Rapid Estimate of Adult Literacy in Medicine short form
S-TOFHLA: Short Test of Functional Health Literacy in Adults

Edited by G Eysenbach; submitted 07.12.16; peer-reviewed by C Kruse, C Cutler, C Lyles PhD, A Paglialonga; comments to author 08.02.17; revised version received 19.02.17; accepted 20.02.17; published 30.03.17.

Please cite as:
Irizarry T, Shoemake J, Nilsen ML, Czaja S, Beach S, DeVito Dabbs A
Patient Portals as a Tool for Health Care Engagement: A Mixed-Method Study of Older Adults With Varying Levels of Health Literacy and Prior Patient Portal Use
J Med Internet Res 2017;19(3):e99
URL: http://www.jmir.org/2017/3/e99/
doi:10.2196/jmir.7099
PMID:28360022

Original Paper

Assessing the Quality of Mobile Exercise Apps Based on the American College of Sports Medicine Guidelines: A Reliable and Valid Scoring Instrument

Yi Guo[1,2*], PhD; Jiang Bian[2], PhD; Trevor Leavitt[3], BS; Heather K Vincent[3], FACSM, PhD; Lindsey Vander Zalm[3], BS; Tyler L Teurlings[3], BS; Megan D Smith[2], BHS; François Modave[2*], PhD

[1]Institute of Natural Resources and Environmental Audits, Nanjing Audit University, Nanjing, China

[2]Department of Health Outcomes and Policy, University of Florida, Gainesville, FL, United States

[3]Department of Orthopaedics and Rehabilitation, University of Florida, Gainesville, FL, United States

[*]these authors contributed equally

Corresponding Author:
Yi Guo, PhD
Department of Health Outcomes and Policy
University of Florida
2004 Mowry Road, Suite 2251 PO Box 100165
Gainesville, FL, 32610
United States
Phone: 1 352 294 5969
Fax: 1 352 265 8047
Email: yiguo@ufl.edu

Abstract

Background: Regular physical activity can not only help with weight management, but also lower cardiovascular risks, cancer rates, and chronic disease burden. Yet, only approximately 20% of Americans currently meet the physical activity guidelines recommended by the US Department of Health and Human Services. With the rapid development of mobile technologies, mobile apps have the potential to improve participation rates in exercise programs, particularly if they are evidence-based and are of sufficient content quality.

Objective: The goal of this study was to develop and test an instrument, which was designed to score the content quality of exercise program apps with respect to the exercise guidelines set forth by the American College of Sports Medicine (ACSM).

Methods: We conducted two focus groups (N=14) to elicit input for developing a preliminary 27-item scoring instruments based on the ACSM exercise prescription guidelines. Three reviewers who were no sports medicine experts independently scored 28 exercise program apps using the instrument. Inter- and intra-rater reliability was assessed among the 3 reviewers. An expert reviewer, a Fellow of the ACSM, also scored the 28 apps to create criterion scores. Criterion validity was assessed by comparing nonexpert reviewers' scores to the criterion scores.

Results: Overall, inter- and intra-rater reliability was high with most coefficients being greater than .7. Inter-rater reliability coefficients ranged from .59 to .99, and intra-rater reliability coefficients ranged from .47 to 1.00. All reliability coefficients were statistically significant. Criterion validity was found to be excellent, with the weighted kappa statistics ranging from .67 to .99, indicating a substantial agreement between the scores of expert and nonexpert reviewers. Finally, all apps scored poorly against the ACSM exercise prescription guidelines. None of the apps received a score greater than 35, out of a possible maximal score of 70.

Conclusions: We have developed and presented valid and reliable scoring instruments for exercise program apps. Our instrument may be useful for consumers and health care providers who are looking for apps that provide safe, progressive general exercise programs for health and fitness.

(J Med Internet Res 2017;19(3):e67) doi:10.2196/jmir.6976

KEYWORDS

mHealth; mobile apps; physical activity; measures

Introduction

Physical Activity

Regular physical activity provides many health benefits [1-3] and is one of the recommendations made to address epidemic lifestyle-related diseases in the United States [4,5]. Participation in regular physical activity can lower the risk of early death and many diseases, including coronary heart disease, stroke, adverse blood lipid profile, type 2 diabetes, metabolic syndrome, some cancers, obesity, hypertension, bone and joint diseases, some autoimmune conditions, and depression [1,6-9]. Prospective epidemiological studies have documented a causal relationship between physical inactivity and heart disease, the leading cause of death in the United States [10-16]. These studies show that individuals who are more physically active have lower rates of heart disease, and the most physically active group develops heart disease at rates half of that of the most sedentary group [17,18]. However, despite the many benefits of regular physical activity, the majority of US adults do not meet the national physical activity guidelines [6,19,20]. It is estimated that only 20.9% of US adults meet the recommendations for both aerobic and muscle-strengthening activities [6].

In the past decade, there has been a tremendous increase in the availability and use of mobile phones and mobile phone apps [21,22]. A 2015 survey shows that 64% of US adults and 82% of US adults aged between 18 and 49 years own an app-enabled mobile phone [21]. This increase in mobile phones use has allowed for a growth in mobile phone apps related to physical activity and exercise. There is an estimated 100,000 health- and fitness-related apps in the Apple store alone, and over 165,000 apps when including the Android's Google Play store [23]. These apps include fitness and exercise trackers such as heart rate monitors, step counters, exercise programs, and coaching apps. There is preliminary evidence that these apps can be used effectively to improve health-related behaviors for a variety of chronic conditions [22,24-40]. However, there are very few exercise program apps that are evidence-based and follow the exercise guidelines set forth by the American College of Sports Medicine (ACSM) [41,42].

Three Components of Exercise Programs

The US Department of Health and Human Services recommends that adults should perform at least 150 min of moderate-intensity aerobic activity or 75 minutes of vigorous-intensity activity per week in addition to performing muscle-strengthening activities at least two times per week [43]. Further, based on the cumulative evidence pertaining to health and fitness, the ACSM recommends that the frequency, intensity, time, and type (FITT) principle should be followed for any exercise program to have health benefits while avoiding injuries and other adverse events. Exercise sessions should include components of safety precautions, warm-up, conditioning including strengthening, and cool-down. In addition, exercise programs should progress safely at a rate that is appropriate for the individual's fitness level and goals. The ACSM principles of exercise prescription recommend that optimal exercise programs include 3 main components: aerobic exercise, strength and resistance exercise, and flexibility. These components improve cardiovascular fitness, strength, neuromuscular fitness, and overall health [42].

Presently, the quality and accuracy of available mobile exercise program apps and the theoretical foundations that underpin these apps are not clear. Modave et al [41] used an initial scoring system to determine whether the content of free mobile apps related to exercise programming were evidence-based and adhered to the ACSM guidelines for aerobic exercise, strength and resistance exercise, and flexibility. These findings revealed a significant gap in the app content with the ACSM guidelines. A standardized instrument developed from the system in the first study that determines the quality of exercise programming apps could be of widespread benefit to clinicians and consumers who must make informed decisions about which apps to choose. We address this significant problem by developing a reliable and valid scoring instrument that can evaluate the quality of fitness and exercise-prescriptive apps with respect to the highest standards set forth by the ACSM. To our knowledge, there are no fitness app scoring instruments developed based on the ACSM exercise guidelines. Stoyanov et al developed a rating scale for assessing the quality of mobile health apps in general [44]. But this scale assesses domains of engagement, functionality, aesthetics, and information quality, rather than evidence-based exercise principles. In this paper, we describe our process of instrument development and present data demonstrating the inter- and intra-rater reliability and criterion validity of the instrument. Finally, the most popular free exercise prescriptive apps are scored and ranked using the developed instrument.

Methods

Focus Groups and Instrument Development

Our study included two focus groups (N=14) from whom we elicited input and guidance on survey item refinement and questionnaire design. Specifically, we asked the focus group participants to identify unclear words or sentences, suggest alternative ways of phrasing a question, recommend response formats, and consider how the questions would have worked in eliciting responses. University of Florida (UF) College of Medicine employees and students were recruited to participate in these focus groups.

Before the focus groups, the study investigators wrote survey questions according to the published ACSM exercise principles to create an initial version of the scoring instrument. (ACSM Guidelines, 9th ed.) For example, the single training session principle for aerobic exercise "Warm-up: 5-10 minutes of light/moderate intensity cardiovascular exercise" was written as "Does the app advise you to warm up for 5-10 minutes with light or moderate cardiovascular exercise before starting any aerobic exercise?" The responses were initially written on a 3-point scale with 0 indicating "Principle missing," 1 indicating "Principle present but unclear (or not 100% aligned with ACSM guidelines)," and 2 indicating "Principle present and clear." The scoring instrument was divided into 3 sections: (I) aerobic exercise, (II) strength and resistance, and (III) flexibility. Sections I and II were further divided into 3 subsections: (1) safety, (2) program principles, and (3) single training session

principles. Section III was divided into 2 subsections: (1) safety and (2) program principles.

Next, we provided the participants with the initial instrument and asked them to independently rate 5 fitness apps randomly selected from the apps evaluated in [41]. We instructed the participants to list the problems of the instrument during app scoring. All participants were allowed at least one day to thoroughly examine the instrument. During the focus groups, our moderator went through each question in the instrument and asked the participants to discuss potential issues such as ambiguity, excessive complexity, or inaccuracy with the text, phrasing, and format of the questions and accompanying responses. For each item, we asked the participants if they could paraphrase the question or if they thought the question should be asked in another way. Following completion of the focus groups, the investigators met to review the findings and to develop the final scoring instrument. The final instrument consisted of 27 questions, with 10 questions on aerobic exercise, 12 questions on strength and resistance, and 5 questions on flexibility (see Multimedia Appendix 1).

Scoring Strategy

The 3 main components of the ACSM exercise principles (aerobic exercise, strength and resistance, and flexibility) were weighted 3:3:1 based on the time allocated by the ACSM within a standard exercise program for health and fitness. For each of the 3 principles, the subsections (safety, program principles, and single training session principles) were allocated the same weight due to the lack of evidence that the subsections should be emphasized differently. Therefore, the overall quality score was scaled to have a highest possible score of 70, with 30 points in aerobic exercise, 30 points in strength and resistance, and 10 points in flexibility.

Data Collection and Psychometric Analysis

Sample

For testing the psychometric properties of the scoring instrument, we chose the same fitness apps evaluated in [41]. The apps were selected by searching the Apple store with keywords “workout” and “training” in the “health and fitness” category, and selecting the top 50 apps, based on their popularity. After removing duplicate apps from these two search terms, a list of 83 apps was generated. The investigators then evaluated and removed apps that did not provide exercise prescriptive programs. Finally, 30 apps were selected for scoring in [41]. During the development phase of the new instrument, two of the apps were no longer available in the app store. Therefore, we used the remaining 28 apps for scoring and psychometric analysis.

Inter- and Intra-Rater Reliability

We assessed the inter-rater and intra-rater (test-retest) reliability of the instrument. To assess inter-rater reliability, we asked 3 reviewers to score all 28 apps using the instrument concurrently but independently. The reviewers included 1 staff member and 2 college students from the UF Orthopaedics and Sports Medicine Institute. To assess intra-rater reliability, 5 apps were randomly selected from the 28 apps (7 Minute Workout, Body Space, FitStar, JEFIT, and Sworkit). Then, the same 3 reviewers rated the 5 apps again using the instrument approximately 1 month later. The Spearman correlation coefficient *r* among pairs of scores was computed as a measure of reliability.

Criterion Validity

We asked the director of UF Human Performance Laboratory and UF Health Sports Performance Center (expert reviewer, Fellow of the ACSM) to rate all 28 apps on a scale of 1-10 (criterion scores). In addition, the expert reviewer rated 5 apps (7 Minute Workout, Body Space, FitStar, JEFIT, and Sworkit) using our scoring instrument. Criterion validity was then assessed by calculating the Spearman correlation coefficient between the criterion scores and those obtained with the instrument, given by the nonexpert reviewers. For the 5 apps rated using the instrument, we computed the Spearman correlation coefficient between scores from the expert and nonexpert reviewers. In addition, we computed a weighted kappa statistic to assess the agreement between the expert and nonexpert reviewers for each app. We used the following guidelines for interpreting kappa statistics suggested by Landis and Koch: <0=poor agreement, 0-.2=slight agreement, .2-.4=fair agreement, .4-.6=moderate agreement, .6-.8=substantial agreement, and .8-1=almost perfect agreement [45].

Results

Instrument Item Development

The focus groups identified multiple challenges when using the initial instrument for scoring. The biggest challenge was the need to clarify meanings and definitions of domain-specific words and phrases. For instance, one of the questions asked was “Does the app advise you to train each major muscle group 2-3 times per week?” Focus group participants did not know what the “major muscle groups” were or what “advise” represented and thus could not judge whether the app provided a comprehensive strength training program or not. In addition, some participants did not understand some of the technical terms such as metabolic equivalent (MET) and Proprioceptive Neuromuscular Facilitation (PNF) stretching. To improve comprehension, participants were asked to suggest alternative strategies for phrasing the questions. Eventually, we removed terms that were too technical and added footnotes to most questions explaining the words that were unclear and phrases compiled from the focus groups.

Participants’ preferences of response options were tested. The majority of the participants felt that the initial 3-point response scale did not fully differentiate the apps on quality since there was only one middle category “Principle present but unclear (or not 100% aligned with ACSM guidelines).” As a result, final response options were revised to include a 5-point Likert scale with 1 being “No,” 3 being “Partially,” and 5 being “Yes.”

Inter- and Intra-Rater Reliability

Levels of inter-rater reliability for the principle section, subsection, and overall scores for each pair of reviewers are summarized in Table 1. Reliability was high overall, with most coefficients being greater than .7. The average inter-rater reliability was .88 for R1-R2, .81 for R1-R3, and .73 for R2-R3.

The reliability coefficients range from .59 to .99, and all coefficients were statistically significant. The least reliable question pertained to safety warning for aerobic exercise and asked, "Does the app provide safety warnings about health conditions or advise you to consult a doctor before starting any aerobic exercise?" (Section I.1.). The average inter-rater reliability across the rater pairs was .69 for this safety warning question. Overall, the aerobic exercise (average *r*=.75) and strength and resistance (average *r*=.79) sections had lower reliability than the flexibility section (average *r*=.88).

Table 1. Inter-rater reliability for section, subsection, and overall scores for each rater pair.

Section	R1-R2	*P* value	R1-R3	*P* value	R2-R3	*P* value	Average
I. Aerobic exercise	.80	<.001	.76	.009	.70	.011	.75
1. Safety	.77	.002	.68	.01	.62	.012	.69
2. Program principles	.76	.003	.89	<.001	.82	.001	.82
3. Single training session principles	.92	<.001	.92	<.001	.81	.001	.88
II. Strength and resistance	.94	<.001	.73	<.001	.69	<.001	.79
1. Safety	.92	<.001	.77	<.001	.65	<.001	.78
2. Program principles	.95	<.001	.92	<.001	.84	<.001	.90
3. Single training session Principles	.84	<.001	.78	<.001	.59	.003	.74
III. Flexibility	.97	<.001	.87	.009	.81	.009	.88
1. Safety	.99	<.001	.90	<.001	.88	.002	.92
2. Program principles	.87	.005	.70	.02	.67	.03	.75
Overall score	.92	<.001	.85	<.001	.79	<.001	.85
Average	.88		.81		.73		

Results from the intra-rater (test-retest) reliability analysis are summarized in Table 2. Overall, intra-rater reliability was high with most coefficients being greater than .7. The reliability coefficients ranged from .47 to 1.00, and all reliability coefficients were statistically significant. JEFIT had the lowest intra-rater reliability among the apps (average *r*=.66). The reliability for Reviewer 2 was only 0.47 for JEFIT. On the other hand, the intra-rater reliability was high for FitStar. All raters were able to provide the same scores for FitStar 1 month later.

Table 2. Intra-rater (test-retest) reliability by rater.

App	R1	*P* value	R2	*P* value	R3	*P* value	Average
7 Minute Workout	1.00	-	.85	<.001	.74	.006	.86
Body Space	.64	<.001	.76	<.001	.85	<.001	.75
FitStar	1.00	-	1.00	-	1.00	-	1.00
JEFIT	.75	<.001	.47	.01	.76	<.001	.66
Sworkit	.69	<.001	.73	<.001	.76	<.001	.73

Criterion Validity

In evaluating the validity of the instrument, the Spearman correlation coefficient between the criterion scores and those obtained with the instrument was .70 (*P*<.001) for Reviewer 1, .69 (*P*<.001) for Reviewer 2, and .78 (*P*<.001) for Reviewer 3. The Spearman correlation coefficient between the criterion scores and average nonexpert reviewer scores obtained with the instrument was .72 (*P*<.001). The high correlations provided support for the validity of our scoring instrument.

Table 3 summarizes weighted kappa statistic κ and Spearman correlation coefficient *r* from our agreement analysis. The weighted kappa statistics ranged from .67 to .99, indicating a substantial (κ is between .6 and .8) or almost perfect (κ is between .8 and 1) agreement between the scores given by the expert and nonexpert reviewers. The correlations between the scores from the nonexpert reviewers and those from the expert reviewer were high, ranging from .58 to .99, with most coefficients being greater than .7. The agreement was relatively poorer for FitStar (average κ=.73; average *r*=.76) and Sworkit (average κ=.79; average *r*=.76) when compared with the other apps.

Table 3. Validity measures using weighted kappa statistic κ and Spearman correlation coefficient *r*.

App	R1-Expert		R2-Expert		R3-Expert		Average	
	κ (*P* value)	*r* (*P* value)	κ (*P* value)	*r* (*P* value)	κ (*P* value)	*r* (*P* value)	κ	*r*
7 Minute Workout	.94 (<.001)	.74 (.005)	.84 (.002)	.58 (.04)	.94 (<.001)	.74 (.008)	.91	.69
Body Space	.81 (<.001)	.67 (.005)	.84 (<.001)	.78 (<.001)	.77 (<.001)	.75	.81	
.81 (<.001)								
FitStar	.68 (.003)	.75 (<.001)	.83 (<.001)	.79 (<.001)	.68 (.001)	.75 (<.001)	.73	.76
JEFIT	.98 (<.001)	.99 (<.001)	.91 (<.001)	.94 (<.001)	.88 (<.001)	.88 (<.001)	.92	.94
Sworkit	.78 (.001)	.69 (<.001)	.86 (<.001)	.82 (<.001)	.72 (.001)	.76 (<.001)	.79	.76

Overall and Section Quality Sores

The overall and section quality scores for the 28 apps evaluated using the scoring instrument are summarized in Table 4. The apps are presented in order from high to low overall scores. None of the apps had an overall quality score higher than 35 out of a highest possible score of 70 points. This confirms the results reported in [41] that the most popular apps for exercise prescription do not meet the standards set forth by the ACSM. Among the 28 apps, only 4 apps had an overall score higher than 30 (The Johnson and Johnson Official 7-Minute Workout, 33.5; Nike+ Training Club, 32.6; Running for Weight Loss: Interval Training, 32.6; Fitness Buddy Free, 31.0). For the aerobic exercise principle section, Running for Weight Loss: Interval Training (section score=26.7) was the only app that had a quality score higher than 20, out of a highest possible score of 30 points. For the strength and resistance principle section, 3 apps had a quality score higher than 20, out of a highest possible score of 30 points (StrongLifts 5x5, 24.0; Fitness Buddy Free, 22.9; FitnessBuilder, 21.2). For the flexibility principle section, 3 apps had a quality score higher than 5, out of a highest possible score of 10 points (FitnessBuilder, 7.1; Simply Yoga Free, 6.4; Daily Yoga-Lose Weight, Get Relief, 6.2). The average overall quality score was 17.8 (SD 8.9). The average section quality scores were 11.7 (SD 5.9) for aerobic exercise, 13.9 (SD 5.2) for strength and resistance, and 4.0 (SD 1.9) for flexibility.

Table 4. Final scores of the apps evaluated with the scoring instrument. All values are expressed in points.

Rank	App	Overall[a]	Section I[b]	Section II[c]	Section III[d]
1	The Johnson and Johnson Official 7-Minute Workout	33.5	16.3	17.2	–
2	Nike+ Training Club	32.6	10.0	19.5	3.1
3	Running for Weight Loss: Interval Training	32.6	26.7	–	3.5
4	Fitness Buddy Free	31.0	8.1	22.9	–
5	FitnessBuilder	28.3	–	21.2	7.1
6	JEFIT	27.5	8.4	16.3	2.8
7	Body Space	26.2	9.0	14.9	2.4
8	Daily Workouts Free	25.5	13.0	12.5	–
9	StrongLifts 5x5	24.0	–	24.0	–
10	Jillian Michaels Slim Down	19.9	–	19.9	–
11	Fitness Point-Workout Exercise	19.0	–	19.0	–
12	C25K-5K Trainer Free	17.7	17.7	.	–
13	7 Minute Workout	17.5	8.7	8.8	–
14	Sworkit	17.5	7.0	8.0	2.5
15	7-Minute Workout-Fitness for Women	16.9	7.0	9.9	–
16	Abs Workout: Get Your Six Pack	14.9	–	14.9	–
17	Daily Butt Workout Free	14.9	–	14.9	–
18	Instant Abs Trainer	12.2	–	12.2	–
19	FitStar	11.3	–	8.9	2.4
20	Daily Ab Workout Free	11.3	–	11.3	–
21	Workout Trainer	10.0	–	10.0	–
22	Runtastic Six Pack Abs Trainer	9.1	–	9.1	–
23	The 7 Minute Workout-Get Fit	8.4	–	8.4	–
24	Cardio-Heart Rate Monitor+7 Minute Workout	8.4	–	8.4	–
25	Belly Fat Workout Free	8.3	–	8.3	–
26	Strava Running and Cycling	8.0	8.0	–	–
27	Simply Yoga Free	6.4	–	–	6.4
28	Daily Yoga-Lose Weight, Get Relief	6.2	–	–	6.2
	Mean (SD)	17.8 (8.9)	11.7 (5.9)	13.9 (5.2)	4.0 (1.9)

[a]The highest possible overall score is 70; the highest possible sections I and II score is 30; the highest possible section III score is 10.

[b]Section I is aerobic exercise.

[c]Section II is strength and resistance.

[d]Section III is flexibility.

All subsection scores for the 28 apps evaluated in this study are summarized in Multimedia Appendix 2. Although the scores were low in general, the apps performed better when providing safety precautions and describing program principles for strength and resistance training than for aerobic exercise. Among apps with a strength and resistance component, more than half (12 out of 23, 52%) of them scored at least 5 points out of a maximal of 10 points in the safety subsection, and 9 of them (39%) scored at least 5 points out of a maximal of 10 points in the program principles subsection. On the other hand, among apps with an aerobic exercise component, 3 out of 12 apps (25%) scored higher than 5 points in the safety subsection, and only 2 apps (17%) scored higher than 5 points in the program principles subsection.

Discussion

Principal Findings

We developed and presented a novel scoring instrument for evaluating the quality of exercise program apps with respect to the ACSM exercise prescription guidelines. Inter- and intra-rater reliability and criterion validity were assessed using the Spearman correlation coefficients and weighted kappa statistics. Our results showed excellent reliability and validity of our

instrument, indicating that it provides accurate and stable measurement of the quality of exercise apps. In addition, our findings confirmed the preliminary study of Modave et al that very few of the most popular exercise prescriptive apps are of sufficient quality to provide evidence-based exercise prescription. This is a significant problem, particularly for novice exercisers who do not have the necessary expertise to develop their own program or assess whether a program is well-designed. In addition, the subscores pertaining to the aerobic, strength and resistance, and flexibility components suggest that most apps are of substandard quality even if we assume that they were designed to address a single aspect of exercise prescription.

For inter-rater reliability, the least reliable question asked whether an app provides any safety warning before starting aerobic exercise. One potential reason for the low reliability is that fitness apps have a variety of ways of delivering safety warnings. Current fitness apps are very sophisticated in terms of user interface design and content presentation. Safety warnings may be presented as part of the introductory text that users need to read or delivered as a push notification that users must read and close before starting a training program. Furthermore, safety warnings may be presented as part of the exercise instructional videos, as many apps are delivering training programs in videos. This diversity in safety warning delivery methods creates some difficulty for app reviewers to find and judge the appropriateness of safety warnings, especially in sophisticated fitness apps.

Although the overall intra-rater reliability is high, we have observed a few relatively lower reliability coefficients. Given that the test and retest were conducted 1 month apart, it is possible that some of the app contents and functions were updated, introducing additional unexpected measurement errors from the app reviewers. In order to survive and stay competitive in the crowded health and fitness apps market, app companies are constantly redesigning their product. It is not surprisingly to see fitness apps redesigning user interface, adding or removing functions, or disappearing from the app stores.

The overall quality scores and principle section quality scores are low for the apps evaluated in this study. It indicates that most popular fitness apps do not fully follow the evidence-based exercise principles set forth by the ACSM. On the other hand, one reason for the low overall quality scores is that not all apps have all 3 components of aerobic exercise, strength and resistance exercise, and flexibility. As the number of fitness apps continue to grow, they have become more specialized and focused in functions, with many apps providing exercise programs only for one specific type of exercise. For instance, with a score of 24.0 (out of 30.0 points) for the strength and resistance section, StrongLifts 5x5 is an app of good quality for strength training. However, it does not provide sufficient prescriptive principles for the aerobic exercise and flexibility, making the overall quality score relatively low for this app. Therefore, we recommend that the section quality scores always be evaluated with the overall quality score, which provides a more comprehensive evaluation of the app quality. In addition, app designers may need to consider all 3 main components of the ACSM exercise prescription principles when designing exercise apps, which can improve the degree to which apps adhere to the ACSM standards.

Limitations

One limitation of the study is that evaluating the quality of fitness apps using the instrument in its current form can be time consuming. Our reviewers have reported spending on average 30-40 min to score one app. It can be tedious when a large number of apps need to be evaluated. However, the reviewers have also noted that the scoring process becomes fairly automatic after having reviewed a few apps. Thus, the scoring process can be shortened with some practice. However, future research is needed to develop reliable and valid short forms of the instrument. Another limitation of the study is the small sample size for intra-rater reliability testing. More apps need to be reviewed for computing and analyzing section-specific intra-rater reliability. Nonetheless, reliability coefficients from this study are all statistically significant, providing initial evidence that the instrument is reliable. Finally, we acknowledge that this scoring instrument was designed to capture app quality for overall exercise programs, not for specialized programs that emphasize one type of activity such as weight lifting, yoga, or running. Additional scoring methods are needed to determine quality for these specialized exercise programs.

Conclusions

We have developed a reliable and valid instrument for evaluating the quality of exercise program apps according to the ACSM exercise prescription guidelines. This instrument can be used to determine which apps are useful for novice exercisers and health care providers who are looking for apps that provide safe, progressive general exercise programs for health and fitness. Given that the instrument provides separate scores for the 3 main components of ACSM principles, it can also be used to determine which apps are suitable for each of the components of an exercise routine.

In future research, we plan to examine the relationship between scores from our instrument and those from other instruments for evaluating app quality, including the instrument published by Stoyanov et al. This would provide a more comprehensive evaluation of exercise apps and thus help users to select apps based on not only evidence-based exercise principles but also functionality and aesthetics. In addition, we plan to evaluate the relationship between app quality scores and behavioral outcomes (eg, adherence, exercise, and health outcomes) among app users. It is important to quantify the differential impacts of app quality aspects on these outcomes to design better exercise apps.

Authors' Contributions

FM was responsible for the conception of the study, writing, and editing of the manuscript. YG was responsible for the study design, data collection and analysis, and writing and editing of the manuscript. HKV was the expert reviewer who participated in apps scoring. JB, TL, and HV assisted in data collection and editing of the manuscript. MS helped with writing and editing.

Conflicts of Interest

None declared.

Multimedia Appendix 1

[PDF File (Adobe PDF File), 171KB - jmir_v19i3e67_app1.pdf]

Multimedia Appendix 2

[PDF File (Adobe PDF File), 51KB - jmir_v19i3e67_app2.pdf]

References

1. US Department of Health and Human Services. 2008 physical activity guidelines for Americans. Hyattsville, MD: US Department of Health and Human Services; 2008.
2. Office of the Surgeon General (US), Office of Disease Prevention and Health Promotion (US), Centers for Disease Control and Prevention (US), National Institutes of Health (US). The surgeon General's call to action to prevent and decrease overweight and obesity. Rockville (MD) 2001:1-39. [Medline: 20669513]
3. Reiner M, Niermann C, Jekauc D, Woll A. Long-term health benefits of physical activity--a systematic review of longitudinal studies. BMC Public Health 2013;13:813 [FREE Full text] [doi: 10.1186/1471-2458-13-813] [Medline: 24010994]
4. Ogden CL, Carroll MD, Flegal KM. Prevalence of obesity in the United States. J Am Med Assoc 2014 Jul;312(2):189-190. [doi: 10.1001/jama.2014.6228] [Medline: 25005661]
5. Ogden CL, Carroll MD, Kit BK, Flegal KM. Prevalence of childhood and adult obesity in the United States, 2011-2012. J Am Med Assoc 2014 Feb 26;311(8):806-814. [doi: 10.1001/jama.2014.732] [Medline: 24570244]
6. Ward B, Clarke T, Nugent C, Schiller J. CDC. Early release of selected estimates based on data from the 2015 National Health Interview Survey URL: https://www.cdc.gov/nchs/nhis.htm [accessed 2017-02-07] [WebCite Cache ID 6o5S0JQaw]
7. de Rezende LF, Rodrigues LM, Rey-López JP, Matsudo VK, Luiz Odo C. Sedentary behavior and health outcomes: an overview of systematic reviews. PLoS One 2014;9(8):e105620 [FREE Full text] [doi: 10.1371/journal.pone.0105620] [Medline: 25144686]
8. González-Gross M, Meléndez A. Sedentarism, active lifestyle and sport: impact on health and obesity prevention. Nutr Hosp 2013 Sep;28(Suppl 5):89-98 [FREE Full text] [doi: 10.3305/nh.2013.28.sup5.6923] [Medline: 24010748]
9. Hruby A, Hu FB. The epidemiology of obesity: a big picture. Pharmacoeconomics 2015 Jul;33(7):673-689 [FREE Full text] [doi: 10.1007/s40273-014-0243-x] [Medline: 25471927]
10. Thompson PD, Buchner D, Pina IL, Balady GJ, Williams MA, Marcus BH, American Heart Association Council on Clinical Cardiology Subcommittee on Exercise, Rehabilitation, and Prevention, American Heart Association Council on Nutrition, Physical Activity, and Metabolism Subcommittee on Physical Activity. Exercise and physical activity in the prevention and treatment of atherosclerotic cardiovascular disease: a statement from the Council on Clinical Cardiology (Subcommittee on Exercise, Rehabilitation, and Prevention) and the Council on Nutrition, Physical Activity, and Metabolism (Subcommittee on Physical Activity). Circulation 2003 Jun 24;107(24):3109-3116 [FREE Full text] [doi: 10.1161/01.CIR.0000075572.40158.77] [Medline: 12821592]
11. CDC. 2016. Leading causes of death URL: https://www.cdc.gov/nchs/fastats/leading-causes-of-death.htm [accessed 2017-02-07] [WebCite Cache ID 6o5S7OxCT]
12. Allison DB, Fontaine KR, Manson JE, Stevens J, VanItallie TB. Annual deaths attributable to obesity in the United States. J Am Med Assoc 1999 Oct 27;282(16):1530-1538. [Medline: 10546692]
13. Finkelstein EA, Brown DS, Wrage LA, Allaire BT, Hoerger TJ. Individual and aggregate years-of-life-lost associated with overweight and obesity. Obesity (Silver Spring) 2010 Feb;18(2):333-339 [FREE Full text] [doi: 10.1038/oby.2009.253] [Medline: 19680230]
14. Finkelstein EA, Trogdon JG, Cohen JW, Dietz W. Annual medical spending attributable to obesity: payer-and service-specific estimates. Health Aff (Millwood) 2009;28(5):w822-w831 [FREE Full text] [doi: 10.1377/hlthaff.28.5.w822] [Medline: 19635784]
15. Manson JE, Bassuk SS, Hu FB, Stampfer MJ, Colditz GA, Willett WC. Estimating the number of deaths due to obesity: can the divergent findings be reconciled? J Womens Health (Larchmt) 2007 Mar;16(2):168-176. [doi: 10.1089/jwh.2006.0080] [Medline: 17388733]
16. Morris JN, Heady JA, Raffle PA, Roberts CG, Parks JW. Coronary heart-disease and physical activity of work. Lancet 1953 Nov 28;265(6796):1111-1120. [Medline: 13110075]

17. Lee IM, Paffenbarger RS, Hennekens CH. Physical activity, physical fitness and longevity. Aging (Milano) 1997;9(1-2):2-11. [Medline: 9177581]
18. Blair SN, Jackson AS. Physical fitness and activity as separate heart disease risk factors: a meta-analysis. Med Sci Sports Exerc 2001 May;33(5):762-764. [Medline: 11323545]
19. Haskell WL, Lee I, Pate RR, Powell KE, Blair SN, Franklin BA, et al. Physical activity and public health: updated recommendation for adults from the American College of Sports Medicine and the American Heart Association. Med Sci Sports Exerc 2007 Aug;39(8):1423-1434. [doi: 10.1249/mss.0b013e3180616b27] [Medline: 17762377]
20. Donnelly JE, Blair SN, Jakicic JM, Manore MM, Rankin JW, Smith BK, American College of Sports Medicine. American College of Sports Medicine Position Stand. Appropriate physical activity intervention strategies for weight loss and prevention of weight regain for adults. Med Sci Sports Exerc 2009 Feb;41(2):459-471. [doi: 10.1249/MSS.0b013e3181949333] [Medline: 19127177]
21. Smith A. Pew Research Center. 2015. U.S. Smartphone Use in 2015 URL: http://www.pewinternet.org/2015/04/01/us-smartphone-use-in-2015/ [WebCite Cache ID 6o5SmMur3]
22. Pelletier S. AAMC. 2012. Explosive Growth in Health Care Apps Raises Oversight Questions URL: https://www.aamc.org/newsroom/reporter/october2012/308516/health-care-apps.html [WebCite Cache ID 6YSWW7CBr]
23. IMS Institute for Healthcare Informatics. IMShealth. 2015. Patient Adoption of mHealth URL: http://1.http://www.imshealth.com/files/web/IMSH%20Institute/Reports/Patient%20Adoption%20of%20mHealth/IIHI_Patient_Adoption_of_mHealth.pdf [accessed 2017-03-01] [WebCite Cache ID 6odLeI6Hv]
24. Azar KM, Lesser LI, Laing BY, Stephens J, Aurora MS, Burke LE, et al. Mobile applications for weight management: theory-based content analysis. Am J Prev Med 2013 Nov;45(5):583-589. [doi: 10.1016/j.amepre.2013.07.005] [Medline: 24139771]
25. Bailey J, Mann S, Wayal S, Abraham C, Murray E. Digital media interventions for sexual health promotion-opportunities and challenges: a great way to reach people, particularly those at increased risk of sexual ill health. Br Med J 2015;350:h1099. [Medline: 25736806]
26. Cavallo DN, Tate DF, Ries AV, Brown JD, DeVellis RF, Ammerman AS. A social media-based physical activity intervention: a randomized controlled trial. Am J Prev Med 2012 Nov;43(5):527-532 [FREE Full text] [doi: 10.1016/j.amepre.2012.07.019] [Medline: 23079176]
27. Cowan LT, Van Wagenen SA, Brown BA, Hedin RJ, Seino-Stephan Y, Hall PC, et al. Apps of steel: are exercise apps providing consumers with realistic expectations?: a content analysis of exercise apps for presence of behavior change theory. Health Educ Behav 2013 Apr;40(2):133-139. [doi: 10.1177/1090198112452126] [Medline: 22991048]
28. Hall CS, Fottrell E, Wilkinson S, Byass P. Assessing the impact of mHealth interventions in low- and middle-income countries--what has been shown to work? Glob Health Action 2014;7:25606 [FREE Full text] [Medline: 25361730]
29. Hswen Y, Viswanath K. Beyond the hype: mobile technologies and opportunities to address health disparities. J Mob Technol Med 2015 Jan;4(1):39-40. [doi: 10.7309/jmtm.4.1.9]
30. Marcus BH, Bock BC, Pinto BM, Forsyth LH, Roberts MB, Traficante RM. Efficacy of an individualized, motivationally-tailored physical activity intervention. Ann Behav Med 1998;20(3):174-180. [Medline: 9989324]
31. Middelweerd A, Mollee JS, van der Wal CN, Brug J, Te Velde SJ. Apps to promote physical activity among adults: a review and content analysis. Int J Behav Nutr Phys Act 2014;11:97 [FREE Full text] [doi: 10.1186/s12966-014-0097-9] [Medline: 25059981]
32. Payne HE, Lister C, West JH, Bernhardt JM. Behavioral functionality of mobile apps in health interventions: a systematic review of the literature. JMIR Mhealth Uhealth 2015;3(1):e20 [FREE Full text] [doi: 10.2196/mhealth.3335] [Medline: 25803705]
33. Smith JJ, Morgan PJ, Plotnikoff RC, Dally KA, Salmon J, Okely AD, et al. Smart-phone obesity prevention trial for adolescent boys in low-income communities: the ATLAS RCT. Pediatrics 2014 Sep;134(3):e723-e731 [FREE Full text] [doi: 10.1542/peds.2014-1012] [Medline: 25157000]
34. Cobb NK, Graham AL. Health behavior interventions in the age of facebook. Am J Prev Med 2012 Nov;43(5):571-572. [doi: 10.1016/j.amepre.2012.08.001] [Medline: 23079184]
35. Hwang KO, Farheen K, Johnson CW, Thomas EJ, Barnes AS, Bernstam EV. Quality of weight loss advice on internet forums. Am J Med 2007 Jul;120(7):604-609 [FREE Full text] [doi: 10.1016/j.amjmed.2007.04.017] [Medline: 17602934]
36. McCully SN, Don BP, Updegraff JA. Using the Internet to help with diet, weight, and physical activity: results from the Health Information National Trends Survey (HINTS). J Med Internet Res 2013;15(8):e148 [FREE Full text] [doi: 10.2196/jmir.2612] [Medline: 23906945]
37. Modave F, Shokar NK, Peñaranda E, Nguyen N. Analysis of the accuracy of weight loss information search engine results on the internet. Am J Public Health 2014 Oct;104(10):1971-1978. [doi: 10.2105/AJPH.2014.302070] [Medline: 25122030]
38. Moller AC, Majewski S, Standish M, Agarwal P, Podowski A, Carson R, et al. Active fantasy sports: rationale and feasibility of leveraging online fantasy sports to promote physical activity. JMIR Serious Games 2014 Nov 25;2(2):e13 [FREE Full text] [doi: 10.2196/games.3691] [Medline: 25654304]
39. Tate DF, Wing RR, Winett RA. Using Internet technology to deliver a behavioral weight loss program. J Am Med Assoc 2001 Mar 7;285(9):1172-1177. [Medline: 11231746]

40. Direito A, Jiang Y, Whittaker R, Maddison R. Apps for IMproving FITness and increasing physical activity among young people: The AIMFIT pragmatic randomized controlled trial. J Med Internet Res 2015;17(8):e210 [FREE Full text] [doi: 10.2196/jmir.4568] [Medline: 26316499]
41. Modave F, Bian J, Leavitt T, Bromwell J, Harris IC, Vincent H. Low quality of free coaching apps with respect to the American College of Sports Medicine Guidelines: a review of current mobile apps. JMIR Mhealth Uhealth 2015;3(3):e77 [FREE Full text] [doi: 10.2196/mhealth.4669] [Medline: 26209109]
42. American College of Sports Medicine. In: Pescatello LS, Arena R, Riebe D, Thompson PD, editors. ACSM's Guidelines for Exercise Testing and Prescription. Philadelphia, PA: Lippincott Williams and Wilkins; 2014.
43. health. 2008. 2008 Physical Activity Guidelines for Americans URL: https://health.gov/paguidelines/pdf/paguide.pdf [accessed 2017-03-01] [WebCite Cache ID 6odM8XIou]
44. Stoyanov SR, Hides L, Kavanagh DJ, Zelenko O, Tjondronegoro D, Mani M. Mobile app rating scale: a new tool for assessing the quality of health mobile apps. JMIR Mhealth Uhealth 2015;3(1):e27 [FREE Full text] [doi: 10.2196/mhealth.3422] [Medline: 25760773]
45. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977 Mar;33(1):159-174. [Medline: 843571]

Abbreviations

ACSM: American College of Sports Medicine
FITT: frequency, intensity, time, and type
MET: metabolic equivalent
PNF: proprioceptive neuromuscular facilitation
SD: standard deviation
UF: University of Florida

Edited by G Eysenbach; submitted 10.11.16; peer-reviewed by A Keath, S McCully; comments to author 01.12.16; revised version received 19.01.17; accepted 11.02.17; published 07.03.17.

Please cite as:
Guo Y, Bian J, Leavitt T, Vincent HK, Vander Zalm L, Teurlings TL, Smith MD, Modave F
Assessing the Quality of Mobile Exercise Apps Based on the American College of Sports Medicine Guidelines: A Reliable and Valid Scoring Instrument
J Med Internet Res 2017;19(3):e67
URL: http://www.jmir.org/2017/3/e67/
doi:10.2196/jmir.6976
PMID:28270378

Original Paper

Design and Implementation of a Novel Web-Based E-Learning Tool for Education of Health Professionals on the Antibiotic Vancomycin

Stuart Evan Bond[1,2,3], B Pharm, Dip Pharm Prac; Shelley P Crowther[1], B Pharm, M Public Health; Suman Adhikari[4,5], B Pharm; Adriana J Chubaty[6], BSc (Pharm); Ping Yu[7], PhD; Jay P Borchard[8], BSc (Hons); Craig Steven Boutlis[9], MBBS, PhD; Wilfred Winston Yeo[2,3,10], MD; Spiros Miyakis[2,3,9], MD, PhD

[1]Wollongong Hospital, Department of Pharmacy, Illawarra Shoalhaven Local Health District, Wollongong, Australia
[2]School of Medicine, Faculty of Science, Medicine and Health, University of Wollongong, Wollongong, Australia
[3]Illawarra Health and Medical Research Institute, University of Wollongong, Wollongong, Australia
[4]St George Hospital, Department of Pharmacy, South Eastern Sydney Local Health District, Kogarah, Australia
[5]St George Clinical School, Faculty of Medicine, University of New South Wales, Kogarah, Australia
[6]Prince of Wales Hospital, Department of Pharmacy, South Eastern Sydney Local Health District, Randwick, Australia
[7]School of Computing and Information Technology, Faculty of Engineering and Information Sciences, University of Wollongong, Wollongong, Australia
[8]Research Central, Wollongong Hospital, Illawarra Shoalhaven Local Health District, Wollongong, Australia
[9]Wollongong Hospital, Department of Infectious Diseases, Illawarra Shoalhaven Local Health District, Wollongong, Australia
[10]Wollongong Hospital, Division of Medicine, Illawarra Shoalhaven Local Health District, Wollongong, Australia

Corresponding Author:
Stuart Evan Bond, B Pharm, Dip Pharm Prac
Wollongong Hospital
Department of Pharmacy
Illawarra Shoalhaven Local Health District
Loftus St
Wollongong, 2500
Australia
Phone: 61 4255 1530
Fax: 61 4222 5430
Email: stuart.bond@health.nsw.gov.au

Abstract

Background: Traditional approaches to health professional education are being challenged by increased clinical demands and decreased available time. Web-based e-learning tools offer a convenient and effective method of delivering education, particularly across multiple health care facilities. The effectiveness of this model for health professional education needs to be explored in context.

Objectives: The study aimed to (1) determine health professionals' experience and knowledge of clinical use of vancomycin, an antibiotic used for treatment of serious infections caused by methicillin-resistant *Staphylococcus aureus* (MRSA) and (2) describe the design and implementation of a Web-based e-learning tool created to improve knowledge in this area.

Methods: We conducted a study on the design and implementation of a video-enhanced, Web-based e-learning tool between April 2014 and January 2016. A Web-based survey was developed to determine prior experience and knowledge of vancomycin use among nurses, doctors, and pharmacists. The Vancomycin Interactive (VI) involved a series of video clips interspersed with question and answer scenarios, where a correct response allowed for progression. Dramatic tension and humor were used as tools to engage users. Health professionals' knowledge of clinical vancomycin use was obtained from website data; qualitative participant feedback was also collected.

Results: From the 577 knowledge survey responses, pharmacists (n=70) answered the greatest number of questions correctly (median score 4/5), followed by doctors (n=271; 3/5) and nurses (n=236; 2/5; P<.001). Survey questions on target trough concentration (75.0%, 433/577) and rate of administration (64.9%, 375/577) were answered most correctly, followed by timing of first level (49%, 283/577), maintenance dose (41.9%, 242/577), and loading dose (38.0%, 219/577). Self-reported "very" and

"reasonably" experienced health professionals were also more likely to achieve correct responses. The VI was completed by 163 participants during the study period. The rate of correctly answered VI questions on first attempt was 65% for nurses (n=63), 68% for doctors (n=86), and 82% for pharmacists (n=14; *P*<.001), reflecting a similar pattern to the knowledge survey. Knowledge gaps were identified for loading dose (39.2% correct on first attempt; 64/163), timing of first trough level (50.3%, 82/163), and subsequent trough levels (47.9%, 78/163). Of the 163 participants, we received qualitative user feedback from 51 participants following completion of the VI. Feedback was predominantly positive with themes of "entertaining," "engaging," and "fun" identified; however, there were some technical issues identified relating to accessibility from different operating systems and browsers.

Conclusions: A novel Web-based e-learning tool was successfully developed combining game design principles and humor to improve user engagement. Knowledge gaps were identified that allowed for targeting of future education strategies. The VI provides an innovative model for delivering Web-based education to busy health professionals in different locations.

(J Med Internet Res 2017;19(3):e93) doi:10.2196/jmir.6971

KEYWORDS

nursing education; pharmacy education; medical education; continuing education; vancomycin; survey methods; anti-bacterial agents

Introduction

Internet-Based Learning

The development of Internet-based learning (IBL) for health care professionals has increased in recent years [1]. One reason for advancement of IBL is the existence of barriers associated with implementation of face-to-face health professional education, including increased clinical demands and decreased available time [2]. These barriers become more evident where education is required across multiple facilities that are separated by long distances. Consequently, there is a requirement for more effective and accessible ways of improving knowledge and competence in health professionals [1]. To date, IBL approaches have shown positive effects on health education outcomes through overcoming the above barriers [3].

Serious games have been defined as "interactive computer applications, with or without significant hardware components" that are designed to entertain while achieving changes in knowledge or skills [4,5]. Methods to improve their entertainment value include dramatic tension, humor, and challenge [5]. User engagement can also be improved through the inclusion of a narrative [6]. Humor as an aid to nursing and medical education has been described in the literature [7,8], where the use of games as a medium for humor may increase learners' interest and motivation to learn [7]. As distinct from e-learning with limited user interaction [9], serious games can provide greater engagement with the educational content.

Use of serious game methodologies to deliver health professional education has been reported in previous studies [10-12]. Educational and design frameworks are recommended for the development of games for health professional education [9,13]. Strategies include application of knowledge in a safe environment that resembles real life [13], a degree of interactivity [14], and entertainment [5,9]. These topics were considered in development and assessment of the e-learning tool in this study.

Most e-learning tools in health care have targeted specific groups, such as medical or nursing students, physicians, or nurses [15-18]. We developed the Vancomycin Interactive (VI) to target nurses, doctors, and pharmacists, the three main groups involved in use of medicines in hospitals. The specific educational content of the VI was clinical use of the glycopeptide antibiotic, vancomycin, given intravenously in hospitals for treatment of infections caused by methicillin-resistant *Staphylococcus aureus* (MRSA). MRSA infections have high mortality and are resistant to conventional treatment with safer antibiotics such as penicillins, which usually do not require such specific administration and monitoring. Vancomycin is a commonly used antibiotic for treatment of MRSA infections [19], but there are problems associated with its use. Those include the requirement for a loading dose (initial higher single dose) in serious infections, side effects when administered too rapidly, and the need to monitor vancomycin plasma levels [20]. As part of our antimicrobial stewardship program [21], local quality improvement activities identified gaps in competence around clinical use of vancomycin. Three main topics were identified from those local activities and from previous studies: (1) dosing, including loading and maintenance [22-24]; (2) administration, such as compatible fluids and rate of infusion [24,25]; and (3) therapeutic drug monitoring (TDM), including appropriate timing of blood sampling, target trough levels, and required actions based on reported levels [24,26,27].

Aims of This Study

The aims of this study were to (1) report the design and implementation of a Web-based, interactive e-learning tool providing education on the dosing, administration, and TDM of vancomycin, (2) assess health professionals' preintervention knowledge of vancomycin use in order to inform development of the e-learning tool, and (3) assess health professionals' initial acceptance of the VI.

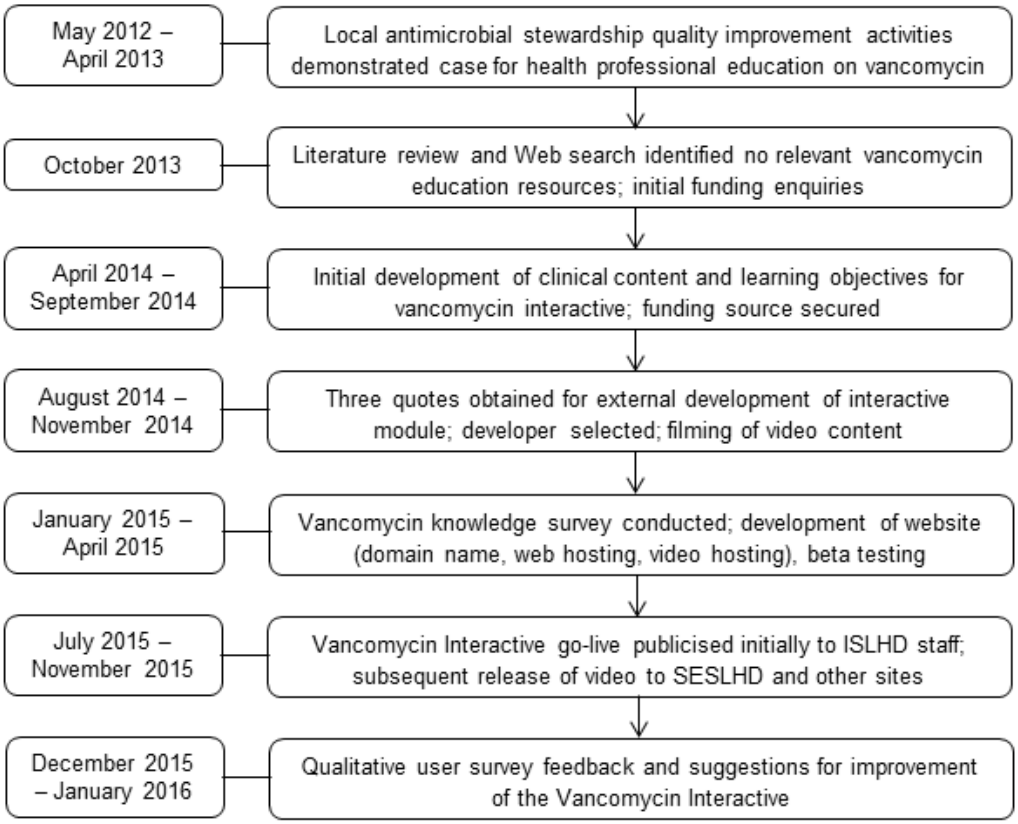
Methods

Setting

This prospective design and implementation study of a video-enhanced, Web-based e-learning tool took place in Illawarra Shoalhaven Local Health District (ISLHD) and South Eastern Sydney Local Health District (SESLHD), located in New South Wales (NSW), Australia. These health districts cover

a geographic area of 6331 km^2 and have an estimated population of 1.17 million, reaching from central Sydney to 3 h drive south [28]. The districts' 14 hospitals contain a total of 2500 beds and range from small rural facilities to large tertiary metropolitan hospitals. A timeline of design, implementation, and evaluation is shown in Figure 1.

Figure 1. Timeline of Vancomycin Interactive design, implementation, and evaluation. ISLHD: Illawarra Shoalhaven Local Health District; SESLHD: South Eastern Sydney Local Health District.



Web-Based Vancomycin Knowledge Survey

An anonymous open Web-based survey was created using SurveyMonkey (SurveyMonkey Inc, Palo Alto, CA) to determine confidence, experience, and knowledge of vancomycin before the VI. The survey was developed locally by the antimicrobial stewardship and educator pharmacists as part of routine activities, with input from the infectious diseases team. Clinical content was based on the Australian Therapeutic Guidelines: Antibiotic, Version 15, 2014 [20] and the Australian Injectable Drugs Handbook, Version 6, 2015 [29]. Use of these references was required as part of the Australian hospital accreditation standards [30]. Survey participants were nurses, doctors, and pharmacists from the two health districts. A 4-point Likert scale was used to determine levels of experience, confidence and knowledge on dosing, administration, and TDM of vancomycin (see Multimedia Appendix 1). The survey was advertised using email and the districts' fortnightly newsletters. The survey link was open from February 1, 2015 to June 30, 2015, and participation was voluntary. Only one attempt was allowed on each question and users were directed to further reading material at completion of the survey. Nurses were expected to correctly answer questions on fluid compatibility and administration rate, since they were mainly responsible for administration of medicines in hospitals. Doctors were anticipated to correctly answer questions relating to dosing and TDM, arising from their role as prescribers. Pharmacists were expected to have a working knowledge of all aspects of clinical vancomycin use. The response rate to the survey was calculated from the number of respondents and the number of recipients on staff email groups.

Design and Implementation of the Vancomycin Interactive

Similar to the survey, clinical content of the VI was developed locally, based on the Therapeutic Guidelines: Antibiotic [20] and the Australian Injectable Drugs Handbook [29]. The vancomycin knowledge survey informed in part the VI's educational content in the postproduction phase, allowing finalization of the multiple choice questions. An entertaining Web-based educational tool was selected in the early development stage (mid-2013; Figure 1) for two reasons. First, there was already a health district requirement for staff to complete between 10 and 20 h mandatory training per year on other topics, and study investigators did not wish to contribute to the burden of further Web-based mandatory training. Rather, a brief, targeted, and light-hearted educational tool was thought to be more acceptable and beneficial for staff. Second, large distances between hospital sites meant that face-to-face education of health professionals was very resource intensive. The learning objectives of the VI for target users (nurses, doctors, and pharmacists) were to improve knowledge of vancomycin dosing, administration, and TDM. The VI did not address clinical indications for vancomycin, dosing in specialist areas such as intensive care and renal dialysis, use of continuous infusions, or surgical prophylaxis.

A single interactive video was produced due to financial constraints; there was the expectation that all professional groups should have rudimentary knowledge of clinical vancomycin use. The VI (ISLHD) was hosted on an open website [31]. Using the serious game design concepts of interactivity and entertainment, we presented a case study resembling real-life interaction between a patient and a health professional. Dramatic tension between the two characters created the basis for the plot, along with the unprofessional behavior of the modeled health professional. The interaction was also designed to be humorous, particularly through the special effect of "shrinking" the health professional, and references to William Shakespeare's plays (Figures 2 and 3). The concept was intended to appeal to health professionals who may feel that they are at the mercy of their patients, a theme that emerged during the script-writing process.

The user interface consisted of video clips interspersed with interactive question and answer scenarios placed at the specific points, so that technical content felt organic to the narrative (Figure 4; Multimedia Appendix 2). A correct answer allowed progression to the next section, whereas an incorrect answer resulted in a shaking screen and a sound effect. Data captured from answers to the interactive questions allowed for subsequent analysis. Only data from the targeted health professionals were included in the analysis; students and other participants were excluded. Additional questions in the VI, as distinct from the survey, related to compatibility of vancomycin with various fluids and clinical actions in response to different trough levels. Completion of the VI took approximately 10 min based on user testing.

Quotes for production were obtained from three developers in accordance with NSW Health policy, with financial support provided internally by the Clinical Governance Unit of the health district. Content development began in April 2014, and the video was filmed using professional actors in November 2014. Postproduction modifications were made to the video until release in July 2015. In early 2015, the website was established to promote improved access to the VI, and to include additional clinical content not contained in the VI. Testing of content and usability was performed by pharmacists and infectious diseases doctors (N=8) at the study site, with feedback provided by email to the study investigators. Feedback from testers predominantly related to accuracy of the clinical content in the context of the narrative, and informed the final iteration of the VI. The first phase of dissemination and advertisement (email, newsletters, link on intranet home page) to ISLHD staff occurred on July 27, 2015 (Figure 1), with the initial target audience estimated from organizational records to be 1000 staff. General release of the VI outside of ISLHD occurred on November 17, 2015. The final production cost was Aus $15,000; time devoted to content development, testing, advertising, implementation, and evaluation was not included in those costs as it fell within usual activities for the pharmacy and infectious diseases department staff members involved in development of the VI.

Figure 2. Filming of the Vancomycin Interactive (VI).

Figure 3. Dramatic tension created the basis for the Vancomycin Interactive's plot.

Figure 4. Example of user interface for an interactive question from the Vancomycin Interactive (VI).

User Acceptance Evaluation

Following release of the VI, qualitative survey responses were assessed to inform the investigators about user acceptability and suggestions for improvement. The qualitative survey was open between December 1, 2015 and January 31, 2016, in order to conclude before the annual intake of new junior doctors in February 2016 (Figure 1).

Outcome Measures

The primary outcome measure was comparative vancomycin knowledge between health professions and self-reported levels of confidence and experience. Vancomycin knowledge responses from website data (not linked at a participant level) were also assessed and compared with the knowledge survey. In addition, qualitative feedback on the VI was evaluated using a 5-point Likert scale and free text responses that were grouped into key themes. Assessments were derived from survey responses and VI website data. Technical issues around compatibility with desktop and mobile operating systems and Web browsers were also assessed. Reporting of outcomes on quantitative postintervention survey data, clinical measures of quality vancomycin use such as therapeutic vancomycin plasma levels, and clinical outcomes related to vancomycin treatment was beyond the scope of this study.

Statistical Analyses

Chi-square and Fisher exact tests were used for proportions. Chi-square for trend was used to determine trend between professions for knowledge questions. Kruskal-Wallis and Mann-Whitney *U* tests were used to examine total survey scores. For continuous data, normality was assessed using the Shapiro-Wilk statistic. A skewed distribution was denoted by $P<.05$. Kruskal-Wallis and follow-up Wilcoxon rank-sum tests were used to investigate between subjects effects with nonnormal distributions. A multivariate analysis was performed to examine influential factors on correct survey responses. For each item, a logistic regression was conducted followed by a multiple regression on the total score. For profession, nurses were allocated to the reference group, and self-reported "no experience or confidence" was used at the reference for the experience analysis. Statistical significance was accepted as $P<.05$. Additionally, a mediation analysis [32] was carried out to explore the mediating effects of vancomycin experience on the association between profession and knowledge (reflected by the total number of correct responses). For the mediation analysis, significance was determined by the 95% CI of the regression coefficient, *b*. If the 95% CI did not contain 0, it was considered significant. The extent of mediation was reported as a percentage, where a higher percentage reflects greater mediation. Statistical analyses were performed using Stata statistical software: Release 14 (Statacorp LP).

Ethics

Ethics approval was granted by the Joint University of Wollongong and Illawarra Shoalhaven Local Health District Health and Medical Human Research Ethics Committee (EC00150; approval number HE15/005). The VI website contained a disclaimer that anonymous data collected from the video could be used for research purposes.

Results

Vancomycin Knowledge Survey Before Release of the Vancomycin Interactive

The response rate to the survey was 26.87% (577 responses from 2147 email recipients). The response rates by profession were 24.4% (236/967) for nurses, 25.33% (271/1070) for doctors, and 63.6% (70/110) for pharmacists ($P<.001$).

As shown in Table 1, the median knowledge survey score for nurses was 2/5 (interquartile range, IQR 1-3), compared with 3/5 (IQR 3-4) for doctors and 4/5 (IQR 3-4) for pharmacists ($P<.001$). Pharmacists had greater total scores than doctors ($P<.001$) and nurses ($P<.001$), whereas doctors had greater total scores than nurses ($P<.001$). For nurses, the most correctly answered questions were on administration rate (64.4% correct; 152/236) and target trough range (57.6% correct; 136/236), whereas only 19.5% (46/236) of nurses answered the loading dose question correctly. The most correctly answered question by doctors was on target trough range (86.3% correct; 234/271). Pharmacists answered all responses correctly greater than 80% of the time.

Table 1. Number of correct responses to Web-based vancomycin knowledge survey.

Survey question	Nurse (n=236) n (%)	Doctor (n=271) n (%)	Pharmacist (n=70) n (%)	*P* value	Total (N=577) n (%)
Loading dose	46 (19.5)	112 (41.3)	59 (84)	<.001	217 (37.6)
Maintenance dose	58 (24.6)	126 (46.5)	58 (83)	<.001	242 (41.9)
Administration rate	152 (64.4)	160 (59.0)	62 (89)	<.001	374 (64.8)
First level timing	70 (29.7)	155 (57.2)	59 (84)	<.001	284 (49.2)
Target trough range	136 (57.6)	234 (86.3)	65 (93)	<.001	435 (75.4)
Median total score (IQR[a])	2 (1-3)	3 (3-4)	4 (3-4)	<.001	3 (2-4)

[a]IQR: interquartile range.

Multivariate analysis (Table 2) showed that for the loading dose question, pharmacists and doctors were more likely to achieve a correct response than nurses. A smaller variation between professions was seen for the administration rate question, with the comparison between pharmacists and nurses reaching significance. In addition, self-reported "very" and "reasonably" experienced health professionals were more likely to achieve a correct response. Similar associations between professions and experience levels were seen for maintenance dose, first level timing, and trough level range (Table 2). Pharmacists self-reported more experience and confidence than doctors or nurses, which influenced the likelihood of a correct response.

Table 2. Multivariate analysis of vancomycin knowledge survey responses (N=577).

Topic		Profession			Experience or confidence[a]			
		Nurse ref[b]	Doctor	Pharmacist	1 ref[b]	2	3	4
Loading dose	OR[c] (95% CI)	-	2.6 (1.7-4.1)	16.8 (7.9-35.7)	-	1.4 (0.8-2.7)	4.6 (2.4-8.6)	11.1 (3.3-36.9)
	P value	-	<.001	<.001	-	.25	<.001	<.001
Maint dose[d]	OR (95% CI)	-	2.5 (1.7-3.8)	12.1 (5.9-24.7)	-	1.0 (0.6-1.8)	2.2 (1.2-3.8)	3.3 (1.1-9.6)
	P value	-	<.001	<.001	-	.85	.006	.03
Admin rate[e]	OR (95% CI)	-	1.0 (0.6-1.4)	2.9 (1.3-6.4)	-	2.1 (1.4-3.3)	4.7 (2.9-7.6)	5.7 (0.2-1.1)
	P value	-	.82	.01	-	.001	<.001	<.001
Level timing[f]	OR (95% CI)	-	2.7 (1.9-4.1)	8.5 (4.0-17.7)	-	3.6 (2.1-5.9)	6.8 (4.0-11.6)	4.0 (1.4-10.9)
	P value	-	<.001	<.001	-	<.001	<.001	.008
Trough range[g]	OR (95% CI)	-	3.9 (2.5-6.1)	5.6 (2.1-15.1)	-	3.4 (2.1-5.6)	5.7 (3.2-10.0)	5.7 (1.2-26.5)
	P value	-	<.001	.001	-	<.001	<.001	.03
Total correct[h]	*b*[i] (95% CI)	-	.8 (0.6-1.1)	1.7 (1.4-2.1)	-	.9[j] (1.4-2.1)		
	P value	-	<.001	<.001	-	<.001		

[a]Experience or confidence: 1, none; 2, a little; 3, moderate; 4, very experienced or confident.

[b]ref: reference group for multivariate analysis.

[c]OR: odds ratio.

[d]maint dose: maintenance dose.

[e]admin rate: administration rate.

[f]level timing: timing of first level.

[g]trough range: target range for plasma trough level.

[h]total correct: nurse (all levels), doctor (all levels), pharmacist (all levels).

[i]*b*: regression coefficient.

[j]Average of responses to three vancomycin experience or confidence questions, therefore a multiple regression was performed for total correct.

Subsequent mediation analysis revealed that vancomycin experience significantly mediated the effect of profession on total score (total indirect effect: *b*=.63, bias-corrected 95% CI 0.44-0.85). Approximately 58% of the profession effect was mediated by experience, where a higher percentage value indicates greater mediation.

Vancomycin Interactive

Responses to the VI were analyzed using background website data received from July 27, 2015 to November 14, 2015, with ISLHD as the target population group. The initial dropdown question asking the user's profession was answered by 389 participants, of which 163 health professionals (41.9% of those answering the initial profession question) completed all 10 questions (Table 3). The rate of correctly answered questions on first attempt was 65% for nurses, 68% for doctors, and 82% for pharmacists, significantly higher in the pharmacist group (*P*<.001). Notably low numbers of correct responses were identified for the following three questions, averaged over the three professional groups: loading dose (39% correct), timing of first level (50%), and timing of subsequent levels (48%).

Table 3. Number (%) of correct answers on first attempt by nurses, doctors, and pharmacists from VI data.

Question[a]	Nurse n=63	Doctor n=86	Pharmacist n=14	*P* value[b]	Total N=163
1. Loading dose	19 (30)	36 (42)	9 (64)	.05	64 (39)
2. Maintenance dose	50 (79)	59 (69)	11 (79)	.32	120 (74)
3. Compatible fluids	53 (84)	67 (78)	11 (79)	.76	131 (80)
4. Administration rate	56 (89)	55 (64)	14 (100)	<.001	125 (77)
5. Timing of first level	20 (32)	49 (57)	12 (86)	<.001	81 (50)
6. Target trough level	47 (75)	72 (84)	12 (86)	.34	131 (80)
7. Level of 35 mg/L	43 (68)	68 (79)	14 (100)	.02	125 (77)
8. Level of 20 mg/L	49 (78)	81 (94)	13 (93)	.01	143 (88)
9. Level of 26 mg/L	46 (73)	55 (64)	12 (86)	.20	113 (69)
10. Subsequent levels	27 (43)	45 (52)	7 (50)	.52	79 (48)
Average score	65%	68%	82%	<.001	68%

[a]VI questions are shown in Multimedia Appendix 2.

[b]*P* values obtained using chi-square for trend.

Comparison of Responses Between Vancomycin Interactive (VI) and Web-Based Survey

The rates of correct response from the VI were significantly higher than the knowledge survey for maintenance dose (74% VI vs 42% survey; *P*<.001) and administration rate questions (77% VI vs 65% survey; *P*=.004). There was a slightly higher correct response rate for the question on target trough level (80% VI vs 75% survey; *P*=.19). Uniformly low correct response rates were observed for the questions on loading dose (39% for VI vs 38% for survey; *P*=.70) and the timing of first level (50% VI vs 49% survey; *P*=.89). The question on timing of levels subsequent to the first level in the VI was answered correctly in 48% of cases; there was no equivalent question in the survey.

User Acceptance Evaluation of the Vancomycin Interactive

Among the 163 VI participants, 51 (31%) responses were received. Responses were predominantly positive, as shown in Table 4.

Table 4. Qualitative responses (%) following participation in the Vancomycin Interactive (VI).

Survey statement or question	Strongly agree	Agree	Neutral	Disagree	Strongly disagree
Using the VI[a] has enhanced my knowledge (n=51)	11 (22)	29 (57)	8 (16)	3 (6)	0 (0)
Using the VI has improved my performance (n=50)	8 (16)	28 (56)	12 (24)	2 (4)	0 (0)

[a]VI: Vancomycin Interactive.

When users were asked, “What’s good about the VI in comparison to other e-learning modules?” 28 free text responses were received. It was found that 4 responses (14%) were related to not being able to load the video. Key themes from the remaining 24 responses (86%) were “entertaining,” “engaging,” “a lighter approach to learning,” “more real life,” and “held attention.” The question, “Does the training provided by the VI meet your needs? If not, what can be improved?” received 23 free text responses. A total of 16 respondents (70%) reported, “yes it met needs;” 2 users (10%) stated issues loading VI; 3 users (13%) requested printable resources; 1 user was “not sure;” and 1 user requested more information to be available when answering questions. All qualitative survey responses are provided in Multimedia Appendix 3.

Discussion

Principal Findings

We have reported on the design, implementation, and user evaluation of a novel Web-based e-learning tool for education of health professionals on clinical use of the antibiotic vancomycin. The VI was developed for noncommercial use and targeted three health professional groups across multiple hospital sites. Responses from the survey that preceded the VI demonstrated a global lack of knowledge on the safe and effective use of vancomycin among nurses and doctors, justifying a Web-based learning approach that was suitable for disparate geographical locations. Pharmacists were shown to be more knowledgeable on clinical vancomycin dosing, administration, and TDM.

As expected, self-reported levels of confidence and experience were correlated with increased likelihood of correct responses

to the knowledge survey questions. Responses from the Web-based knowledge survey and VI data were only similar for three of the five common questions, loading dose, timing of first level, and the target trough level (see Multimedia Appendices 1 and 2). This may suggest that respondents equally understood those three questions in the VI and the knowledge survey. Responses to two questions, maintenance dose and administration rate, were significantly better in the knowledge survey compared with the VI. This could be caused by the respondents' different understanding about the survey questions presented in the two media or difference in knowledge level between the participants in the two surveys. Following implementation, qualitative survey responses demonstrated that the VI was well-received by users, and was considered to be an engaging and entertaining method of improving knowledge. A small number of responses highlighted technical issues relating to not being able to load the video content, which were generally resolved through software upgrades.

Numerous studies have reported the development and evaluation of serious games for training health professionals, but few have targeted multiple professions [11,33-35]. One study reported development of a serious game on appropriate antibiotic use, but this was not specific to any particular antibiotic [12]. Vancomycin was chosen as the topic for our Web-based tool due to its frequency of use, and complexities associated with treatment of serious MRSA infections, the requirement for loading doses, TDM, and subsequent dose adjustment. The VI in this study adopted some principles of serious game design [9], including interactivity and entertainment, and combined those with humor [8] to engage multiprofessional users. Knowledge responses from the VI are promising, and further research is needed to determine the reasons for difference in responses to questions between the classical knowledge survey and knowledge responses from VI website data.

Feedback from the majority of the participants suggested that the VI enhanced their vancomycin knowledge (79%) and improved their performance (72%). This supports the VI as a Web-based resource to provide health care professionals with training on clinical use of vancomycin. Qualitative responses were generally positive, further supporting the use of the VI for health professional education. The main challenges for implementation of the VI related to developing clinical content for the video that would remain applicable to all three professional groups, without creating a tool that would take too long to complete. Advertising the tool using different media was also challenging, as the tool was made available across two health districts with multiple hospitals, and the target professional groups may have preferred to receive alerts regarding content in different ways.

The creation of a brief, Web-based, entertaining educational tool was the purpose of the project, whereby no further mandatory training burden was placed on staff. As distinct from existing local mandatory learning modules, the VI was intended for use among clinical staff involved in vancomycin use. Employing serious game design concepts may provide greater educational benefit than traditional computer-based learning methods through the use of greater interactivity, entertainment, and scoring; however, further published comparisons are required [11]. Our results suggest that pharmacists have the greatest level of knowledge on clinical vancomycin use. Therefore, to deliver the best learning outcomes for health professionals in this area, it is recommended to combine face-to-face teaching with VI learning, using pharmacists as educators in the future.

Limitations of This Study

We studied the logistics and design of a Web-based e-learning tool incorporating interactive video content for health professional education relating to clinical use of the antibiotic vancomycin. Postintervention knowledge and clinical outcomes were not reported here; these form the basis of ongoing research that will be reported separately. The use of an open Internet site allowed for potential diffusion worldwide, since users outside our organizations may have found the VI using an Internet search engine. In August 2015, the website was also shared on a professional network with members outside the targeted health district. As a result, there was some unintended use of the video before its general release. However, the greatest number of Web sessions was from ISLHD, and employees of the target ISLHD hospitals may not have been physically located in the region while completing the VI.

Question design within the VI was limited to multiple choice and multiple answer questions. Further variation in question types such as open questions, as previously reported [16], could be made in future versions to improve immersion and interactivity. The inclusion of a formal testing process immediately before and after the e-learning tool may also have added some informative value on its effect and could inform future improvements. In addition, further scoring methodology, such as time limitation, competition, and increasing difficulty could improve the robustness of the design [11]. The Hawthorne effect may have introduced bias into the study, whereby participants' behavior may have been altered through their awareness of being measured. This bias may have been limited by participation being anonymous, and the primary intended aim for users being to further their vancomycin knowledge, rather than participation in a research project. Detailed economic analysis of the study was limited by the project forming part of usual educational activities for study investigators. As such, the total project cost was likely greater than the reported production cost.

There was relatively low uptake of the VI among clinical staff during the study period despite broad advertisement; this limited the statistical power of the study and highlighted the challenge of using a new e-learning tool for delivery of noncompulsory training material to health professionals. Reasons for this probably related to the following: (1) the VI was not mandatory learning, so health professionals who did not regularly use vancomycin may not have been motivated to participate; (2) competing education priorities in those health professionals not otherwise intrinsically motivated to participate; (3) lack of time out from clinical responsibilities; (4) the likelihood that multiple staff completed the VI together, meaning that the VI's reach might have been greater than the results demonstrated; (5) the tool was not targeted toward a specific profession; and (6) not being able to access the VI using hospital computers, which

may have hampered widespread use by health professionals during office hours. However, there were only four reports of the VI not loading from 51 survey responses, suggesting that the majority of participants could access the VI. Although the free access website allowed for participation during working hours, there may have been less motivation to perform work-related education in this setting. It was expected that the greatest amount of participation would occur during working hours on hospital computers. Clinical indications for vancomycin were not addressed by the VI, as its primary purpose was to improve knowledge once the decision to prescribe had been made.

Our study presented a model for adopting serious game concepts in combination with humor to develop and conduct Web-based health professional education in a light-hearted, interactive, and entertaining way. This model may be useful in settings where use of face-to-face education is limited by resources and geography. As the VI learning material was made available around the world, it showcased another significant benefit of open Web-based education resources. Health professionals and health care organizations with the same learning needs can reuse the material we have published rather than expending resources to develop similar material.

Conclusions

We demonstrated a novel Web-based e-learning tool that used humor and some game design principles to deliver health professional education on the commonly used antibiotic vancomycin. The VI was well accepted by users, and it was thus useful for delivering the intended health professional education. Future learning needs for different professional groups were identified through both the Web-based knowledge survey and VI data. This will allow tailoring of face-to-face education programs, in addition to subsequent versions of the VI that will embed robust gaming methodology. Further research will be aimed at measuring the effect on knowledge of the VI compared with a traditional email intervention and examining the impact of the VI on clinical vancomycin use.

Acknowledgments

The authors would like to thank Michael Boland from Digital League and team for production and programming of the VI and its website. We also thank the pharmacy and infectious diseases teams at ISLHD and SESLHD for user testing of the survey and VI. No financial support was reported. The filming and production of the VI was funded internally by Illawarra Shoalhaven Local Health District. Preliminary results from this study were presented in abstract form at Medicines Management 2015: the 41st Society of Hospital Pharmacists of Australia National Conference, Melbourne, Australia, December 3-6, 2015.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Vancomycin Web-based knowledge survey questions.

[PDF File (Adobe PDF File), 17KB - jmir_v19i3e93_app1.pdf]

Multimedia Appendix 2

Vancomycin Interactive questions.

[PDF File (Adobe PDF File), 18KB - jmir_v19i3e93_app2.pdf]

Multimedia Appendix 3

Qualitative survey feedback on the Vancomycin Interactive.

[PDF File (Adobe PDF File), 26KB - jmir_v19i3e93_app3.pdf]

References

1. Liu Q, Peng W, Zhang F, Hu R, Li Y, Yan W. The effectiveness of blended learning in health professions: systematic review and meta-analysis. J Med Internet Res 2016;18(1):e2 [FREE Full text] [doi: 10.2196/jmir.4807] [Medline: 26729058]
2. Cook DA, Levinson AJ, Garside S. Time and learning efficiency in internet-based learning: a systematic review and meta-analysis. Adv Health Sci Educ Theory Pract 2010 Dec;15(5):755-770. [doi: 10.1007/s10459-010-9231-x] [Medline: 20467807]
3. Cook DA, Levinson AJ, Garside S, Dupras DM, Erwin PJ, Montori VM. Internet-based learning in the health professions: a meta-analysis. JAMA 2008 Sep 10;300(10):1181-1196. [doi: 10.1001/jama.300.10.1181] [Medline: 18780847]
4. Bergeron B. Developing serious games. Hingham, MA: Charles River Media; 2006.
5. Thompson D. Designing serious video games for health behavior change: current status and future directions. J Diabetes Sci Technol 2012 Jul 01;6(4):807-811 [FREE Full text] [Medline: 22920806]

6. Lu AS, Baranowski T, Hong SL, Buday R, Thompson D, Beltran A, et al. The narrative impact of active video games on physical activity among children: a feasibility study. J Med Internet Res 2016 Oct 14;18(10):e272 [FREE Full text] [doi: 10.2196/jmir.6538] [Medline: 27742605]
7. Baid H, Lambert N. Enjoyable learning: the role of humour, games, and fun activities in nursing and midwifery education. Nurse Educ Today 2010 Aug;30(6):548-552. [doi: 10.1016/j.nedt.2009.11.007] [Medline: 20044181]
8. Ziegler JB. Humour in medical teaching. Med J Aust 1999;171(11-12):579-580. [Medline: 10721335]
9. Graafland M, Dankbaar M, Mert A, Lagro J, De WL, Schuit S, et al. How to systematically assess serious games applied to health care. JMIR Serious Games 2014;2(2):e11 [FREE Full text] [doi: 10.2196/games.3825] [Medline: 25654163]
10. Graafland M, Vollebergh MF, Lagarde SM, van HM, Bemelman WA, Schijven MP. A serious game can be a valid method to train clinical decision-making in surgery. World J Surg 2014 Dec;38(12):3056-3062. [doi: 10.1007/s00268-014-2743-4] [Medline: 25159118]
11. Wang R, DeMaria S, Goldberg A, Katz D. A systematic review of serious games in training health care professionals. Simul Healthc 2016 Feb;11(1):41-51. [doi: 10.1097/SIH.0000000000000118] [Medline: 26536340]
12. Castro-Sanchez E, Charani E, Moore L, Gharbi M, Holmes A. “On call: antibiotics”- development and evaluation of a serious antimicrobial prescribing game for hospital care. In: Schouten B, Fedtke S, Schijven M, Vosmeer M, Gekker A, editors. Games for Health 2014. Wiesbaden: Springer Vieweg; 2014:1-7.
13. Akl EA, Kairouz VF, Sackett KM, Erdley WS, Mustafa RA, Fiander M, et al. Educational games for health professionals. Cochrane Database Syst Rev 2013 Mar 28(3):CD006411. [doi: 10.1002/14651858.CD006411.pub4] [Medline: 23543543]
14. Cain J, Piascik P. Are serious games a good strategy for pharmacy education? Am J Pharm Educ 2015 May 25;79(4):47 [FREE Full text] [doi: 10.5688/ajpe79447] [Medline: 26089556]
15. Boeker M, Andel P, Vach W, Frankenschmidt A. Game-based e-learning is more effective than a conventional instructional method: a randomized controlled trial with third-year medical students. PLoS One 2013;8(12):e82328 [FREE Full text] [doi: 10.1371/journal.pone.0082328] [Medline: 24349257]
16. Johnsen HM, Fossum M, Vivekananda-Schmidt P, Fruhling A, Slettebø A. A serious game for teaching nursing students clinical reasoning and decision-making skills. Stud Health Technol Inform 2016;225:905-906. [Medline: 27332402]
17. Williams JG. Are online learning modules an effective way to deliver hand trauma management continuing medical education to emergency physicians? Plast Surg (Oakv) 2014;22(2):75-78 [FREE Full text] [Medline: 25114616]
18. Youngblood P, Harter PM, Srivastava S, Moffett S, Heinrichs WL, Dev P. Design, development, and evaluation of an online virtual emergency department for training trauma teams. Simul Healthc 2008;3(3):146-153. [doi: 10.1097/SIH.0b013e31817bedf7] [Medline: 19088658]
19. Rybak MJ, Lomaestro BM, Rotschafer JC, Moellering RC, Craig WA, Billeter M, et al. Therapeutic monitoring of vancomycin in adults summary of consensus recommendations from the American Society of Health-System Pharmacists, the Infectious Diseases Society of America, and the Society of Infectious Diseases Pharmacists. Pharmacotherapy 2009 Nov;29(11):1275-1279. [Medline: 19873687]
20. Antibiotic Expert Groups. Therapeutic guidelines: Antibiotic, Version 15. Melbourne: Therapeutic Guidelines Limited; 2014.
21. Duguid M, Cruickshank ME, editors. Antimicrobial Stewardship in Australian Hospitals. Sydney: Australian Commission on Safety and Quality in Health Care; 2010.
22. Swartling M, Gupta R, Dudas V, Guglielmo BJ. Short term impact of guidelines on vancomycin dosing and therapeutic drug monitoring. Int J Clin Pharm 2012 Apr;34(2):282-285. [doi: 10.1007/s11096-012-9614-6] [Medline: 22331444]
23. Lomaestro BM. Vancomycin dosing and monitoring 2 years after the guidelines. Expert Rev Anti Infect Ther 2011 Jun;9(6):657-667. [doi: 10.1586/eri.11.46] [Medline: 21692671]
24. Phillips CJ, Wisdom AJ, Eaton VS, Woodman RJ, McKinnon RA. The impact of a pilot continuing professional development module on hospital pharmacists' preparedness to provide contemporary advice on the clinical use of vancomycin. Springerplus 2016;5:331 [FREE Full text] [doi: 10.1186/s40064-016-1966-2] [Medline: 27064954]
25. Crowley RK, Fitzpatrick F, Solanki D, FitzGerald S, Humphreys H, Smyth EG. Vancomycin administration: the impact of multidisciplinary interventions. J Clin Pathol 2007 Oct;60(10):1155-1159 [FREE Full text] [doi: 10.1136/jcp.2006.044727] [Medline: 17293388]
26. Coleman LK, Wilson AS. Impact of nursing education on the proportion of appropriately drawn vancomycin trough concentrations. J Pharm Pract 2016 Oct;29(5):472-474. [doi: 10.1177/0897190014568381] [Medline: 25613053]
27. Cardile AP, Tan C, Lustik MB, Stratton AN, Madar CS, Elegino J, et al. Optimization of time to initial vancomycin target trough improves clinical outcomes. Springerplus 2015;4:364 [FREE Full text] [doi: 10.1186/s40064-015-1146-9] [Medline: 26203410]
28. NSW Health. Services and Facilities Annual Report 2009-2010. Sydney: NSW Health; 2010.
29. SHPA Publications Reference Group. Australian Injectable Drugs Handbook, 6th edition. Collingwood: The Society of Hospital Pharmacists of Australia; 2015.
30. Australian Commission on Safety and Quality in Health Care. National Safety and Quality Health Service Standards. Sydney: ACSQHC; 2011.

31. Bond SE, Crowther SP. Vancomycin. 2015. Vancomycin Interactive URL: http://www.vancomycin.com.au/ [accessed 2017-02-02] [WebCite Cache ID 6nzGDot6m]
32. Preacher KJ, Hayes AF. Asymptotic and resampling strategies for assessing and comparing indirect effects in multiple mediator models. Behav Res Methods 2008 Aug;40(3):879-891. [Medline: 18697684]
33. Graafland M, Schraagen JM, Schijven MP. Systematic review of serious games for medical education and surgical skills training. Br J Surg 2012 Oct;99(10):1322-1330. [doi: 10.1002/bjs.8819] [Medline: 22961509]
34. Bergeron BP. Learning & retention in adaptive serious games. Stud Health Technol Inform 2008;132:26-30. [Medline: 18391250]
35. Taekman JM, Shelley K. Virtual environments in healthcare: immersion, disruption, and flow. Int Anesthesiol Clin 2010;48(3):101-121. [doi: 10.1097/AIA.0b013e3181eace73] [Medline: 20616640]

Abbreviations

IBL: Internet-based learning
IQR: interquartile range
ISLHD: Illawarra Shoalhaven Local Health District
MRSA: methicillin-resistant Staphylococcus aureus
NSW: New South Wales
SESLHD: South Eastern Sydney Local Health District
TDM: therapeutic drug monitoring
VI: Vancomycin Interactive

Edited by G Eysenbach; submitted 13.11.16; peer-reviewed by E Castro-Sánchez, A Antoniades, M Schijven, R Buday; comments to author 05.01.17; revised version received 02.02.17; accepted 22.02.17; published 30.03.17.

Please cite as:
Bond SE, Crowther SP, Adhikari S, Chubaty AJ, Yu P, Borchard JP, Boutlis CS, Yeo WW, Miyakis S
Design and Implementation of a Novel Web-Based E-Learning Tool for Education of Health Professionals on the Antibiotic Vancomycin
J Med Internet Res 2017;19(3):e93
URL: http://www.jmir.org/2017/3/e93/
doi:10.2196/jmir.6971
PMID:28360025

Original Paper

InsuOnline, an Electronic Game for Medical Education on Insulin Therapy: A Randomized Controlled Trial With Primary Care Physicians

Leandro Arthur Diehl[1*], MD, MSc, PhD; Rodrigo Martins Souza[2*], BArch, MBA in IT; Pedro Alejandro Gordan[1*], MD, PhD, MEd; Roberto Zonato Esteves[3,4*], MD, PhD; Izabel Cristina Meister Coelho[4*], MD, PhD

[1]Internal Medicine Department, Health Sciences Center, Londrina State University (UEL), Londrina PR, Brazil

[2]Games Division, Oniria Software Industry, Londrina PR, Brazil

[3]Medicine Department, Maringá State University (UEM), Maringá PR, Brazil

[4]Pró-Ensino na Saúde, Pequeno Príncipe College, Pequeno Príncipe Complex, Curitiba PR, Brazil

[*]all authors contributed equally

Corresponding Author:
Leandro Arthur Diehl, MD, MSc, PhD
Internal Medicine Department
Health Sciences Center
Londrina State University (UEL)
Depto Clinica Medica/CCS
Av Robert Koch 60
Londrina PR, 86038350
Brazil
Phone: 55 4333712234
Fax: 55 4333712234
Email: drgaucho@yahoo.com

Abstract

Background: Most patients with diabetes mellitus (DM) are followed by primary care physicians, who often lack knowledge or confidence to prescribe insulin properly. This contributes to clinical inertia and poor glycemic control. Effectiveness of traditional continuing medical education (CME) to solve that is limited, so new approaches are required. Electronic games are a good option, as they can be very effective and easily disseminated.

Objective: The objective of our study was to assess applicability, user acceptance, and educational effectiveness of InsuOnline, an electronic serious game for medical education on insulin therapy for DM, compared with a traditional CME activity.

Methods: Primary care physicians (PCPs) from South of Brazil were invited by phone or email to participate in an unblinded randomized controlled trial and randomly allocated to play the game InsuOnline, installed as an app in their own computers, at the time of their choice, with minimal or no external guidance, or to participate in a traditional CME session, composed by onsite lectures and cases discussion. Both interventions had the same content and duration (~4 h). Applicability was assessed by the number of subjects who completed the assigned intervention in each group. Insulin-prescribing competence (factual knowledge, problem-solving skills, and attitudes) was self-assessed through a questionnaire applied before, immediately after, and 3 months after the interventions. Acceptance of the intervention (satisfaction and perceived importance for clinical practice) was also assessed immediately after and 3 months after the interventions, respectively.

Results: Subjects' characteristics were similar between groups (mean age 38, 51.4% [69/134] male). In the game group, 69 of 88 (78%) completed the intervention, compared with 65 of 73 (89%) in the control group, with no difference in applicability. Percentage of right answers in the competence subscale, which was 52% at the baseline in both groups, significantly improved immediately after both interventions to 92% in the game group and to 85% in control (P<.001). After 3 months, it remained significantly higher than that at the baseline in both groups (80% in game, and 76% in control; P<.001). Absolute increase in competence score was better with the game (40%) than with traditional CME (34%; P=.01). Insulin-related attitudes were improved both after the game (significant improvement in 4 of 9 items) and after control activity (3 of 9). Both interventions were very well accepted, with most subjects rating them as "fun or pleasant," "useful," and "practice-changing."

Conclusions: The game InsuOnline was applicable, very well accepted, and highly effective for medical education on insulin therapy. In view of its flexibility and easy dissemination, it is a valid option for large-scale CME, potentially helping to reduce clinical inertia and to improve quality of care for DM patients.

Trial Registration: Clinicaltrials.gov NCT001759953; https://clinicaltrials.gov/ct2/show/NCT01759953 (Archived by WebCite at http://www.webcitation.org/6oeHoTrBf)

(J Med Internet Res 2017;19(3):e72) doi:10.2196/jmir.6944

KEYWORDS

diabetes mellitus; insulin; video games; medical education; continuing medical education; educational technology

Introduction

Diabetes mellitus (DM) is a main public health problem of 21st century, affecting 422 million adults worldwide [1]. In most countries, most patients with DM are followed and treated by doctors who are not specialists in Endocrinology or diabetes, mainly primary care physicians (PCPs) [2]. However, only 24-56% of patients with DM present good glycemic control [3,4], defined as plasma levels of glycated hemoglobin A1c<7% [5].

Many factors may be implied in the low frequency of good metabolic control among patients with diabetes, but it is likely that one of the main reasons may be PCPs' lack of knowledge and confidence on several aspects of DM management [6], specially regarding insulin use [7].

This gap in PCPs' competence to treat diabetes with insulin contributes to the problem known as clinical inertia, "the failure to advance therapy when indicated," [8] resulting in underuse of insulin [9] and poor glycemic control. In fact, previous reports show that there is an average delay of about 3-5 years between the first demonstrations that a patient with DM requires insulin and the actual initiation of insulin therapy [10,11].

Continuing medical education (CME) on DM and insulin is often advocated as a solution to optimize the knowledge and the practice of PCPs [12]; however, traditional CME activities (such as lectures and group discussions) have small and short-lasting efficacy [13]. Thus, new educational methods are urgently required. Electronic games are a powerful tool for education [14], as they "create a tight marriage among content, game play, and valued ways of thinking and acting" [15]. Reasons for using games include the familiarity of most college students with this medium [16] and their favorable views on the matter [17,18]. Nevertheless, the 2 most compelling arguments sustaining the adoption of games for learning are (1) their potential educational effectiveness [15,16] and (2) their flexibility and easy dissemination [19].

Good learning games are usually built following the same rules that guide the design of effective learning activities, which include stimulus to players' intrinsic motivation, practice and repetition, effective feedback, arousal of positive feelings, intensity of the experience, and learner choice and involvement [20,21]. In the medical area, allowing students to practice their skills in a game may increase the safety for real patients [22].

Also, electronic games are learning resources much more flexible than traditional onsite educational activities, as they can be used in learners' own equipment, location, and time schedule, rendering them more scalable [23].

In the field of diabetes, some games for education of patients [24-28] and a few technology-based initiatives for education of health professionals [13,29-34] have been described, but to our knowledge, no game has been previously reported for education of health professionals on diabetes or insulin.

We have previously described the process of design and development of InsuOnline, the first electronic serious game intended for medical education on insulin therapy for diabetes [35] and a formative assessment of usability and playability of its prototype [36].

A previous pilot study with medical students and residents has shown that InsuOnline was applicable and well accepted as a tool for medical education on insulin therapy, in what we called "real-world conditions" for self-directed distance learning activities: as a stand-alone resource, played in learners' own equipment, in learners' own time schedule, with minimal or no external guidance [36,37].

The next step in InsuOnline validation process is assessing its educational effectiveness.

Thus, the objectives of this study were to assess applicability, user acceptance, and effectiveness of InsuOnline for education of PCPs on insulin therapy for diabetes, as compared with a traditional onsite CME session with the same content and duration.

Methods

Trial Design

We performed an open-label randomized controlled trial to assess the effectiveness of InsuOnline as a method for education of PCPs on insulin therapy for diabetes, as compared with a "traditional" onsite educational activity with the same content and same duration.

Eligibility Criteria

Subjects were eligible if they were medical doctors with an active register at a regional Council of Medicine in Brazil, were not specialists in Endocrinology or diabetes, were currently working at a public health care unit as a primary care physician (PCP), and were directly involved in the treatment of patients with diabetes in those facilities, with any degree of computer or gaming literacy.

Recruitment

Researchers contacted local public health authorities of some cities from the states of Paraná (Londrina and other cities from the 17th Health Regional, Curitiba and São José dos Pinhais) and Santa Catarina (Blumenau), who agreed to help with recruitment. Those authorities have enlisted PCPs in their area to attend a training session about diabetes, and then we contacted those PCPs via phone, mail, or email to invite them to participate in the study.

Those who agreed to participate were informed of the research procedures and were sent by email a link to an online informed consent form stored on Google Drive (Multimedia Appendix 1). Respondents were allocated to groups by simple randomization using an online random number generator [38] by the first author. Due to our expectancy of a higher attrition rate in the game group, allocation was made at a 2:1 ratio.

Settings and Locations Where Data Were Collected

After filling the informed consent form, participants were randomized and informed about the group they were allocated to and received an email with a link to the baseline questionnaire (stored on Google Drive). Also, participants allocated to control group were sent an email with information about the time and location of their scheduled onsite learning activity. A printed questionnaire was applied to those participants of the onsite learning activity immediately after its ending.

Participants allocated to game group received an email with instructions on how to download and install the game, after filling the baseline questionnaire, and then were contacted (weekly, if needed) to check if they have finished the game. After that, they were sent the link to another Web-based questionnaire.

After 3 months, a link to the third questionnaire was sent to participants of both groups.

If the participants did not answer the questionnaire after a few days, reminders were sent at 3-7–day intervals (initially by email, then by text message, and finally by phone). If they did not answer after up to 6 consecutive reminders, they were considered as loss of follow-up.

Interventions

Game

InsuOnline was developed by transdisciplinary collaboration from a team composed by clinical endocrinologists, game designers, and experts in medical education, with the help of other professionals (programmers, graphic designers, sound editors, etc) when needed at specific points of the process, using the methodology of iterative prototyping [39]. The game was designed to be a tool for education of PCPs on how to best use insulin in the treatment of patients with DM, in a primary health care setting. Educational objectives of the game are presented in Textbox 1. InsuOnline was developed as a 3D app, with simple commands (all player actions were made using a mouse); game engine was built using Unity, and visual elements (scenarios, characters, animations) were designed on Blender 3D creation suite. A detailed description of InsuOnline design and development process can be found in [35]. The version used in this study was a code release (version 1.6.1); screenshots (in Portuguese) are shown in Figures 1 and 2.

Participants allocated to the game group received a personal login and password to access a website from where the game could be downloaded and installed in players' own desktops or notebooks with Windows or MacOS, without any cost for study subjects.

During the game, players take on the role of a young doctor in a primary health care unit, whose mission is to improve glycemic control of an increasing-complexity series of 19 patients with diabetes, usually by means of initiating or adjusting insulin. Patients and situations presented in the game were designed to be similar to scenarios most commonly seen in a primary health care unit. The game gives immediate feedback after each player's decision, and the player's progression in the game is only possible if the most adequate decisions are made.

Several pedagogical elements were included in the game, aiming for the best educational effects [21,40]. These were based on the principles of adult learning and problem-based learning, including motivation, goal-orientation, relevancy-orientation, self-pacing, timely and appropriate feedback, contextualization, and practical (ie, hands-on) approach with active participation of the learner [41-43].

Recommendations about how to use insulin were selected from main clinical guidelines [5,44-49] and adapted to the reality of Brazilian public primary health care. Only the insulins commonly available in Brazilian public primary health care units (NPH and regular) were made available in the game. The software recorded players' progress in the game. Participants were allowed to play the game in their preferred time and place, in the number of sessions they wanted. The amount of time necessary to finish the game was about 4 h in a previous study [37]. Usability, playability, and preliminary educational effectiveness of previous versions of the game were previously assessed and reported [36,37]; development was "frozen" during the trial.

Players were sent minimal written instructions on how to download and play the game, and most did so with minimal or no external guidance. If subjects had any difficulties, researchers were available to give remote assistance by email, text message, or phone, at any moment.

Textbox 1. Educational objectives of InsuOnline.

- Recognize the goals of glycemic control in adults with DM
- Know when to start insulin in type 2 DM
- Know how to start insulin in type 2 DM (“bedtime” scheme)
- Know how to orientate proper insulin use (storing, injection technique, and devices)
- Know how to prevent, recognize, and treat hypoglycemia
- Know how to orientate and to interpret self-monitoring of blood glucose
- Know when and how to adjust insulin dosage
- Compare types of insulin (NPH, regular) and know when to use which
- Know when to intensify insulin therapy in type 2 DM
- Know how to intensify insulin therapy in type 2 DM (“basal-plus” scheme)
- Know how to prescribe intensive insulin therapy (“basal-bolus” scheme)
- Know how to recognize type 1 DM and how to start treatment
- Know how to recognize diabetic ketoacidosis and to start treatment
- Recognize the main barriers to insulin initiation and know how to address them
- Know how to orientate lifestyle changes and how to manage oral antidiabetic drugs in association with insulin
- Recognize the main factors that require insulin dosage adjustment and how to manage them

Figure 1. InsuOnline main characters: the young Dr Lucas (in the left) is the player’s avatar; the nurse Mariana (in the middle) gives useful hints; and the experienced Dr Braga (in the right) is a “mentor” who helps the player in critical moments.

Figure 2. The player's avatar, Dr Lucas, interacts with one of the patients in the game (dialogue in Portuguese).

Onsite Learning Activity

Participants allocated to the "control" group (ie, the active comparator) were scheduled to join a traditional-format onsite learning session, comprising a series of 3 interactive expositive lectures and 2 group case discussions (the first with 3, and the second with 4 cases, which were identical to corresponding InsuOnline levels). PowerPoint slides for the lectures and printed case presentations were previously prepared by the main author, and their content was exactly identical to the recommendations included in the game. To avoid potential biases, lectures were given by a clinical endocrinologist not linked to the research team, previously trained by the main author and familiarized with the didactic material. Also, the onsite learning activity was designed to have the same duration of the game (about 4 h).

Outcomes

Outcomes were self-assessed using written questionnaires which were applied to participants of both groups at 3 time points. The questionnaires were composed by different subscales at each time point: (1) at baseline, "competence" and "attitudes" subscales, plus demographic and professional data (for both groups); (2) immediately post intervention, the same "competence" and "attitudes" subscales for both groups, plus "game evaluation" subscale (for game group) or "onsite activity" subscale (for control group); (3) 3 months post intervention, "competence" and "attitudes" again, plus "importance for professional practice" subscale (for both groups).

The subscale of "competence" included 9 multiple-choice questions to assess factual knowledge about insulin and 11 clinical case vignettes with multiple-choice questions to assess problem-solving skills regarding insulin initiation and adjustment, which were created by our research team based on InsuOnline's educational objectives (Textbox 1) and content. The subscale of "attitudes" included 9 Likert-type scale questions to assess attitudes and beliefs regarding diabetes and insulin, which were adapted from previous surveys [7,50].

The subscale of "game evaluation" had 16 Likert-type scale questions to assess playability, user satisfaction, and perceived educational utility of the game, with 1 additional free text item for additional comments from the participants, which were freely adapted from a questionnaire used to assess another game [51].

The subscale of "onsite activity evaluation" was composed by 10 Likert-type scale questions regarding methodology, user satisfaction, and perceived educational utility of the activity, and 1 free text item to collect additional comments, which were based on the ones from the "game evaluation" subscale, but adapted to an onsite CME session.

Finally, the subscale of "importance for professional practice" included 5 Likert-type scale questions and 2 free text items to assess participants' opinion on the actual impact of the educational activity (game or onsite learning session) on their current professional practice, which were also created by authors.

All subscales were previously tested and refined in our pilot study with medical students and residents [36,37].

An additional subscale with 5 multiple-choice questions and 7 free text items was applied only to participants allocated to the game group who did not finish the game 60 days after receiving instructions for download or who did not respond to post-intervention questionnaires, to assess reasons why they were unable to play the game, and their perceptions about it.

All instruments used in this study are available in Multimedia Appendix 2.

Applicability of each intervention was defined as the proportion of participants allocated to each group that have received the intervention and did finish all tasks required in each intervention.

The primary outcome of this study was the score in the competence subscale in the immediately post-intervention self-assessment. Secondary outcomes were: score in competence subscale 3 months post-intervention; participants' attitudes regarding diabetes and insulin post-intervention; participants' acceptance and satisfaction with learning activities (game or onsite CME session); and applicability of the intervention (game or onsite CME session).

Statistical Analysis

Only subjects who filled the baseline and the immediate post-intervention questionnaires were included in the analysis. Data from questionnaires were downloaded from Google Drive online forms or typed from printed questionnaires to worksheets on Excel (Microsoft Corporation) and then analyzed on Epi-Info 7 (Centers for Disease Control, Atlanta, Georgia) for descriptive analysis. Demographic data were compared among groups at baseline using chi-square test for proportions or Student *t* or Kruskal-Wallis test for continuous variables, as suited. Correlation among baseline scores on the competence subscale and demographic baseline data was assessed using Pearson *r* or Student *t* test on SPSS 14.0 (SPSS Inc). Comparison of the scores on competence subscale by group and by time point was performed by repeated-measures analysis of variance (ANOVA), with the Tukey post hoc test when a significant difference was detected, and by simple comparison between 2 groups at each time point using Student *t* test, on the statistical software R (The R Foundation). Proportion of subjects with scores of 90% or more on competence subscale and the frequency of "agree" answers on the Likert-type scale items for assessing attitudes were compared by group and by time point using chi-square test (followed by post hoc pairwise comparisons using Bonferroni correction when significant). Absolute increase in competence score from baseline to immediately post intervention was estimated on Epi-Info 7. The effect size of the group variable on the competence score was estimated using Cohen *d* test on R software. Reliability (internal consistency) of the subscales used in this study was evaluated using Cronbach alpha. Comments made by subjects in the free text items were reviewed by content analysis. Statistical significance was defined as $P<.05$.

Sample Size

In order to detect a minimum standard deviation of 0.5 on the score (percentage of right answers) in competence subscale, with 80% of statistical power at 5% of significance level, we estimated a sample size of 128 subjects (64 in each group). Taking into account an expected attrition rate of 40% in the game group (about twice the 17% attrition rate observed in the previous pilot study with students and residents [36,37]) and of 20% in the control group, we decided to enroll at least 90 subjects in the game group and 77 subjects in the control group.

Ethics and Informed Consent

Participation was anonymous and voluntary, and all subjects filled an informed consent form (Multimedia Appendix 1), in accordance with Brazilian Health Ministry's regulations for research on human beings. The study protocol was previously approved by Londrina State University Research Ethics Committee (UEL, #051/2011 and #051/2012), and by all local public health authorities from the cities where the study was performed.

Results

User Statistics

A total of 257 primary care physicians were contacted and assessed for eligibility during the recruitment period (July 2014 in Curitiba and São José dos Pinhais, Paraná; August 2014 in the cities from the 17th Health Regional of Paraná State except Londrina; September to October 2014 in Londrina, Paraná; and October 2014 in Blumenau, Santa Catarina); 170 were randomized, and 134 were included in the final analysis (however, 4 subjects from the game group and 7 from the control did not answer the third questionnaire, 3 months post-intervention). The CONSORT diagram [52] for participant flow is shown in Figure 3.

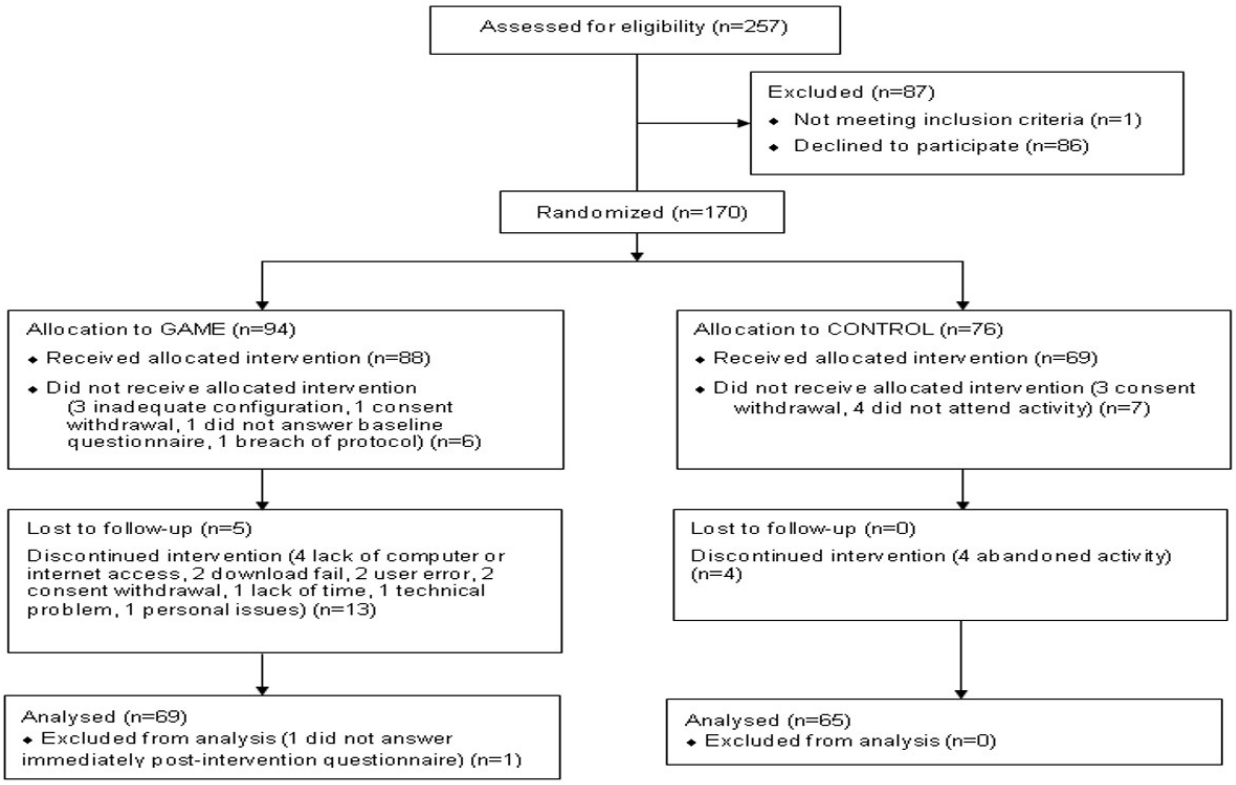
Baseline characteristics of subjects in both intervention groups were comparable, with no significant differences (Table 1).

Table 1. Characteristics of subjects allocated to each intervention group (game or control) at the baseline (n=134).

Characteristics of subjects		Game group (n=69)	Control group (n=65)
Gender, n (%)			
	Male	35 (50)	34 (52)
Age in years, mean (SD)		37.5 (11)	38.5 (12.7)
Location, n (%)			
	Curitiba	34 (49)	25 (38)
	17th HR[a]	20 (29)	19 (29)
	Londrina	15 (22)	18 (28)
	Blumenau	0	3 (5)
Years from graduation, median (range)		6.5 (0-41)	9.0 (0-41)
Years of experience in primary care, median (range)		4.0 (0-38)	5.0 (0-41)
Did residency, n (%)		27 (38)	25 (39)
Residency area, n (%)			
	Family and community health	8 (30)	9 (36)
	Internal medicine	8 (30)	4 (16)
	Ob&Gyn	4 (15)	2 (8)
	Surgery	4 (15)	4 (16)
	Pediatrics	2 (7)	3 (12)
	Other	1 (4)	3 (12)
Self-referred number of patients seen per month with: median (range)			
	type 2 DM	50 (4-300)	50 (1-500)
	type 2 DM on insulin	20 (0-150)	10 (1-120)
	type 1 DM	2 (0-100)	2 (0-80)

[a]17th HR: 17th Health Regional of Paraná state (group of 21 cities around Londrina, with the exception of Londrina).

Figure 3. CONSORT flow diagram.



Outcomes

Applicability

Applicability, defined as the percentage of subjects who received the intervention and finished it, was 78% (69 of 88 subjects) for the game, and 89% (65 of 73 subjects) for the control onsite learning activity, with no significant difference between the two interventions (*P*=.11, chi-square test). Among the subjects who did not finish the game, 13 responded to our contact; reasons alleged for not finishing InsuOnline were mostly nonrelated to the game, such as lack of access to the Internet, lack of computer, or lack of time for playing it. Only 2 subjects were unable to access or download the game and 1 was unable to visualize it correctly in her computer due to incompatibility with other software (antivirus).

Among the game group subjects who did finish playing the game, the mean time required for finishing it after download instructions were sent was 21 days (SD 12, range 2-59). No difference was found in game applicability or in the time for finishing the game by age group, location, or gender of the participants.

Evaluation of the Interventions by Participants

Both interventions were very well rated by participants, as revealed by their responses to the subscales of "game evaluation" (in the game group) and "onsite activity evaluation" (in the control group), applied immediately after the interventions (most relevant data are summarized in Table 2).

Comparing similar questions among both groups, significantly more subjects have found the game "fun" (*P*=.04, chi-square test), without any other difference. The onsite activity was deemed "pleasant" by 83% of control subjects.

About two-thirds (68%) of game subjects strongly agreed that they have "learned more about insulin with this game than they would learn from a lecture," and 68 of 69 (98%) would recommend the game for their friends.

In the free text field for comments, some subjects expressed their intense satisfaction with the game ("it is always better to learn by playing"), and even physicians who had never played an electronic game were able to play it and enjoyed the experience. Also, some criticisms were reported, related to game content (some repetitive or too extensive dialogues, soundtrack too loud) or to the software itself (some "freezes" when the player clicks too fast, incompatibility with some antivirus software).

Regarding the onsite activity, most comments were also highly positive ("excellent lecture," "very practical approach," "very dynamic and practical activity"). Criticisms were about the short duration of the activity ("too much information for one session only," "it would be better addressed in two consecutive days").

Table 2. Evaluation of interventions (game and onsite learning activity) by participants from each study group.

Items	Game group (n=69)		Control group (n=65)	
	Strongly agree, n (%)	Partially agree, n (%)	Strongly agree, n (%)	Partially agree, n (%)
The activity was fun	48 (70)[a]	18 (26)	32 (49)	21 (32)
The activity captured my attention all the time	43 (62)	25 (36)	47 (72)	15 (23)
I would join similar activities in the future	62 (90)	6 (9)	63 (97)	1 (2)
Patients presented in the activity were similar to the ones I see in my practice	43 (62)	22 (32)	46 (71)	16 (25)
The activity increased my knowledge about diabetes	66 (96)	3 (4)	60 (92)	5 (8)
The activity will influence the way I treat patients with diabetes	67 (97)	2 (3)	59 (91)	6 (9)

[a]*P*=.04 (χ^2 test), compared to control group.

Competence for Prescribing Insulin

In relation to the primary outcome of this study, the percentage of right answers in the competence subscale (factual nowledge + problem-solving skills related to insulin) was about 52% at baseline in both groups, and it was significantly increased post-intervention (*P*<.001, ANOVA for repeated measures) in both groups. In the post hoc Tukey test, this increase was significant in the comparison between the baseline and immediate post-intervention time points (*P*<.001), between immediate post-intervention and 3 months post-intervention (*P*<.001), and between 3 months and the baseline (*P*<.001), which means that scores increased significantly in the immediate post-intervention time compared with baseline, then decreased significantly after 3 months in comparison with immediate post-test, but even with that decrease, 3-month scores remained significantly higher than at baseline.

When the effects of both factors “time” and “group” were taken into account, a significant difference was found related to “time” (*P*<.001) and to the iteration “group per time” (*P*=.02), but no difference was found related to the “group” isolatedly (*P*=.27, ANOVA for repeated measures). A simple comparison of competence score between the 2 groups at each time point showed no significant difference at baseline (51.5 [SD 15.6%] in the game vs 51.7 [SD 16.6%] in the control group; *P*=.95) and at 3 months (79.8 [SD 14.4%] in the game vs 76.2 [SD 16.9%] in the control group; *P*=.20), but a significantly higher score in the game group than in the control group at the immediate post-intervention assessment (respectively: 91.7 [SD 8.9%] vs 85.5 [SD 13.4%]; *P*=.001, Student *t* test; Figure 4).

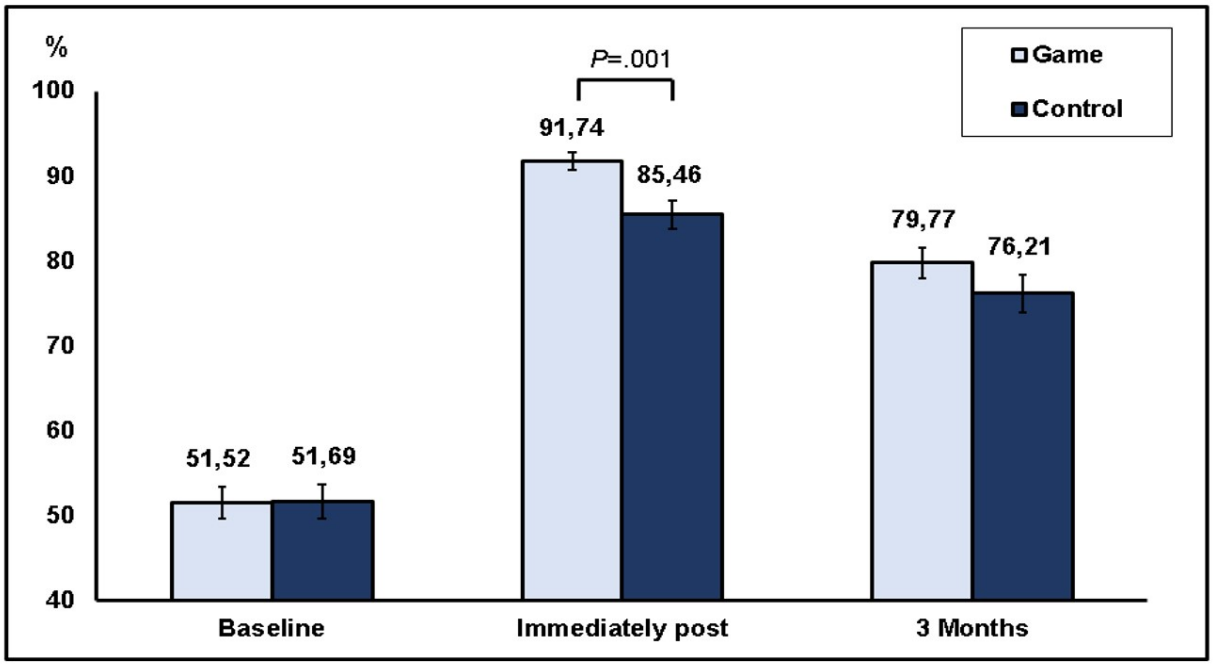
In fact, absolute increase in competence score (from baseline to immediately post intervention) was higher in the game group (40 [SD 15%]) than in the control group (34 [SD 15%]; *P*=.01); that difference was marginal at 3 months (28 [SD 14] in the game vs 23 [SD 17] points in the control group; *P*=.06).

Also, the frequency of subjects who achieved a 90% or higher score in competence subscale at immediate post-intervention assessment was higher in the game group (53 subjects, or 77%) than in the control group (35 subjects, or 54%; *P*<.001), with no difference at baseline (0 in game vs 3% in the control group) or 3 months post-intervention (32% in game vs 26% in the control group).

Cohen *d* size effect of the game group compared with the control group was 0.4583 (0.093-0.8327) in the immediate post-intervention time [53].

No differences were observed in competence score by gender, location, residency status, years of experience in primary care, or self-reported mean number of patients with diabetes seen per month; a significant correlation was only observed with participants’ age (Pearson *r*=−.3314 at the baseline; *P*<.001; and Pearson *r*=−.4616 at immediate post-intervention time; *P*<.001).

Figure 4. Score in competence subscale in both groups (game and control) in the 3 time points (baseline, immediately post-intervention, and 3 months post-intervention; mean and standard error).



Additional Outcomes

Significant improvement in diabetes and insulin-related attitudes from the baseline to post-intervention was observed in 4 of the 9 items in the game group, and in 3 of 9 items in the control. In post hoc analyses, most differences occurred between the baseline and immediate posttest, and a few between the baseline and 3 months. An additional statistical difference was observed between 3 months and immediate posttest in the control group in an item that assessed a personal opinion. Results for attitudes subscale are in Multimedia Appendix 3.

Three months after the interventions, about 80% of the subjects from both the groups stated that the activity had had real impact on their practice and that they knew better what to do when seeing a real patient with diabetes. About 62-75% of participants in both groups also said they were feeling more secure, and found it easier to help patients to improve their glycemic control. In the game group, significantly more subjects said that "it got easier" to orientate DM patients about their therapy, compared with control group (85% vs 66%; *P*=.03). Results from the subscale of "importance for professional practice" are presented in Multimedia Appendix 4.

Reliability of the subscales used in this study was estimated to be about 0.7-0.8 for most subscales (Cronbach alpha=.715 for competence subscale, .739 for game evaluation subscale, .649 for onsite activity evaluation, and .850 for importance for professional practice), except for attitudes subscale (Cronbach alpha=.323) [54].

No harms or adverse effects were reported.

Discussion

Principal Findings

Our results have shown that the game InsuOnline, used in real-world conditions (in players' own computers and in their own flexible time schedule) was highly effective for education of PCPs on insulin therapy for diabetes, compared with an onsite CME activity with the same content and same approximate duration. In fact, competence improvement was better with the game than with the onsite learning session. InsuOnline was also applicable, with about 80% of the subjects with the more varied degrees of computer or gaming literacy being able to finish it with little or no external help. Both interventions were very well-rated by participants, regarding engagement, realism, and perceived educational value. The good evaluation of the onsite CME session proves that InsuOnline was compared with a valid control activity (a gold-standard active comparator). Three months after both the interventions, doctors from both the groups said that they were feeling more secure and more prepared to help real patients with diabetes in their daily professional practice, with even better results in the game group.

Games and simulators are being increasingly used for education of health care professionals on various topics [55-57], but, to our knowledge, this is the first report on the educational effectiveness of an electronic game for education of medical doctors about insulin therapy for diabetes.

Games are a promising way to deliver CME, for many reasons. One reason is their educational adequacy: well-designed games incorporate all the main principles of Adult Education, such as individualized, self-pacing, contextualized learning, with active experimentation and appreciation of previous knowledge [15,16,40-43]. The inclusion of game elements adds

entertainment to the learning experience, increasing learners' motivation to practice and to learn, which renders the learning experience more enjoyable, more engaging, and potentially more effective [40,58]. Games can be designed to simulate an infinite number of medical problems, and are easily updated as needed [15,40]. Finally, games are more flexible than traditional CME, respecting learners' own time availability and learning pace, and are much more scalable [23]. So, we believe that well-designed and validated games will soon become the gold standard option for delivering large-scale CME.

Validation of games for health professional education is a recent and growing research area. Most available games for health professionals' education have not passed through an objective validation process, and many of the published studies on the field are biased [59]. So, there is a great need of good quality studies which can provide solid evidence to support the fast-growing science of games for health. Our randomized-controlled trial was designed to contribute to this body of evidence by following the guidelines for research on health games effectiveness as proposed by Kato, who suggested that studies in this area should attempt to apply the same scientific rigor typical of health sciences [60].

As there is a variety of proposed methods for validating games for health, we opted to use multiple methods for validation of InsuOnline. One of the simplest validation approaches was proposed by Olsen [61], and includes only 3 aspects: usability, playability, and educational effectiveness. The first two aspects were very well rated for InsuOnline, as previously reported [36], and our current results support the third one. However, the best method to validate a game that was designed for CME is the approach to validation of CME activities described by Moore [62], that proposes a continuum of 7 levels: (1) applicability (user participation), (2) user satisfaction, (3) gain of factual and procedural knowledge (learning), (4) gain of competence, (5) improvement of learner's performance, (6) improvement of patient health, and (7) improvement of community health. InsuOnline has already met the first 4 criteria. The following 3 steps (change in PCP's professional performance and improvement of their patients' glycemic control, and, hopefully, improvement of populational risk of DM complications) will be the focus of future research.

Attitudes and beliefs of PCPs in our sample were similar, at the baseline, to the ones of family doctors in United States [7] and in Arab-speaking countries in the Middle East [50], although the Brazilian doctors who participated in our study seem to be more often inclined to perceive the initiation of insulin therapy and the training of the patient with DM for using insulin as difficult and complicated tasks (78% and 54% of our subjects, respectively, compared with 66% and 42% in United States, and 48% and 25% in Middle East), which underscores the urgent need for effective educational interventions.

Reasons for the better results observed with the game than with the onsite learning activity, in this study, can be related to characteristics of the games that foster learning and attitude (and behavioral) changes [63]:

1. The inclusion of behavioral change procedures into the game, such as: goal setting, decision-making, problem solving, goals re-evaluation, social rewards; and

2. The use of a story (the game plot), and the inclusion of behavioral change concepts into that story. Players' identification and empathy for the story protagonist may contribute to change players' attitudes, reflecting the attitudinal change that occurs with the protagonist character in the game, in a process known as "modeling" [63].

Also, the higher emotional involvement of the learner with a game than with a traditional CME activity is another likely factor, as affection is a powerful determinant for changing attitudes and behavior [64]. Intrinsic motivation, a very effective learning catalyst, probably is one of the biggest advantages of games in relation to more traditional modalities of education [65]. Our game was designed to produce a high level of players' engagement, by careful disposition of different game elements (realism, identification with context and characters, increasing-difficulty challenges, humor, clear objectives, immediate and intense feedback for each players' action, progress monitoring, and rewards). With that, the majority of participants stated that the game was "fun" and that it captured their attention all the time.

Limitations

Some limitations of our study should be pointed out. First, blinding of the participants (and researchers) was not feasible due to the nature of the interventions, and that may have affected some of our results, especially the subscales of evaluation of the activities, as many subjects may have found it exciting or unusual to play an electronic game to learn about a medical topic. However, we think that our primary outcome (competence for prescribing insulin) was not compromised, as it was objectively assessed by standardized multiple-choice questions.

Second, the instruments used in this study were not previously validated, because they were developed by our team to assess the specific outcomes we addressed in this study. The use of customized instruments is strongly recommended by Moreno-Ger [66], who argues that generic questionnaires are usually not useful for assessing games that can be very different in their objectives, target audiences, and needs. At our favor, we can say that our instruments were extensively reviewed by our team of experts in endocrinology and medical education, which warrants their face and content validity, and our subscales have shown acceptable reliability, as measured by their Cronbach alpha in the range of .7 to .8, with an exception of the attitudes subscale, which had very poor internal consistency [54]. The authors decided to maintain this subscale because it had questions similar to the ones used in previous surveys [7,50], allowing their comparison with our results.

In third place, this study shows only intermediate outcomes (PCPs' knowledge, skills, attitudes, competence regarding insulin initiation, and adjustment for diabetes), but we don't know yet if the game will have any impact on actual health professionals' performance or on their patients' measures or outcomes ("hard" outcomes), which we expect to assess in future studies.

Finally, although PCPs' competence for prescribing insulin is essential, it is very likely that a simple intervention on the education of medical doctors may be insufficient to induce a significant change in the process of care delivered to patients with DM in primary care. A multifaceted intervention, aiming at several aspects of DM management, should probably be more effective to improve care and control of DM on a primary health care level, but it would demand conjugated efforts from government and society to improve issues such as availability of multiprofessional teams, availability of better insulins and oral drugs, availability of reagent strips for self-monitoring of blood glucose, better access to diabetes education, and others.

Generalizability

We think our results can be generalized for most primary care physicians worldwide, as our sample included medical doctors from different geographic areas, different age groups, different levels of specialization and experience, and any degree of computer or gaming literacy. Also, our game was tested in what we believe to be the "real-world conditions" for the use of an electronic game for distance CME: in players' own computers (with a wide range of possible equipment configurations), with very little or no external help from our team, and in their own spare time. Our results show that InsuOnline is a flexible, applicable and scalable option for large-scale CME on diabetes.

Conclusions

In this "real-world conditions" randomized controlled trial, InsuOnline, the first electronic serious game designed for medical education on insulin therapy for diabetes, was applicable, very well accepted, and highly effective for education, with even better results for improving primary care physicians' competence and attitudes related to diabetes and insulin than a gold-standard onsite CME activity with the same content and duration. For this reason, we believe InsuOnline is a valid tool for large-scale CME on DM, with advantages manifested in its easy Web dissemination, customizable content, and accordance with Adult Learning principles. We hope it can contribute to improve PCPs' performance and optimize the quality of care offered to patients with DM in primary care, eventually improving patients' glycemic control and reducing the risk of DM complications.

Acknowledgments

All authors helped with the writing and review of this manuscript.

These data (preliminary results of baseline and immediately post-intervention assessments only) were partly presented at the 75th American Diabetes Association Scientific Sessions, at Boston Convention Center, Boston, MA, June 5-9, 2015.

The authors acknowledge the support of public health authorities in all the cities where this study was performed. We also thank Dr Victor Hugo Ferreira for his invaluable help by giving the lectures and leading the case discussions with the subjects of control group. We thank our student collaborators (Tayná Rolim Galvão Pereira and Caio César Secci), who helped so much with recruitment, contact with study participants, and data collection. We are in debt with Nicholas Bender Haydu and Juliano Barbosa Alves for their invaluable support. Finally, our thanks to all the participants who kindly agreed to join this adventure with us and contributed a few hours of their busy time.

Conflicts of Interest

InsuOnline is a copyrighted game. Its design and development was entirely founded by personal resources from the authors LAD and PAG, and from Oniria Software Industry, who are copyright holders. All authors contribute to design and evaluation of the game.

Multimedia Appendix 1

Informed consent form.

[PDF File (Adobe PDF File), 320KB - jmir_v19i3e72_app1.pdf]

Multimedia Appendix 2

Instruments (questionnaires) used in this study.

[PDF File (Adobe PDF File), 322KB - jmir_v19i3e72_app2.pdf]

Multimedia Appendix 3

Results from the "attitudes" subscale.

[PDF File (Adobe PDF File), 369KB - jmir_v19i3e72_app3.pdf]

Multimedia Appendix 4

Results from the "importance for professional practice" subscale.

[PDF File (Adobe PDF File), 357KB - jmir_v19i3e72_app4.pdf]

Multimedia Appendix 5

CONSORT eHealth checklist V1.6.1.

[PDF File (Adobe PDF File), 871KB - jmir_v19i3e72_app5.pdf]

References

1. World Health Organization. Global report on diabetes. Geneva: World Health Organization; 2016.
2. DeWitt DE, Hirsch IB. Outpatient insulin therapy in type 1 and type 2 diabetes mellitus: scientific review. J Am Med Assoc 2003 May 7;289(17):2254-2264. [Medline: 12734137]
3. Mendes AB, Fittipaldi JA, Neves RC, Chacra AR, Moreira ED. Prevalence and correlates of inadequate glycaemic control: results from a nationwide survey in 6,671 adults with diabetes in Brazil. Acta Diabetol 2010 Jun;47(2):137-145 [FREE Full text] [doi: 10.1007/s00592-009-0138-z] [Medline: 19655083]
4. Ford ES, Li C, Little RR, Mokdad AH. Trends in A1C concentrations among U.S. adults with diagnosed diabetes from 1999 to 2004. Diabetes Care 2008 Jan;31(1):102-104. [doi: 10.2337/dc07-0565] [Medline: 17934146]
5. American Diabetes Association. Standards of medical care in diabetes--2014. Diabetes Care 2014 Jan;37(Suppl 1):S14-S80. [doi: 10.2337/dc14-S014] [Medline: 24357209]
6. Rubin DJ, Moshang J, Jabbour SA. Diabetes knowledge: are resident physicians and nurses adequately prepared to manage diabetes? Endocr Pract 2007;13(1):17-21. [doi: 10.4158/EP.13.1.17] [Medline: 17360296]
7. Hayes RP, Fitzgerald JT, Jacober SJ. Primary care physician beliefs about insulin initiation in patients with type 2 diabetes. Int J Clin Pract 2008 Jun;62(6):860-868 [FREE Full text] [doi: 10.1111/j.1742-1241.2008.01742.x] [Medline: 18393965]
8. Peyrot M, Rubin RR, Lauritzen T, Skovlund SE, Snoek FJ, Matthews DR, International DAWN Advisory Panel. Resistance to insulin therapy among patients and providers: results of the cross-national Diabetes Attitudes, Wishes, and Needs (DAWN) study. Diabetes Care 2005 Nov;28(11):2673-2679. [Medline: 16249538]
9. Riddle MC. The underuse of insulin therapy in North America. Diabetes Metab Res Rev 2002;18(Suppl 3):S42-S49. [doi: 10.1002/dmrr.277] [Medline: 12324985]
10. Nichols GA, Koo YH, Shah SN. Delay of insulin addition to oral combination therapy despite inadequate glycemic control: delay of insulin therapy. J Gen Intern Med 2007 Apr;22(4):453-458 [FREE Full text] [doi: 10.1007/s11606-007-0139-y] [Medline: 17372792]
11. Rubino A, McQuay LJ, Gough SC, Kvasz M, Tennis P. Delayed initiation of subcutaneous insulin therapy after failure of oral glucose-lowering agents in patients with Type 2 diabetes: a population-based analysis in the UK. Diabet Med 2007 Dec;24(12):1412-1418. [doi: 10.1111/j.1464-5491.2007.02279.x] [Medline: 18042083]
12. Peyrot M, Rubin RR, Khunti K. Addressing barriers to initiation of insulin in patients with type 2 diabetes. Prim Care Diabetes 2010 Apr;4(Suppl 1):S11-S18. [doi: 10.1016/S1751-9918(10)60004-6] [Medline: 20394886]
13. Sharp LK, Lipsky MS. Continuing medical education and attitudes of health care providers toward treating diabetes. J Contin Educ Health Prof 2002;22(2):103-112. [doi: 10.1002/chp.1340220206] [Medline: 12099119]
14. Johnson L, Adams S, Cummins M. NMC. Austin: The New Media Consortium; 2012. The NMC Horizon Report: Higher Education Edition URL: http://www.nmc.org/publications/horizon-report-2012-higher-ed-edition [accessed 2012-12-21] [WebCite Cache ID 6D4zxQo5K]
15. Gee JP. What Video Games Have to Teach Us About Learning and Literacy. New York: Palgrave Macmillan; 2004.
16. Prensky M. Digital game-based learning. New York: McGraw-Hill; 2001.
17. Kron FW, Gjerde CL, Sen A, Fetters MD. Medical student attitudes toward video games and related new media technologies in medical education. BMC Med Educ 2010;10:50 [FREE Full text] [doi: 10.1186/1472-6920-10-50] [Medline: 20576125]
18. Diehl LA, de Souza RM, Gordan PA, Esteves RZ, Coelho IC. Gaming habits and opinions of Brazilian medical school faculty and students: what´s next? Games Health J 2014 Apr;3(2):79-85. [doi: 10.1089/g4h.2013.0069] [Medline: 26196048]
19. Mayo MJ. Games for science and engineering education. Commun ACM 2007 Jul 01;50(7):30. [doi: 10.1145/1272516.1272536]
20. Murphy C. Goodgamesbydesign. 2011. Why games work and the science of learning URL: http://www.goodgamesbydesign.com/Files/WhyGamesWork_TheScienceOfLearning_CMurphy_2011.pdf [accessed 2012-12-21] [WebCite Cache ID 6D50AGPB3]
21. Yusoff A. Eprints.soton. Southampton: University of Southampton, Faculty of Engineering, Sciences and Mathematics, Doctoral Thesis; 2010. A conceptual framework for serious games and its validation URL: http://eprints.soton.ac.uk/171663/1/Thesis_017b.pdf [accessed 2016-10-28] [WebCite Cache ID 6IbUQ9E0d]
22. Ziv A, Small SD, Wolpe PR. Patient safety and simulation-based medical education. Med Teach 2000;22(5):489-495. [doi: 10.1080/01421590050110777] [Medline: 21271963]
23. Demirbilek M. The use of electronic games in distance learning as a tool for teaching-learning. In: Rogers PL, Berg GA, Boettcher JV, Howard C, Justice L, Schenk KD, editors. Encyclopedia of distance learning. Hershey: IGI Global; 2009:2209-2223.

24. Brown SJ, Lieberman DA, Germeny BA, Fan YC, Wilson DM, Pasta DJ. Educational video game for juvenile diabetes: results of a controlled trial. Med Inform (Lond) 1997;22(1):77-89. [Medline: 9183781]
25. Patterson N, Wolfenstein M, Millar S, Halverson R, Squire K. Wcer.wisc.edu. Madison: University of Wisconsin-Madison; 2011. Games and Simulations for Diabetes Education URL: http://wcer.wisc.edu/publications/abstract/wcer-working-paper-no.-2011-1 [accessed 2012-12-21] [WebCite Cache ID 6D538StfS]
26. DeShazo J, Harris L, Pratt W. Effective intervention or child's play? A review of video games for diabetes education. Diabetes Technol Ther 2010 Oct;12(10):815-822. [doi: 10.1089/dia.2010.0030] [Medline: 20807119]
27. Lehmann ED. Interactive educational simulators in diabetes care. Med Inform (Lond) 1997;22(1):47-76. [Medline: 9183780]
28. Thompson D, Baranowski T, Buday R. Conceptual model for the design of a serious video game promoting self-management among youth with type 1 diabetes. J Diabetes Sci Technol 2010 May 01;4(3):744-749 [FREE Full text] [doi: 10.1177/193229681000400331] [Medline: 20513343]
29. O'Connor PJ, Sperl-Hillen JM, Johnson PE, Rush WA, Asche SE, Dutta P, et al. Simulated physician learning intervention to improve safety and quality of diabetes care: a randomized trial. Diabetes Care 2009 Apr;32(4):585-590 [FREE Full text] [doi: 10.2337/dc08-0944] [Medline: 19171723]
30. Harris SB, Leiter LA, Webster-Bogaert S, Van DM, O'Neill C. Teleconferenced educational detailing: diabetes education for primary care physicians. J Contin Educ Health Prof 2005;25(2):87-97. [doi: 10.1002/chp.13] [Medline: 16078807]
31. Tatti P, Lehmann ED. Utility of the AIDA diabetes simulator as an interactive educational teaching tool for general practitioners (primary care physicians). Diabetes Technol Ther 2001;3(1):133-140. [doi: 10.1089/152091501750220118] [Medline: 11469705]
32. Miller D, Roberts P. Using computer simulation as an insulin prescribing virtual preceptorship. In: Diabetes. Alexandria: American Diabetes Association; 2011 Jul Presented at: 71st American Diabetes Association Scientific Sessions; June 24-28, 2011; San Diego p. A197.
33. Simation Medical. Learntotreat. Victoria: Simation Medical Ltd; 2012. Learn to Treat Diabetes URL: http://learntotreat.com [accessed 2012-12-21] [WebCite Cache ID 6D5AeMpBa]
34. American Diabetes Association, Therasim. Therasim. Durham: Therasim, Inc; 2010. Interactive Patient Simulation URL: http://therasim.net/ada01/login.php [accessed 2012-12-21] [WebCite Cache ID 6D5AhLo43]
35. Diehl LA, Souza RM, Alves JB, Gordan PA, Esteves RZ, Jorge ML, et al. InsuOnline, a serious game to teach insulin therapy to primary care physicians: design of the game and a randomized controlled trial for educational validation. JMIR Res Protoc 2013 Jan 21;2(1):e5 [FREE Full text] [doi: 10.2196/resprot.2431] [Medline: 23612462]
36. Diehl LA, de Souza RM, Gordan PA, Esteves RZ, Coelho ICM. User assessment of InsuOnLine, a game to fight clinical inertia in diabetes: a pilot study. Games Health J 2015 Oct;4(5):335-343. [doi: 10.1089/g4h.2014.0111] [Medline: 26287924]
37. Diehl LA, Gordan PA, Esteves RZ, Coelho IC. Effectiveness of a serious game for medical education on insulin therapy: a pilot study. Arch Endocrinol Metab 2015 Oct;59(5):470-473 [FREE Full text] [doi: 10.1590/2359-3997000000118] [Medline: 26537410]
38. Randomness and Integrity Services Ltd. Random. Dublin: Randomness and Integrity Services Ltd; 2016. What's this fuss about true randomness? URL: https://www.random.org/ [accessed 2016-10-27] [WebCite Cache ID 6la0TCBJ5]
39. Bates B. Game design. Boston: Thomson Course Technology PTR; 2004.
40. Aldrich C. The complete guide to simulations and serious games: how the most valuable content will be created in the age beyond Gutenberg to Google. San Francisco: Pfeiffer; 2009.
41. Collins J. Education techniques for lifelong learning: principles of adult learning. Radiographics 2004 Sep;24(5):1483-1489. [doi: 10.1148/rg.245045020] [Medline: 15371622]
42. Barrows H. Problem-based learning applied to medical education. Springfield: Southern Illinois University School of Medicine; 2000.
43. Knowles M. The modern practice of adult education: from pedagogy to andragogy. New York: Association Press; 1980.
44. American Diabetes Association. Practical Insulin: A Handbook for Prescribing Providers. Alexandria: American Diabetes Association; 2011.
45. Beaser RS. Joslin´s insulin deskbook: designing and initiating insulin treatment programs. Boston: Joslin Diabetes Center; 2008.
46. Brazilian Diabetes Society. Guidelines from the Brazilian Diabetes Society 2013-2014. São Paulo: AC Farmacêutica; 2014.
47. Inzucchi SE, Bergenstal RM, Buse JB, Diamant M, Ferrannini E, Nauck M, American Diabetes Association (ADA), European Association for the Study of Diabetes (EASD). Management of hyperglycemia in type 2 diabetes: a patient-centered approach: position statement of the American Diabetes Association (ADA) and the European Association for the Study of Diabetes (EASD). Diabetes Care 2012 Jun;35(6):1364-1379 [FREE Full text] [doi: 10.2337/dc12-0413] [Medline: 22517736]
48. Lerario AC, Chacra AR, Pimazoni-Netto A, Malerbi D, Gross JL, Oliveira JE, et al. Algorithm for the treatment of type 2 diabetes: a position statement of Brazilian Diabetes Society. Diabetol Metab Syndr 2010 Jun 08;2(1):35 [FREE Full text] [doi: 10.1186/1758-5996-2-35] [Medline: 20529311]
49. Nathan DM, Buse JB, Davidson MB, Ferrannini E, Holman RR, Sherwin R, American Diabetes Association, European Association for Study of Diabetes. Medical management of hyperglycemia in type 2 diabetes: a consensus algorithm for the initiation and adjustment of therapy: a consensus statement of the American Diabetes Association and the European

Association for the Study of Diabetes. Diabetes Care 2009 Jan;32(1):193-203 [FREE Full text] [doi: 10.2337/dc08-9025] [Medline: 18945920]
50. Lakkis NA, Maalouf GJ, Mahmassani DM, Hamadeh GN. Insulin therapy attitudes and beliefs of physicians in Middle Eastern Arab countries. Fam Pract 2013 Oct;30(5):560-567 [FREE Full text] [doi: 10.1093/fampra/cmt022] [Medline: 23729488]
51. Reichlin L, Mani N, McArthur K, Harris AM, Rajan N, Dacso CC. Assessing the acceptability and usability of an interactive serious game in aiding treatment decisions for patients with localized prostate cancer. J Med Internet Res 2011 Jan 12;13(1):e4 [FREE Full text] [doi: 10.2196/jmir.1519] [Medline: 21239374]
52. Schulz KF, Altman DG, Moher D. CONSORT 2010 statement: updated guidelines for reporting parallel group randomised trials. Br Med J 2010;340:c332 [FREE Full text] [Medline: 20332509]
53. Cohen J. Statistical power analysis for the behavioral sciences. Hillsdale: L Erlbaum Associates; 1988.
54. Bland JM, Altman DG. Cronbach's alpha. Br Med J 1997 Feb 22;314(7080):572 [FREE Full text] [Medline: 9055718]
55. Akl EA, Sackett KM, Erdley WS, Mustafa RA, Fiander M, Gabriel C, et al. Educational games for health professionals. Cochrane Database Syst Rev 2013;1:CD006411. [doi: 10.1002/14651858.CD006411.pub3] [Medline: 23440807]
56. Abdulmajed H, Park YS, Tekian A. Assessment of educational games for health professions: a systematic review of trends and outcomes. Med Teach 2015 Apr;37(Suppl 1):S27-S32. [doi: 10.3109/0142159X.2015.1006609] [Medline: 25803590]
57. Nicolaidou I, Antoniades A, Constantinou R, Marangos C, Kyriacou E, Bamidis P, et al. A virtual emergency telemedicine serious game in medical training: a quantitative, professional feedback-informed evaluation study. J Med Internet Res 2015;17(6):e150 [FREE Full text] [doi: 10.2196/jmir.3667] [Medline: 26084866]
58. Koster R. A theory of fun for game design. Sebastopol: O'Reilly Media; 2013.
59. Segev A, Rovner M, Appel DI, Abrams AW, Rotem M, Bloch Y. Possible biases of researchers´ attitudes toward video games: publication trends analysis of the medical literature. J Med Internet Res 2016 Jul 18;18(7):e196 [FREE Full text] [doi: 10.2196/jmir.5935] [Medline: 27430187]
60. Kato PM. Evaluating efficacy and validating games for health. Games Health J 2012 Feb;1(1):74-76. [doi: 10.1089/g4h.2012.1017] [Medline: 26196436]
61. Olsen T, Procci K, Bowers C. Serious games usability testing: how to ensure proper usability, playability, and effectiveness. In: Design, user experience, and usability: theory, methods, tools and practice : first international conference, DUXU 2011, held as part of HCI International 2011, Orlando, FL, USA, July 9-14, 2011 : proceedings. Berlin: Springer; 2011 Presented at: HCI International; July 9-14, 2011; Orlando p. 625-634.
62. Moore DE, Green JS, Gallis HA. Achieving desired results and improved outcomes: integrating planning and assessment throughout learning activities. J Contin Educ Health Prof 2009;29(1):1-15. [doi: 10.1002/chp.20001] [Medline: 19288562]
63. Baranowski T, Buday R, Thompson DI, Baranowski J. Playing for real: video games and stories for health-related behavior change. Am J Prev Med 2008 Jan;34(1):74-82 [FREE Full text] [doi: 10.1016/j.amepre.2007.09.027] [Medline: 18083454]
64. Olson JM, Zanna MP. Attitudes and attitude change. Annu Rev Psychol 1993 Jan;44(1):117-154. [doi: 10.1146/annurev.ps.44.020193.001001]
65. Malone TW, Lepper MR. Making learning fun: a taxonomy of intrinsic motivations for learning. In: Snow RE, Farr MJ, editors. Aptitude, learning, and instruction. Volume 3: conative and affective process analyses. Hillsdale: L. Erlbaum Associates; 1987.
66. Moreno-Ger P, Torrente J, Hsieh YG, Lester WT. Usability testing for serious games: making informed design decisions with user data. Advances in Human-Computer Interaction 2012;2012:1-13. [doi: 10.1155/2012/369637]

Abbreviations

A1c: glycated hemoglobin A1c
CME: continuing medical education
DM: diabetes mellitus
PCP: primary care physician
RCT: randomized controlled trial

Edited by G Eysenbach; submitted 04.11.16; peer-reviewed by E da Silva, A Segev; comments to author 08.12.16; revised version received 14.12.16; accepted 13.02.17; published 09.03.17.

Please cite as:
Diehl LA, Souza RM, Gordan PA, Esteves RZ, Coelho ICM
InsuOnline, an Electronic Game for Medical Education on Insulin Therapy: A Randomized Controlled Trial With Primary Care Physicians
J Med Internet Res 2017;19(3):e72
URL: http://www.jmir.org/2017/3/e72/
doi:10.2196/jmir.6944
PMID:28279950

Original Paper

Online Self-Tracking Groups to Increase Fruit and Vegetable Intake: A Small-Scale Study on Mechanisms of Group Effect on Behavior Change

Jingbo Meng[1], PhD; Wei Peng[2], PhD; Soo Yun Shin[1], MA; Minwoong Chung[1], MA

[1]Michigan State University, Department of Communication, East Lansing, MI, United States

[2]Michigan State University, Department of Media and Information, East Lansing, MI, United States

Corresponding Author:
Jingbo Meng, PhD
Michigan State University
Department of Communication
404 Wilson Road, CAS Building
East Lansing, MI, 48824
United States
Phone: 1 517 355 3480
Fax: 1 517 432 1192
Email: jingbome@msu.edu

Abstract

Background: Web-based interventions with a self-tracking component have been found to be effective in promoting adults' fruit and vegetable consumption. However, these interventions primarily focus on individual- rather than group-based self-tracking. The rise of social media technologies enables sharing and comparing self-tracking records in a group context. Therefore, we developed an online group-based self-tracking program to promote fruit and vegetable consumption.

Objective: This study aims to examine (1) the effectiveness of online group-based self-tracking on fruit and vegetable consumption and (2) characteristics of online self-tracking groups that make the group more effective in promoting fruit and vegetable consumption in early young adults.

Methods: During a 4-week Web-based experiment, 111 college students self-tracked their fruit and vegetable consumption either individually (ie, the control group) or in an online group characterized by a 2 (demographic similarity: demographically similar vs demographically diverse) × 2 (social modeling: incremental change vs ideal change) experimental design. Each online group consisted of one focal participant and three confederates as group members or peers, who had their demographics and fruit and vegetable consumption manipulated to create the four intervention groups. Self-reported fruit and vegetable consumption were assessed using the Food Frequency Questionnaire at baseline and after the 4-week experiment.

Results: Participants who self-tracked their fruit and vegetable consumption collectively with other group members consumed more fruits and vegetables than participants who self-tracked individually (P=.01). The results did not show significant main effects of demographic similarity (P=.32) or types of social modeling (P=.48) in making self-tracking groups more effective in promoting fruit and vegetable consumption. However, additional analyses revealed the main effect of performance discrepancy (ie, difference in fruit and vegetable consumption between a focal participant and his/her group members during the experiment), such that participants who had a low performance discrepancy from other group members had greater fruit and vegetable consumption than participants who had a high performance discrepancy from other group members (P=.002). A mediation test showed that low performance discrepancy led to greater downward contrast (b=–0.78, 95% CI –2.44 to –0.15), which in turn led to greater fruit and vegetable consumption.

Conclusions: Online self-tracking groups were more effective than self-tracking alone in promoting fruit and vegetable consumption for early young adults. Low performance discrepancy from other group members lead to downward contrast, which in turn increased participants' fruit and vegetable consumption over time. The study highlighted social comparison processes in online groups that allow for sharing personal health information. Lastly, given the small scale of this study, nonsignificant results with small effect sizes might be subject to bias.

(J Med Internet Res 2017;19(3):e63) doi:10.2196/jmir.6537

KEYWORDS

online support group; quantified self; fruit and vegetable consumption; social comparison; similarity; social modeling

Introduction

Extensive evidence suggests that fruit and vegetable consumption prevents obesity [1,2], reduces cardiovascular disease risk [3,4], and decreases the risk of certain cancers [5,6]. Although a growing body of literature has examined effective strategies to increase fruit and vegetable consumption in children and adolescents [7,8], young adults have been relatively understudied [9]. Early young adults aged 18 to 22 years are at an age of transitioning from parental supervision to independent living, which is an important stage of developing food patterns that will affect their future [10]. A recent report showed that adults aged 18 to 34 years consumed the least fruits and vegetables across all age groups, including adults, children, and adolescents, and their fruit and vegetable consumption had a significant decline from age groups younger than 18 years [11]. Therefore, it is worth studying effective intervention strategies to help young adults consume at least five servings of fruits and vegetables per day (ie, 5 A Day [12]).

Web-based interventions have been found to be effective in promoting fruit and vegetable consumption in adults [13], children, and adolescents [14]. However, those existing online interventions primarily focus on individual-based behavior change, missing the opportunity to leverage online groups. Approximately 62% of US adult Internet users have used health-related online groups to find experiences of others who have similar health interests [15]. The literature has suggested that social support and social influence are pathways through which such online groups may be effective for behavior change [16]. When similar people interact to increase their fruit and vegetable consumption, social support can reduce the uncertainty and costs of behavior change by providing information and companionship [17]. Social influence may also increase fruit and vegetable consumption through observational learning from behavioral models in online groups or complying with normative behavior emerged in such groups [18,19]. Moreover, approximately 70% of US adults track a health indicator, with diet and exercise routines being the most frequently monitored [18]. Given the potential of social influence in online groups and the prevalence of self-tracking behavior, online group-based programs that allow for sharing and comparing self-tracked diet may present a new intervention opportunity to increase fruit and vegetable consumption for young adults, who are characterized as a tech-savvy population [20].

This paper defines online groups as small social networks that contain three or more individuals with similar health conditions who interact via computer networks to achieve a common health goal [21,22]. Although previous research has shown effects of health-related online groups operating via social support exchanges [23,24] and social influence through social networks [25,26], few have examined the impact of groups and group characteristics on individuals' health outcomes during the dynamic group process [27]. In the context of increasing fruit and vegetable consumption, this study aims to directly test the effectiveness of a group-based online health program and examine group characteristics (ie, demographic similarity and types of social modeling) that make a group more successful in promoting individuals' fruit and vegetable consumption.

Individual- Versus Group-Based Self-Tracking

Self-tracking is a way that people monitor and record specific features of their lives [28]. It allows people to reflect on the specific aspects of their health and make improvements accordingly to achieve a health goal [29]. In health behavior change literature, self-tracking is often exchangeable with self-monitoring, which refers to an intervention technique that asks people to keep a record of a target behavior [30]. Meta-analyses of behavior change interventions found that self-monitoring explained a great amount of intervention success for physical activity and healthy eating [30,31].

Self-tracking has been around for a long time. For instance, people with diabetes keep a diary about their blood sugar via glucose meters, whereas overweight people journal about their exercise activities. What is new about self-tracking is the rise of social media technologies that enable sharing and comparing personal records [32]. Social media make two modes of self-tracking convenient in practice: individual- and group-based self-tracking. In individual-based self-tracking, people collect and view information on themselves to increase self-awareness and improve their health, whereas in group-based self-tracking, people collect and share personal information with one another who self-track the same health aspect collectively [28]. Although self-tracking has been documented to have a positive effect on enhancing individual health [32], little research has compared the effectiveness of different modes of self-tracking.

This study seeks to fill this void by comparing the effectiveness of self-tracking alone or in a group wherein other group members consistently increase their fruit and vegetable consumption over time. Other group members' fruit and vegetable consumption may exert a normative influence on the focal person. Particularly computer-mediated communication (CMC) can strengthen normative influence among members in groups [19]. According to the social identity model of deindividuation effects, interactions via computer networks accentuate group influence because of the relative anonymity that can actually facilitate individual members' submergence in the group [33] and thus make individual members more susceptible to the influence of normative behaviors [19].

The group members not only demonstrate normative behavior of fruit and vegetable consumption, but also serve as social models to the focal person in a group. Social cognitive theory argues that observing others performing a recommended behavior is a powerful means of learning [18]. Social modeling has been frequently used in the design of dietary interventions [9,34]. Recent reviews showed that social modeling has a robust and powerful influence on food intake and choice, such that participants ate more when their modeling companions ate more [35], and that participants tended to choose the same food selected by their modeling companions [36]. Social models not only facilitate learning how to do a behavior, but also show the

availability of resources in an environment for consumption (eg, access to healthy food).

Therefore, our first hypothesis is that individuals in a self-tracking group composed of members with increasing fruit and vegetable consumption over time will have greater fruit and vegetable consumption than individuals who self-track alone.

Characteristics of Self-Tracking Groups

Groups can vary in their demographic composition, behavioral patterns, and interaction processes, which all contribute to different group dynamics and performance in achieving specific goals [37]. What are the features of a self-tracking group that make the group more effective in promoting fruit and vegetable consumption? In terms of group composition, group members who serve as social models can be similar or diverse in their demographic characteristics to the focal person. Research on social cognitive theory points out that people are more likely to learn and enact behaviors modeled by someone with whom they identify [18]. In CMC, age, gender, and ethnicity serve as important cues about the identities of group members when other social cues are filtered out [38]. People may easily use the demographic cues to categorize themselves and others into social groups [39]. Especially at the beginning of a group's existence, when demographic characteristics are salient, individuals tend to use these characteristics in the identification process [40]. Moreover, according to the similarity-attraction principle applied in CMC [41], similarity in demographic characteristics may increase attraction to other group members, leading to a stronger group salience and susceptibility to group influence. Empirical studies have shown that demographic similarity among group members is positively associated with group performance and completing group tasks in virtual settings [42,43] in that individual group members are more motivated in high similarity groups than in diverse groups [44]. In online health groups, people consider group members as relevant references if they are demographically similar [26] and, thus, are more susceptible to their influence.

Therefore, our second hypothesis is that demographically similar online groups will have a greater effect on an individual's fruit and vegetable consumption than demographically diverse online groups.

Although social modeling has a robust positive effect on promoting individual health behavior, very little research examines the effect of modeling the progress of behavioral change over time. Previous interventions on food consumption have focused on inducing characteristics of social models, including familiarity between the focal person and the models, models' weight status (eg, slim, normal weight or obese), and live versus remote models [35]. Unlike those interventions in which models are simply superior in performing a recommended behavior, the models are highly likely to show different paces at which they move toward the health-related goal over time. In the context of increasing fruit and vegetable consumption, some models may make an incremental progress to reach the 5-A-Day goal (ie, incremental-change model), such as increasing one serving of fruits and vegetables per week for several weeks, whereas other models may make an ideal change to 5-A-Day immediately (ie, ideal-change model), such as increasing to eat five servings of fruits and vegetables within a week. Incremental-change models might be more effective than ideal-change models in that making a gradual progress seems more realistic and less daunting.

Particularly in online health-related groups, participants expect to meet others with similar health conditions and to compare their own situation to that of others [45]. Social comparison processes naturally occur in such online settings. The discrepancy in health behavior between group members may moderate the social comparison mechanism underlying group influence [46]. An individual's group members, who model the recommended behavior, have to be considered encouraging and achievable to be motivating. If the group members perform much better, the individual tends to stop comparing oneself to those group members and, thus, avoid emulating the group members' behaviors [47]. In online groups that promote fruit and vegetable consumption, performance discrepancy is the difference between the fruit and vegetable consumption of a focal person and of social models at different points of time during the process of behavioral change. On average, the focal person may have a lower performance discrepancy with the presence of incremental-change models than with the presence of ideal-change models. Incremental-change models could be more encouraging, whereas the ideal-change models might make the focal person frustrated in pursuing the health goal.

Therefore, our third hypothesis is that incremental-change models will have a greater effect on an individual's fruit and vegetable consumption than ideal-change models in online groups.

As discussed in the rationale for the preceding hypothesis, social comparisons could be one psychological mechanism underlying the hypothesized positive effect of incremental modeling on increasing individuals' fruit and vegetable consumption [48]. Two types of social comparisons are downward social comparisons that concern comparisons with others not doing better, and upward social comparisons that concern comparisons with others doing better [49]. Buunk and Ybema [50] argued that downward and upward comparison could be further segmented depending on whether individuals contrast themselves to or identify themselves with comparison targets. When comparing with someone worse off, individuals feel relieved and comfortable to be in a good position (ie, downward contrast), but feel anxious to be in the same situation in the future (ie, downward identification). When comparing with someone better off, individuals feel frustrated to be in a worse position (ie, upward contrast), but feel optimistic and hopeful to improve (ie, upward identification) [50]. The instances from the literature present preferences for downward contrast and upward identification because they are associated with better psychological well-being in general [45,51]. In the incremental modeling condition, social models show consistent small increases in their fruit and vegetable consumption and the focal person could perform better or worse than the models during the intervention. In the ideal modeling condition, social models quickly change their fruit and vegetable consumption and reach the ideal amount of fruit and vegetable consumption at the beginning and maintain it throughout the intervention. Then,

the focal person probably performs worse than the models for most time during the intervention, although the performance discrepancy is expected to reduce over time. It is possible that both downward and upward social comparisons take place in the incremental condition: the focal person might outperform social models sometimes and when the models outperform the focal person, the low performance discrepancy may facilitate the focal person's identification with the models. However, given the lack of previous studies on establishing the connections between types of social modeling and different social comparison processes, this study will explore the potential of this mechanism.

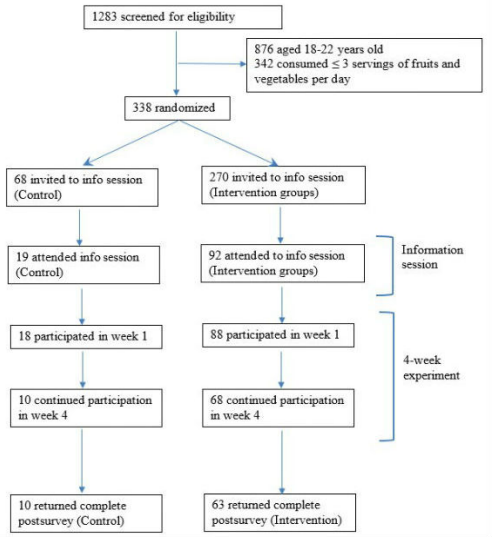
Therefore, our research question is will different social comparison processes mediate the positive effect of social modeling on an individual's fruit and vegetable consumption in online groups?

Methods

Overview

This 4-week Web-based experiment featured a 2 (demographic similarity: demographically similar vs demographically diverse) × 2 (social modeling: incremental change vs ideal change) between-subjects design plus one control group. Participants in the control group accessed the webpage and reported self-tracking of their fruit and vegetable consumption three times a week without the presence of group members. Participants in the intervention groups were asked to join a four-person group wherein every group member posted self-tracking of their fruit and vegetable consumption three times a week. Three group members in each group were confederates. Their demographics (ie, age, gender, and ethnicity) and the fruit and vegetable consumption as indicated in self-tracking posts were manipulated to create the four intervention groups. The intervention included several steps, including screening, information sessions, a 4-week Web-based experiment, and a postexperiment survey. The study design and participant flow are illustrated in Figure 1.

Figure 1. Flow diagram of the participants.



Participants

Participants were recruited via messages sent to undergraduate students through the registrar's office and two participant pools at Michigan State University, East Lansing, MI. The Human Research Protection Program at Michigan State University approved this study. Participants were first invited to take a 10-minute online survey for screening, in which they were instructed that the purpose of the survey was to assess their food consumption. Based on responses in the screening survey, the selection criteria for eligible participants for the 4-week intervention study were that the participant must (1) be age 18 to 22 years and white to achieve a clean manipulation of demographic similarity, (2) currently consume less than three servings of fruits and vegetables per day, and (3) have daily Internet access via a computer and/or a mobile phone. The screening yielded 338 eligible participants who were then sent an invitation via email to participate in the 4-week intervention. In the invitation email, they were instructed to physically attend an information session to participate in the 4-week intervention study. Then, 111 participants attended the information session and began the study. Each participant was rewarded with US $30 for completing the study.

Procedures

Once identified as eligible, the 338 participants were randomly assigned into control and experimental conditions because the online group page needed customization for each participant based on their assigned conditions. These online group pages were created before the information sessions because participants needed to complete a few important tasks on their group pages in the information sessions. Randomization was conducted using computer-generated random digits.

A week before the experiment, 111 eligible participants attended a 30-minute information session. At the beginning of the information session, participants provided written informed consent to join the study. Then, each participant received an invitation email to join an online group page ostensibly premade based on participants' demographic information (ie, age, gender, and ethnicity) collected from the screening survey. The starting numbers of fruit and vegetable servings of the three confederates were set up to be equal to, 0.5 above, and 0.5 below, with a mean equal to the current servings of the participant as indicated in the screening survey. After creating an online profile and joining the assigned group page, a participant was able to see the information of him- or herself and the three confederates. Participants were informed that group members' information was obtained from the screening survey. We conducted separate information sessions for the control group and intervention groups. A premade video illustrated steps to create online profiles and join an online group page, and instructed participants to post self-tracking messages to report their fruit and vegetable consumption for 3 days each week during the 4-week experiment.

Experiment Conditions

A healthy eating community was created on Ning.com, with by-invitation-only group pages. Although the participants in the intervention groups were made to believe they were in a

group with three other participants; in fact, the information of these three other participants were experimentally manipulated. Each group page consisted of three modules: (1) group goal (ie, every group member will eat five servings of fruits and vegetables per day at the end of the program) and group members' demographics, (2) self-track message wall that allowed group members to post their fruit and vegetable consumption, and (3) a bar graph illustrating weekly summaries of fruit and vegetable consumption for each group member. Figures 2 and 3 present examples of the group pages.

Figure 2. Visual example of the demographically diverse and ideal-change model intervention group page.

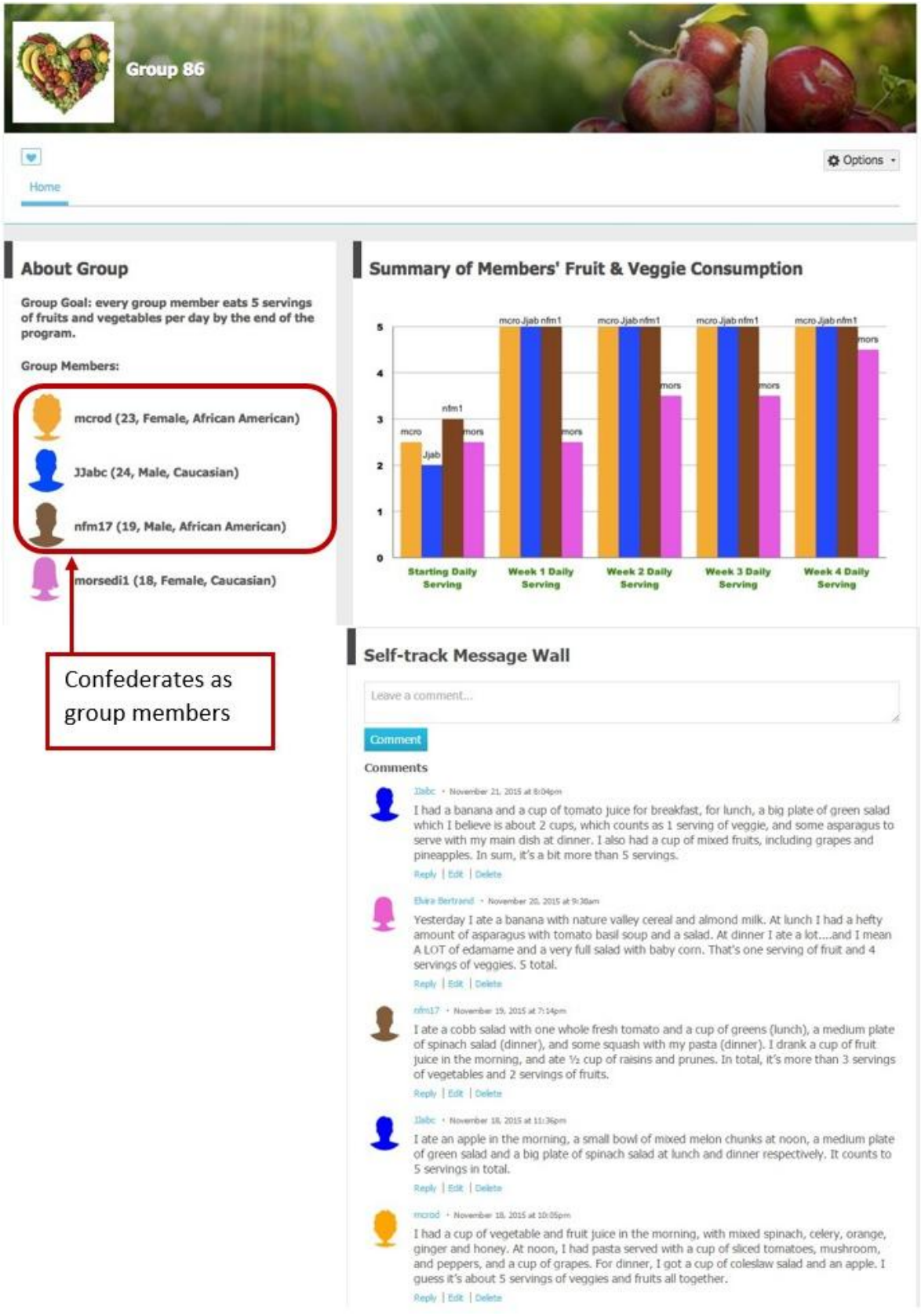


Figure 3. Visual example of the demographically similar and incremental-change model intervention group page.

Demographically Similar Versus Diverse Condition

Demographic similarity was manipulated by varying confederates' displayed ages, gender, and ethnicity. In the demographically similar condition, the three confederates' ages, gender, ethnicities were identical to the participant. In the demographically diverse condition, each confederate was different from the participant on two of the three demographic factors. For example, when a participant was an 18-year-old white female, the three confederates were a 26-year-old

African-American female (ie, different age and ethnicity), an 18-year-old African-American male (ie, different ethnicity and gender), and a 25-year-old white male (ie, different age and gender).

Incremental-Change Versus Ideal-Change Model Condition

The type of model was manipulated by varying the confederates' fruit and vegetable consumption over the 4-week experiment. In the incremental-change model condition, the three confederates posted prewritten messages that indicated each confederate's average fruit and vegetable consumption to be 3, 4, 4.5, and 5 servings in the 4 weeks. In the ideal-change model condition, the three confederates posted prewritten messages that indicated each confederate's average fruit and vegetable consumption to be five servings throughout the 4 weeks. Essentially, this manipulated the confederates' rate of progress toward the 5-A-Day goal: incremental progress versus instant progress to achieve the ideal goal. Because we could not control a participant's actual fruit and vegetable consumption during the 4-week experiment, the participant might sometimes outperform, underperform, or perform equally compared to others in the incremental-change condition. Textbox 1 presents examples of prewritten self-tracking messages.

Textbox 1. Message examples of self-tracking posts by confederates on group pages.

5 servings: "I had a cup of coleslaw salad, one baked sweet potato and a handful of baby carrots. I ate a big orange and a banana while I was in a meeting. It's about 5 servings of veggies and fruits all together."

4.5 servings: "For breakfast this morning I had a bowl of oatmeal with a half-cup of dried apricots mixed in and a glass of orange juice. At lunch I had a cup of chicken soup and a medium-sized salad, with an apple for an afternoon snack. I then ate two slices of pizza and about three carrot sticks for dinner. This should make three servings of fruits and one-and-a-half servings of vegetables. So, 4.5 total. Not bad today."

4 servings: "I ate a cobb salad with one fresh tomato and two cups of greens (lunch), and I drank a cup of fruit juice in the morning, and ate ½ cup of raisins and prunes. It's about 2 servings of veggies and 2 serving of fruit."

3 servings: "I had a ½ cantaloupe today. I love melons. For veggies, I had a cup of coleslaw at noon. It's about 1 serving of veggies and 2 servings of fruits."

Measurement

The primary outcome fruit and vegetable consumption was measured using the Food Frequency Questionnaire (FFQ) [52]. The FFQ asked about the amount and the frequency of consumption of 23 kinds of vegetables (eg, string beans, green beans, cooked greens such as spinach, mustard greens) and 12 kinds of fruits (eg, nectarines, plums, watermelon, honeydew) over the past 4 weeks. For the amount of consumption, participants could select the serving size (small, medium, large). The pictorial examples of different serving sizes were provided along with the questions. For the frequency of consumption, participants were asked to rate on an 8-point scale: (1=never or less than once per month, 8=2 or more per day).

Social comparison processes were measured by four different indexes each consisting of three items [49]. Participants were asked to rate their feelings on a 5-point scale (1=not at all, 5=very much) when reading other group members' self-tracking messages and weekly summaries of fruit and vegetable consumption. One index measured downward contrast (eg, "...I am happy that I am doing well myself," Cronbach alpha=.92; mean 3.14, SD 1.04), one index measured upward contrast (eg, "...I feel frustrated about my own situation," Cronbach alpha=.91; mean 2.12, SD 1.08), the other two indexes measured upward identification (eg, "...I realize that it is possible to improve," Cronbach alpha=.77; mean 3.70, SD 0.93) and downward identification (eg "...I fear that my future will be similar," Cronbach alpha=.83; mean 1.77, SD 0.86).

In addition to age and gender, participants' baseline fruit and vegetable consumption was measured in the screening survey. Self-reported height and weight were used to compute body mass index (BMI) scores. Enrollment in a campus meal plan (1=yes, 0=no) was measured because meal plan participants tended to have the recommended daily fruit and vegetable consumption [53].

Analytic Plan

First, a series of independent-sample *t* tests and chi-square tests were conducted to check the equivalence in terms of age, gender composition, BMI, etc, between participants who attended the information session and those who did not attend, between participants who began the study (ie, attended the information session) but dropped out later and those who completed the study, and between participants assigned into the different experimental conditions. The same methods of analyses were used to check the success of our manipulations.

To test our hypotheses, using intention-to-treat analysis [54], differences in fruit and vegetable consumption after the intervention were assessed with linear mixed-effects models, including conditions and time (ie, the time at which fruit and vegetable consumption was measured: baseline vs posttest) as independent variables, and a random intercept to account for missing data. Specifically, to test our first hypothesis that individuals in a self-tracking group composed of members with fruit and vegetable consumption that increases over time will have greater fruit and vegetable consumption than individuals who self-track alone, the study condition (control vs intervention condition) was entered together with time as the independent variables. To test the hypotheses that demographically similar online groups and incremental-change models will have a greater effect on an individual's fruit and vegetable consumption than demographically diverse online groups and ideal-change models, demographic similarity (demographically similar vs diverse) and types of social modeling (incremental change vs ideal change) together with time were entered as the independent variables. The time factor was introduced to include dropouts in the analysis and to have a grouping factor for mixed-effects

analysis. As random effects, we allowed each participant to have differing intercepts. The *P* values for testing statistical significance of each independent variable were obtained by *F* tests with the Kenward-Roger approximation [55]. These analyses were conducted using the R package "lme4" [56]. Then, our research question was tested in a parallel mediation model using bootstrapping [57]. Bootstrapping has been recommended for testing indirect effects on small to moderate samples because the Baron and Kenny method [58] is one of the least powerful approaches to testing mediation due to its reliance on a number of inferential procedures [59].

Results

Sample Characteristics

A total of 73 participants completed the 4-week intervention and the postsurvey. Among them, 49 were female, and the mean age was 19.86 (SD 1.65) years. The 73 participants consumed a mean 2.03 (SD 0.72) servings per day at baseline and a mean 3.05 (SD 2.01) servings per day after the 4-week experiment. To compare eligible participants who attended and who did not attend the information session. The results show that they did not differ on age (t_{333}=–1.61, *P*=.11), baseline fruit and vegetable consumption (t_{333}= 1.29, *P*=.19), or gender (χ^2_1=0.3, *P*=.66), although participants who attended the information session reported more interest on a 5-point scale (mean 2.36, SD 1.25) in participating in an online support group for healthy eating than those who did not attend the information session (mean 1.94, SD 1.13; t_{333}=2.49, *P*=.01).

For the 111 participants who attended the information session, 73 (65.8%) completed the 4-week intervention and the postsurvey. Analyses showed that participants who began the study but dropped out did not differ from those who completed the study on age (t_{109}=1.53, *P*=.13), gender (χ^2_1=0.1, *P*=.81), BMI (t_{109}=–0.69, *P*=.49), baseline fruit and vegetable consumption (t_{109}=0.24, *P*=.81), or interest in participating in an online support group for heathy eating (t_{109}=1.52, *P*=.13).

Table 1 presents descriptive statistics of the samples split by conditions, including baseline fruit and vegetable consumption, BMI, age, gender distribution, and completion rate. Participants across conditions did not significantly differ on any of these characteristics.

Table 1. Baseline fruit and vegetable consumption, BMI, age, gender, meal plan, and completion rate for the intervention conditions (N=73).

Sample characteristics	Condition					Total sample (N = 73)	$F_{4,68}$	χ^2_4	*P*
	Similar/incremental (n=17)	Diverse/incremental (n=14)	Similar/ideal (n=16)	Diverse/ideal (n=16)	Control (n=10)				
Baseline fruit and vegetable consumption, mean (SD)	1.92 (0.76)	2.25 (0.67)	2.17 (0.66)	2.07 (0.78)	1.94 (0.76)	2.04 (0.73)	0.93		.45
BMI (kg/m^2), mean (SD)	23.05 (3.32)	22.06 (2.26)	22.43 (2.15)	22.96 (3.60)	21.47 (2.91)	22.74 (2.98)	1.55		.20
Age (years), mean (SD)	19.76 (1.79)	19.17 (1.53)	19.75 (1.07)	20.27 (1.83)	20.40 (2.01)	19.86 (1.65)	1.06		.39
Gender (female), n (%)	13 (77)	11 (79)	10 (63)	9 (56)	6 (60)	49 (67)		2.75	.60
Meal plan (yes), n (%)	12 (71)	9 (64)	11 (69)	11 (69)	7 (70)	50 (69)		2.59	.63
Completion rate, n (%)[a]	24 (71)	21 (67)	23 (70)	24 (67)	19 (53)	111 (65.8)		3.53	.47

[a] Completion rate was calculated using the number of participants who completed the postintervention survey divided by the number of participants who attended the information session.

Manipulation Check

To check whether the manipulation for group demographic similarity was successful, in the postexperiment survey we asked participants to indicate which of the following statements best described their group members: "We are diverse in terms of age, gender, and ethnicity," (coded as 0) and "We are similar in terms of age, gender, and ethnicity" (coded as 1). A chi-square test showed a significant difference between demographically similar groups (mean 1.81, SD 0.40) and diverse groups (mean 1.00, SD 0.01; χ^2_1=11.2, *P*<.001).

To check whether the manipulation for incremental- versus ideal-change models was successful, in a pilot test with 10 participants we asked them to indicate whether other group members' fruit and vegetable consumption increased gradually (coded as 0) or stayed the same (coded as 1). A chi-square test showed that participants in the incremental modeling condition were significantly more likely to choose 0, whereas participants in the ideal modeling condition were more likely to choose 1 (χ^2_1=10.0, *P*=.008).

We also included two indirect manipulation checks for the incremental- versus ideal-change models: actual and perceived performance discrepancy. Actual performance discrepancy was calculated based on participants' wall posts on the group pages. It was the mean difference between manipulated serving sizes of social models and actual serving sizes reported by participants in their self-tracking messages across the first 3 weeks during the 4-week intervention. Perceived performance discrepancy was measured in the postexperiment survey by asking participants to rate whether other group members performed worse than (=1), similar to (=2), a little bit better (=3), or much better (=4) than them in achieving 5 A Day. The *t* tests revealed no significant difference in actual performance discrepancy

between the incremental-change (mean 0.82, SD 0.75) and ideal-change models (mean 1.18, SD 1.04; t_{71}=–1.48, P=.15), or in perceived performance discrepancy between the incremental-change (mean 3.35, SD 0.79) and ideal-change models (mean 3.29, SD 0.69; t_{71}=–0.34, P=.74). Therefore, incremental versus ideal modeling did not necessarily induce low versus high performance discrepancy.

Individual- Versus Group-Based Self-Tracking

The results showed that participants assigned to intervention groups (ie, group-based self-tracking) consumed more fruits and vegetables than those assigned to the control group (ie, individual-based self-tracking; beta=.20, $F_{1,88.25}$=6.83, P=.01). Participants who self-tracked their fruit and vegetable consumption collectively with other group members (mean 3.37, SD 2.01) consumed more fruits and vegetables than participants who self-tracked their fruit and vegetable consumption alone (mean 1.37, SD 1.44).

Group-Based Self-Tracking

Demographic similarity did not show a significant main effect on fruit and vegetable consumption (beta=–.06, $F_{1,76.92}$=1.02, P=.32). Types of social modeling did not show a significant effect either (beta=.04, $F_{1,75.88}$=0.50, P=.48). There was no interaction effect between demographic similarity and types of social modeling (P=.46). In other words, there was no difference between the demographically similar groups (mean 3.40, SD 2.55) and the demographically diverse groups (mean 3.19, SD 1.45) in fruit and vegetable consumption after the 4-week experiment. Similarly, no difference was found between the groups with incremental-change models (mean 3.58, SD 2.30) and the groups with ideal-change models (mean 3.18, SD 1.70) in terms of fruit and vegetable consumption. Moreover, the results showed a significant main effect of time (beta=.34, $F_{1,77.52}$=24.82, P<.001), such that participants' fruit and vegetable consumption increased over the course of the intervention. However, there were no significant interaction effects between time and the experimental conditions.

Post Hoc Analysis: Performance Discrepancy

Our indirect manipulation check demonstrated that the incremental-change and ideal-change models did not necessarily create the performance discrepancy as we expected; therefore, we explored the effect of performance discrepancy as the independent variable in the post hoc analysis. Using the mean of actual performance discrepancy (mean 0.96, SD 0.92), participants were split into low (coded as 0, n=32) versus high (coded as 1, n=31) actual performance discrepancy groups. A linear mixed-effects model with a random intercept to account for missing data was used to assess the difference in fruit and vegetable consumption after the intervention, including performance discrepancy, demographic similarity, and time as the independent variables. The results revealed a significant main effect of actual performance discrepancy on participants' fruit and vegetable consumption at the end of the 4-week experiment (beta=–.29, $F_{1,56}$=12.64, P=.002). Specifically, participants who had a low performance discrepancy from other group members had greater fruit and vegetable consumption (mean 4.11, SD 2.28) than participants who had a high performance discrepancy from other group members (mean 2.56, SD 1.32). Using a similar analytic method, an additional analysis was conducted by using actual values of performance discrepancy (ie, without dichotomizing performance discrepancy) as one of the independent variables. The result showed the same result, such that performance discrepancy had a negative effect on fruit and vegetable consumption (beta=–.35, P=.001). Given the significant main effect of actual performance discrepancy, a series of post hoc analyses were conducted to test if different social comparison processes mediated the effect of performance discrepancy on fruit and vegetable consumption. Indirect effects were tested in a parallel mediation model using bootstrapping [57]. The existence of an indirect effect was determined by the following two criteria. First, the total effect should be statistically significant, and the direct effect should become statistically nonsignificant. Second, the indirect effect should be statistically significant [57]. As presented in Table 2, the indirect effect through downward contrast was significant with a coefficient of –0.78 (95% CI –2.44 to –0.15). It showed that low performance discrepancy led to greater downward contrast, which in turn led to greater fruit and vegetable consumption. Therefore, downward contrast mediated the relationship between performance discrepancy and fruit and vegetable consumption.

Table 2. Tests of indirect effect of performance discrepancy on fruit and vegetable consumption through social comparisons.

Mediation tests	b (SE)	*P*
Performance discrepancy to social comparisons		
Performance discrepancy to upward contrast	0.64 (0.26)	.03
Performance discrepancy to downward contrast	–0.85 (0.25)	.001
Performance discrepancy to upward identification	–0.45 (0.24)	.09
Performance discrepancy to downward identification	0.25 (0.22)	.12
Social comparisons to fruit and vegetable consumption		
Upward contrast to fruit and vegetable consumption	0.58 (0.29)	.07
Downward contrast to fruit and vegetable consumption	0.84 (0.29)	.01
Upward identification to fruit and vegetable consumption	–0.37 (0.29)	.67
Downward identification to fruit and vegetable consumption	–0.20 (0.32)	.58
Total effect of performance discrepancy on fruit and vegetable consumption	–1.46 (0.49)	.01
Direct effect of performance discrepancy on fruit and vegetable consumption	–1.03 (0.53)	.08
Indirect effects of performance discrepancy on fruit and vegetable consumption through social comparisons		
Through upward contrast (bias-corrected 95% CI –0.12 to 1.53)	0.37 (0.37)	
Through downward contrast (bias-corrected 95% CI –2.44 to –0.15)	–0.78 (0.39)	
Through upward identification (bias-corrected 95% CI –0.21 to 0.47)	0.05 (0.15)	
Through downward identification (bias-corrected 95% CI –0.77 to 0.13)	0.00 (0.16)	

Discussion

This study built an online group-based program that allows for self-tracking dietary behavior for young adults. A significantly greater fruit and vegetable consumption was evident when participants self-tracked in groups wherein other group members showed consistent increases in fruit and vegetable consumption than when participants self-tracked alone. Therefore, our first hypothesis was supported. The finding suggests the effectiveness of using online self-tracking groups with social models for young adults. Although self-tracking helps to increase self-awareness of one's fruit and vegetable consumption [28], people need a larger context, such as a group environment, where they could observe and compare with others' performances to make more significant increases in fruit and vegetable consumption.

Theoretical Implications

The lack of differences between demographically similar and diverse groups in increasing fruit and vegetable consumption revealed that demographic similarity was not a determining group characteristic that made a self-tracking group more successful in behavior change. Social cognitive theory states that social learning and behavior change are more likely to take place when a person perceives a strong identification with a model [18]. In this study, young adults may not necessarily identify more with others sharing similar demographic characteristics because young adults live in a multiracial and multicultural society where demographic diversity is natural in everyday encounters [60,61]. Moreover, participants were put in groups with anonymous strangers. Although the manipulation of demographic similarity may foster some degree of identification with group members, groups consisting of already-known people may present a completely different level of identification, likely to be stronger than what we found in this study. Future research is encouraged to study different group compositions such that group members have preexisting connections, and to assess the effect of speed at which models progress toward a health goal in a naturalistic experiment.

An alternative explanation is that eating five servings of fruits and vegetables per day is a healthy behavior that applies to all [62], regardless of age, gender, or ethnicity. Therefore, demographic characteristics of the models may not play a critical role here. In addition, although participants in self-tracking groups collectively worked toward the 5-A-Day goal, they were not interdependent on one another to achieve the goal. Group literature that showed significant effects of demographic similarity on performance often involved higher levels of interdependency and interactions among group members [40]. For members who work coactively rather than interdependently, the effect of demographic similarity is probably minimal.

Incremental-change models did not show an advantage in increasing individuals' fruit and vegetable consumption when compared with ideal-change models. The finding suggested that the speed at which social models achieved the health goal did not affect individuals' behavior change. The manipulation check showed that, contrary to our hypothesis, incremental-change models did not necessarily lead to a lower actual or perceived performance discrepancy than ideal-change models. The participants' actual performance discrepancy was consistent with their perceived performance discrepancy, and that actual performance discrepancy had a significant effect on increasing individuals' fruit and vegetable consumption. Taken together, what makes an online group more effective is not the models'

progress of behavior change but the amount of difference from the models' performance in behavior change. Participants with a low performance discrepancy from models in the same online groups consumed more fruits and vegetables than participants with a high performance discrepancy from models. High performance discrepancy may make the focal person believe that he or she can never match or exceed perpetually superior group members [63]. High performance discrepancy then may weaken individuals' efficacy to increase their fruit and vegetable consumption persistently over time. Interestingly, it may also be that consistent upward or downward comparisons are less effective in behavior change because they present less motivating or challenging scenarios than an upward-downward combined comparison. Participants with a low performance discrepancy from models sometimes outperformed, underperformed, or equally performed compared to models (discrepancy ranged from –0.48 to 0.96 servings), whereas participants with a high performance discrepancy always underperformed compared to models (discrepancy ranged from 1.00 to 3.29 servings). Future research should test this speculation and increase our understanding of effective social comparison mechanisms for behavior change.

In addition, our post hoc analysis revealed that downward contrast mediated the effect of performance discrepancy on fruit and vegetable consumption outcomes. When comparing with group members who performed worse or similarly now and then, the focal person would feel good and relieved that they were doing well. However, upward identification was not a significant mediator. When comparing with group members who performed better, upward identification was the mechanism that could lead to positive behavior change but this was not observed in this study. Therefore, in online groups, in addition to matching group members with similar performance (ie, sometimes worse, sometimes a bit better) to facilitate downward contrast, it is critical to activate upward identification so that participants can emulate better models.

Practical Implications

This study provides several insights into the design of effective online groups to promote fruit and vegetable consumption for young adults. With the rapid growth of various self-tracking technologies [29], virtual connections among self-trackers allow for sharing of personal health information. Collective self-tracking in a group is more effective than self-tracking alone, when group members show positive progress toward the health goal. Existing self-tracking mobile apps and online communities (eg, MyFitnessPal, FatSecret) may leverage this insight to virtually connect self-trackers into small groups. For people who need to increase their fruit and vegetable consumption, such as patients with diabetic or cardiovascular diseases, health care providers may want to prescribe beyond self-tracking practice and encourage them to get connected with other self-trackers via online or offline support groups or via mobile networks.

Moreover, previous studies have found that people in online health social networks tend to connect with others with similar demographic backgrounds and similar progress toward a shared health goal [26]. This study found that it was the similarity in health progress rather than similarity in demographic background that made online groups more effective in promoting fruit and vegetable consumption behavior. Therefore, in creating online groups or online social networks for increasing people's fruit and vegetable consumption, algorithms may be developed to recommend teaming up with others who have similar health progress toward the goal.

Limitations

There were a few limitations in this study, beyond the reasonable attrition. First, for the 338 participants who were eligible and invited to participate in the 4-week intervention, only 111 (32.8%) participants attended the information session of the study. The reason for the loss of eligible participants in the information session was that student participants recruited from the registrar's office and participant pools typically had choices to participate in other studies. Student participants tended to choose studies that cost minimal effort, such as an online survey. Our study involved physical attendance to the information session and a 4-week intervention, which was less attractive to student participants. However, our analysis showed that there were not any significant differences between eligible participants who attended and who did not attend the information session in terms of age, gender, and baseline fruit and vegetable consumption.

Second, in post hoc power analyses, we had a 74% observed power to detect a significant difference between the control versus intervention groups with Cohen d=0.63, an 80% power to detect a significant main effect of performance discrepancy on postintervention fruit and vegetable consumption (post hoc analysis) with Cohen d=0.61. However, we only had a 13% power to detect a significant main effect of social modeling on postintervention fruit and vegetable consumption with Cohen d=0.12, and a 10% power for demographic similarity with Cohen's d=0.08. Therefore, the small sample size might have contributed to the null findings for our hypotheses that demographically similar online groups and incremental-change models will have a greater effect on an individual's fruit and vegetable consumption than demographically diverse online groups and ideal-change models, with very small effect sizes. Similarly, group equivalence tests in sample characteristics may be biased by the small sample size in each condition. Readers are cautioned to interpret these results with the small sample size and lack of power in mind.

Third, this study was a short-term behavior change (ie, 4 weeks). We have no knowledge whether fruit and vegetable consumption would continue to increase or remain at five servings. Future studies should examine the potential long-term effect of online groups for changing fruit and vegetable intake. Lastly, the self-tracking method used in this intervention was traditional self-reported food journaling, which imposes a high burden on participants [64]. Recent automated food recognition technologies rely on wearable cameras or phones to capture food photos and leverage computer vision techniques to analyze food ingredients [65,66]. Another alternative method is to use an in-the-moment photo as a lightweight food journal to reduce user effort [67]. Future research should consider these more advanced self-tracking methods to reduce participation attrition

as well as increase the accuracy and usefulness of self-tracked information.

Conclusion

This study is one of the first attempts to test the effects of online self-tracking groups in increasing fruit and vegetable consumption for young adults. The 4-week experiment showed that online self-tracking groups with models consistently increasing their fruit and vegetable consumption were more effective than self-tracking alone in promoting fruit and vegetable consumption. We also found that low performance discrepancy from models would lead to downward contrast, which in turn increases participants' fruit and vegetable consumption over time. The study highlighted social comparison processes in online groups that allow for sharing personal health information.

Acknowledgments

This project received financial support from the College of Communication Arts and Sciences at Michigan State University, Lansing, MI, USA.

Conflicts of Interest

None declared.

References

1. Casazza K, Fontaine KR, Astrup A, Birch LL, Brown AW, Bohan Brown MM, et al. Myths, presumptions, and facts about obesity. N Engl J Med 2013 Jan 31;368(5):446-454 [FREE Full text] [doi: 10.1056/NEJMsa1208051] [Medline: 23363498]
2. Rolls BJ, Ello-Martin JA, Tohill BC. What can intervention studies tell us about the relationship between fruit and vegetable consumption and weight management? Nutr Rev 2004 Jan;62(1):1-17. [Medline: 14995052]
3. Bazzano LA, He J, Ogden LG, Loria CM, Vupputuri S, Myers L, et al. Fruit and vegetable intake and risk of cardiovascular disease in US adults: the first National Health and Nutrition Examination Survey Epidemiologic Follow-up Study. Am J Clin Nutr 2002 Jul;76(1):93-99 [FREE Full text] [Medline: 12081821]
4. Dauchet L, Amouyel P, Hercberg S, Dallongeville J. Fruit and vegetable consumption and risk of coronary heart disease: a meta-analysis of cohort studies. J Nutr 2006 Oct;136(10):2588-2593 [FREE Full text] [Medline: 16988131]
5. Steinmetz KA, Potter JD. Vegetables, fruit, and cancer prevention: a review. J Am Diet Assoc 1996 Oct;96(10):1027-1039. [doi: 10.1016/S0002-8223(96)00273-8] [Medline: 8841165]
6. Turati F, Rossi M, Pelucchi C, Levi F, La VC. Fruit and vegetables and cancer risk: a review of southern European studies. Br J Nutr 2015 Apr;113 Suppl 2:S102-S110. [doi: 10.1017/S0007114515000148] [Medline: 26148912]
7. Mathias KC, Rolls BJ, Birch LL, Kral TV, Hanna EL, Davey A, et al. Serving larger portions of fruits and vegetables together at dinner promotes intake of both foods among young children. J Acad Nutr Diet 2012 Feb;112(2):266-270 [FREE Full text] [Medline: 22741168]
8. Thompson D, Bhatt R, Vazquez I, Cullen KW, Baranowski J, Baranowski T, et al. Creating action plans in a serious video game increases and maintains child fruit-vegetable intake: a randomized controlled trial. Int J Behav Nutr Phys Act 2015;12:39 [FREE Full text] [doi: 10.1186/s12966-015-0199-z] [Medline: 25890060]
9. Peng W. Design and evaluation of a computer game to promote a healthy diet for young adults. Health Commun 2009 Mar;24(2):115-127. [doi: 10.1080/10410230802676490] [Medline: 19280455]
10. Richards A, Kattelmann KK, Ren C. Motivating 18- to 24-year-olds to increase their fruit and vegetable consumption. J Am Diet Assoc 2006 Sep;106(9):1405-1411. [doi: 10.1016/j.jada.2006.06.005] [Medline: 16963345]
11. Produce for Better Health Foundation. State of the Plate: 2015 Study on America's Consumption of Fruit & Vegetables. 2015. URL: http://www.pbhfoundation.org/pdfs/about/res/pbh_res/State_of_the_Plate_2015_WEB_Bookmarked.pdf [WebCite Cache ID 6jt0sqQ8d]
12. Casagrande SS, Wang Y, Anderson C, Gary TL. Have Americans increased their fruit and vegetable intake? The trends between 1988 and 2002. Am J Prev Med 2007 Apr;32(4):257-263. [doi: 10.1016/j.amepre.2006.12.002] [Medline: 17383556]
13. Buller DB, Woodall WG, Zimmerman DE, Slater MD, Heimendinger J, Waters E, et al. Randomized trial on the 5 a day, the Rio Grande Way Website, a web-based program to improve fruit and vegetable consumption in rural communities. J Health Commun 2008;13(3):230-249 [FREE Full text] [doi: 10.1080/10810730801985285] [Medline: 18569356]
14. Hamel LM, Robbins LB. Computer- and web-based interventions to promote healthy eating among children and adolescents: a systematic review. J Adv Nurs 2013 Jan;69(1):16-30. [doi: 10.1111/j.1365-2648.2012.06086.x] [Medline: 22757605]
15. Fox S. The Social Life of Health Information. Washington, DC: Pew Internet & American Life Project; 2011 May 12. URL: http://www.pewinternet.org/files/old-media//Files/Reports/2011/PIP_Social_Life_of_Health_Info.pdf [WebCite Cache ID 6jYx4fFRX]
16. Berkman LF, Glass T, Brissette I, Seeman TE. From social integration to health: Durkheim in the new millennium. Soc Sci Med 2000 Sep;51(6):843-857. [Medline: 10972429]

17. Cavallo DN, Brown JD, Tate DF, DeVellis RF, Zimmer C, Ammerman AS. The role of companionship, esteem, and informational support in explaining physical activity among young women in an online social network intervention. J Behav Med 2014 Oct;37(5):955-966 [FREE Full text] [doi: 10.1007/s10865-013-9534-5] [Medline: 24081454]
18. Bandura A. Social Foundations of Thought and Action: A Social Cognitive Theory. Upper Saddle River, NJ: Prentice Hall; 1986.
19. Postmes T, Spears R, Lea M. Breaching or building social boundaries?: SIDE-effects of computer-mediated communication. Commun Res 1998 Dec 01;25(6):689-715. [doi: 10.1177/009365098025006006]
20. Duggan M, Ellison NB, Lampe C, Lenhart A, Madden M. Pew Research Center. 2015 Jan 09. Social media update 2014 URL: http://www.pewinternet.org/2015/01/09/social-media-update-2014/ [WebCite Cache ID 6n7dRRFbd]
21. Centola D. Social media and the science of health behavior. Circulation 2013 May 28;127(21):2135-2144 [FREE Full text] [doi: 10.1161/CIRCULATIONAHA.112.101816] [Medline: 23716382]
22. Maloney-Krichmar D, Preece J. A multilevel analysis of sociability, usability, and community dynamics in an online health community. ACM Trans Comput-Hum Interact 2005 Jun 01;12(2):201-232. [doi: 10.1145/1067860.1067864]
23. Meng J, Martinez L, Holmstrom A, Chung M, Cox J. Research on social networking sites and social support from 2004 to 2015: a narrative review and directions for future research. Cyberpsychol Behav Soc Netw 2016 Dec 21:e1. [doi: 10.1089/cyber.2016.0325] [Medline: 28002686]
24. Meng J, Chung M, Cox J. Linking network structure to support messages: effects of brokerage and closure on received social support. J Commun 2016 Dec 19;66(6):982-1006. [doi: 10.1111/jcom.12268]
25. Centola D. The spread of behavior in an online social network experiment. Science 2010 Sep 3;329(5996):1194-1197 [FREE Full text] [doi: 10.1126/science.1185231] [Medline: 20813952]
26. Meng J. Your health buddies matter: preferential selection and social influence on weight management in an online health social network. Health Commun 2016 Dec;31(12):1460-1471. [doi: 10.1080/10410236.2015.1079760] [Medline: 27055008]
27. Kreps GL, Neuhauser L. New directions in eHealth communication: opportunities and challenges. Patient Educ Couns 2010 Mar;78(3):329-336. [doi: 10.1016/j.pec.2010.01.013] [Medline: 20202779]
28. Lupton D. Self-tracking cultures: towards a sociology of personal informatics. In: Australian Computer-Human Interaction Conference on Designing Futures: The Future of Design. 2014 Presented at: 26th Australian Computer-Human Interaction Conference on Designing Futures: The Future of Design; Dec 2-5, 2014; Sydney, Australia p. 2-5. [doi: 10.1145/2686612.2686623]
29. Swan M. Emerging patient-driven health care models: an examination of health social networks, consumer personalized medicine and quantified self-tracking. Int J Environ Res Public Health 2009 Feb;6(2):492-525 [FREE Full text] [doi: 10.3390/ijerph6020492] [Medline: 19440396]
30. Dusseldorp E, van Genugten L, van Buuren S, Verheijden MW, van Empelen P. Combinations of techniques that effectively change health behavior: evidence from Meta-CART analysis. Health Psychol 2014 Dec;33(12):1530-1540. [doi: 10.1037/hea0000018] [Medline: 24274802]
31. Michie S, Abraham C, Whittington C, McAteer J, Gupta S. Effective techniques in healthy eating and physical activity interventions: a meta-regression. Health Psychol 2009 Nov;28(6):690-701. [doi: 10.1037/a0016136] [Medline: 19916637]
32. Wiederhold BK. Self-tracking: better medicine through pattern recognition. Cyberpsychol Behav Soc Netw 2012 May;15(5):235-236. [doi: 10.1089/cyber.2012.1545] [Medline: 22582982]
33. Lea M, Spears R. Computer-mediated communication, de-individuation and group decision-making. Int J Man Mach Stud 1991 Feb;34(2):283-301. [doi: 10.1016/0020-7373(91)90045-9]
34. Baranowski T, Cullen KW, Nicklas T, Thompson D, Baranowski J. Are current health behavioral change models helpful in guiding prevention of weight gain efforts? Obes Res 2003 Oct;11 Suppl:23S-43S. [doi: 10.1038/oby.2003.222] [Medline: 14569036]
35. Vartanian LR, Spanos S, Herman CP, Polivy J. Modeling of food intake: a meta-analytic review. Soc Influ 2015 Feb 24;10(3):119-136. [doi: 10.1080/15534510.2015.1008037]
36. Cruwys T, Bevelander KE, Hermans RC. Social modeling of eating: a review of when and why social influence affects food intake and choice. Appetite 2015 Mar;86:3-18. [doi: 10.1016/j.appet.2014.08.035] [Medline: 25174571]
37. Poole M, Hollingshead A. Theories of Small Groups: Interdisciplinary Perspectives. Thousand Oaks, CA: Sage Publications; 2005.
38. Walther J, Parks M. Cues filtered out, cues filtered in: computer-mediated communication and relationships. In: Knapp ML, Daly JA, editors. Handbook of Interpersonal Communication. Thousand Oaks, CA: Sage; 2002:529-563.
39. Turner J. Rediscovering the Social Group: A Self-Categorization Theory. Oxford, UK: Blackwell Publishers; 1987.
40. Price KH, Harrison DA, Gavin JH, Florey AT. Time, teams, and task performance: changing effects of surface- and deep-level diversity on group functioning. Acad Manage J 2002 Oct 01;45(5):1029-1045. [doi: 10.2307/3069328]
41. Nass C, Lee KM. Does computer-synthesized speech manifest personality? Experimental tests of recognition, similarity-attraction, and consistency-attraction. J Exp Psychol Appl 2001 Sep;7(3):171-181. [Medline: 11676096]
42. Horwitz SK, Horwitz IB. The effects of team diversity on team outcomes: a meta-analytic review of team demography. J Manage 2007 Dec 01;33(6):987-1015. [doi: 10.1177/0149206307308587]

43. van der Land SF, Schouten AP, Feldberg F, Huysman M, van den Hooff B. Does Avatar appearance matter? How team visual similarity and member-Avatar similarity influence virtual team performance. Hum Commun Res 2014 Aug 19;41(1):128-153. [doi: 10.1111/hcre.12044]
44. Watson WE, Johnson L, Kumar K, Critelli J. Process gain and process loss: comparing interpersonal processes and performance of culturally diverse and non-diverse teams across time. Int J Intercult Rel 1998 Nov;22(4):409-430. [doi: 10.1016/S0147-1767(98)00016-9]
45. Batenburg A, Das E. Virtual support communities and psychological well-being: the role of optimistic and pessimistic social comparison strategies. J Comput-Mediat Comm 2015 Aug 24;20(6):585-600. [doi: 10.1111/jcc4.12131]
46. Kerr NL, Messé LA, Seok D, Sambolec EJ, Lount RB, Park ES. Psychological mechanisms underlying the Köhler motivation gain. Pers Soc Psychol Bull 2007 Jun;33(6):828-841. [doi: 10.1177/0146167207301020] [Medline: 17475617]
47. Locke E. Motivation by goal setting. In: Golembiewski RT, editor. Handbook of Organizational Behavior. New York: Marcel Dekker; 2001:43-54.
48. Festinger L. A theory of social comparison processes. Hum Relat 1954 May 01;7(2):117-140. [doi: 10.1177/001872675400700202]
49. Van der Zee K, Buunk B, Sanderman R, Botke G, van den Bergh F. Social comparison and coping with cancer treatment. Pers Indivl Differ 2000 Jan;28(1):17-34. [doi: 10.1016/S0191-8869(99)00045-8]
50. Buunk B, Ybema J. Social comparisons and occupational stress: the identification-contrast model. In: Buunk B, Gibbons FX, editors. Health, Coping and Well-being: Perspectives from Social Comparison Theory. Mahwah, NJ: Lawrence Erlbaum; 1997:359-388.
51. Dibb B, Yardley L. How does social comparison within a self-help group influence adjustment to chronic illness? A longitudinal study. Soc Sci Med 2006 Sep;63(6):1602-1613. [doi: 10.1016/j.socscimed.2006.03.031] [Medline: 16716478]
52. Thompson FE, Byers T. Dietary assessment resource manual. J Nutr 1994 Nov;124(11 Suppl):2245S-2317S. [Medline: 7965210]
53. Brown LB, Dresen RK, Eggett DL. College students can benefit by participating in a prepaid meal plan. J Am Diet Assoc 2005 Mar;105(3):445-448. [doi: 10.1016/j.jada.2004.12.030] [Medline: 15746834]
54. Gupta SK. Intention-to-treat concept: a review. Perspect Clin Res 2011 Jul;2(3):109-112 [FREE Full text] [doi: 10.4103/2229-3485.83221] [Medline: 21897887]
55. Kenward MG, Roger JH. Small sample inference for fixed effects from restricted maximum likelihood. Biometrics 1997 Sep;53(3):983. [doi: 10.2307/2533558]
56. Bates D, Mächler M, Bolker B, Walker S. Fitting linear mixed-effects models using lme4. J Stat Soft 2015;67(1):1-48. [doi: 10.18637/jss.v067.i01]
57. Hayes A. Introduction to Mediation, Moderation, and Conditional Process Analysis: A Regression-Based Approach. New York: Guilford Press; 2013.
58. Baron RM, Kenny DA. The moderator-mediator variable distinction in social psychological research: conceptual, strategic, and statistical considerations. J Pers Soc Psychol 1986 Dec;51(6):1173-1182. [Medline: 3806354]
59. Mackinnon DP, Lockwood CM, Williams J. Confidence limits for the indirect effect: distribution of the product and resampling methods. Multivariate Behav Res 2004 Jan 1;39(1):99 [FREE Full text] [doi: 10.1207/s15327906mbr3901_4] [Medline: 20157642]
60. Parker K, Morin R, Horowitz J, Lopez M, Rohal M. Pew Research Center. 2015 Jun 11. Multiracial in America: proud, diverse and growing in numbers URL: http://www.pewsocialtrends.org/2015/06/11/multiracial-in-america/ [WebCite Cache ID 6jYxVkDc5]
61. United States Census Bureau. 2014 Dec 04. New Census Bureau statistics show how young adults today compare with previous generations in neighborhoods nationwide URL: http://www.census.gov/newsroom/press-releases/2014/cb14-219.html [WebCite Cache ID 6jYxgG3x6]
62. Centers for Disease Control and Prevention. 5 A Day Works!. Washington, DC: US Department of Health and Human Services; 2005. URL: http://www.cdc.gov/nccdphp/dnpa/nutrition/health_professionals/programs/5aday_works.pdf [WebCite Cache ID 6ja75Tf0p]
63. Kerr NL, Forlenza ST, Irwin BC, Feltz DL. "… been down so long …": Perpetual vs intermittent inferiority and the Köhler group motivation gain in exercise groups. Group Dyn-Theor Res 2013;17(2):67-80. [doi: 10.1037/a0031588]
64. Baranowski T. 24-hour recall and diet record methods. In: Willett W, editor. Nutritional Epidemiology. Third edition. Oxford, UK: Oxford University Press; 2012:49-69.
65. Bettadapura V, Thomaz E, Parnami A, Abowd G, Essa I. Leveraging context to support automated food recognition in restaurants. In: Proceedings of WACV 2015: IEEE Winter Conference on Applications of Computer Vision. 2015 Presented at: IEEE Winter Conference on Applications of Computer Vision; Jan 6-9, 2015; Waikoloa Beach, HI. [doi: 10.1109/WACV.2015.83]
66. Merler M, Wu H, Uceda-Sosa R, Nguyen Q, Smith J. Snap, Eat, RepEat: a food recognition engine for dietary logging. In: Proceedings of the 2nd International Workshop on Multimedia Assisted Dietary Management. 2016 Presented at: 2nd International Workshop on Multimedia Assisted Dietary Management; Oct 16, 2016; Amsterdam, The Netherlands p. 31-40. [doi: 10.1145/2986035.2986036]

67. Cordeiro F, Bales E, Cherry E, Fogarty J. Rethinking the mobile food journal: Exploring opportunities for lightweight photo-based capture. In: Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems. 2015 Presented at: 33rd Annual ACM Conference on Human Factors in Computing Systems; Apr 18-23, 2015; Seoul, Korea p. 3207-3216. [doi: 10.1145/2702123.2702154]

Abbreviations

BMI: body mass index
CMC: computer-mediated communication
FFQ: Food Frequency Questionnaire

Edited by G Eysenbach; submitted 23.08.16; peer-reviewed by A DeSmet, K Schneider, H Enwald; comments to author 26.11.16; revised version received 04.01.17; accepted 08.02.17; published 06.03.17.

Please cite as:
Meng J, Peng W, Shin SY, Chung M
Online Self-Tracking Groups to Increase Fruit and Vegetable Intake: A Small-Scale Study on Mechanisms of Group Effect on Behavior Change
J Med Internet Res 2017;19(3):e63
URL: http://www.jmir.org/2017/3/e63/
doi:10.2196/jmir.6537
PMID:28264793

Original Paper

With Some Help From My Network: Supplementing eHealth Literacy With Social Ties

Tsahi (Zack) Hayat[1], PhD; Esther Brainin[2], PhD; Efrat Neter[2], Ph.D.

[1]Interdisciplinary Center, Herzliya, Israel

[2]Ruppin Academic Center, Emek Hefer, Israel

Corresponding Author:
Tsahi (Zack) Hayat, PhD
Interdisciplinary Center
PO Box 167
Herzliya, 4610101
Israel
Phone: 972 99602821
Fax: 972 99527650
Email: tsahi.hayat@gmail.com

Abstract

Background: eHealth literacy is defined as the ability to seek, find, understand, and appraise health information from electronic sources and apply knowledge gained to addressing or solving a health problem. Previous research has shown high reliance on both online and face-to-face interpersonal sources when sharing and receiving health information.

Objective: In this paper, we examine these interpersonal sources and their interplay with respondents' eHealth literacy and perceived health outcomes. Specifically, we look at how the relationship between eHealth literacy and health outcomes is moderated by (1) finding help while performing online activities, (2) finding others with similar health concerns online, and (3) the importance of finding others with similar health concerns for people from ethnic minorities, specifically Palestinian citizens of Israel versus Israeli Jews.

Methods: We used a nationally representative random-digit dial telephone household survey of an Israeli adult population (age ≥21 years, N=819). The collected data were analyzed using two regression models. The first examined how the correlation between eHealth literacy and perceived outcomes was moderated by the availability of help. The second examined how the correlation between eHealth literacy and perceived outcomes was moderated by finding others with similar health concerns and by ethnicity.

Results: Respondents with low eHealth literacy who were able to recruit help when performing online activities demonstrated higher perceived health outcomes compared to similar respondents who did not find help. Respondents with low eHealth literacy, who were able to find others with similar health concerns (online), demonstrated higher perceived health outcomes when compared to similar respondents who did not find others with similar health concerns. Finally, finding similar others online was more helpful in enhancing health outcomes for ethnic minorities; Palestinian citizens of Israel gained more health benefits by finding similar others compared to Israeli Jews.

Conclusions: Although the availability of help and the notion of ethnicity have been discussed extensively within the context of social capital and health, our findings offer initial evidence for the relevancy of these concepts when studying individuals' eHealth literacy. Specifically, our findings enable a better understanding of the role of social ties and ethnicity in moderating the interplay between eHealth literacy and perceived health outcomes. Given the increased importance of eHealth information, our findings enhance understanding of how social ties can potentially compensate for low eHealth literacy.

(J Med Internet Res 2017;19(3):e98) doi:10.2196/jmir.6472

KEYWORDS

eHealth literacy; consumer health information; outcomes assessment; ethnicity

Introduction

The term eHealth refers to "the use of emerging information and communications technology to improve or enable health and health care" [1]. In the rapidly developing use of the Internet in society, eHealth literacy has become an important prerequisite for promoting healthy behavior [2,3]. Previous research has shown high reliance on both online and face-to-face interpersonal sources in searching for relevant health information and on forming decisions regarding different health-related topics [4,5]. Within the Israeli context, it was shown that consulting with other online users on health-related topics led to increased levels of health information and increased usage levels of online health services [6,7].

Despite the central role of interpersonal communication in accessing health information, research investigating how interpersonal communication interplays with eHealth literacy is scarce. In this study, we examine the impact of social ties on eHealth literacy among ethnic groups in Israel. Specifically, we investigate to what extent the availability of offline help when searching for health information online and finding other online individuals with similar health conditions can supplement individuals' low eHealth literacy, which in turn can help health consumers achieve positive outcomes when using the Internet for health purposes.

Related Work

This section outlines the theoretical concepts used to frame the paper and provides a summary of past and present research. The conceptual framework offers a brief perspective on the association between eHealth literacy and perceived health outcomes. This section also introduces the concepts of interpersonal communication, social networks, and social capital.

eHealth Literacy

Health literacy is defined as the ability of an individual to obtain, process, and understand basic health information in order to make appropriate decisions concerning health [8,9], whereas eHealth literacy is defined as "the ability to seek, find, understand and appraise health information from electronic sources and apply knowledge gained to addressing or solving a health problem" [10]. According to this definition, eHealth literacy encompasses six types of literacy: traditional (literacy and numeracy), information, media, health, computer, and scientific. Of these, media and computer literacies are unique to the Internet context, with eHealth media literacy being the awareness of media bias or perspective, the ability to discern both explicit and implicit meaning from media messages, and to derive meaning from media messages. The literature includes other definitions of perceived media capability or efficacy, but these were not specific to health information on the Internet [11].

Having the composite skills of eHealth literacy allows health consumers to achieve positive outcomes from using the Internet for health purposes. eHealth literacy has the potential to both protect consumers from harm and empower them to fully participate in informed health-related decision making [10]. People with high levels of eHealth literacy are also more aware of the risk of encountering unreliable information on the Internet [7,12]. On the other hand, the extension of digital resources to the health domain in the form of eHealth literacy can also create new gaps between health consumers [11]. eHealth literacy hinges not on the mere access to technology, but rather on the skill to apply the accessed knowledge [13], thus lending support to the hypothesis that information technology is creating a new social inequality rather than leveling social discrepancies [14].

The importance of assessing eHealth literacy is highlighted by recent findings suggesting that instead of basing health care treatment decisions on professional weighing of risks and benefits for different treatments and outcomes, individual preferences regarding health treatments are based in reality on limited information influenced by prior individual and collective experience [15]. This can be partially understood by findings indicating that most interpersonal communication sources for health information are lay people, such as family and peers within one's social network, rather than professional sources such as health care providers [16,17].

The examination of individuals' social ties is gaining increased prominence in studies dealing with patients' attainment of health information [18]. Specifically, more attention is being paid to the impact technologies such as email and the Internet have on the way individuals gather and share health information through their social ties [18]. Although these ties are important conduits for shared resources [18], much of the public health literature focuses either on patient-provider communication [19,20] or identifying and training people to serve as opinion leaders, peer leaders, or community health workers to aid in the implementation of designated health promotion/prevention disease interventions [21-23]. Less attention has been paid to identifying how similar health conditions are constructive in facilitating effective communication, for gaining access to information, for applying health information, and for increasing the benefits associated with such interpersonal health information sources [20]. More specifically, to the best of our knowledge, no attention has been paid in the interpersonal communication and health communication literature to study how social ties can potentially compensate for low eHealth literacy. Given the increased importance of social ties in attaining health information, and in light of the importance of eHealth literacy in understanding and appraising health information, we believe that by studying these two domains together we can better understand their joint potential contribution for enhancing health outcomes. Ways for bridging these two domains through combining emerging findings from the interpersonal ties literature with established literature in the health communication field are addressed in the next sections.

Social Ties

Social ties among individuals are often important conduits for sharing resources and can be described in such terms as their density, range, boundedness, and homogeneity [18]. Although social ties describe the structure of social interaction, social support describes the resources that are shared through these ties. Social support exists in a number of forms, including emotional support (eg, love, caring, and sympathy), instrumental support (eg, assistance with tangible needs), and informational

support (eg, provision of advice or information) [24]. Social networks and social support can have a profound impact on health by improving health behaviors (eg, [25]) and allowing for the provision of health-conserving resources (eg, [26]).

Social ties are also a means for information exchange among individuals. For example, previous studies have shown that belonging to two or more community organizations is associated with increased knowledge of screening for colon cancer, and increased knowledge of levels of exercise and fruit and vegetable intake recommended to reduce cancer risk [24]. These results support the intuitive notion that the more social ties individuals have the more likely they are to be exposed to health-enhancing information. On the other hand, although social ties may encourage some to stop smoking [27], they may encourage others to take up the practice [27] or grow obese together [28].

As the previous examples indicate, the relationship between social ties and health communication are not as straightforward as they initially appear. One contentious issue involves the distinction between strong ties, defined as those that are frequent and multifaceted, and weak ties, defined as infrequent and unidimensional [29]. Conceptually, one might argue that strong ties are less likely to lead to the distribution of new information compared with weak ties because individuals tend to form stronger relationships with those who are more similar to themselves and thus strong ties may not be able to offer many innovative ideas in the course of a social exchange [29]. Thus, weak ties may be more important for exposing people to innovative ideas, particularly in terms of transferring health information [30]. Findings on this issue are mixed: one study from the United States found that people reported strong ties as the most important source of health information [31], whereas other research emphasized the importance of both strong and weak ties [32-34]. This finding was also documented in the Israeli context [6].

Social ties through which health information can be accessed and applied include contacts with professionals such as medical health providers and/or lay interpersonal communication sources, including other service consumers, family and friends, neighbors, and religious leaders [16,17]. Using the Internet to access health information, where it is widely prevalent, is steadily rising [7,35,36]. A recent study by Mesch [37], based on a representative sample, indicates a reliance on online communication for health purposes within the Israeli population both in accessing medical information as well as for communication on health-related topics. These findings on increased reliance on online sources of information among the general population are reiterated in other studies [32,38]. Due to the increasingly important role of computer-mediated communication (CMC) in the formation and preservation of social ties, as well as for information sharing, we focus in this paper on social ties facilitated by both CMC and face-to-face interactions, a field that has not been explored as a medium that can potentially supplement eHealth literacy.

Assistance Provision Through Social Capital

A concept that may further explain the role social ties play within the context of health communication is that of social capital [39-41]. There is no firm, unanimous definition of social capital, and the particular definition adopted by any given study is dependent on the discipline and level of investigation. In its simplest form, social capital can be defined as the social ties or connections through which one gains access to resources [42]. Coleman [43] defines social capital as a function of social structure producing advantage, whereas Bourdieu [42] defines social capital as "the aggregate of the actual or potential resources that are linked to possession of a durable network of more or less institutionalized relationships of mutual acquaintance and recognition" (p 248).

Despite the fact that social capital is a contentious and slippery term, the preceding definitions emphasize the notion that social capital resides not within the individual but rather in the relationships that an individual or group has with others. For the purpose of this study, social capital theory is defined as the sum of the actual and potential resources embedded within, available through, and derived from an individual's network of relationships. Social capital comprises both the network and the assets that may be mobilized through it [44]. Such value can also be associated and facilitate the flow of information [45].

The importance of these social capital within the context of health is seminal, given that such resources has been found to play a central role in enhancing health [46,47]. Previous work have shown that social capital has a positive impact on participants' emotional and physical well-being, either through gratifying the health information needs of individuals or through communication with health care providers and with other lay people [46,47]. Although the effect of social capital on health has been extensively studied in the past three decades [48], relatively little work has been done on the role social ties plays for individuals with low eHealth literacy. Such individuals can potentially access resources in their social networks (ie, other individuals who can guide them when searching for health-related information online); such assistance can supplement their low eHealth literacy and enhance their perceived health outcomes. This study will not measure the social capital of individuals, although it is guided by the social capital literature in studying the benefits social ties can offer for individuals' perceived health outcomes. Specifically, in light of this literature, we are interested in examining whether the services and resources, available through the social ties of individuals, can help individuals to seek, find, understand, and appraise health information from electronic sources. Thus, our first hypothesis is that at low levels of eHealth literacy, finding others who can help perform online activities is associated with higher perceived health outcomes, whereas at high levels of eHealth literacy, perceived health outcomes do not vary as a function of finding others who can help.

Information Provision Through Similar Others

Interpersonal communications are central to social capital [20]. Interpersonal communication is the medium through which individuals and groups create, foster, alter, and terminate their social ties. Expressing needs and negotiating assistance constitute key aspects of social support that draw on interpersonal communications. Through the iterative communication between individuals, societies collectively create

(or fail to create) the social participation, norms of reciprocity, and group trust that are the hallmarks of social capital. Without interpersonal communication, these social processes would lose their meaning and indeed cease to exist [20].

Despite the tremendous amount of information available on the Web, research has shown that users continue to rely on people in their networks when seeking various types of information. Such work has examined different domains of information search, ranging from recreational activities [49] to cultural content [50], and has found that users supplement online sources with advice they obtain from friends and family. Furthermore, research on information flow and attitudes within social networks suggests that ties between individuals and similar others can promote the exchange of relevant information among peers [51] and affect their attitudes toward that information [52,53]. These findings are attributed to the fact that most people tend to rely on the subjective evaluation of friends, family, and trusted others rather than on scientific evidence to form an opinion and make a decision about something [18]. Moreover, individuals are more likely to be receptive to information shared by others who are similar to them [54]. Although recent studies have documented the tendency of Internet users to seek others who might share the same health concerns they have [5], studies thus far have not explicitly addressed the role of information received from similar others as a mechanism that can compensate for low eHealth literacy. Previous studies that looked at interpersonal health communication (eg, [55]) have shown that such communication tends to use lay terminology and plain language, thus ‘‘translating’’ complicated health information for one another. Additionally, these interactions provide opportunities to hear about the personal experiences of other network members, who are perceived as both dealing with similar concerns and as unbiasedly sharing relevant information [56,57]. Thus, our second hypothesis is that at low levels of eHealth literacy, finding others with similar health concerns is associated with higher perceived health outcomes, whereas at high levels of eHealth literacy, perceived health outcomes do not vary as a function of finding others with similar health concerns.

Information Provision Through Similar Others and Ethnicity

The findings regarding the association between social ties and health outcomes have been inconsistent. Some studies have found a beneficial effect of social ties over health outcomes both internationally [58] and in Israel [6,37], whereas others have demonstrated either null or negative associations [59]. These inconsistencies may reflect the reality that social ties can be both an asset and a liability for health. Social ties are believed to benefit health through access to resources such as emotional and material support and health information generated by social networks [60]. Detrimental health effects of social ties may occur because of unmanageable demands of networks or exposure to unhealthy behaviors such as smoking [61]. Social ties may provide different positive and negative resources, and there is some evidence that the potential benefits of different types of social networks for health may vary, for instance by social classes [48,62]. Among ethnic minorities, local social ties might exclude people from access to health information. Furthermore, mechanisms of control and social pressure, resulting from these local social ties, can cause social exclusion [62] and lead to further deteriorations in health.

In Israel, disparities in health exist between the three main population groups: nonimmigrant Jews, immigrants from the former Soviet Union (arriving in Israel since 1990), and Palestinian citizens of Israel (PCI) [37,63]. Specifically, inequalities in the use of health care services based on ethnicity are also evident: PCI are more likely to visit a general practitioner, less likely to visit a specialist, and more likely to be hospitalized compared to Jewish Israelis [64]. These differences cannot be attributed to insurance factors because all Israeli residents have access to universal health care coverage. However, several structural barriers may explain differences in access to services. Although primary services are available in Arab localities (villages and midsize towns), specialist clinics are more likely to be located in large cities. Thus, PCI must overcome both language and geographic barriers to access specialists: sometimes they have to travel an hour each way to Jewish cities where specialized clinics are located, missing work and incurring transportation costs [65]. These challenges may also explain why PCI are more likely than the ethnic majority to access health information online because such information is accessible in Arabic and it can potentially eliminate transportation costs [37].

Barriers in communication with health care specialists have been documented in low-literate and minority communities; these individuals are more likely to seek advice from friends and family than from trained health care providers or peer-reviewed journals [65]. It is not surprising that studies both in Israel [6] and the United States [66,67] documented that members of ethnic minorities are likely to rely more on their social ties as sources for health information compared to their ethnic majority counterparts, who tend to rely more on mass media sources.

Nevertheless, the benefits of advice from friends and family might not be available or beneficial to everyone in the same way. For instance, social ties might benefit those who are better off in society, but constrain or even exclude people with a lower socioeconomic status or in a minority position [62]. A recent study offered a systematic review of 60 studies examining the interactions between the benefits associated with social ties and socioeconomic-disadvantaged groups [48]. The article reported findings indicating a greater health benefit of social ties for people with a disadvantaged position in society, and no effects or limited health benefits for those with a position higher up the social ladder. People with high perceived availability of social ties can turn to these social ties, and use the resources available through these ties to enjoy greater perceived health, than expected considering their low socioeconomic status [48]. We hypothesize that similar findings will be evident when studying perceived health outcomes among PCI. Thus, our third hypothesis is that the interaction between eHealth literacy, finding others with similar health concerns, and perceived outcomes is stronger for PCI compared to Jewish Israelis.

Methods

Data Collection and Sample Characteristics

Data were collected in a nationally representative random-digital dial (RDD) telephone household survey of Israeli adult population (aged 21 years and older) conducted in November 2014 (landlines and mobile combined). The following RDD-based sampling procedure was used: statistical areas were divided into four strata layers according to (1) population group (Jews, PCI, and mixed localities), (2) seven geographical districts, (3) size of settlements (large cities, small towns and villages), and (4) the locality's socioeconomic status index. These strata were based on the Israeli Central Bureau of Statistics classification. Sampling employed a dual-frame design, incorporating two selection stages. The first frame was designed to provide national coverage of the eligible population and was guided by the statistical areas defined by the Israeli Central Bureau of Statistics; the second frame sampled households within each statistical area. Calls were placed to 1789 residential households; 1628 eligible potential respondents were identified (ineligible numbers included eight fax numbers and 153 disconnected phones), of whom 819 agreed to be interviewed, representing a 50.31% response rate. The interviews were conducted in Hebrew, Russian, or Arabic by professional interviewers who underwent a special training session to familiarize them with the questionnaire's terminology. The interviewers also read the interviewees the consent form (adapted from [12]) and invited the interviewee to ask any questions she or he might have. The interviewers conducted the telephone survey using computer-assisted telephone interviewing software.

Measures

Perceived health outcomes of seeking health information on the Internet, the dependent variable in our study, was measured using the item "Do you agree or disagree that seeking health information on the Internet...?" followed by a list of nine outcomes (adapted from [4,68,69]). Responses were expressed on a five-point scale from 1 (strongly agree) to 5 (strongly disagree). The respondents also had the option to indicate as a response to each item that they either "don't know" or that the question is "irrelevant." For each respondent, a total mean score of perceived health-related outcomes was computed (alpha=.87, mean 1.48, SD 1.64) (see Multimedia Appendix 1 for a full list of the items used in the survey).

The independent variables in this study included perceived ease of obtaining face-to-face help in performing online activities, eHealth literacy, finding others online with similar health concerns, and ethnicity. Each of these variables were measured using the following items. Ease of obtaining help in performing online activities was assessed by responses to the question "When you need advice or help surfing the Internet, for example help in finding a particular site or service, how easy is it for you to find someone who will help you?" on a five-point Likert scale ranging from 1 (very difficult) to 5 (very easy) [70] (mean 1.94, SD 2.12). This item assessed the help attained through face-to-face interactions. Perceived eHealth literacy was assessed using the eHealth Literacy Scale (eHEALS) [10], a scale consisting of eight items on a five-point Likert scale ranging from 1 (strongly disagree) to 5 (strongly agree). The scale was previously translated to Hebrew, Arabic, and Russian [7] (alpha=.78, mean 1.64, SD 1.77). Finding others with similar health concerns was assessed by responses to the question "During the past 12 months, how often have you used the Internet for finding others with health issues or concerns similar to the ones you are facing?" on a five-point Likert scale ranging from 1 (very rarely) to 5 (very often) [5] (mean 0.75, SD 0.84). This item assessed only the online ties with similar others. Ethnicity was measured by self-identification (Jewish / Arab Muslim / Arab Christian / Arab Druze).

Our control variables were Internet activities (Web 1.0 and Web 2.0) [71] assessed by reported frequency of 15 digital activities on a five-point Likert scale from 1 (never) to 5 (very often). The Web 1.0 activity index was assessed using a scale consisting of eight items (eg, searches for driving instructions, searches for a product/service) (alpha=.81, mean 2.38, SD 0.71). The Web 2.0 activity index was assessed using a scale consisting of seven items (eg, manages a personal site, uploads photos/video or responds to photos by others) (alpha=.83, mean 1.59, SD 0.71) [72]. Sociodemographic information on age, gender, and education were obtained as part of the background variables. Sociodemographic information, specifically age and education, are known to be strong predictors of online and offline health information behavior [7] and needed to be controlled in our analysis.

Data Analysis

We employed two linear regression models to study the interplay between eHealth literacy and perceived outcomes of seeking health information on the Internet. The first regression model, a three-step hierarchical multiple regression model, assessed the interaction effect of help availability and eHealth literacy on the perceived health outcomes of information search. Our second regression model, a four-step hierarchical multiple regression model, assessed the interaction effect of finding others with similar health concerns and eHealth literacy on the perceived health outcomes of an individual. The second model also assessed a three-way interaction between finding others with similar health problems, eHealth literacy, and ethnicity on the perceived health outcomes of an individual, for PCI, and for Jewish Israelis.

For both regression models, the Durbin-Watson statistic was used to investigate the assumption of independence. Normal probability plots were used to investigate the normality of error terms and homoscedasticity was tested by observing the scatterplot of the residuals and the predicted value. These checks identified no violations of multiple regression assumptions. All statistical tests were one-tailed and a significance level of $P<.001$ was set for all analyses.

To facilitate the interpretation of the interactions, all continuous variables used in our models were standardized [73]. To calculate the statistical power of this study to reject false null hypotheses, we conducted a post hoc statistical power test [74,75]. With 11 predictors in the regression analysis, an

observed R^2 of .88, a sample size of 487, and alpha=.05, the test results indicated an observed power of 1.0.

Results

Among the Israeli Jew respondents, the mean age was 51.1 (SD 17.2) years, and 549 of 683 (80.4%) had Internet access. Among the PCI respondents, the mean age was 42.5 (SD 13.6) years, and 112 of 136 (82.4%) had Internet access. Table 1 provides the gender and education distribution, respectively, within our sample. Furthermore, given that this study focused on individuals who use the Internet for acquiring health information, it is important to note that 49.1% of our sample indicated that they use the Internet as a source for health information. This figure is consistent with current studies conducted in the Israeli context [37].

Table 1. Demographic distribution among the respondents (N=819).

Demographics	Israeli Jews (n=683)	PCI (n=136)
Age (years), mean (SD)	51.1 (17.2)	42.5 (13.6)
Gender, n (%)		
Male	328 (48.0)	67 (49.3)
Female	355 (51.9)	69 (50.7)
Education, n (%)		
High school or less	303 (44.2)	66 (48.6)
Professional degree	125 (18.3)	16 (11.8)
Partial academic degree	19 (2.8)	7 (5.1)
Bachelor's degree	147 (21.5)	35 (25.7)
Master's degree or above	89 (13.0)	12 (8.8)

A three-step hierarchical multiple regression model was conducted to examine our first hypothesis, which stated that the association between eHealth literacy and perceived outcomes is moderated by availability of help (see Multimedia Appendix 2 for the distribution of the following variables: ease of obtaining help in performing online activities, perceived eHealth literacy, finding others with similar health concerns, and the outcome measures).

As can be seen in Table 2, age, gender, education, and Internet activity were entered in the first step. eHealth literacy and help availability were entered in the second step and their interaction term in the third step. The overall model was significant (step 3: ΔR^2=.009; $F_{7,479}$=352.22, P<.001). To test the appropriateness of our steps, we assessed the R^2 increase in step 2 relative to step 1, as well as for step 3 relative to step 2, with an F test. The results of the F test show that the respective F changes of step 2 and step 3 were 1042.56 (P<.001) and 27.81 (P<.001), respectively. Regression coefficients and significance values are presented in Table 2. As indicated by our findings, there was a significant main effect for eHealth literacy (beta=2.35, SE=0.31; t_{480}=46.29, P<.001), a significant main effect for help availability (beta=0.15, SE=0.30; t_{480}=3.17, P=.003), and a significant interaction (beta=–0.45, SE=0.51; t_{479}=–5.27, P<.001). The interaction plot, depicted in Figure 1, suggests that high help availability yielded higher perceived outcomes when eHealth literacy was low, as compared to lower perceived outcomes when there was low help availability. Simple slopes tests, following Cohen et al [76], were conducted at one standard deviation above and below the mean of help availability. Both slopes were significant (P<.001). Note that the background variables, which significantly predicted the perceived outcomes in the first step, did not contribute to the prediction of perceived outcomes once eHealth literacy and the availability of help were included as predictors. Gender was not associated with perceived outcomes even in the first step. Thus, our first hypothesis was supported.

Table 2. Standardized variables included in the hierarchical regression model predicting perceived outcomes.

Independent variables	Step 1 (n=508)			Step 2 (n=487)			Step 3 (n=487)		
	β[a]	t_{507}	*P*	β[a]	t_{486}	*P*	β[a]	t_{486}	*P*
Age	–2.59	–3.69	<.001	0.02	0.07	.23	0.12	0.39	.22
Gender	0.64	0.95	.08	–0.28	–0.95	.13	–0.30	–1.03	.10
Education	2.81	3.68	<.001	0.09	0.27	.55	0.14	0.43	.75
Internet activity	1.82	2.57	.006	–0.60	–1.92	.09	–0.44	–1.45	.10
eHealth literacy				4.21	44.81	<.001	2.35	46.29	<.001
Availability of help				0.21	3.52	<.001	0.15	3.17	.004
eHealth literacy × availability of help							–0.45	–5.27	<.001

[a] Because all continuous variables were standardized, betas for continuous predictors correspond to standardized regression coefficients.

A four-step hierarchical multiple regression model was used to examine our second and third hypotheses. Our second hypothesis stated that the association between eHealth literacy and perceived outcomes is moderated by finding others with similar health concerns, whereas our third hypothesis posited that this interaction between eHealth literacy and finding others with similar health concerns on perceived outcomes is stronger for PCI (compared to Jewish Israelis). As can be seen in Table 3, age, gender, education, and Internet activity were entered in the first step. eHealth literacy, finding others with similar health concerns, and ethnicity (as a dummy variable; Israeli Jews were coded as 0), were entered in the second step, their interaction terms in the third step, and the three-way interaction (eHealth literacy × finding others with similar health concerns × ethnicity) was entered in the fourth step. The overall model was significant (step 4: ΔR^2=.02; $F_{7,485}$=377.97, *P*<.001). To test the appropriateness of our steps, we assess the R^2 increase in step 2 relative to step 1, for step 3 relative to step 2, as well as for step 4 relative to step 3 with an *F* test. The results of the *F* test show that the respective *F* changes of step 2, step 3, and step 4 were 1030.14 (*P*<.001), 64.29 (*P*<.001), and 58.34 (*P*<.001). Regression coefficients and significances are presented in Table 3. Our findings (see step 3 in Table 3) indicated a significant main effect for eHealth literacy (beta=2.04, SE=0.50; t_{486}=24.99, *P*<.001), a significant main effect for finding others with similar health concerns (beta=.30, SE=0.48; t_{486}=0.87, *P*<.001), and a significant interaction (beta=–0.55, SE=0.53; t_{485}=–0.23, *P*<.001). The interaction plot, depicted in Figure 2, suggests that finding others with similar health concerns yielded higher perceived outcomes when eHealth literacy was low, as compared to lower perceived outcomes when the rate of finding others with similar health concerns was low. Simple slopes tests at one standard deviation above and below the mean of finding others with similar health concerns were conducted. Both slopes were significant (*P*<.001). Thus, our second hypothesis was supported.

Table 3. Standardized variables included in the hierarchical regression model predicting perceived outcomes.

Independent variables	Step 1 (n=508)			Step 2 (n=493)			Step 3 (n=493)			Step 4 (n=493)		
	β[a]	t_{507}	*P*	β[a]	t_{492}	*P*	β[a]	t_{492}	*P*	β[a]	t_{492}	*P*
Age	–2.61	–3.73	<.001	0.19	0.61	.30	0.24	0.80	.31	0.24	0.60	.31
Gender (0=male)	0.65	0.95	.08	–0.08	–0.25	.15	–0.06	–0.20	.17	–0.08	–0.32	.19
Education	2.80	3.68	<.001	–0.05	–0.14	.37	0.11	0.35	.21	0.08	0.28	.20
Internet activity	1.87	2.66	.006	–0.48	–0.15	.07	–0.19	–0.64	.08	–0.27	0.57	.09
eHealth literacy				2.78	25.86	<.001	2.04	24.99	<.001	1.78	21.42	<.001
Finding others with similar health concerns				0.21	0.68	.008	0.3	0.87	<.001	0.28	0.85	<.001
Ethnicity (0=Jewish Israelis)				–0.24	–0.18	.004	–0.17	–0.22	.003	–0.18	0.25	.007
eHealth literacy × finding others with similar health concerns							–0.55	–0.23	<.001	–0.48	–0.31	<.001
eHealth literacy × ethnicity							–0.23	0.63	.007	–0.16	0.41	.95
Ethnicity × finding others with similar health concerns							–0.41	0.28	.004	–0.32	0.38	.62
eHealth literacy × finding others with similar health concerns × ethnicity										–0.54	–0.72	.008

[a] Because all continuous variables were standardized, betas for continuous predictors correspond to standardized regression coefficients.

As depicted in Table 3, step 4 of our analysis revealed a significant three-way interaction (finding similar others × eHealth literacy × ethnicity) for perceived health outcomes (β=–0.54, SE=0.51; t_{482}=–0.72, *P*=.003). This significant three-way interaction is depicted in Figure 3 and Figure 4 indicating that the interaction between eHealth literacy and finding others with similar problems is weaker for Jewish Israelis (β=–0.53; t_{482}=–0.64, *P*=.02) than for PCI (β=0.69, t_{482}=–0.75, *P*=.03). In either case, finding others with similar health concerns yielded higher perceived outcomes when eHealth literacy was low as compared to lower perceived outcomes when the rates of findings others with similar health concerns was lower. This pattern diminishes under higher levels of eHealth literacy. We conducted simple slopes tests, separately for PCI and Jewish Israelis, at one standard deviation above and below the mean of finding similar others. Both slopes were significant for Jewish Israelis (*P*=.03) and for PCI (*P*=.02). Thus, our third hypothesis was supported.

Figure 1. Interaction effect of help availability and eHealth literacy on the perceived health outcomes of information search (n=487).

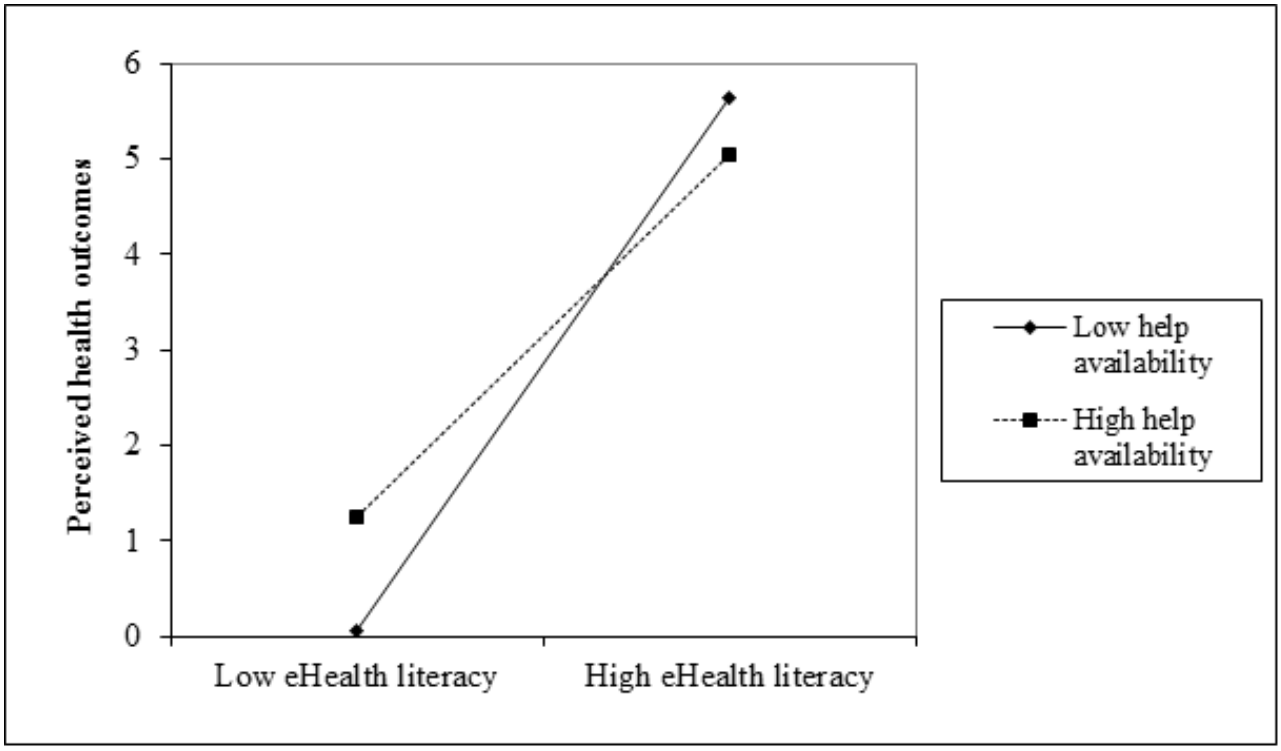


Figure 2. Interaction effect of finding others with similar health concerns and eHealth literacy on the perceived health outcomes of an individual (n=493).

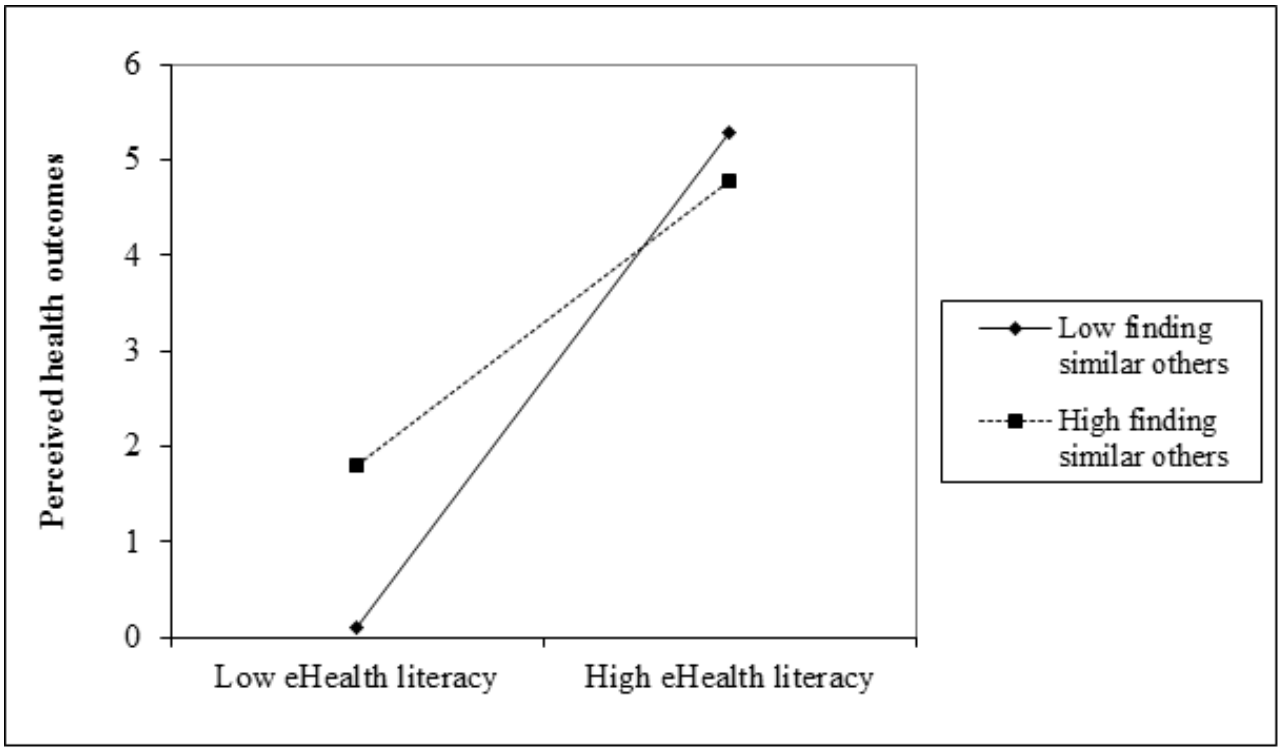


Figure 3. Three-way interaction effect of finding others with similar health problems, eHealth literacy, and ethnicity with the perceived health outcomes for PCI.

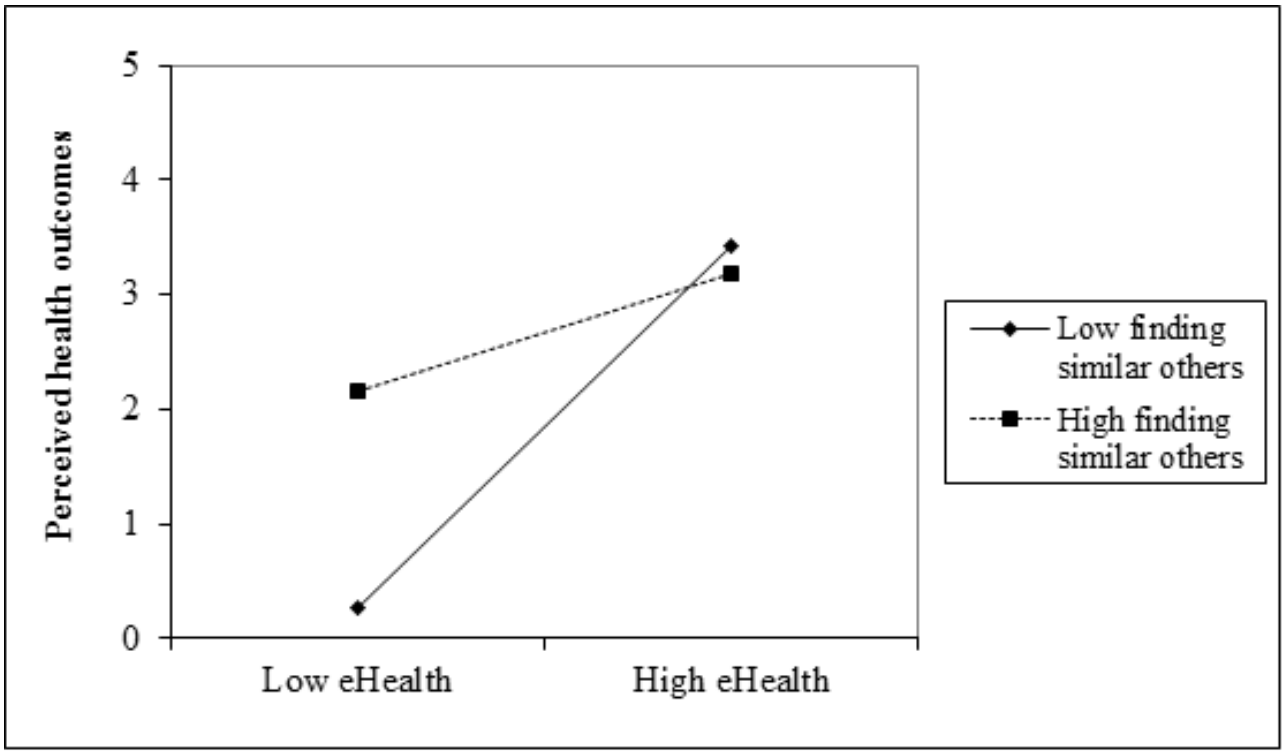
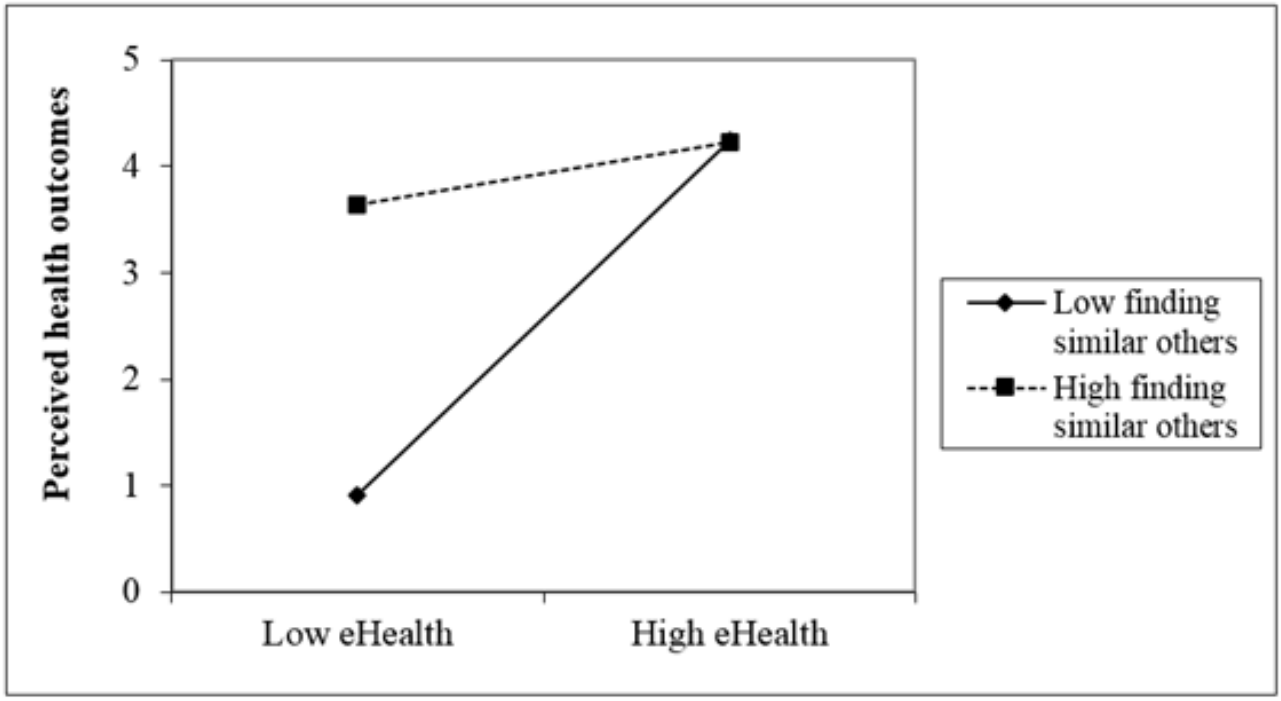


Figure 4. Three-way interaction effect of finding others with similar health problems, eHealth literacy, and ethnicity with the perceived health outcomes for Jewish Israelis.



Discussion

Main Findings

The contribution of this paper to the field of eHealth literacy is threefold. First, we showed how the availability of help when searching for health information online can enhance the perceived health outcomes of individuals with low eHealth literacy. We then showed that finding others with similar health concerns online can also enhance the perceived health outcomes of individuals with low eHealth literacy. Finally, we showed that finding similar others online is more helpful in enhancing perceived health outcomes for ethnic minorities; in our case, PCI gained more by finding similar others when compared with Israeli Jews. Although the availability of help and the notion of ethnicity have been discussed extensively within the context of social capital and health, our findings offer initial evidence for the relevance of these concepts for eHealth literacy as well. It is important to note that the differences in perceived health outcomes are very large between people with high and low eHealth literacy (despite the influence of help available and the ability to find others with the same health problems) as depicted by the main effect of eHealth literacy on perceived health outcomes (see Tables 2 and 3). Thus, availability of help and finding similar others can elevate individuals with low eHealth literacy, and increase their perceived health outcomes, but not to eradicate the gap between people with low and high eHealth literacy.

Comparison With Prior Work

Previous studies have stressed the notion that some Internet usage activities are more beneficial or advantageous for Internet users than others. Some activities offer users more chances and resources in moving forward in their career, work, education, and societal position than others that are mainly consumptive or entertaining [77]. People's use of the Internet as an important source for making health-related decisions is an indication of their having adopted an "Internet-oriented lifestyle," which leaves different "footprints" on the users' lives, manifested in an increase of social and financial returns [77]. Within the context of health, previous studies have also highlighted the importance of enhancing eHealth literacy to improve individuals' perceived health outcomes [10].

The vast majority of attention in eHealth literacy research has been focused on information accessibility, namely the delivery and readability of health-related information [78]. Accessible information that one understands is a necessary but not sufficient condition for addressing eHealth literacy. One's ability to apply the information in making health care decisions based on the information accessed is also an important part of eHealth literacy. Yet, even if one assumes that information is written at a reading level that can be understood by its readers, translating it into a culturally appropriate vernacular and delivering it via a communication channel that is accepted and easily accessed does not guarantee that the information will be utilized as it was intended [79]. At best, having accurate information will lead to a basic understanding of what the message sender desires the recipient both to know and to do, which is what Nutbeam characterized as functional health literacy [80]. Functional health literacy is a vital first step to realizing improvements in many health-related outcomes. Failing to move beyond functional health literacy can be likened to a health care system that is only concerned with emergency medicine, rather than addressing problems at their source. Nutbeam [80] has gone on to describe a second level of health literacy; namely, interactive health literacy. This level refers to the personal capacity to build skills and "act independently" (p 266 [80]) when armed with factual information. Finally, critical health literacy is the empowerment

of an individual to promote his or her own health outcomes despite difficult economic or social situations.

Addressing the interactive and critical eHealth literacy needs of a population presents an increasingly complex set of issues for health providers and researchers. Similar to health promotion efforts, the scope of the problem becomes exponentially large when one considers the need to equip and empower people to be their own educated, capable health promoters. Today's researchers and health care providers can benefit from the current proliferation of eHealth literacy research. Some researchers have begun to consider how such large-scale efforts could be conceptualized. For example, Ratzan [81] suggested a four-pronged approach, including integrated marketing communication (drawing on public relations and social marketing experts), health education (using the Internet and other multimedia channels), shared decision making (building partnerships with key groups), and efforts to increase the social capital and social ties of disenfranchised groups (providing social and relationship resources that are more scarce for underprivileged people). Kickbusch [82] has suggested that by considering health literacy broadly (as opposed to isolating diseases or specific health risks) and working to increase a population's social capital, health care advocates can achieve a more integrated and sustained program of health and social change. Simply put, increasing a population's health literacy across multiple health contexts will result in that population being empowered to take more control when addressing future health-related challenges.

In an attempt to expand this eHealth literacy agenda, our findings enable a better understanding of the role social ties and ethnicity play in moderating the interplay between eHealth literacy and perceived health outcomes. Specifically, we highlight the potential benefit of social ties in compensating for low eHealth literacy. In order to achieve this goal, this paper utilized the social capital and eHealth literacy frameworks for studying the role social ties play within the process of accessing and utilizing health information. As our findings indicate, social ties can enhance both the interactive and critical eHealth literacy needs (as evident from their importance in moderating the association between eHealth literacy and perceived health outcomes). Thus, we suggest that both researchers and practitioners will incorporate social ties into the study and implementation of eHealth literacy enhancement. Although many large-scale efforts to enhance eHealth literacy implicitly discuss the importance of social ties and social capital within the context of enhancing eHealth literacy, our findings provide a clear indication for the potential role social ties can play within this context.

Our work joins the extensive evidence indicating that social ties have implications in terms of both exposure to diverse information and the ability to utilize it [32,83]. In this work, we add to the literature on social ties and social capital by gaining a better understanding of the mechanisms through which social ties can moderate the interplay between eHealth literacy and perceived health outcome. We join previous work that highlight the notion that social capital is not based solely on face-to-face relationships, but also on online relationships (eg, [84]). Such a conceptualization allows us to investigate whether and to what extent people who encounter obstacles in obtaining health information through the Internet turn to their social ties for help and how these ties benefit people from different ethnic backgrounds. The fundamental notion of this study is that people who encounter barriers in obtaining health information and services through the Internet turn to their social ties for help [1]. The contribution of social ties in the context of eHealth is less widely acknowledged and this study addresses this gap.

Limitations

Our findings are hampered by three major limitations. First, the cross-sectional design of the study precludes causal conclusions and allows us to draw conclusions regarding only correlated relationships. For example, we can assume only that finding others with similar health concerns online mediates the interplay between eHealth literacy and perceived health outcomes, not that it affects eHealth literacy. Second, we did not measure performed eHealth literacy and health outcomes, but rather perceived efficacy of searching and using health information on the Internet. Although previous studies have found an association between perceived health information (eg, control of illness and perceived understanding of the illness) and actual health outcomes (eg, [85]), future studies should use measures of performed eHealth literacy and health outcomes. Indeed, measures for performed digital literacy [86] and measures for health literacy [87] exist. These measures may serve as inspiration for a measure that captures performed eHealth literacy and health outcomes. Future studies, possibly based on "big data" that records and monitors actual activities, may shed more light on the association between online usage and gains, at least measuring the usage more accurately. Thirdly, although this study points to the potential importance of social ties in moderating the association between eHealth literacy and perceived health outcomes, future studies could explore the extent to which people are indeed able to turn to their social ties, ask others for help or find online peers, and whether the interplay of eHealth literacy and social ties is more important for certain types of health information seeking (eg, whether this interplay is more important for enhancing information about prevention, symptom identification and self-diagnosis, or for learning about potential treatments). Finally, in this study, we did not look at the specific features of individuals' social ties (eg, strength and diversity of social ties) and how such features might affect eHealth literacy. More specifically, no attention was directed to an exploration of the types of social ties (eg, family, local friends, neighbors, online contacts) that are more effective in compensating for low eHealth literacy. Given the documented importance of interpersonal ties in the attainment of health information [16,17], we suggest that future studies should address this theoretical and empirical gap.

Conclusions

One of the most powerful trends is the increasing penetration of new CMC and the role they play in health. This paper suggests an additional contribution CMC offers: potential enhancement of eHealth literacy through interpersonal sources of health information. This study is one of the first empirical works that systematically investigates the role of interpersonal ties in eHealth literacy. More specifically, previous studies in

the interpersonal communication and health communication literature have not identified how the availability of other people while searching for health information, and their attributes, may promote perceived health outcomes. In this paper, we address this theoretical and empirical gap by combining the literature on social ties with established literature in the health communication field. By doing so, we were able to offer a new perspective on the role of social ties in compensating for low eHealth literacy and in increasing perceived health outcomes.

Although this study is anchored in the Israeli context, given that ethnic minorities in other countries rely on interpersonal sources for attaining health information (eg, [18]), and the evident increase among Internet users in turning to find others who might share their health concerns (eg, [5]), we believe that our findings are also relevant for other cultural contexts.

Acknowledgments

The study was supported by Israel's Ministry of Science, Technology & Space (grant # 3-10840). The authors thank the study participants for their time and for sharing their experiences and perspectives. We appreciate the work of Cohen Institute for Public Opinion Research at TAU in conducting the survey.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Items used for assessing respondents' perceived health outcomes from using the Internet for health purposes.

[PDF File (Adobe PDF File), 39KB - jmir_v19i3e98_app1.pdf]

Multimedia Appendix 2

Descriptive statistics of the ease of obtaining help in performing online activities, perceived eHealth literacy, finding others with similar health concerns, and the perceived outcome of Internet use measures.

[PDF File (Adobe PDF File), 40KB - jmir_v19i3e98_app2.pdf]

References

1. Eng T. The eHealth Landscape: A Terrain Map of Emerging Information and Communication Technologies in Health and Health Care. Princeton, NJ: Robert Wood Johnson Foundation; 2001.
2. Mitsutake S, Shibata A, Ishii K, Oka K. Associations of eHealth literacy with health behavior among adult Internet users. J Med Internet Res 2016;18(7):e192. [doi: 10.2196/jmir.5413]
3. Suri V, Majid S, Chang Y, Foo S. Assessing the influence of health literacy on health information behaviors: a multi-domain skills-based approach. Patient Educ Couns 2016 Dec;99(6):1038-1045. [doi: 10.1016/j.pec.2015.12.017] [Medline: 26794667]
4. Baker L, Wagner TH, Singer S, Bundorf MK. Use of the Internet and e-mail for health care information: results from a national survey. JAMA 2003 May 14;289(18):2400-2406. [doi: 10.1001/jama.289.18.2400] [Medline: 12746364]
5. Fox S. Pew Research Center. 2011 Feb 28. Peer-to-peer health care URL: http://www.pewinternet.org/2011/02/28/peer-to-peer-health-care-2/ [accessed 2016-07-31] [WebCite Cache ID 6jPoFjTXP]
6. Hayat Z, Nakash O, Ben-David B, Hamburger Y, Regev T, Sagi T. The role of social networks in enhancing mental health among elderly migrants: preliminary results. 2016 Feb Presented at: The 21st bi-annual conference of The Israel Gerontological Society; February, 2016; Tel-Aviv, Israel.
7. Neter E, Brainin E. eHealth literacy: extending the digital divide to the realm of health information. J Med Internet Res 2012 Jan 27;14(1):e19 [FREE Full text] [doi: 10.2196/jmir.1619] [Medline: 22357448]
8. Cutilli C. Health literacy in geriatric patients: an integrative review of the literature. Orthop Nurs 2007;26(1):43-48. [Medline: 17273109]
9. Parker R, Ratzan S, Lurie N. Health literacy: a policy challenge for advancing high-quality health care. Health Affair 2003;22(4):147-153.
10. Norman CD, Skinner HA. eHEALS: The eHealth Literacy Scale. J Med Internet Res 2006;8(2):e9. [doi: 10.2196/jmir.8.2.e9]
11. Miller E, West D. Characteristics associated with use of public and private web sites as sources of health care information: results from a national survey. Med Care 2007 Mar;45(3):245-251. [doi: 10.1097/01.mlr.0000244509.60556.49] [Medline: 17304082]
12. Bodkin C, Miaoulis G. eHealth information quality and ethics issues: an exploratory study of consumer perceptions. Intl J Pharm & Health Mrkt 2007 Apr 10;1(1):27-42. [doi: 10.1108/17506120710740261]
13. Viswanath K, Finnegan J. The knowledge gap hypothesis: twenty-five years later. Ann Int Commun Assoc 2016 May 18;19(1):187-228. [doi: 10.1080/23808985.1996.11678931]

14. Sassi S. Cultural differentiation or social segregation? Four approaches to the digital divide. New Media Soc 2005 Oct;7(5):684-700. [doi: 10.1177/1461444805056012]
15. Jimenez DE, Cook B, Bartels SJ, Alegría M. Disparities in mental health service use of racial and ethnic minority elderly adults. J Am Geriatr Soc 2013 Jan;61(1):18-25 [FREE Full text] [doi: 10.1111/jgs.12063] [Medline: 23252464]
16. Bonds DE, Foley KL, Dugan E, Hall MA, Extrom P. An exploration of patients' trust in physicians in training. J Health Care Poor Underserved 2004 May;15(2):294-306. [Medline: 15253380]
17. Smith K, Christakis N. Social networks and health. Annu Rev Sociol 2008 Aug;34(1):405-429. [doi: 10.1146/annurev.soc.34.040507.134601]
18. Ackerson LK, Viswanath K. The social context of interpersonal communication and health. J Health Commun 2009;14 Suppl 1:5-17. [doi: 10.1080/10810730902806836] [Medline: 19449264]
19. Brundage MD, Feldman-Stewart D, Tishelman C. How do interventions designed to improve provider-patient communication work? Illustrative applications of a framework for communication. Acta Oncol 2010;49(2):136-143. [doi: 10.3109/02841860903483684] [Medline: 20100151]
20. Kontos E, Emmons K, Puleo E, Viswanath K. Communication inequalities and public health implications of adult social networking site use in the United States. J Health Comm 2010;15(3):216-235. [Medline: 21154095]
21. Kontos EZ, Emmons KM, Puleo E, Viswanath K. Determinants and beliefs of health information mavens among a lower-socioeconomic position and minority population. Soc Sci Med 2011 Jul;73(1):22-32 [FREE Full text] [doi: 10.1016/j.socscimed.2011.04.024] [Medline: 21683493]
22. Valente TW, Pumpuang P. Identifying opinion leaders to promote behavior change. Health Educ Behav 2007 Dec;34(6):881-896. [doi: 10.1177/1090198106297855] [Medline: 17602096]
23. Berkman L, Glass T. Social integration, social networks, social support,health. In: Berkman LF, Kawachi I, editors. Social Epidemiology. New York: Oxford University Press; 2000:137-173.
24. Cantor D, Covell J, Davis T, Park I, Rizzo L. Health Information National Trends Survey 2005 (HINTS 2005): Final Report. Bethesda, MD: National Cancer Institute; 2005.
25. Väänänen A, Kouvonen A, Kivimäki M, Pentti J, Vahtera J. Social support, network heterogeneity, and smoking behavior in women: the 10-town study. Am J Health Promot 2008;22(4):246-255. [doi: 10.4278/0701094R1.1] [Medline: 18421889]
26. Sayers SL, Riegel B, Pawlowski S, Coyne JC, Samaha FF. Social support and self-care of patients with heart failure. Ann Behav Med 2008 Feb;35(1):70-79. [doi: 10.1007/s12160-007-9003-x] [Medline: 18347906]
27. Christakis NA, Fowler JH. The collective dynamics of smoking in a large social network. N Engl J Med 2008 May 22;358(21):2249-2258. [doi: 10.1056/NEJMsa0706154] [Medline: 18499567]
28. Christakis N, Fowler J. The spread of obesity in a large social network over 32 years. N Engl J Med 2007 Jul 26;357(4):370-379. [doi: 10.1056/NEJMsa066082] [Medline: 17652652]
29. Granovetter M. Strength of weak ties. Am J Sociol 1973;78(6):1360-1380.
30. Pettigrew K. Lay information provision in community settings: How community health nurses disseminate human services information to the elderly. Libr Quart 2000;70(1):47-85.
31. Fisher K, Naumer C, Durrance J, Stromski L, Christiansen T. Something old, something new: preliminary findings from an exploratory study about people's information habits and information grounds. Inform Res 2005;10(2):53-74 [FREE Full text]
32. Ellison N, Steinfield C, Lampe C. The benefits of Facebook ?friends:? Social capital and college students? use of online social network sites. J Compute-Mediat Comm 2007;12(4):1143-1168. [doi: 10.1111/j.1083-6101.2007.00367.x]
33. Burke M, Kraut R. Using Facebook after losing a job: Differential benefits of strong and weak ties. 2013 Feb Presented at: CSCW '13 2013 conference on Computer Supported Cooperative Work; Feb 23-27, 2013; San Antonio, TX.
34. Heidelberger C, El-Gayar O, Sarnikar S. Online health social networks and patient health decision behavior: A research agenda. 2011 Presented at: HICSS 44th Hawaii International Conference on System Sciences; Jan 4-7, 2011; Waikoloa, HI.
35. Choi N. Relationship between health service use and health information technology use among older adults: analysis of the US National Health Interview Survey. J Med Internet Res 2011;13(2):e33. [doi: 10.2196/jmir.1753]
36. Flash Eurobarometer 404: European Citizens' Digital Health Literacy. 2014. URL: http://ec.europa.eu/public_opinion/flash/fl_404_en.pdf [accessed 2017-03-17] [WebCite Cache ID 6p2GKIWkA]
37. Mesch GS. Ethnic origin and access to electronic health services. Health Inform J 2015 Aug 10;22(4):791-803. [doi: 10.1177/1460458215590863]
38. Ellison N, Vitak J, Gray R, Lampe C. Cultivating social resources on social network sites: Facebook relationship maintenance behaviors and their role in social capital processes. J Comput Mediat Comm 2013;19(4):855-870.
39. Kawachi I, Berkman L. Social ties and mental health. J Urban Health 2001;78(3):458-467. [doi: 10.1093/jurban/78.3.458]
40. Putnam R. Bowling Alone: America's declining social capital. J Democr 1995;6(1):65-78.
41. Putnam R. Bowling Alone: The Collapse and Revival of the American Community. New York: Simon & Schuster; 2000.
42. Bourdieu P. The forms of capital. In: Richardson JG, editor. Handbook of Theory and Research for the Sociology of Education. Westport, CT: Greenwood Press; 1986:241-258.
43. Coleman JS. Social capital in the creation of human capital. Am J Sociol 1988 Jan;94:S95-S120. [doi: 10.1086/228943]

44. Nahapiet J, Ghoshal S. Social capital, intellectual capital, and the organizational advantage. Acad Manage Rev 1998 Apr;23(2):242-266. [doi: 10.2307/259373]
45. Lin N. Social Capital: A Theory of Social Structure and Action. Cambridge, MA: Cambridge University Press; 2001.
46. Cohen S. Issues in the study and application of social support. In: Social Support and Health. San Francisco, CA: Academic Press; 1985:3-22.
47. Cohen S. Social relationships and health. Am Psychol 2004;59(8):676-684. [doi: 10.1037/0003-066x.59.8.676]
48. Uphoff E, Pickett K, Cabieses B, Small N, Wright J. A systematic review of the relationships between social capital and socioeconomic inequalities in health: a contribution to understanding the psychosocial pathway of health inequalities. Int J Equity Health 2013 Jul 19;12:54 [FREE Full text] [doi: 10.1186/1475-9276-12-54] [Medline: 23870068]
49. Kayahara J, Wellman B. Searching for culture-high and low. J Comput-Mediat Comm 2007;12(3):824-845. [doi: 10.1111/j.1083-6101.2007.00352.x]
50. Tepper S, Hargittai E, Touve D. Music, mavenstechnology. In: Ivey B, Tepper T, editors. Engaging Art: The Next Great Transformation in America's Cultural Life. New York: Routledge; 2008:199-200.
51. Wu F, Huberman B, Adamic L, Tyler J. Information flow in social groups. Physica A 2004 Jun;337(1-2):327-335. [doi: 10.1016/j.physa.2004.01.030]
52. Aral S, Muchnik L, Sundararajan A. Distinguishing influence-based contagion from homophily-driven diffusion in dynamic networks. Proc Natl Acad Sci U S A 2009 Dec 22;106(51):21544-21549 [FREE Full text] [doi: 10.1073/pnas.0908800106] [Medline: 20007780]
53. Woodside A, Davenport J. The effect of salesman similarity and expertise on consumer purchasing behavior. J Marketing Res 1974;11(2):198-202.
54. McPherson M, Smith-Lovin L, Cook J. Birds of a feather: homophily in social networks. Annu Rev Sociol 2001 Aug;27(1):415-444. [doi: 10.1146/annurev.soc.27.1.415]
55. Potvin-Boucher J, Szumilas M, Sheikh T, Kutcher S. Transitions: A Mental Health Literacy Program for Postsecondary Students. J Coll Student Dev 2010;51(6):723-727. [doi: 10.1353/csd.2010.0014]
56. Freedman A, Miner K, Echt K, Parker R, Cooper HL. Amplifying diffusion of health information in low-literate populations through adult education health literacy classes. J Health Commun 2011;16 Suppl 3:119-133. [doi: 10.1080/10810730.2011.604706] [Medline: 21951247]
57. Magnezi R, Grosberg D, Novikov I, Ziv A, Shani M, Freedman L. Characteristics of patients seeking health information online via social health networks versus general Internet sites: a comparative study. Inform Health Soc Care 2015 Mar;40(2):125-138. [doi: 10.3109/17538157.2013.879147] [Medline: 24475937]
58. Kawachi I, Subramanian SV, Kim D. Social Capital and Health. New York: Springer; 2008.
59. Murayama H, Fujiwara Y, Kawachi I. Social capital and health: a review of prospective multilevel studies. J Epidemiol 2012;22(3):179-187 [FREE Full text] [Medline: 22447212]
60. Poortinga W. Social capital: an individual or collective resource for health? Soc Sci Med 2006 Jan;62(2):292-302. [doi: 10.1016/j.socscimed.2005.06.008] [Medline: 16024152]
61. Ziersch A, Baum F. Involvement in civil society groups: is it good for your health? J Epidemiol Community Health 2004 Jun;58(6):493-500 [FREE Full text] [Medline: 15143118]
62. Arneil B. Diverse Communities: The Problem with Social Capital. Cambridge, UK: Cambridge University Press; 2006.
63. Baron-Epel O, Kaplan G. Can subjective and objective socioeconomic status explain minority health disparities in Israel? Soc Sci Med 2009 Nov;69(10):1460-1467. [doi: 10.1016/j.socscimed.2009.08.028] [Medline: 19765878]
64. Baron-Epel O, Weinstein R, Haviv-Mesika A, Garty-Sandalon N, Green M. Individual-level analysis of social capital and health: a comparison of Arab and Jewish Israelis. Soc Sci Med 2008 Feb;66(4):900-910. [doi: 10.1016/j.socscimed.2007.10.025] [Medline: 18158209]
65. Earp JA, Flax VL. What lay health advisors do: an evaluation of advisors' activities. Cancer Pract 1999;7(1):16-21. [Medline: 9892999]
66. Cheong PH. Health communication resources for uninsured and insured Hispanics. Health Commun 2007;21(2):153-163. [doi: 10.1080/10410230701307188] [Medline: 17523861]
67. Vanderpool R, Kornfeld J, Rutten L, Squiers L. Cancer information-seeking experiences: the implications of Hispanic ethnicity and Spanish language. J Cancer Educ 2009;24(2):141-147 [FREE Full text] [doi: 10.1080/08858190902854772] [Medline: 19431032]
68. Choi NG, Dinitto DM. The digital divide among low-income homebound older adults: Internet use patterns, eHealth literacy, and attitudes toward computer/Internet use. J Med Internet Res 2013 May 02;15(5):e93 [FREE Full text] [doi: 10.2196/jmir.2645] [Medline: 23639979]
69. Kontos E, Blake KD, Chou WS, Prestin A. Predictors of eHealth usage: insights on the digital divide from the Health Information National Trends Survey 2012. J Med Internet Res 2014 Jul 16;16(7):e172 [FREE Full text] [doi: 10.2196/jmir.3117] [Medline: 25048379]
70. Mano RS. Social media and online health services: a health empowerment perspective to online health information. Comput Hum Behav 2014 Oct;39:404-412. [doi: 10.1016/j.chb.2014.07.032]

71. van Deursen AJ, van Dijk JA. The digital divide shifts to differences in usage. New Media Soc 2013 Jun 07;16(3):507-526. [doi: 10.1177/1461444813487959]
72. Schradie J. The digital production gap: the digital divide and Web 2.0 collide. Poetics 2011 Apr;39(2):145-168. [doi: 10.1016/j.poetic.2011.02.003]
73. Jaccard J, Wan CK, Turrisi R. The detection and interpretation of interaction effects between continuous variables in multiple regression. Multivariate Behav Res 1990 Oct 01;25(4):467-478. [doi: 10.1207/s15327906mbr2504_4] [Medline: 26820822]
74. Faul F, Erdfelder E, Buchner A, Lang A. Statistical power analyses using G*Power 3.1: tests for correlation and regression analyses. Behav Res Methods 2009 Nov;41(4):1149-1160. [doi: 10.3758/BRM.41.4.1149] [Medline: 19897823]
75. Cohen J. Statistical power analysis. Curr Dir Psychol Sci 1992;1(3):98-101.
76. Cohen J, Cohen P, West S, Aiken L. Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences. New York: Routledge; 2013.
77. Witte J, Mannon SE. The Internet Social Inequalities. New York: Routledge; 2010.
78. Bodie GD, Dutta MJ. Understanding health literacy for strategic health marketing: eHealth literacy, health disparities, and the digital divide. Health Mark Q 2008;25(1-2):175-203. [doi: 10.1080/07359680802126301] [Medline: 18935884]
79. Egbert N, Nanna KM. Health literacy: challenges and strategies. Online J Iss Nurs 2009;14(3):e1 [FREE Full text] [doi: 10.3912/OJIN.Vol14No03Man01]
80. Nutbeam D. Health literacy as a public health goal: a challenge for contemporary health education and communication strategies into the 21st century. Health Promot Int 2000;15(3):259-267. [doi: 10.1093/heapro/15.3.259]
81. Ratzan SC. Health literacy: communication for the public good. Health Promot Int 2001 Jun;16(2):207-214. [Medline: 11356759]
82. Kickbusch IS. Health literacy: addressing the health and education divide. Health Promot Int 2001 Sep;16(3):289-297. [Medline: 11509466]
83. Park N, Lee S, Kim JH. Individuals' personal network characteristics and patterns of Facebook use: A social network approach. Comput Hum Behav 2012 Sep;28(5):1700-1707. [doi: 10.1016/j.chb.2012.04.009]
84. Wellman B. Computer networks as social networks. Science 2001 Sep 14;293(5537):2031-2034. [doi: 10.1126/science.1065547] [Medline: 11557877]
85. Hagger M, Orbell S. A meta-analytic review of the common-sense model of illness representations. Psychol Health 2003 Jan;18(2):141-184. [doi: 10.1080/0887044031000081321]
86. van Deursen AJ, Van Dijk J. Internet skills and the digital divide. New Media Soc 2010;13(6):893-911. [doi: 10.1177/14614448103867740]
87. McCray AT. Promoting health literacy. J Am Med Inform Assoc 2005 Nov;12(2):152-163 [FREE Full text] [doi: 10.1197/jamia.M1687] [Medline: 15561782]

Abbreviations

CMC: computer-mediated communication
PCI: Palestinian citizens of Israel
RDD: random-digital dial

Edited by A Moorhead; submitted 09.08.16; peer-reviewed by R van der Vaart, R Haase, E Chavarria, A Nunn, J Nguyen, T Ginossar, B Chaney; comments to author 03.09.16; revised version received 07.12.16; accepted 28.02.17; published 30.03.17.

Please cite as:
Hayat T(, Brainin E, Neter E
With Some Help From My Network: Supplementing eHealth Literacy With Social Ties
J Med Internet Res 2017;19(3):e98
URL: http://www.jmir.org/2017/3/e98/
doi:10.2196/jmir.6472
PMID:28360024

Original Paper

Impact of Social Processes in Online Health Communities on Patient Empowerment in Relationship With the Physician: Emergence of Functional and Dysfunctional Empowerment

Gregor Petrič[1], PhD; Sara Atanasova[1], MA; Tanja Kamin[2], PhD

[1]Centre for Methodology and Informatics, Faculty of Social Sciences, University of Ljubljana, Ljubljana, Slovenia

[2]Centre for Social Psychology, Faculty of Social Sciences, University of Ljubljana, Ljubljana, Slovenia

Corresponding Author:
Gregor Petrič, PhD
Centre for Methodology and Informatics
Faculty of Social Sciences
University of Ljubljana
Kardeljeva ploščad 5
Ljubljana, 1000
Slovenia
Phone: 386 15805361
Fax: 386 15805101
Email: gregor.petric@fdv.uni-lj.si

Abstract

Background: Substantial research demonstrates the importance of online health communities (OHCs) for patient empowerment, although the impact on the patient-physician relationship is understudied. Patient empowerment also occurs in relationship with the physician, but studies of OHCs mostly disregard this. The question also remains about the nature and consequences of this empowerment, as it might be based on the limited validity of some information in OHCs.

Objective: The main purpose of this study was to examine the impact of social processes in OHCs (information exchange with users and health professional moderators, social support, finding meaning, and self-expressing) on functional and dysfunctional patient empowerment in relationship with the physician (PERP). This impact was investigated by taking into account moderating role of eHealth literacy and physician's paternalism.

Method: An email list–based Web survey on a simple random sample of 25,000 registered users of the most popular general OHC in Slovenia was conducted. A total of 1572 respondents completed the survey. The analyses were conducted on a subsample of 591 regular users, who had visited a physician at least once in the past 2 years. To estimate the impact of social processes in OHC on functional and dysfunctional PERP, we performed a series of hierarchical regression analyses. To determine the moderating role of eHealth literacy and the perceived physician characteristics, interactions were included in the regression analyses.

Results: The mean age of the respondents in the sample was 37.6 years (SD 10.3) and 83.3% were females. Factor analyses of the PERP revealed a five-factor structure with acceptable fit (root-mean-square error of approximation =.06). Most important results are that functional self-efficacy is positively predicted by information exchange with health professional moderators (beta=.12, *P*=.02), information exchange with users (beta=.12, *P*=.05), and giving social support (beta=.13, *P*=.02), but negatively predicted with receiving social support (beta=−.21, *P*<.001). Functional control is also predicted by information exchange with health professional moderators (beta=.16, *P*=.005). Dysfunctional control and competence are inhibited by information exchanges with health professionals (beta=−.12, *P*=.03), whereas dysfunctional self-efficacy is inhibited by self-expressing (beta=−.12, *P*=.05). The process of finding meaning likely leads to the development of dysfunctional competences and control if the physician is perceived to be paternalistic (beta=.14, *P*=.03). Under the condition of high eHealth literacy, the process of finding meaning will inhibit the development of dysfunctional competences and control (beta=−.17, *P*=.01).

Conclusions: Social processes in OHCs do not have a uniform impact on PERP. This impact is moderated by eHealth literacy and physician paternalism. Exchanging information with health professional moderators in OHCs is the most important factor for stimulating functional PERP as well as diminishing dysfunctional PERP. Social support in OHCs plays an ambiguous role, often making patients behave in a strategic, uncooperative way toward physicians.

(J Med Internet Res 2017;19(3):e74) doi:10.2196/jmir.7002

KEYWORDS

physician-patient relations; patient empowerment; patient compliance; conflict; online health community; eHealth literacy; cross-sectional survey

Introduction

State of Research on Role of Online Health Communities in Patient Empowerment in Relationship With the Physician

Online health communities (OHCs), one of the most important eHealth Internet applications [1], play an important role in the process of transformation from traditional physician-centered relationships to patient-centered relationships [2,3]. This process is often hailed as the emergence of the empowered patient [4,5]. OHCs such as PatientsLikeMe, WebMD, and MedHelp are environments that can provide more relevant health information than search engines [6]. OHCs allow users to share health-related experiences, exchange social support, compare information on health issues, find meaning, and have discussions with online health professional moderators who provide health consultations and clinical expertise to the users [7-11]. These processes lead to individual and collective empowerment [10,12,13] and may consequently change the dynamics of patient-physician relationships [4,13-15]. In OHCs, users can engage in physician previsit activities, such as reading posts and asking questions [16]. There exists an association among exchanging social support, comparing with others, sharing experiences and expressing greater confidence in relationships with physicians [4,13,15]. In one study, 60.3% of active users reported that their use of OHCs affected their relationships with their physicians [4].

Although the above-mentioned studies provide us valuable information regarding the impacts of OHCs on patient empowerment in relationship with the physician (PERP), little is known about how some potentially problematic facets of patient empowerment in OHCs [6,17,18] are associated with patient's relationship with the physician. Even though concerns have been expressed about access to misinformation and patients' inability to understand, evaluate and process relevant information under conditions of low eHealth literacy [19], studies have shown only positive effects on patient-physician relationship from using OHCs. This is due to the fact that previous studies understand PERP as a somewhat narrow concept and measure it mostly with single items. Being prepared for the visit [16] and having the intention to actively communicate with a physician [4,15] are important facets of PERP; however, this concept needs to be expanded to more systematic and thorough observation of the positive and negative facets of participation, control, confidence, efficacy, and patient skills in this relationship. In addition, the following factors need to be considered in the scope of the concept such as patient empowerment does not necessarily lead to better communication with the physician [20], has limitations [5], can be unproductive for the relationship, can lead to negative encounters [6,21] and nonadherence [18], may induce conflicts, and can enable manipulations and even revenge [22,23].

Functional and Dysfunctional Empowerment of Patient in Relationship With Physician

In this study, we propose that the positive and negative aspects of PERP can be systematically conceptualized by leaning on the theory of psychological empowerment by Zimmerman [24] and the distinction between communicative and strategic orientation toward partner in interaction by Habermas [25]. In general, we can differentiate between two types of patient empowerment. One focuses on the individual patient and his or her personal transformation, whereas the other occurs in the context of the patient-physician relationship [26]. This distinction can be observed in many conceptualizations and measurements of patient empowerment, but is usually not manifested as such [27]. Zimmerman dimensions of psychological empowerment (gaining self-efficacy, sense of control, and competences) were understood in subsequent research exclusively as individual characteristics, yet he claimed that they also refer to qualities of social interaction and communication [24]. Therefore, we propose that these components can also be applied at the level of communicative behavior of the patient in relation to his or her physician. We can synthesize that self-efficacy refers to a patient's confidence and ability to achieve his or her goals in an encounter with the physician and patient's awareness of the outcomes he or she wants from the interaction [28]. Patient's control in a relationship with the physician signifies the intention and ability to participate in shared decision making [26,27,29] and the development of tailored treatment plans [30]. Competences pertain to a self-assessed mastery in accomplishing tasks and coping with role-related situations [24,31], as well as gaining the skills and abilities needed to have a meaningful discussion with a physician [7,15,30,32].

Therefore, PERP is a form of communicative orientation and behavior that is an important determinant of a relationship with the physician and of high-quality health care [20,33]. This communicative behavior, however, is not always positive or productive for the relationship, as it is often reminded that changing power positions and associated processes can result in the emergence of conflicts between a patient and his or her physician [9,19,34]. This can result in patient-physician distrust, the patient being perceived as difficult, the patient becoming overconfident, or the patient provoking and taking an aggressive stance toward the physician [22,23,35,36]. We suggest that there are essentially two different forms of orientation toward the physician, which can result in PERP that can have functional or dysfunctional consequences.

This assumption is based on the main premise of one of the most influential sociological theories: the theory of communicative action [25]. This theory suggests that a person in an intersubjective situation can generally relate to other person in a strategic or a communicative way. A strategic orientation pursues maximizing the effectiveness of influencing the actions,

decisions, and expectations of other involved in an interaction. Conversely, a communicative orientation denotes social actions in which actors coordinate themselves and attain their goals based on mutual efforts to achieve understanding [25]. In a strategic orientation, patient acts exclusively toward his or her own gains and considers the physician as means for accomplishing his or her own goals. Such a patient does not acknowledge the interests of the physician, which results in interactions that can be misleading, insincere, disrespectful, and dominating. The power that patient gains from strategic (inter)action is manifested in forms of domination, force, and coercion and is thus an illegitimate power with dysfunctional consequences for the relationship [12]. In contrast, the power emanating from communicative (inter)action can be described as a legitimate form of power that is directed at reaching agreement.

Table 1. Components of patient empowerment in a relationship with the physician (PERP).

Orientation to physician	Self-efficacy	Control	Competence
Communicative orientation	Prepared to have a collaborative and efficient encounter; aims for efficient communication; is able to attract attention	Has the confidence to make propositions and express doubts; willing to communicate information; willing to participate and decide	Is able to understand physician; is able to retain and repeat information; is able to describe symptoms
Strategic orientation	Deceives physician with intention of getting desired medicine, getting more sick leave than needed or simulating disease	Expresses aggression and hostility; disregards advice; intentionally undermines authority	Provokes with information; exhibits overconfidence; disregards professional knowledge; assumes authoritative role

If we merge Zimmerman dimensions of empowerment and the two types of patient orientations, we arrive at the typology of 6 components of the PERP, of which 3 are forms of functional empowerment and 3 are forms of dysfunctional empowerment (Table 1). A functional PERP appears when patients exert self-efficacy, control, and competences in relationship with the physician based on a communicative orientation, which will likely lead to collaborative, mutual, and open communication that is beneficial for all the parties involved [33,37]. Empowerment in such cases enables a productive partnership with a physician that leads to positive health outcomes [4]. Conversely, a dysfunctional PERP appears when self-efficacy, control, and competences are exerted in a strategic way in which the physician is no more than a means to the patient's end [38]. This opens the possibilities for the patient to perform manipulative tasks to obtain the treatment that he or she wants without negotiating with the physician such as (1) to exert anger and rejection [39], (2) to intentionally undermine authority [40], (3) to simulate disease, or (4) to assume an authoritative role [5]. A dysfunctional PERP is characterized by the patient's strategic orientation, which may lead to disruptions in the relationship with the physician and even its breakdown. It may also lead eventually to severe, problematic health outcomes for the patient.

Aim of Research

The main goal of this study was to investigate how social processes, in which patients are involved in OHCs (such as exchanging information with other users and health professional moderators, exchanging social support, finding meaning and recognition, self-expression), are associated with patient's functional and dysfunctional PERP. An investigation of this association needs to consider the moderating role of dominant factors of the patient-physician relationship, such as the physician's willingness to give up the paternalistic role [1,9,19] and the level of eHealth literacy, which limits patient empowerment and elicit conflicts with physicians [5]. The research questions of this study are as follows:

RQ_1: What is the impact of social processes in OHCs on functional and dysfunctional PERP?

RQ_2: Do eHealth literacy and perceived physician characteristics moderate the impact of social processes in OHCs on functional and dysfunctional PERP?

Methods

Procedure and Participants

The data for this study came from a Web survey of users of Med.Over.Net (MON), the largest OHC in Slovenia. MON was established in 2000 and offers around 200 online discussion forums. Most of these are moderated by different types of voluntary moderators, among whom are around 150 health-related professionals. In general, the studied OHC covers the following three types of online interactional spaces: (1) online counseling forums in which health professional moderators answer user queries, (2) social support groups forums focused on specific symptoms or health conditions, and (3) general social forums dedicated to topics that are indirectly associated with health issues (parenting, food, relationships, etc). MON registers more than 400,000 visits monthly and has more than 70,000 registered users.

This study was conducted in collaboration with the providers of MON as a part of their annual survey on user experiences and satisfaction with the OHC. The survey, in which respondents participate voluntarily and anonymously, was administered during June 2016 by the OHC provider, who followed ethical standards for administering scientific surveys. The OHC provider invited potential respondents to participate in the Web survey via its email newsletter service. After clicking the link for the Web survey in the email, potential respondents were taken to an informed consent Web page with information about

the purpose of research and the length of the survey, an assurance that the data would be dealt with in accordance with national and EU laws, information on who the investigator was, a contact information and a statement that they were under no obligation to participate, and that the aggregated results may be published.

After giving their informed consent and clicking 'Next' button respondents could start to fill in the survey. The survey was conducted on platform english.1ka.si, which has mechanisms that disallow multiple entries by the same users. MON is a reputable Web service that treats all personal information (emails) in accordance with national and EU laws and protects data with standard security procedures, which include the deidentification of locally held data files, physical protection of hardware and, strong password protection. The authors of this study had no access to the emails of respondents and received an anonymized dataset containing no identifiable personal information. As per the code of ethics for researchers at the University of Ljubljana [41], no institutional ethics approval was needed for this retrospective type of study. All research was conducted in line with the WMA Declaration of Helsinki on ethical principles for medical research involving human subjects.

The OHC provider first designed a random sample of 40,000 users from the list of all registered users who visited MON at least once in past 6 years. Approximately, 25,000 of these users were randomly assigned to the first Web survey used for this study, whereas the remaining 15,000 users were assigned to a second survey, which mostly focused on service quality and did not provide data for this study. Of approximately 25,000 potential respondents, 4106 (16.42%) clicked the link to the Web survey and 2587 (10.35% participation rate) viewed the informed consent page and clicked the button to start the survey. Of these 1572 finished the survey, which lead to a 60.77% completion rate. The survey took 15 minutes in average. The total response rate of 6.29% (1572/25,000) is small, but not uncommon for Web surveys of this length [42]. The analyses were performed on a subsample of respondents (n=656), who had an encounter with a physician in an ordination at least once in the past 2 years and had visited the forums at MON at least once in the past month. After the exclusion of unit nonresponse, the final sample for analyses contained 591 respondents.

The sample consisted of 16.7% men and 83.3% women (Table 2), whereas the gender structure of the whole portal that hosts the OHC is 70.5% female and 29.5% male, according to Google Analytics. The sample is overrepresented by females, which is likely due to health-related online support groups—which present an important part of the studied OHC—being used predominantly by females, as reported in a review study [43]. Respondents ranged in age from 14 to 74 years (mean 37.6, SD 10.3). More than half (62.9%) of the respondents had at least a college degree, a large majority (74.6%) was married or de facto married and 66.3% were employed or self-employed. In the past 2 years, 36.9% of respondents had visited a physician 7 or more times and 29.9% had had up to 3 visits. In total, 64.0% of respondents reported most often visiting a family or personal physician and 36.0% of respondents reported most often visiting a specialist. Additionally, 41.6% of respondents claimed to have a chronic or acute disease.

Table 2. Sample characteristics.

Variable		n (%)
Gender		
	Male	492 (16.7)
	Female	99 (83.3)
Education		
	Lower	45 (7.6)
	Middle	174 (29.5)
	Higher	372 (62.9)
Labor market status		
	School-age youth	53 (9.0)
	Worker, farmer	392 (66.3)
	Retired, unemployed, disabled	131 (22.2)
	Other	15 (2.5)
Marital status		
	Married or de facto married	441 (74.6)
	Single, divorced, widowed	150 (25.4)
Chronic or acute disease		
	Yes	246 (41.6)
	No	345 (58.4)
Physician most visited in past 2 years		
	Family or personal physician	378 (64.0)
	Specialist	213 (36.0)
Frequency of visiting physician in past 2 years		
	Up to 3 visits	177 (29.9)
	3-7 visits	196 (33.2)
	More than 7 visits	218 (36.9)
Total		591 (100)

Measures

Functional and Dysfunctional PERP

The theoretical background for developing items is summarized in Table 1. The initial pool of items was developed by the authors of this study by drawing broadly on two sorts of studies. One on hand, we adopted items from existing measurement instruments that tap aspects of self-efficacy [28,44], competence [3,29], and control [15,29,30,33,45] in patient-physician relationship. On the contrary, several items were newly developed, especially for dysfunctional components of PERP. In developing these items, we strove to achieve high content validity by relying on studies that discuss at least implicitly essential elements of functional and dysfunctional self-efficacy [5,27,34], competence [5,7,15,24,30,31,32], and control [26,27,29,30,40] in relationship with the physician.

Following the standard procedure for scale construction [46], 3 experts (1 in social science methodology, 1 in health communication, 1 one in Internet studies) evaluated an initial pool of 101 items for content validity. On this basis, a refined set of 40 items was selected. These items were put into a survey system and evaluated for clarity, readability, and sensitivity by 5 postgraduate students in Social Informatics trained in survey design and item development. Upon receiving their feedback, we further reduced the item set to 30 altogether. After excluding highly skewed items and items with very low communality (<.2), our exploratory factor analyses on 22 items unveiled 5 latent factors that overlap highly with theoretical components, and in total explain 42.67% of variability in items. Only the dimensions of dysfunctional competences and dysfunctional control were confounded in a single factor. Confirmatory factor analysis demonstrated an acceptable fit for the five-factor model (Root-mean-square error of approximation=.06, standardized root-mean-square residual=.06, Comparative Fit Index=.9). Table 3 presents the factor loadings and Cronbach alphas as measures of reliability. Whereas two dimensions demonstrate satisfactory reliability, the reliability of 3 factors is somewhat below the desired .7. We nevertheless used the scales for further analyses because they are novel and composed of a small

number of items, in which case an internal consistency above .6 is also acceptable [47].

Table 3. Exploratory factor analysis of the functional and dysfunctional patient empowerment in relationship with physician.

Scale items[a]	Fac 1[b]	Fac 2[c]	Fac 3[d]	Fac 4[e]	Fac 5[f]
I'm prepared for the meeting with my physician so that I get as much as possible out of it.	.70[g]				
When having an encounter with the physician, I make an effort to increase the efficiency of the meeting.	.70				
I am able to attract my physician's attention if I notice that he/she has become distracted with something else.	.45				
I understand information that I receive from my physician.		.80			
I am able to recollect what my physician said during the encounter.		.87			
I can describe my symptoms to the physician in a very clear way.		.40			
Occasionally I have trouble understanding my physician's instructions (reversed item)		-.54			
I have the confidence to express possible doubts about the therapy that a physician would recommend			.41		
I do not need to just listen to the doctor; I can also suggest something.			.51		
If I see or read important health-related information, I usually mention it to the physician.			.40		
Even though I do not read medical journals, I believe that I have more knowledge about my health problems than my physician does.				.70	
I like to provoke the physician with information that he/she may not be aware of.				.50	
I diagnose my condition with the help of the Internet, and I go to the physician just to get his/her confirmation of my diagnosis.				.51	
I do not need to ask my physician about instructions for medication or therapy because I am more knowledgable about this than he/she.				.62	
I would rather not ask questions of my physician because I am sure to find better explanations elsewhere (like on the Internet, from my friends, etc)				.66	
Sometimes I do not listen to my physician because I know in advance what he/she is about to say.				.67	
Sometimes I disagree with my physician just to show him/her that he/she is not always right.				.52	
I began to visit my physician more frequently just to complain about his/her previous procedures.				.43	
I know that I could ask my physician more questions, but I probably would not get any useful advice.				.52	
I can convince my physician to approve a longer sick leave for me if necessary.					.70
A physician could not stop me from getting medicine if I really wanted to get it.					.45
I can persuade physician for appointment with (another) specialist even if not needed					.63
Cronbach alpha	.66	.76	.66	.81	.64

[a]All items are on scale ranging from 1-completely disagree to 5-completely agree.

[b]Fac 1 corresponds to the factor functional self-efficacy.

[c]Fac 2 corresponds to functional competences.

[d]Fac 3 corresponds to functional control.

[e]Fac 4 corresponds to dysfunctional competences and control.

[f]Fac 5 corresponds to dysfunctional self-efficacy.

[g]Only factor weights of absolute value equal or larger than .40 are reported.

Social Processes in OHCs

Items were developed on the basis of an empowerment processes scale [10,13] which delineates the main social processes that are important for patient empowerment. The original scale with 29 items measuring 5 dimensions (exchanging informational support, receiving emotional support, giving support, self-expressing, and finding meaning) was supplemented with a measure of information exchange with health professional moderators in the OHC. We identified the latter dimension as OHCs are increasingly involving online health professionals, who might strengthen and qualitatively enrich OHCs' informational support as they filter health-related information against scientific knowledge and thus improve the reliability and objectivity of experiential knowledge in OHCs [8,48,49]. The reliability of the scale was satisfactory because the Cronbach alphas ranged from .80 to .91.

The eHealth Literacy Measure

This eHealth Literacy measure was adopted from the eHEALS Scale [50]. As Norman [51] identified several issues with the eHEALS Scale, we slightly modified some items and reversed a few of them to make the scale less prone to social desirability. The scale failed to demonstrate unidimensionality, but for the analysis we decided to retain a five-item factor, which seemed most representative of the eHealth literacy construct (example item: "I have difficulties separating quality health information from less-quality ones on the Internet" [reversed item]). The scale was on the margin of acceptability (.71).

The Perceived Paternalism of Physician Measure

The perceived paternalism of physician measure was adapted from the STAR Scale [52], which relates to what extent a physician is perceived to be paternalistic or cooperative. A five-item scale (example item: "My physician does not allow me to express my thoughts and opinions") demonstrated satisfactory reliability (.84). Higher values represent a more paternalistic physician, whereas lower values represent a more cooperative one.

Analyses

A series of exploratory factor analyses were conducted to explore the factor structure of the scale to measure functional and dysfunctional PERP and to determine what items of the scale should be retained. Factors were extracted using Principal Axis Factoring with oblimin rotation as we didn't expect orthogonal factor solution. The number of factors was selected on the basis of eigenvalues higher than 1. This decision was also supported by inspection of the scree plot. The obtained factor solution was put into a confirmatory factor procedure, which resulted in several statistics that estimate goodness of fit of the factor model to the study data. Since the statistics showed good fit of the model, no modifications were needed.

To analyze RQ_1 and RQ_2, we conducted a multiple regression analysis. As our factor analysis revealed 5 dimensions (factors) of dependent variable (PERP), a regression analysis had to be conducted for each dimension separately. More precisely, a hierarchical ordinary least squares multiple regression analysis approach [53] was used. This was conducted in such a way that 3 successive linear regression models were estimated for each of the 5 PERP dimensions. In step one, a model with only control variables (sociodemographics, length of relationship with physician, presence of acute or chronic disease) was estimated. In step two, independent variables (social processes in OHCs) were added, and in step three, two moderating variables (eHealth literacy, perceived physician characteristics) and the interactions between them and social processes in OHCs were added. This procedure allows researcher to test if successive model fits better than the previous one. The comparison of models in step two and models in the step one provided insight into RQ_1, whereas comparison of the models in step three and the models in step two allowed us to analyze RQ_2. All variables that appear in interactions were a priori centered to avoid collinearity. A logarithm of a scale for dysfunctional competences and self-efficacy was used because the original variables were highly skewed.

Results

Descriptive Statistics of Variables in the Model

Among the 5 dimensions of the PERP, the predominant one was functional competences (mean 4.09, SD 0.57). We also noticed an above-average presence of functional self-efficacy (mean 3.96, SD 0.65) and functional control (mean 3.49, SD 0.71) in relationships with the physician. All components of dysfunctional PERP were present to a lesser extent. Comparatively, the one predominantly present was dysfunctional self-efficacy (mean 2.39, SD 0.81), followed by dysfunctional competences and control (mean 1.97, SD 0.65). The latter variable was highly skewed because only 5% of all respondents on average responded to items with agree or completely agree values, which shows that only a small proportion of respondents obtained power by taking an explicitly manipulative, disrespectful, and generally negative stance toward the physician.

Among the social processes, the predominant (Table 4) in the studied OHC was exchanging information with health professional moderators (mean 3.66, SD 0.73), followed by exchanging information with users (mean 3.17, SD 0.67). The processes of receiving social support (mean 2.84, SD 0.93) and

sharing experiences (mean 2.62, SD 1.1.8) were present to a lesser extent. Giving social support was a rare practice among the studied sample (mean 1.91, SD 0.91). Descriptive analyses of moderating variables showed that eHealth literacy was quite dispersed among the sample, with the majority being in the middle range (mean 3.31, SD 1.86). Users also reported that their physicians were, on average, more cooperative than paternalistic (mean 2.34, SD 0.88).

Table 4. Descriptive statistics of dependent, independent, and moderating variables in the model.

Type	Variable	Mean (SD)	Min	Max
Dependent	Functional self-efficacy	3.96 (0.65)	1.3	5
	Functional competences	4.09 (0.57)	2.2	5
	Functional control	3.49 (0.71)	1	5
	Dysfunctional self-efficacy	2.39 (0.81)	1	5
	Dysfunctional competence/control	1.97 (0.65)	1	5
Independent	Exchanging information with users	3.17 (0.67)	1	5
	Exchanging information with health professional moderators	3.66 (0.73)	1	5
	Receiving social support	2.84 (0.93)	1	5
	Giving social support	1.91 (0.91)	1	5
	Sharing experiences	2.62 (1.18)	1	5
	Finding meaning	2.84 (0.90)	1	5
Moderating	eHealth literacy	3.31 (1.86)	0	6
	Physician's paternalism	2.34 (0.88)	1	5

Analysis of Research Questions

Five hierarchical regression analyses were conducted for each one of 5 dependent variables that correspond to the 5 dimensions of PERP. Results of hierarchical regression analyses, where functional PERP is the dependent variable, are reported in Table 5; whereas results of hierarchical regression analyses that pertain to dysfunctional PERP as dependent variable are reported in Table 6. To illustrate an example of regression on functional self-efficacy in relationship with the physician (third column in Table 5), we first entered control variables as predictors in the model in step one. This model did not fit to the data (R^2_{adj}=.002, *P*=.78). In step two, predictors that pertain to social processes in OHC were entered and this model significantly fit the data (R^2_{adj}=.05, *P*<.001). Moreover, in comparison with the model in step one, the increase in R^2 was also significant (note that in SPSS, the difference in R^2 is reported and not the difference in R^2_{adj}), thus suggesting that the model in step two is more valid for interpretation. In step three, moderating variables were entered together with interactions with predictors in step two. A significant increase in R^2 (ΔR^2=.116, *P*<.001) was noted, making the model in step three the most valid for interpretation. Such a hierarchical regression analysis approach was repeated for the other 4 dimensions of PERP, and the results consistently showed that the best-fitting models were those in step three (R^2 adjusted, ranging from .19 to .28), except for the model on dysfunctional self-efficacy where the fit was nonsignificant (R^2_{adj}=.01, *P*=.61). We nevertheless considered regression coefficients for interpretation.

For a detailed analysis of RQ_1, the regression coefficients of the predictors of the model in step two needed to be investigated. Among all social processes on OHC, exchanging information with health professional moderators proved to be the most important factor of PERP. It had a weak but statistically significant impact on functional self-efficacy (beta=.12, *P*=.02) and functional control (beta=.16, *P*=.005), and negative impact on dysfunctional competences and control (beta=−.12, *P*=.03), thus enabling a more cooperative relationship with the physician. Exchanging information with users also proved to be a predictor of functional self-efficacy (beta =.12, *P*=.05) but had no impact on the other two components of functional empowerment. In addition, it demonstrated a positive, if marginally significant, impact on dysfunctional competences (beta=.11, *P*=.08). Significant predictors of functional self-efficacy proved to be self-expressing (beta=.13, *P*=.01) and giving social support (beta=.13, *P*=.02). Self-expressing also had marginally significant impact on competences (beta =.09, *P*=.08) and a negative impact on dysfunctional self-efficacy (beta=−.12, *P*=.05). Interestingly, receiving social support had a significant but negative influence on functional self-efficacy (beta=−.21, *P*<.001), whereas it had a positive impact on dysfunctional competences (beta=.11, *P*=.05).

Table 5. Regression coefficients of independent, contextual, and control variables and interactions on dimensions of functional patient empowerment in relationship with the physician (PERP).

	Predictor	Self-efficacy		Competence		Control	
		beta	*P* value	beta	*P* value	beta	*P* value
Step 1							
	Gender	.06	.17	.02	.65	.03	.58
	Age	.06	.17	.10	.03	.08	.09
	Education	.02	.65	.09	.07	.03	.55
	Chronic or acute disease (0=no,1=yes)	−.04	.38	−.02	.62	−.09	.07
	Length of relationship with physician	−.04	.42	.06	.20	.03	.55
Step 2							
	Exchanging information with users	.12	.05	.01	.85	.09	.10
	Exchanging information with HPM[b]	.12	.02	.05	.36	.16	.005
	Receiving social support	−.21	<.001	−.08	.12	−.09	.13
	Giving social support	.13	.02	−.02	.70	.10	.07
	Finding meaning	.02	.77	−.01	.88	−.06	.37
	Self-expressing	.13	.01	.09	.08	.05	.38
Step 3							
	Physician's paternalism	−.23	<.001	−.25	<.001	−.24	<.001
	eHealth literacy	.18	<.001	.34	<.001	.23	<.001
	Finding meaning X[c] Physician's paternalism[a]			−.12	.06		
	Exchange info with HPM X eHealth literacy[a]					.14	.01
	Finding meaning X eHealth literacy[a]						
	Receiving social support X Physician's paternalism[a]						
R^2_{adj}(step 1)		.002	.46	.027	<.001	.010	.20
R^2_{adj} (step 2)		.050	<.001	.050	<.001	.051	<.001
ΔR^2		.053	<.001	.035	<.001	.054	<.001
R^2_{adj} (step 3)		.200	<.001	.279	<.001	.193	<.001
ΔR^2		.116	<.001	.244	<.001	.161	<.001

[a]Only significant interactions are reported.

[b]HPM: health professional moderators.

[c]X denotes interaction between two variables.

Table 6. Regression coefficients of independent, contextual, and control variables and interactions on dimensions of dysfunctional patient empowerment in relationship with the physician (PERP).

	Predictor	Self-efficacy		Competence and Control	
		beta	*P* value	beta	*P* value
Step 1					
	Gender	−.03	.59	.02	.74
	Age	.02	.84	−.06	.23
	Education	−.01	.83	.01	.92
	Chronic or acute disease (0=no,1=yes)	−.01	.80	−.02	.64
	Length of relationship with physician	−.02	.86	−.02	.68
Step 2					
	Exchanging information with users	.03	.67	.11	.08
	Exchanging information with HPM[b]	−.02	.78	−.12	.03
	Receiving social support	.02	.81	.11	.05
	Giving social support	.01	.89	.07	.19
	Finding meaning	.05	.52	.02	.70
	Self-expressing	−.12	.05	−.08	.15
Step 3					
	Physician's paternalism	.01	.79	.40	<.001
	eHealth literacy	.08	.27	.06	.42
	Finding meaning X[c] Physician's paternalism[a]			.14	.03
	Exchange info with HPM X eHealth literacy[a]	.10	.09		
	Finding meaning X eHealth literacy[a]			−.17	.01
,	Receiving social support X Physician's paternalism[a]			−.11	.08
R^2_{adj}(step 1)		.000	.70	.000	.93
R^2_{adj} (step 2)		.000	.81	.041	.001
ΔR^2		.015	.93	.058	<.001
R^2_{adj} (step 3)		.010	.61	.218	<.001
ΔR^2		.037	.10	.194	<.001

[a]Only significant interactions are reported.

[b]HPM: health professional moderators.

[c]X denotes interaction between two variables.

For a detailed analysis of RQ_2, the regression coefficients of the predictors that were entered in the model in step three needed to be investigated. In all models, except for dysfunctional self-efficacy, the perceived paternalism of the physician had a moderate influence on the functional components of empowerment (beta=−.23 to −.25, *P*<.001) and a strong impact on dysfunctional competences and control (beta=.4, *P*<.001). These estimates showed that the more physicians were perceived to be paternalistic, the less the patients were able to develop functional empowerment and the more they developed dysfunctional competences and control in relation to physicians. The eHealth literacy, as another moderating variable,

demonstrated a significant impact on all 3 components of functional empowerment (beta=.18 to .34, *P*<.001) but not on the dysfunctional ones.

Table 5 also reports interactions between eHealth literacy and social processes in OHC and between the perceived physician characteristics and the social processes in OHC. We noted that respondents who found meaning under the condition of a paternalistic physician were less likely to exhibit functional competences in their relationships (beta=−.12, *P*=.06). Conversely, this condition likely leads to the development of dysfunctional competences and control (beta=.14, *P*=.03). High eHealth literacy intensified the effect of exchanging information with health professional moderators on functional control (beta=.14, *P*=.01), but it surprisingly also led to more dysfunctional self-efficacy (beta=.10, *P*=.09). People who found meaning in the OHC and had high eHealth literacy were less likely to develop dysfunctional competences and control (beta=-.17, *P*=.01). Finally, receiving emotional support under the conditions of a paternalistic physician were less likely to lead to the development of dysfunctional competences and control (beta=−.11, *P*=.08).

Discussion

Emergence of Dysfuntional Empowerment

The main goal of this research was to investigate whether social processes in which users are involved in OHCs have any impact on the patient empowerment experienced in relationships with the physician and, if so, what is the nature of such an impact. Beginning from the limitations of previous studies, we conceptualized the PERP and on this basis we developed a relatively valid and reliable scale. As this is a proposal of a new scale, it has some issues with reliability. Nonetheless, it demonstrates a meaningful factor structure, indicating that PERP should be investigated along at least two dividing lines: (1) two types of communicative orientation of a patient toward the physician and the associated functional or dysfunctional outcomes for the patient-physician relationship and (2) 3 components of patient empowerment in relationship with the physician: self-efficacy, control, and competence. Such conceptualization might be useful for further research on the dynamics of patient-physician relationships in the online era.

To claim that empowered patients are emerging solely due to processes in OHCs would be a gross overstatement, as the results clearly show that it is foremost patient's eHealth literacy and physician's personality and communication style that are driving PERP. However, even after controlling for these 2 determinants, we could show that some social processes in which users of OHCs are involved do have an impact on how patients relate to their physicians.

Getting Clinical and Experiential Knowledge as an Empowering Process

The most important factor pertaining to activities in the OHC was the involvement in exchanging information with health professional moderators. This influenced patients to (1) become more efficient in their relationships with the physician, so that they knew what to ask and were motivated to have collaborative and efficient meetings and (2) get more control in terms of having confidence to ask questions and participate in decisions. The impact on control was attenuated under the condition of high eHealth literacy. Furthermore, respondents exchanging information with health professional moderators were less likely to develop overconfident and strategically self-interested stances toward their physicians. The results reveal, however, that exchanging information with health professional moderators for those with high eHealth literacy can be somewhat problematic. People who know a lot and talk a lot with experts in OHCs might learn communication skills [8] to strategically deal with their physician in order to efficiently strive for their own private, often non-legitimate interests, such as getting medicine no matter what and simulating disease, as manifested by a raised dysfunctional self-efficacy.

Exchanging experiential information with users had a conflicting impact on PERP. On one hand, it had a positive impact on self-efficacy, suggesting that interaction with other users helps develop skills for more efficient communication with physicians. However, it can also lead to overconfidence in one's own competences in relation to the professional competences of the physician, as already suggested in previous research [17,20]. The concerns with accuracy and completeness of information in peer exchanges in OHCs have already been raised in previous studies, but we can only conclude, as does Sillence [54], that the exchange of experiential information is a complex process that does not have a simple unidirectional impact on relationship with the physician.

Problematic Facets of Exchanging Social Support

Our results revealed a surprising finding about the role of receiving social support, something often hailed as one of the most important processes in OHCs in terms of leading to individual empowerment [7,15]. In particular, people who are getting advice, consolation, and other types of emotional support from other users likely report less self-efficacy in relationships with their physicians and more dysfunctional competences. Why would people with support in OHCs become less self-efficient in relationships with their physicians, undermining a chance for meaningful and efficient encounters? And why would people who get support from other users in OHCs be more inclined to become overconfident in relation to their physician, start to distrust them, and develop feelings of superiority over professional knowledge? The regression analysis did not give answers to these questions, but these associations could be attributed to the plausible prevalence of "nurturant" social support [55], which is fitted to people who do not search for true causes of their disease but rather for ways to manage negative feelings and emotions in a short-term manner [56]. In this process, they blame external factors (also physicians) for their inability to solve their health problems [56]. Receiving social support in OHCs can also form expectations of physicians providing support [57] and as the latter are perceived to be too busy for supportive discussions [58], it may be that the patients in this process also develop an aversion toward the physician. This way they become less efficient in relation with the physician and develop uncooperative and generally negative attitudes toward the physician and the professional knowledge he or she represents.

Receiving emotional support does not have negative impacts in all cases. The problematic effects of social support are diminished if a patient's physician is perceived as paternalistic, since social support in OHCs in this case lowers dysfunctional competences and control. This might be connected to the finding that the more the patients exchange social support in OHCs, the less likely it is that they will discuss information found online with the health professionals [4]. We could speculate that those patients whose physicians are perceived to be paternalistic and supposedly unsupportive probably satisfy their need for social support in OHCs, consequently diminishing their expectation of getting social support from their physicians. Relieved of the expectation of a supportive physician, a patient is less likely to develop a negative attitude toward the physician.

Although receiving social support seems to have an ambivalent role in the PERP, giving social support demonstrated a positive impact. Providing other users with help and advice bestowed patient's confidence and competence in asking the right questions to physicians and making meetings more efficient. Those who gave support to other OHC users were also more likely to engage in shared decision making with their physicians. This is in line with the general finding from online community research that people who are active contributors in online environments are usually also more proactive in their offline environment [59]. Writing posts for the purpose of self-expression also had a positive impact on the functional components of PERP, which confirms the results of previous studies [4,15,16]. Self-expression stimulates cooperative self-efficacy, which supports findings that writing stories and expressing disease in safe "testing" environments has positive impacts on a patient's competence and efficiency [7], as well as on the level of his or her relationship with the physician.

Does Finding Meaning Diminish Meaning in Interactions With the Physician?

Patients who found meaning in their life and made sense of their situation with the help of other users and who at the same time perceived their physicians as paternalistic through OHC were less likely to develop competences for more understanding relationships with the physician. As such, they will be more likely to develop dysfunctional competences and control, opening possible conflicts between the 2 partners. One possible explanation for this result might lie in the finding from previous studies that information received from external resources like OHC is often perceived to be superior to that provided by the medical staff [17]. This connects to findings that complementary and alternative medicine fills the gaps that evidence-based medicine leaves behind in terms of providing a more holistic, therapeutic approach with treatments that are easier to understand and follow [60]. Moreover, as exchanges of advice in OHCs are often based on similar beliefs [48] that can be based on opposition to evidence-based medicine [36], it is plausible that users find new meanings and purposes that involve a negative and underestimating attitude toward their physicians.

Finding meaning can on the other hand be a positive empowerment process, as under condition of high eHealth literacy, finding meaning had an inhibiting effect on dysfunctional competences and control. More precisely, people who are digitally literate in health issues and use OHCs to find meaning in exchanges with other users will be less likely to develop a problematic stance toward their physicians.

Although not a central focus, it needs to be pointed out that this study confirms an important role of eHealth literacy in the dynamics of patient-physician relationships [61]. People who successfully encompass a constellation of digital literacy, health literacy, media literacy, and other literacies [50] will also develop competences and skills for a successful relationship with the physician, which enables better understanding and retention of information and improved ability to ask meaningful questions. Furthermore, this characteristic will also allow patients to be more effective in their relationships with their physicians and have more control in terms of participation in shared decision making.

Limitations

This study has some methodological limitations that warrant further research. First, because the respondents in this study were recruited from a single OHC, generalizations of any findings should be made with caution because OHCs can vary greatly in their membership, focus, and role structure [62]. Ideally, the research should be repeated in various national contexts and with different types of OHCs, focusing on different diseases and user needs. Second, although the proposed measure of PERP presents an improvement over single-item measures, the scale is novel and consequently did not go through enough testing, thus rendering it marginally valid and reliable. However, we hope that researchers will recognize its value and work further to improve its psychometric qualities. In connection arises a third limitation that the patient's perception of empowerment in the relationship with the physician is only one part of a dyad, or even triad [39]. To observe the relationship completely, the communication orientation of physicians should be considered, but this would demand a much more complex research design.

Finally, in this study, we proposed a theoretical model in which social processes in OHCs have an impact on PERP, but it needs to be stated that, empirically, the association between variables can also go in the opposite direction. Moreover, the 2 phenomena are probably in a recursive relationship, where processes in OHCs are influencing relationships with the physicians and patients' experiences with physicians can impact how they use OHCs and other online tools. To address this issue, we may have to undertake a random control trial type of research design.

Practical Implications

On the basis of these results, several important messages can be discerned for OHC managers, users, patients, and health care practitioners. First, strong empirical evidence for the importance of health professionals being visibly present in OHCs should stimulate online community managers to invest energy in attracting physicians and other health experts to actively participate and moderate discussions in OHCs. They are not only carers of high-quality information in OHCs, but they also stimulate processes that can ease patients' relationships with their physicians (whether family doctors or specialists) [8].

Second, both patients and physicians should be aware of ambiguous processes in OHCs and their impact not only on patients handling their diseases but also on the fact that patients encounter health experts in different venues—physically in ordination and online in OHCs and other platforms. Physicians should be aware of the processes in OHCs and other experiences in which patients are involved, providing material for arriving at decisions that would lead to better health outcomes. If all these encounters are not discussed, conditions might form that open space for conflicts and strategic action, which in the end is harmful for the patient and physician.

We believe that a better understanding of such aspects of patient-physician relationships is required because it has implications for how physicians are trained and health care is organized. We know that if physicians surrender their control, they can create an atmosphere of respect that facilitates the positive aspects of empowerment [9]. Now we also know that if physicians do not surrender their paternalistic role, empowerment can go astray into its dysfunctional mode. Finally, it seems that although OHCs are environments where patients can get emotional support, this does not mean that patients should not expect supportive communication from their physicians. Our results suggest that some patients prefer what we call passive support, which is composed of compassion and consolation and does not imply any action on the patient's part. However, we could speculate that physicians are more inclined to give patients something we call active support—a sort of compassion but one with proactive components—for which the patient is advised to more actively take care of his or her health. Some patients see these two supports as conflicting, and it might be advisable for OHC managers to emphasize such proactive support so that patients do not fall into a bubble where exclusively external factors are blamed for their current health statuses.

Conclusions

To our knowledge, this is the first study with a quantitative approach to empirically demonstrate the problematic aspects of patient empowerment that should be considered when investigating the impact of a patient's engagement in online environments, both as an information seeker and as an active coproducer of online environments. Patients can feel psychologically empowered, and this sort of acquired internal power can translate into a cooperative relationship with the physician, thus bringing positive outcomes for both. We termed such empowerment functional. However, a patient's feelings of empowerment can also be transformed, especially under conditions of low eHealth literacy and the presence of a paternalistic physician, into a manipulative, disrespectful, and generally negative stance toward the physician. Such empowerment is not productive for the relationship and possibly leads to conflicts as well as worsens the health outlooks for the patient.

Online environments can have many benefits for patients, but unchecked information, under conditions of poor literacy, can lead to social contagion processes [63] that can have dangerous implications for patients and for wider societal processes [64]. Once we are aware of such processes, we can start addressing them. Our findings suggest that they can be inhibited by filtering experiential knowledge through health professional moderators and through raising awareness regarding the importance of eHealth literacy when dealing with online health information. In addition, to achieve true empowerment, this concept should not be understood solely from the perspective of an individual benefits. It also needs to include a perspective of working together for our shared interests and to improve our interactions, communities, and institutions [65].

Acknowledgments

The survey data in this study has been acquired in collaboration with A Verovšek (Institute Med.Over.Net).

Conflicts of Interest

None declared.

References

1. McMullan M. Patients using the Internet to obtain health information: how this affects the patient-health professional relationship. Patient Educ Couns 2006 Oct;63(1-2):24-28. [doi: 10.1016/j.pec.2005.10.006] [Medline: 16406474]
2. Keating NL, Green DC, Kao AC, Gazmararian JA, Wu VY, Cleary PD. How are patients' specific ambulatory care experiences related to trust, satisfaction, and considering changing physicians? J Gen Intern Med 2002 Jan;17(1):29-39 [FREE Full text] [Medline: 11903773]
3. Mazor KM, Roblin DW, Greene SM, Lemay CA, Firneno CL, Calvi J, et al. Toward patient-centered cancer care: patient perceptions of problematic events, impact, and response. J Clin Oncol 2012;30(15):1784-1790. [doi: 10.1200/JCO.2011.38.1384] [Medline: 22508828]
4. Bartlett YK, Coulson NS. An investigation into the empowerment effects of using online support groups and how this affects health professional/patient communication. Patient Educ Couns 2011 Apr;83(1):113-119. [doi: 10.1016/j.pec.2010.05.029] [Medline: 20599338]
5. Schulz PJ, Nakamoto K. Patient behavior and the benefits of artificial intelligence: the perils of "dangerous" literacy and illusory patient empowerment. Patient Educ Couns 2013 Aug;92(2):223-228. [doi: 10.1016/j.pec.2013.05.002] [Medline: 23743214]

6. Kanthawala S, Vermeesch A, Given B, Huh J. Answers to health questions: internet search results versus online health community responses. J Med Internet Res 2016 Apr 28;18(4):e95 [FREE Full text] [doi: 10.2196/jmir.5369] [Medline: 27125622]
7. Barak A, Boniel-Nissim M, Suler J. Fostering empowerment in online support groups. Comput Human Behav 2008 Sep;24(5):1867-1883. [doi: 10.1016/j.chb.2008.02.004]
8. Bientzle M, Griewatz J, Kimmerle J, Küppers J, Cress U, Lammerding-Koeppel M. Impact of scientific versus emotional wording of patient questions on doctor-patient communication in an internet forum: a randomized controlled experiment with medical students. J Med Internet Res 2015;17(11):e268 [FREE Full text] [doi: 10.2196/jmir.4597] [Medline: 26607233]
9. Groen WG, Kuijpers W, Oldenburg HS, Wouters MW, Aaronson NK, van Harten WH. Empowerment of cancer survivors through information technology: an integrative review. J Med Internet Res 2015;17(11):e270 [FREE Full text] [doi: 10.2196/jmir.4818] [Medline: 26614438]
10. Mo P, Coulson NS. Empowering processes in online support groups among people living with HIV/AIDS: a comparative analysis of 'lurkers' and 'posters'. Comput Human Behav 2010 Sep;26(5):1183-1193. [doi: 10.1016/j.chb.2010.03.028]
11. Petrovčič A, Petrič G. Differences in intrapersonal and interactional empowerment between lurkers and posters in health-related online support communities. Comput Human Behav 2014 May;34:39-48. [doi: 10.1016/j.chb.2014.01.008]
12. Petrič G, Petrovčič A. Individual and collective empowerment in online communities: the mediating role of communicative interaction in web forums. Inform Soc 2014 May 12;30(3):184-199. [doi: 10.1080/01972243.2014.896683]
13. van Uden-Kraan CF, Drossaert CH, Taal E, Seydel ER, van de Laar MA. Self-reported differences in empowerment between lurkers and posters in online patient support groups. J Med Internet Res 2008;10(2):e18 [FREE Full text] [doi: 10.2196/jmir.992] [Medline: 18653442]
14. Stewart Loane S, D'Alessandro S. Empowered and knowledgeable health consumers: the impact of online support groups on the doctor–patient relationship. Australasian Marketing Journal (AMJ) 2014 Aug;22(3):238-245. [doi: 10.1016/j.ausmj.2014.08.007]
15. Oh HJ, Lee B. The effect of computer-mediated social support in online communities on patient empowerment and doctor-patient communication. Health Commun 2012;27(1):30-41. [doi: 10.1080/10410236.2011.567449] [Medline: 21797714]
16. Hu X, Bell RA, Kravitz RL, Orrange S. The prepared patient: information seeking of online support group members before their medical appointments. J Health Commun 2012;17(8):960-978. [doi: 10.1080/10810730.2011.650828] [Medline: 22574697]
17. Hewitt-Taylor J, Bond CS. What e-patients want from the doctor-patient relationship: content analysis of posts on discussion boards. J Med Internet Res 2012;14(6):e155 [FREE Full text] [doi: 10.2196/jmir.2068] [Medline: 23137788]
18. Seçkin G, Yeatts D, Hughes S, Hudson C, Bell V. Being an informed consumer of health information and assessment of electronic health literacy in a national sample of internet users: validity and reliability of the e-HLS instrument. J Med Internet Res 2016 Jul 11;18(7):e161 [FREE Full text] [doi: 10.2196/jmir.5496] [Medline: 27400726]
19. Gerber BS, Eiser AR. The patient physician relationship in the Internet age: future prospects and the research agenda. J Med Internet Res 2001;3(2):E15 [FREE Full text] [doi: 10.2196/jmir.3.2.e15] [Medline: 11720957]
20. Tran AN, Haidet P, Street Jr RL, O'Malley KJ, Martin F, Ashton CM. Empowering communication: a community-based intervention for patients. Patient Educ Couns 2004;52(1):113-121. [doi: 10.1016/S0738-3991(03)00002-8] [Medline: 14729298]
21. Snow R, Humphrey C, Sandall J. What happens when patients know more than their doctors? Experiences of health interactions after diabetes patient education: a qualitative patient-led study. BMJ Open 2013;3(11):1-8. [doi: 10.1136/bmjopen-2013-003583] [Medline: 24231459]
22. Broom A. Virtually he@lthy: the impact of internet use on disease experience and the doctor-patient relationship. Qual Health Res 2005 Mar;15(3):325-345. [doi: 10.1177/1049732304272916] [Medline: 15761103]
23. Morgan M. The doctor-patient relationship. In: Scambler G, editor. Sociology As Applied to Medicine. Edinburgh, New York: Saunders/Elsevier; 2008:49-65.
24. Zimmerman MA. Psychological empowerment: issues and illustrations. Am J Community Psychol 1995 Oct;23(5):581-599. [Medline: 8851341]
25. Habermas J. The Theory Of Communicative Action. 1st ed. Boston: Beacon Press; 1984.
26. Jacobson P. Empowering the physician-patient relationship: the effect of the Internet. Partnership 2007 May 29;2(1):1-13 [FREE Full text] [doi: 10.21083/partnership.v2i1.244]
27. Barr PJ, Scholl I, Bravo P, Faber MJ, Elwyn G, McAllister M. Assessment of patient empowerment--a systematic review of measures. PLoS One 2015;10(5):e0126553 [FREE Full text] [doi: 10.1371/journal.pone.0126553] [Medline: 25970618]
28. Capone V, Petrillo G. Patient's Communication Perceived Self-efficacy Scale (PCSS): construction and validation of a new measure in a socio-cognitive perspective. Patient Educ Couns 2014 Jun;95(3):340-347. [doi: 10.1016/j.pec.2014.03.006] [Medline: 24690475]
29. Menon ST. Toward a model of psychological health empowerment: implications for health care in multicultural communities. Nurse Educ Today 2002 Jan;22(1):28-39; discussion 40. [doi: 10.1054/nedt.2001.0721] [Medline: 11886225]

30. Ouschan R, Sweeney J, Johnson L. Customer empowerment and relationship outcomes in healthcare consultations. Eur J Mark 2006 Sep;40(9/10):1068-1086. [doi: 10.1108/03090560610681014]
31. Levesque A, Li HZ, Pahal JS. Factors related to patients' adherence to medication and lifestyle change recommendations: data from Canada. Int J Psychol Stud 2012 May 28;4(2):42-55. [doi: 10.5539/ijps.v4n2p42]
32. Rappaport J. Terms of empowerment/exemplars of prevention: toward a theory for community psychology. Am J Community Psychol 1987 Apr;15(2):121-148. [Medline: 3604997]
33. Dutta-Bergman MJ. The relation between health-orientation, provider-patient communication, and satisfaction: an individual-difference approach. Health Commun 2005;18(3):291-303. [doi: 10.1207/s15327027hc1803_6] [Medline: 16187933]
34. Eysenbach G, Jadad AR. Evidence-based patient choice and consumer health informatics in the Internet age. J Med Internet Res 2001 May;3(2):E19 [FREE Full text] [doi: 10.2196/jmir.3.2.e19] [Medline: 11720961]
35. Goodyear-Smith F, Buetow S. Power issues in the doctor-patient relationship. Health Care Anal 2001;9(4):449-462. [doi: 10.1023/A:1013812802937] [Medline: 11874258]
36. Wilson PM, Kendall S, Brooks F. The Expert Patients Programme: a paradox of patient empowerment and medical dominance. Health Soc Care Community 2007 Sep;15(5):426-438. [doi: 10.1111/j.1365-2524.2007.00701.x] [Medline: 17685988]
37. Roter D. The enduring and evolving nature of the patient-physician relationship. Patient Educ Couns 2000 Jan;39(1):5-15. [Medline: 11013543]
38. Kushner T. Doctor-patient relationships in general practice--a different model. J Med Ethics 1981 Sep 01;7(3):128-131. [doi: 10.1136/jme.7.3.128] [Medline: 7277407]
39. Abbott J. Difficult patients, difficult doctors: can consultants interrupt the “blame game”? Am J Bioeth 2012;12(5):18-20. [doi: 10.1080/15265161.2012.665145] [Medline: 22548517]
40. Hart A, Henwood F, Wyatt S. The role of the Internet in patient-practitioner relationships: findings from a qualitative research study. J Med Internet Res 2004 Sep 30;6(3):e36 [FREE Full text] [doi: 10.2196/jmir.6.3.e36] [Medline: 15471762]
41. Uni-Lj. Code of ethics for researchers of University of Ljubljana URL: https://www.uni-lj.si/university/code_of_ethics/ [accessed 2016-11-15] [WebCite Cache ID 6m2ZtYJmq]
42. Callegaro M, Manfreda K, Vehovar V. Web Survey Methodology. London: Sage Publications; 2015:-344.
43. Coulson N, Malik S. Health-related online support communities. In: Yan Z, editor. Encyclopedia of cyber-behavior. Hershey Pennsylvania: IGI Global; 2012:671-688.
44. Rider T, Malik M, Chevassut T. Haematology patients and the Internet--the use of on-line health information and the impact on the patient-doctor relationship. Patient Educ Couns 2014 Nov;97(2):223-238. [doi: 10.1016/j.pec.2014.06.018] [Medline: 25053474]
45. Kim SC, Kim S, Boren D. The quality of therapeutic alliance between patient and provider predicts general satisfaction. Mil Med 2008 Jan;173(1):85-90. [doi: 10.7205/milmed.173.1.85] [Medline: 18251337]
46. DeVellis RF. Scale Development: Theory and Applications. Thousand Oaks, London, New Delhi: Sage publications; 2003.
47. Loewenthal K, Lewis C. An Introduction to Psychological Tests and Scales. New York: Psychology Press; 2015.
48. Hartzler A, Pratt W. Managing the personal side of health: how patient expertise differs from the expertise of clinicians. J Med Internet Res 2011;13(3):e62 [FREE Full text] [doi: 10.2196/jmir.1728] [Medline: 21846635]
49. Vennik FD, Adams SA, Faber MJ, Putters K. Expert and experiential knowledge in the same place: patients' experiences with online communities connecting patients and health professionals. Patient Educ Couns 2014 May;95(2):265-270. [doi: 10.1016/j.pec.2014.02.003] [Medline: 24598314]
50. Norman CD, Skinner HA. eHealth Literacy: essential skills for consumer health in a networked world. J Med Internet Res 2006 Jun;8(2):e9 [FREE Full text] [doi: 10.2196/jmir.8.2.e9] [Medline: 16867972]
51. Norman C. eHealth literacy 2.0: problems and opportunities with an evolving concept. J Med Internet Res 2011 Dec;13(4):e125 [FREE Full text] [doi: 10.2196/jmir.2035] [Medline: 22193243]
52. McGuire-Snieckus R, McCabe R, Catty J, Hansson L, Priebe S. A new scale to assess the therapeutic relationship in community mental health care: STAR. Psychol Med 2007 Jan;37(1):85-95. [doi: 10.1017/S0033291706009299] [Medline: 17094819]
53. Cohen J, Cohen P, West S, Aiken L. Applied Multiple Regression Analysis for the Behavioral Sciences, 3rd edition. Mahwah, NJ: L Erlbaum Associates; 2003.
54. Sillence E. Giving and receiving peer advice in an online breast cancer support group. Cyberpsychol Behav Soc Netw 2013 Jun;16(6):480-485. [doi: 10.1089/cyber.2013.1512] [Medline: 23659724]
55. Cutrona CE, Suhr JA. Controllability of stressful events and satisfaction with spouse support behaviors. Communic Res 1992 Apr;19(2):154-174. [doi: 10.1177/009365092019002002]
56. Bambina A. Online Social Support: The Interplay of Social Networks and Computer-mediated Communication. New York: Cambria Press; 2007.
57. Boutin-Foster C, Charlson ME. Problematic resident-patient relationships. the patient's perspective. J Gen Intern Med 2001 Nov;16(11):750-754. [doi: 10.1046/j.1525-1497.2001.00803.x] [Medline: 11722689]

58. Rupert DJ, Moultrie RR, Read JG, Amoozegar JB, Bornkessel AS, O'Donoghue AC, et al. Perceived healthcare provider reactions to patient and caregiver use of online health communities. Patient Educ Couns 2014 Sep;96(3):320-326. [doi: 10.1016/j.pec.2014.05.015] [Medline: 24923652]
59. Wellman B, Haase AQ, Witte J, Hampton K. Does the internet increase, decrease, or supplement social capital? Social networks, participation, and community commitment. Am Behav Sci 2001 Nov 01;45(3):436-455. [doi: 10.1177/00027640121957286]
60. Ule M. Spregledana Razmerja. Maribor: Aristej; 2003.
61. Jordan JE, Buchbinder R, Osborne RH. Conceptualising health literacy from the patient perspective. Patient Educ Couns 2010 Apr;79(1):36-42. [doi: 10.1016/j.pec.2009.10.001] [Medline: 19896320]
62. Preece J. Online communities: designing usability, supporting sociability. West Sussex: Wiley; 2000.
63. Hodas NO, Lerman K. The simple rules of social contagion. Sci Rep 2014;4:4343 [FREE Full text] [doi: 10.1038/srep04343] [Medline: 24614301]
64. Pariser E. The Filter Bubble: What the Internet is Hiding from You. London: Penguin Books; 2012:0241954525.
65. Perkins DD, Zimmerman MA. Empowerment theory, research, and application. Am J Community Psychol 1995 Oct;23(5):569-579. [Medline: 8851340]

Abbreviations

MON: Med.Over.Net
OHC: online health community
PERP: patient empowerment in relationship with physician

Edited by K Dinakar; submitted 17.11.16; peer-reviewed by N Coulson, J Apolinário-Hagen, M Bradley; comments to author 22.01.17; revised version received 10.02.17; accepted 15.02.17; published 13.03.17.

Please cite as:
Petrič G, Atanasova S, Kamin T
Impact of Social Processes in Online Health Communities on Patient Empowerment in Relationship With the Physician: Emergence of Functional and Dysfunctional Empowerment
J Med Internet Res 2017;19(3):e74
URL: http://www.jmir.org/2017/3/e74/
doi:10.2196/jmir.7002
PMID:28288953

Original Paper

Readability Formulas and User Perceptions of Electronic Health Records Difficulty: A Corpus Study

Jiaping Zheng[1], MS; Hong Yu[2,3], PhD

[1]College of Information and Computer Sciences, University of Massachusetts, Amherst, MA, United States

[2]Bedford Veterans Affairs Medical Center, Center for Healthcare Organization and Implementation Research, Bedford, MA, United States

[3]Department of Quantitative Health Sciences, University of Massachusetts Medical School, Worcester, MA, United States

Corresponding Author:
Jiaping Zheng, MS
College of Information and Computer Sciences
University of Massachusetts
140 Governors Dr
Amherst, MA,
United States
Phone: 1 4135452744
Fax: 1 4135452744
Email: jzheng@cs.umass.edu

Abstract

Background: Electronic health records (EHRs) are a rich resource for developing applications to engage patients and foster patient activation, thus holding a strong potential to enhance patient-centered care. Studies have shown that providing patients with access to their own EHR notes may improve the understanding of their own clinical conditions and treatments, leading to improved health care outcomes. However, the highly technical language in EHR notes impedes patients' comprehension. Numerous studies have evaluated the difficulty of health-related text using readability formulas such as Flesch-Kincaid Grade Level (FKGL), Simple Measure of Gobbledygook (SMOG), and Gunning-Fog Index (GFI). They conclude that the materials are often written at a grade level higher than common recommendations.

Objective: The objective of our study was to explore the relationship between the aforementioned readability formulas and the laypeople's *perceived* difficulty on 2 genres of text: general health information and EHR notes. We also validated the formulas' appropriateness and generalizability on predicting difficulty levels of highly complex technical documents.

Methods: We collected 140 Wikipedia articles on diabetes and 242 EHR notes with diabetes International Classification of Diseases, Ninth Revision code. We recruited 15 Amazon Mechanical Turk (AMT) users to rate difficulty levels of the documents. Correlations between laypeople's perceived difficulty levels and readability formula scores were measured, and their difference was tested. We also compared word usage and the impact of medical concepts of the 2 genres of text.

Results: The distributions of both readability formulas' scores (P<.001) and laypeople's perceptions (P=.002) on the 2 genres were different. Correlations of readability predictions and laypeople's perceptions were weak. Furthermore, despite being graded at similar levels, documents of different genres were still perceived with different difficulty (P<.001). Word usage in the 2 related genres still differed significantly (P<.001).

Conclusions: Our findings suggested that the readability formulas' predictions did not align with perceived difficulty in either text genre. The widely used readability formulas were highly correlated with each other but did not show adequate correlation with readers' perceived difficulty. Therefore, they were not appropriate to assess the readability of EHR notes.

(J Med Internet Res 2017;19(3):e59) doi:10.2196/jmir.6962

KEYWORDS

electronic health records; readability; patients

Introduction

Background

Patient engagement and effective patient-physician communication are essential in patient-centered care, defined by the Institute of Medicine as "respectful of and responsive to individual patient preferences, needs, and values, and ensuring that patient values guide all clinical decisions" [1]. Electronic health records (EHRs) are a rich resource for developing applications to engage the patients and foster patient activation [2-4]. Thus, allowing patients access to their own EHR records holds a strong potential to enhance patient-centered care. It may improve the understanding of their own clinical conditions and treatments, leading to improved health care outcomes (eg, increased medication adherence [4]).

As patients express interests in reading their own EHR data [5], health care institutions have also begun to open up access to the EHR records [6]. However, EHRs are written by physicians to communicate with other health care professionals [7]. Therefore, EHRs are full of medical jargon, abbreviations, and other domain-specific usages and expressions that are ill-suited for the lay people (patients). One study showed that nearly two-thirds of the surveyed patients considered physicians' notes difficult to understand, and radiology reports and nurses' notes were also perceived as difficult [8]. Another study recruited healthy volunteers to read and retell medical documents [9]. Common retelling errors included misunderstanding clinical concepts and physician's findings during a patient's visit. In a study of electronic primary care records, many patients requested explanations of medical terms and abbreviations [10]. A recent patient survey on Web-based access to laboratory results concluded that test result comprehension still needed improvement [11]. Findings from an assessment of lay understanding of medical terms suggested that a substantial proportion of the lay public did not understand phrases often used in cancer consultations and that knowledge of basic anatomy could not be assumed [12]. In the emergency department setting, patients understood less than 30% of commonly used medical terms [13]. Moreover, the vocabulary gap between professionals and laypeople has motivated a thread of research to develop controlled vocabulary resources [14-16].

Merely providing patients with their own EHR records, therefore, does not necessarily help the patients better understand their own conditions. Further complicating the issue, it is estimated in the National Assessment of Adult Literacy that the average American has a reading level between the 7th and 8th grade [17]. It is also reported in the same assessment that about 36% of the US population or 75 million Americans have basic or below basic health literacy. The opaque narratives in the EHR present a challenge to the average patient.

Electronic Health Records and Readability

Measuring the readability of the EHR notes is one important step toward making the notes accessible to the patients. Numerous studies [18-20] have evaluated the difficulty of health information intended for patient consumption using readability formulas. They conclude that the materials are often written at a grade level higher than common recommendations. However, the trust in these formulas to measure difficulty may be overextended. Grade-level readability formulas were originally developed to try to ensure that a school textbook for a particular grade was appropriate for children at that grade level [21]. Their capabilities in measuring documents of a highly technical nature such as health care are not thoroughly validated. There are recent attempts to develop methods for text in the medical domain [22,23]. They have yet to enjoy wide adoption in the community, which may be attributed to the fact that efforts in learning models are inevitable.

Readability Formulas

Numerous readability metrics have been used for the purposes of preparing texts for schoolchildren and language learners and ensuring smooth written communication. These metrics assess the grade level or the number of years of education needed for a person to understand the content. Here we briefly introduce 3 of the metrics. For more discussions on these traditional readability formulas, we refer the reader to the review in [24].

Flesch-Kincaid Grade Level (FKGL) [25] predicts a grade level using the average sentence length and the average word length. Simple Measure of Gobbledygook (SMOG) [26] predicts readability based on the number of polysyllabic words (words with more than 3 syllables) and the number of sentences. Similarly, Gunning-Fog Index (GFI) [27] employs sentence length and the proportion of polysyllabic words. Detailed equations are shown in Multimedia Appendix 1.

These metrics are also used extensively in the health care domain to measure the readability of patient handouts [18,28-30], Web-based health information for patients [19,31,32], medication inserts [33,34], informed consent forms [20,35,36], clinical trial information [37], and Wikipedia medical entries [38,39]. FKGL, in particular, is used in more than half of readability studies compared in one review [40].

In general, these aforementioned metrics rely on the assumption that the longer the words and the sentences, the more difficult the text is. However, this assumption may not hold true for EHR narratives, which contain lists of clinical events (eg, medication list), abbreviations, and incomplete and short sentences, unduly lowering the readability score.

One measurement that tailors to the medical domain was proposed by Kim H et al [22]. This method compared surface text, syntactic, and semantic differences to predefined easy and difficult documents and reported normalized scores instead of grade levels. Another method for health text based on a naive Bayes classifier was developed [23]. The authors collected training documents from Web-based blogs, patient education documents, and medical journal articles. Vocabularies in these documents were used as features for the classifier. Both of the methods relied on manually curated documents. Therefore, different choices in constructing the sets might result in variation in the scores or classification results. Moreover, the classifier was limited, as it assigned only 3 categories—easy, intermediate, and difficult, and did not assign a grade-level scale. Furthermore, the reference document sets were not available.

Less research has been conducted on whether the readability grade levels predicted by these formulas or computational

models agree with actual users' perceptions of text difficulty. The objective characteristics are shown to not always align with user perceptions in other research fields. In one study, user perceptions of computer manufacturers' websites were different from content analysis tools [41]. In this work, we explored the relationship between users' perceptions of text difficulty and the readability formulas' output.

We evaluated FKGL and other widely used traditional readability metrics. These metrics usually hinged on a few textual characteristics and did not take into account the domain of the text. We also explored the effectiveness of the existing readability formulas on predicting the users' *perceptions* of difficulty. We hypothesized that the perceived readability of technical documents on complex topics was dependent on the domain of the text, not an absolute measure of the difficulty of a piece of text.

Methods

Overview

We evaluated existing metrics for assessing EHR readability and investigated their utility in EHR notes. We used the open-source Java library Flesh 2.0 [42] to calculate FKGL. In addition, we used the same program to calculate the number of sentences, words, and syllables, and then applied the other 2 formulas (SMOG and GFI). In the following sections, we first describe the data we used for evaluation, followed by an analysis of this corpus.

Data

We collected documents about diabetes from 2 different resources: English Wikipedia (denoted as wiki) and deidentified EHR notes (denoted as med). In wiki documents, we traversed from the Diabetes category. The EHR notes were selected using the International Classification of Diseases, Ninth Revision, code range 250.00 to 250.93. The 2 sources provided a contrast between texts aimed at the general audience and those written with health care professionals in mind. The statistics of this collection is shown in Table 1 under the columns labeled "all."

Diabetes is a common disease that we can expect a large body of readers to be aware of and can provide reasonable judgments on readability. This is especially important in the EHR collection because randomly selected EHR notes may contain information about rare conditions, which can confuse the readers. The common theme of the content in the 2 sources also helps address the problem of variations of a user's knowledge in different areas. By constraining to a single condition, we can limit the confounding effect of a user's different levels of familiarity in different areas.

Table 1. Document collection statistics.

Genre	Documents		Sentences		Tokens		FKGL[c]	
	All[a]	Paired[b]	All	Paired	All	Paired	All	Paired
Wiki	140	58	5703	1084	142, 106	23, 185	7.33–21.85	7.33–17.82
Med	242	133	8715	4232	120, 315	57, 655	6.48–15.76	6.99–15.76

[a]Columns labeled "all" include all documents.

[b]Columns labeled "paired" include only documents where another one with a similar length and FKGL score is also available.

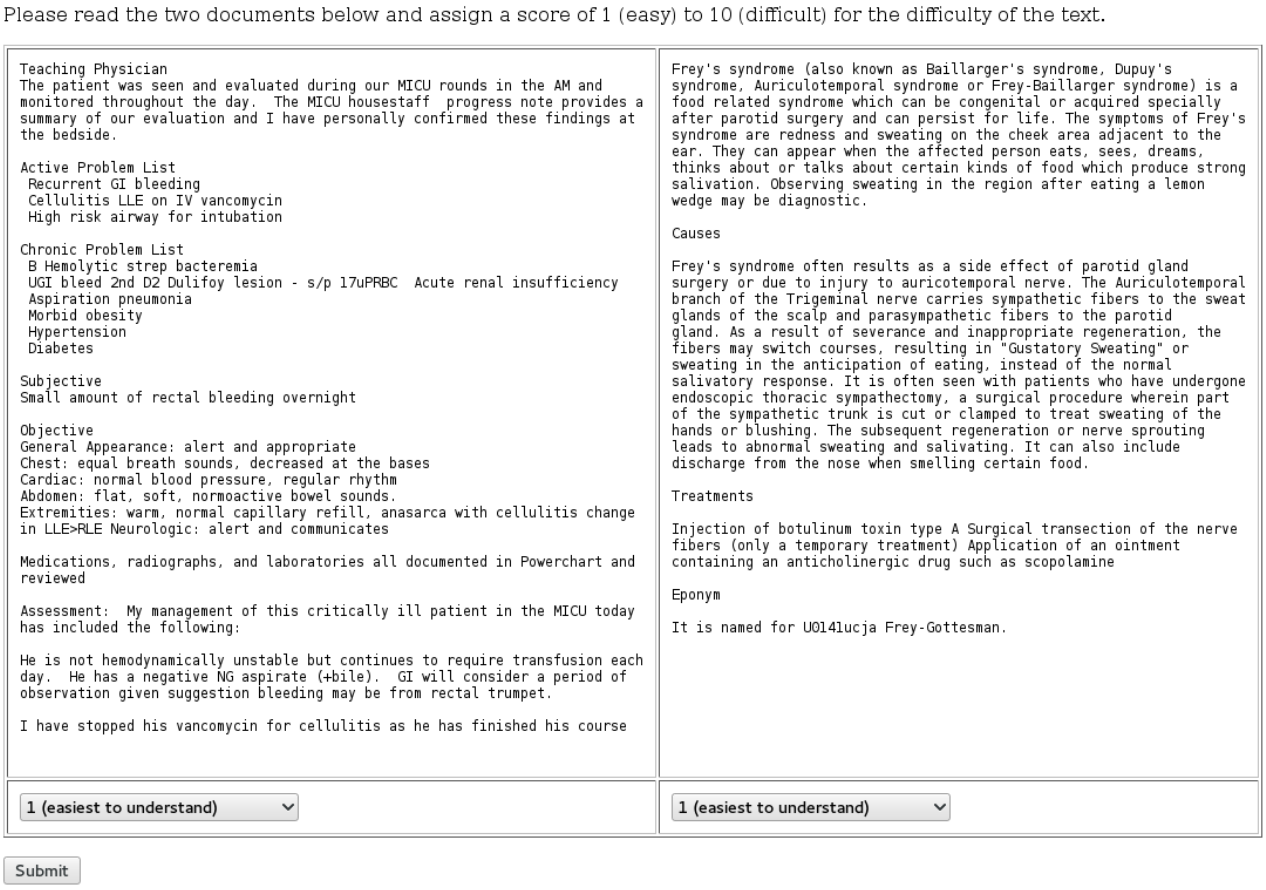
[c]FKGL: Flesch-Kincaid Grade Level.

Amazon Mechanical Turk Annotators

To validate one of the most frequently used readability formulas, FKGL, we paired analogous documents in our collection to ask Amazon Mechanical Turk (AMT) users to compare them. Specifically, documents were paired so that they had similar lengths (within 50-token difference) and comparable readability levels according to FKGL (within 0.5 grade level). The statistics on documents that were paired are shown in Table 1 under the columns labeled "paired."

We recruited 15 AMT subjects to read and rate pairs of documents. The readers were screened to have English as their native language and be AMT master workers. Three readers had a high school diploma, 7 had an associate degree, 4 had a Bachelor's degree, and 1 did not report education level. Each reader was presented with 20 randomly selected pairs of documents side by side on the computer screen. The 20 document pairs consisted of 5 pairs of wiki documents, 5 pairs of med documents, and 10 pairs of mixed-genre documents. The readers were requested to rate the readability of the documents on a scale from 1 (easiest to understand) to 10 (most difficult to understand). Each reader was given 6 hours to complete the task, and was not explicitly prohibited from using external resources. On average, they finished the assignment in 1 hour. Figure 1 is a screenshot of the interface with a mixed-genre pair.

Figure 1. Screenshot of the interface for the Amazon Mechanical Turk (AMT) users.



Corpus Analysis

Readability and User Rating Distributions

We first analyzed the empirical distribution of AMT users' ratings on the text difficulty and compared it with the empirical distribution of the readability formulas' scores.

Correlation Between AMT Users

We next measured correlations between different AMT users. For each user, all the documents that he or she provided a rating were collected. Since the document pairs were randomly assigned, in general no 2 users worked on an identical set of documents. Only a subset of the documents were rated by any 2 users. On average, a document was rated by 2.3 users. Between 2 users, 8.6 documents were on average rated by both.

We calculated correlations for a user's and any other user's ratings on the documents that were rated by both. The average for each user was obtained by first transforming the correlations by Fisher z-transformation, and then back-transformed [43]. Document genres were not separated in the calculation; otherwise, it would result in too few instances.

Correlation Between AMT User and Readability Formulas

To evaluate traditional readability formulas' applicability in technical documents, correlations between each AMT user's ratings and the 3 readability formulas were measured separately for the wiki and med genres. The average over each user's correlations were also obtained by Fisher z-transformation.

Differences in Users' Perceived Difficulty

To validate the generalizability of FKGL to different genres of text, we tested whether users perceived a difference when the readability scores were similar. The AMT users in our experiments were presented with documents of comparable difficulty (within a difference of 0.5) according to FKGL and of similar length (within 50-token difference). We tested the statistical significance of the difference between the difficulty values assigned by the users to 2 similar documents, separately for wiki, med, and mixed pairs. Two statistical tests are employed—Wilcoxon signed-rank test and Kolmogorov-Smirnov test.

We also tested the generalizability of 2 other formulas using the same procedure. Among all of the document pairs, we selected the subset of document pairs in which the SMOG scores were within 0.5 between each pair. The same process was repeated using GFI scores.

Furthermore, we explored the disparity in users' perceived difficulty when a readability formula reported a difference between 2 documents. For each user, we generated pairs of documents from all of the documents he or she rated and then removed the pairs that were presented during the AMT work session. These document pairs were separated into 3 types based on the genres of the documents, as in the previous experiments.

Correlation Between Readability Formulas

Since FKGL, SMOG, and GFI all involved similar variables (sentence length in words or polysyllabic words, word length), we examined the correlations between different readability formulas on the 2 genres of text in our dataset. Many studies adopted more than one of the traditional formulas to ascertain readability grade level on documents intended for patient consumption [44-49]. Analyzing the formulas' correlations would inform us of this approach's utility.

Word Usage

We compared the word usage patterns in the 2 genres of text by examining the common words. First, words in both med and wiki sources were ordered by the frequency in which they appeared in their respective genre. Then, the common words that were in both genres of text in the top frequently used words were counted. The shared vocabulary size might reveal a difference in word usage in different text genres.

Impact of Medical Concepts

Medical jargon is one of the barriers for the patient to understand health information. The eligibility criteria in clinical trials are found to be too difficult for the average American population, mainly due to the frequent use of technical jargon [50]. One study has shown that linking medical terms in EHR notes to Wikipedia pages can improve patient's comprehension [51]. Moreover, many methods have been proposed to identify important or potentially unfamiliar medical terms [52,53].

We explored the effects of the medical concepts by measuring the correlation between users' ratings and the number of concepts. Medical concepts were identified by running MetaMap [54] and excluding concepts from the following semantic groups and types: Activities & Behaviors, Concepts & Ideas, Geographic Areas, Objects, Occupations, Organizations, Age Group, Animal, Family Group, Group, Human, Patient or Disabled Group, Population Group, Professional or Occupational Group, Educational Activity, Health Care Activity, and Research Activity. These semantic groups and types usually do not contain technical medical jargon, and are uncommon in EHR notes. We also excluded Anatomical Structure because in our dataset almost all terms in this category were "body," with the rest being such common body parts as "head" that would not pose difficulty for an average reader.

Results

Readability and User Rating Distributions

Empirical distributions of the FKGL readability scores and users' ratings are shown in Figures 2 and 3. The FKGL histograms (Figure 2) on the 2 genres have clear distinctions. However, contrary to the general belief that EHR notes are more difficult to read, the histogram on the med data peaks to the left of the wiki data histogram. The users' ratings (Figure 3), although to a smaller degree, show a higher difficulty level for the med than for the wiki data.

Table 2 shows the average score of each readability formula and the AMT users' ratings. All of the 3 readability scores suggested that the technical EHR notes were significantly easier than lay language wiki articles, whereas the AMT users rated the opposite—wiki articles were 21.31% harder than EHR notes.

These results suggested that although FKGL might distinguish the readability of different genres, its counterintuitive predictions could lead to underestimation of difficulty levels on highly complex documents.

Table 2. Average readability score and users' ratings.

Genre	Average score or rating			
	FKGL[a]	SMOG[b]	GFI[c]	AMT[d] user rating
Wiki	14.75	11.07	12.33	4.41
Med	9.87	8.74	8.16	5.35
Difference[e] (%)	−33.09	−21.03	−33.76	21.31
P value	<.001	<.001	<.001	.002

[a]FKGL: Flesch-Kincaid Grade Level.

[b]SMOG: Simple Measure of Gobbledygook.

[c]GFI: Gunning-Fog Index.

[d]AMT: Amazon Mechanical Turk.

[e]All differences in scores between the wiki and med genres were statistically significant at level *P*=.01 (Mann-Whitney *U* test). The second to last row shows that the percentage med score was higher than the percentage wiki score.

Figure 2. Histogram of Flesch-Kincaid Grade Level (FKGL).

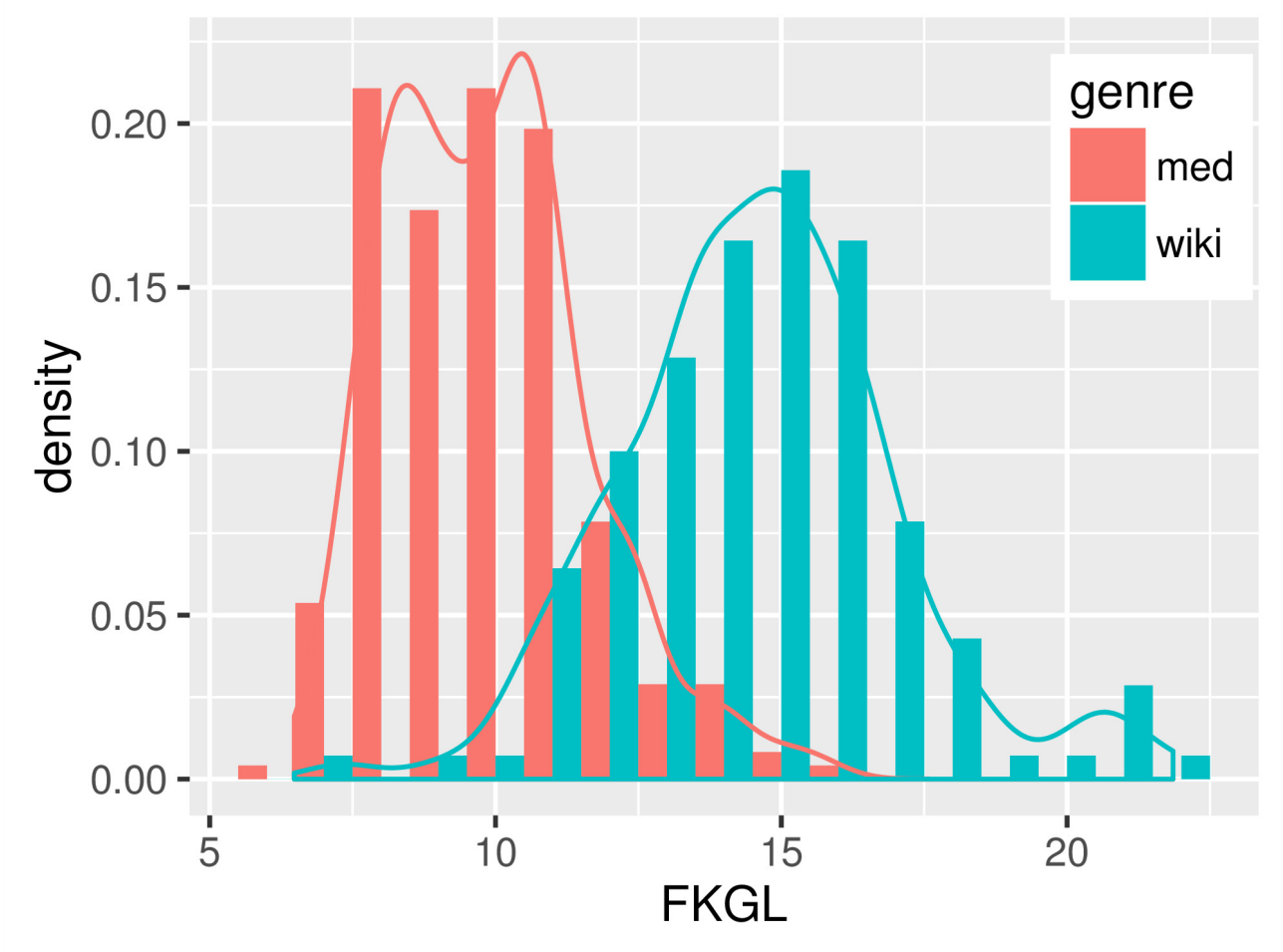
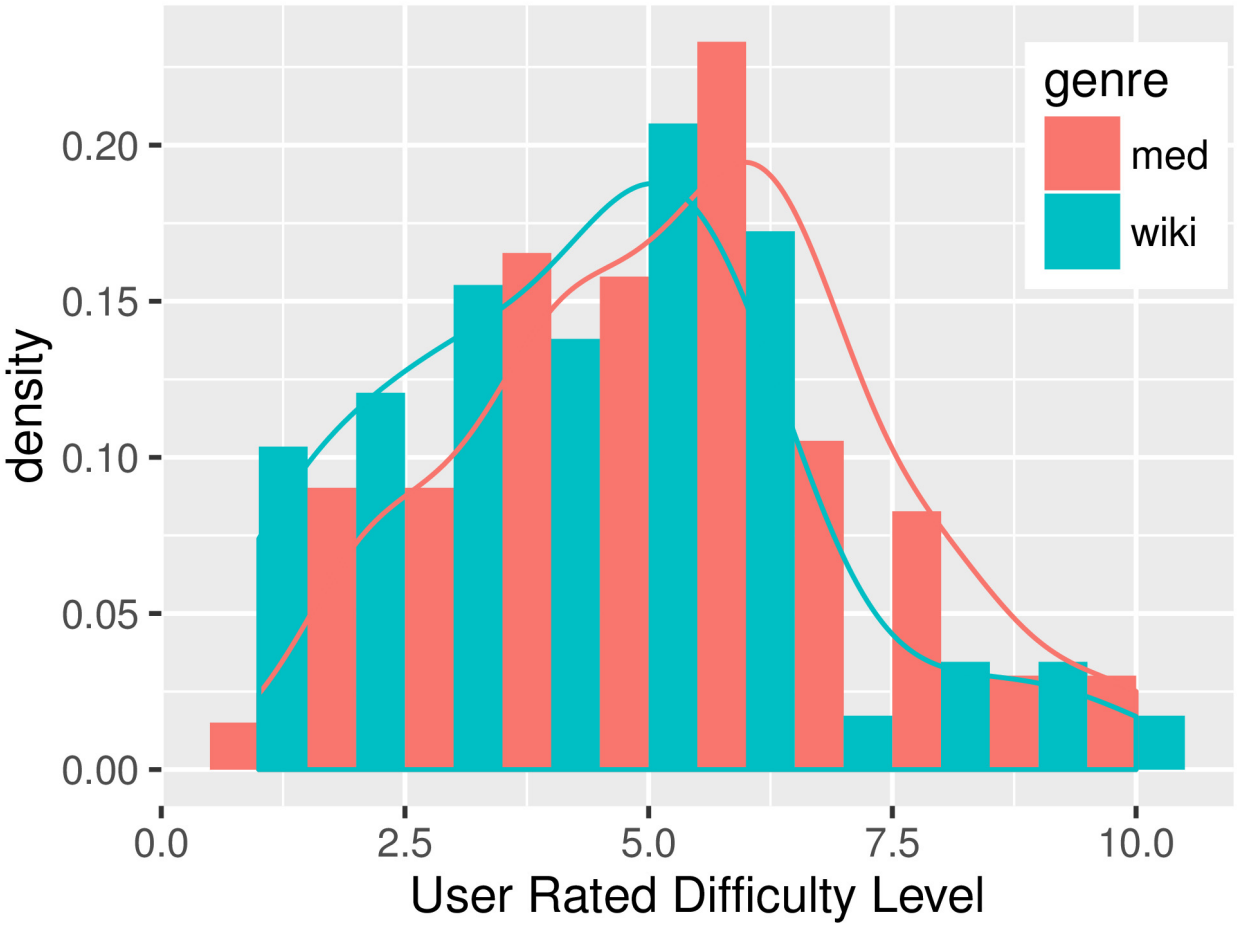


Figure 3. Histogram of Amazon Mechanical Turk (AMT) users' ratings.



Correlation Between AMT Users

Table 3 summarizes the correlations between 2 users' ratings. Most users showed moderate or strong correlation with other users, suggesting that the AMT users' perceptions of difficulty were congruous among themselves.

Table 3. Average correlations between a user and everyone else.

Average correlation	No. of users
<0.4	3
0.4–0.6	5
>0.6	7

Correlation Between AMT User and Readability Formulas

Table 4 shows the average correlation coefficients between an AMT user's ratings and the 3 readability formulas' output. All the correlations were very low, especially in the med category. The SMOG and user rating correlation on wiki data, although slightly higher than that of FKGL and GFI, was barely moderate. The low correlations suggested that users' perceived difficulty levels were inconsistent with the readability formulas' predictions. For example, one user consistently assigned low difficulty levels to documents with FK scores 12–16. However, another user's scores for documents with FK levels approximately 13.5 varied considerably. In contrast, the difficulty perceptions among different users were highly consistent (Table 3).

Table 4. Average correlation between users' ratings and readability formulas.

Readability formula	Wiki	Med
FKGL[a]	0.1758	0.2999
SMOG[b]	0.4134	0.1024
GFI[c]	0.2695	0.1272

[a]FKGL: Flesch-Kincaid Grade Level.

[b]SMOG: Simple Measure of Gobbledygook.

[c]GFI: Gunning-Fog Index.

This pattern of inconsistency highlighted the inadequacy of these formulas' utility in measuring EHR readability. It also highlighted their weakness in testing readability of documents of complex topics such as medicine, as they were developed to help users in the education community to gauge text difficulty below 12 grade. All 3 formulas relied on word counts and sentence counts to estimate text readability. The implicit assumption that longer words were more difficult, however, could often be violated. For instance, abbreviations that were not normally used outside the medical domain, such as "CHF" (Congestive Heart Failure) and "EKG" (electrocardiogram), were prevalent in EHR notes, without full definitions. Because these short abbreviations often comprised very few, if any, syllables, they would have exactly the same impact on the readability score as did the common stop words such as "the." However, the abbreviations were obviously one of the barriers for a patient to understanding an EHR note. Furthermore, many abbreviations were ambiguous. For example, "MI" can be the shorthand for both "myocardial infarction" and "myocardial ischemia," 2 different clinical conditions. In fact, disambiguating these abbreviations has been actively studied [55,56]. Finally, SMOG and GFI's use of polysyllabic words could also exacerbate the problems with abbreviations. For example, "COPD" might be considered a 1-syllable word in calculating FKGL, but it would make no contribution to the calculation of SMOG or GFI.

Differences in Users' Perceived Difficulty

When 2 documents of similar length and FKGL score were shown together, the ratings assigned by the AMT users exhibited different patterns depending on the genres of the 2 documents. Using a Wilcoxon signed-rank test, the *P* values are displayed in Table 5 under "Wilcoxon signed-rank test."

Table 5. Statistical significance of difference in AMT users' perceived difficulty between documents of similar Flesch-Kincaid Grade Level.

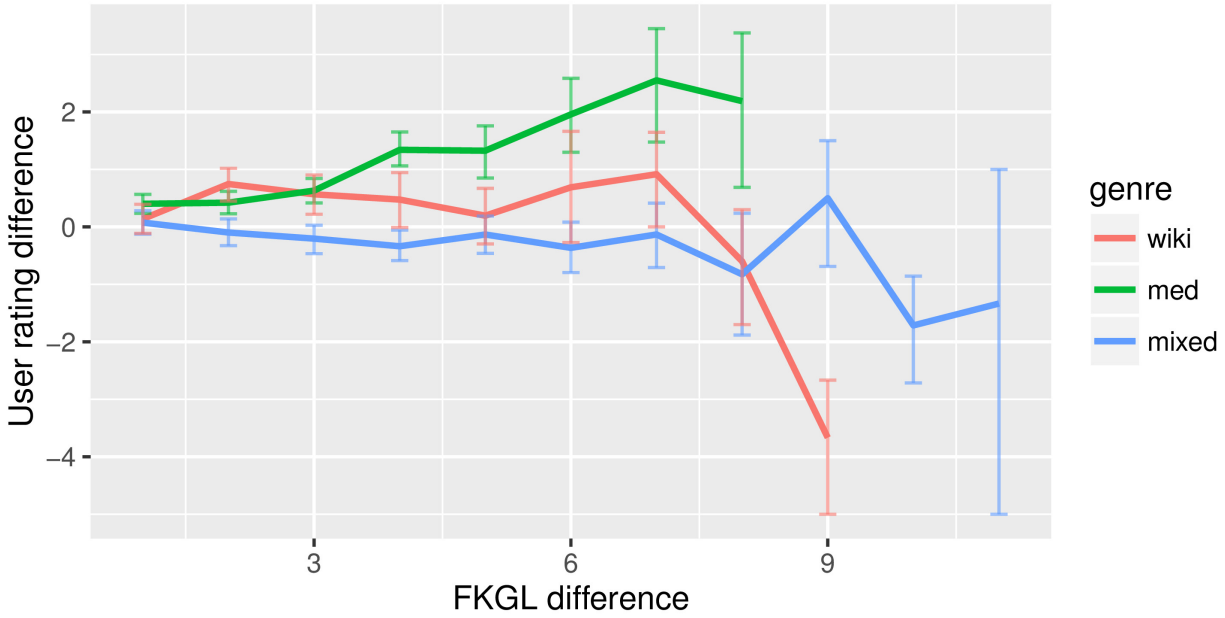
Genre of pair	*P* value	
	Wilcoxon signed-rank test	Kolmogorov-Smirnov test
Wiki	.80	.95
Med	.25	.80
Mixed	<.001	<.001

The *P* values for a pair of same-genre documents showed that the users' assignments were not significantly different, consistent with the traditional formula's assessment. However, the *P* value for a pair of documents from different genres indicated that despite being assessed at similar difficulty, actual users perceived them as significantly different in terms of readability. Kolmogorov-Smirnov test (Table 5) also showed the same trend.

The same tests, when repeated on a subset of document pairs whose SMOG or GFI score difference was within 0.5, confirmed that they were not generalizable to different text domains. Detailed significance test results are displayed in Multimedia Appendix 2.

AMT users' perceptions of difficulty varied depending on the genre of text, even though a readability formula shows no difference. We then explored the disparity in users' perceived difficulty when a readability formula reported a difference between 2 documents. Figure 4 shows the average difference in users' ratings on a pair of documents with varying differences in FKGL scores.

Figure 4. Average user's rating difference on 2 documents of different Flesch-Kincaid Grade Level (FKGL) scores. Error bars are bootstrapped 95% CI.



For a pair of EHR notes, as the difference in FKGL scores widened, AMT users' rating difference also gradually increased. However, the users' ratings were limited to at most 2.5 levels difference even for large FKGL differences. For a pair of Wikipedia documents, AMT users' rating difference varied slightly within 1 level. These patterns suggested that in a wide range of FKGL scores, users' ratings did not agree with FKGL.

A similar trend was present in a mixed pair of documents: AMT users' rating difference stayed close to 0. The limited difference in AMT users' ratings indicated that FKGL scores did not align well with user perceptions across different genres.

Correlation Between Readability Formulas

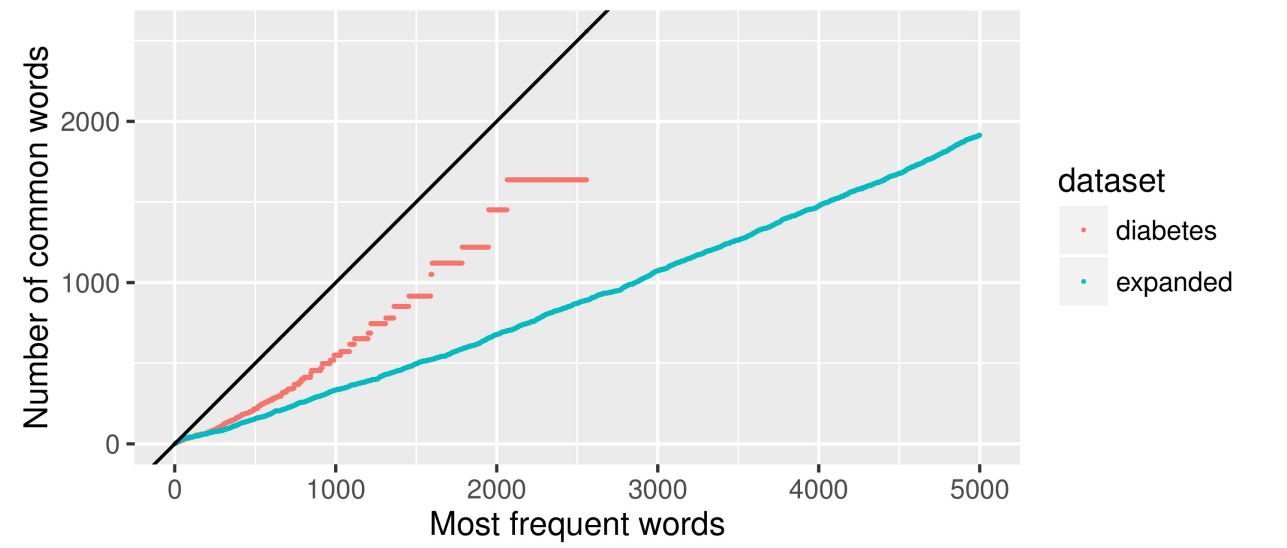
The correlation coefficients between different formulas confirmed that all 3 formulas were strongly correlated on our dataset regardless of text genre, consistent with the findings from previous studies [57,58]. Detailed plots and table showing the correlation are displayed in Multimedia Appendix 3. The substantial correlation implied that there was limited utility in employing multiple formulas, especially those relying on word and sentence lengths, to reduce potential bias of the individual ones when assessing text readability, as is often done in research studies [44-47,59].

Word Usage

In 2 similar corpora, the N most frequent words from each corpus would be similar. Therefore, the number of common words would increase at approximately the same rate as more frequent words were examined from the 2 corpora. Significant deviations from this pattern were indications of different word usage patterns. As shown in Figure 4, in our set of diabetes documents, the rate of increase in common words between wiki and med documents was significantly smaller (at the level $P<.001$) than 1 (shown as the solid line in the figure). This suggested that the word usage patterns in the technical (med) and lay language (wiki) documents on the same topic were different.

Expanding to more topics, we built the same word frequency statistic in all Wikipedia articles and about 100,000 EHR notes. Shown in Figure 5 as the "expanded" collection, the slope of common word count was also significantly smaller than 1 (at the level $P<.001$).

Figure 5. Common words in the med and wiki genre texts.



Impact of Medical Concepts

The correlation coefficients between the number of medical concepts and user difficulty ratings, shown in Table 6, were measured for each user and averaged. The average correlation was again obtained by Fisher *z*-transformation. Correlations with unique concepts were slightly higher than correlations with all concepts in both med and wiki texts. More unique medical concepts were likely to result in more cognitive load for a user to comprehend. It was also worth noting that EHR notes showed a lower level of correlation than the Wikipedia documents. This could be in part attributed to the multitude of complexities of EHR notes not limited to the abundance of technical jargon. Writing style such as choice of words and textual cohesion might also account for some of the variances in the difficulty in perceptions of EHR notes. In spite of the differences, these correlations suggested that medical jargon was a substantial contributor to readers' perceived difficulty of both genres of text.

Table 6. Average correlations between users' ratings and number of medical concepts.

Number of medical concepts	Wiki	Med
Number of all concepts	0.4434	0.3987
Number of unique concepts	0.5041	0.4329

Discussion

Principal Findings

We evaluated 3 widely used readability formulas' predictions of text difficulty and their correlation with users' perceived difficulty. Our results showed that although the formulas' predictions were highly correlated, they did not align well with user perceptions. Furthermore, despite being graded at similar levels by a readability formula, documents of different genres were still perceived with different difficulty.

Numerous studies have used the traditional readability formulas to evaluate a variety of information sources for patients. Many conclude that the materials intended for patients to improve understanding of their health conditions are too complex, exceeding the recommended grade levels for this purpose. Some also suggest areas of improvements to help align the documents to recommendation levels.

Unlike these studies that focus on documents mainly for patient consumption (patient handouts, education materials, Web-based information sources), we aimed to provide insight into users' *perceptions* of EHR notes readability. EHR notes are foremost a tool for physician communication, and a large portion of them are not originally written with easy patient comprehension in mind. However, they are shown to be beneficial to the patients. As more institutions allow patients access to their own EHR records, patients are also interested in reading them. Research has shown that patients may need help in understanding them. An accurate readability metric for the EHR notes can encourage physicians to write notes in a simpler language. It may make patient portals more useful. For example, it can be helpful in EHR record presentation by highlighting the easier ones that are within a patient's reading capabilities and providing comprehension assistance on the difficult ones. Such a metric may also enable the assistance tools to locate education materials that are easier to read than the EHR notes.

We noted that readability was not the only factor affecting patients' comprehension of EHR notes. For example, reader interest and motivation have been pointed out in the literature to be a factor contributing to comprehension [60]. In a more

realistic scenario where patients read their own EHR notes, they are likely to be motivated and show interest in knowing their own health conditions. Comprehension may also depend on a user's existing knowledge of the subject matter. Since we evaluated on many anonymous AMT users, the bias in individual prior knowledge was reduced due to such a random selection of participants. Nonetheless, higher readability could facilitate patient comprehension. Existing formulas are widely used in the health domain, but our analysis suggested that they were not accurate when applied on complex documents. A better metric should incorporate features beyond simple word and sentence length, such as the complexity of the concepts involved in the document.

Developments in computer science have led to new methods that utilize statistical language modeling and machine learning to predict text readability. For example, readability levels are modeled as a linear combination of a unigram language model and a normal distribution-based sentence length model [61]. This approach is expanded by introducing smoothing into the language models [62]. A Support Vector Machine model to assess text readability is proposed in [63] to learn from features including syntactic information such as noun phrases, traditional readability formulas, and language models.

However, language model-based approaches do not receive so widely an adoption in the medical domain as the traditional formulas. This can be in part due to the need to learn a language model, a much more involved task than using a formula. FKGL and other formulas, on the contrary, are readily available, free of charge, and easy to use [64]. We will explore these new models' performance in differentiating user perceptions of difficulty in the future.

Limitations

Our analysis was based on recruiting volunteers from AMT to evaluate readability of EHR records. Having patients directly perform this evaluation might represent a more realistic user experience. The patient, through his or her interactions with a health care professional, might have a better context than an AMT user to rate his or her own record's readability. In our study, to mitigate the problem, we selected documents based on a common condition.

Our sample documents were from one condition, selected to provide some context so that users would not be surprised by an unfamiliar topic. Thus, the user's prior knowledge might affect their ratings of text difficulty.

There are several other avenues we plan to pursue in future work. Clustering users based on their pre-existing knowledge may reveal readability formulas' differing capabilities in predicting users' perceptions of difficulty for different populations. We also plan to develop new methods that can better capture the readability of complex technical documents so that both health care providers and patients can benefit from focusing first on EHR notes that are at an appropriate difficulty level.

Conclusions

Studies have shown that providing patients with access to their own EHR notes may lead to improved health care outcomes. Measuring the readability of the EHR notes is an important step toward making the highly complex and technical narratives accessible to the patients. Despite being widely used in the health care domain, existing readability formulas are not thoroughly validated for their appositeness in this domain. In this study, we evaluated several such formulas' abilities in predicting *perceptions* of difficulty in health-related text from Wikipedia and EHR notes. We collected AMT users' ratings on text difficulty from these 2 different genres. Word usage in the 2 genres differed significantly despite their sharing a common topic. We found that the readability formulas' predictions did not align with perceived difficulty in either text genre. Furthermore, there was significant difference in the user's perceived difficulty in the general English and medical language when similar scores were predicted by readability formulas. Therefore, the widely used and highly correlated FKGL, SMOG, and GFI readability scales did not show adequate agreement with human ratings, and thus were not appropriate to assess the readability of EHR notes.

Acknowledgments

This work was in part supported by the Investigator Initiated Research 1I01HX001457-01 from the Health Services Research & Development Program of the US Department of Veterans Affairs, and in part by the Center for Intelligent Information Retrieval. The content is solely the responsibility of the authors and do not represent the views of the US Department of Veterans Affairs, the US Government, or the Center for Intelligent Information Retrieval.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Details of the readability formulas evaluated.

[PDF File (Adobe PDF File), 29KB - jmir_v19i3e59_app1.pdf]

Multimedia Appendix 2

Statistical significance of the difference in AMT users' perceived difficulty between documents of similar SMOG or GFI levels.

[PDF File (Adobe PDF File), 15KB - jmir_v19i3e59_app2.pdf]

Multimedia Appendix 3

Correlations of grade levels from different readability formulas.

[PDF File (Adobe PDF File), 209KB - jmir_v19i3e59_app3.pdf]

References

1. Institute of Medicine, Committee on Quality of Health Care in America. Crossing the quality chasm: a new health system for the 21st century. Washington, DC: National Academy Press; 2001.
2. Greene J, Hibbard JH. Why does patient activation matter? An examination of the relationships between patient activation and health-related outcomes. J Gen Intern Med 2012 May;27(5):520-526 [FREE Full text] [doi: 10.1007/s11606-011-1931-2] [Medline: 22127797]
3. White A, Danis M. Enhancing patient-centered communication and collaboration by using the electronic health record in the examination room. J Am Med Assoc 2013 Jun 12;309(22):2327-2328 [FREE Full text] [doi: 10.1001/jama.2013.6030] [Medline: 23757080]
4. Delbanco T, Walker J, Bell SK, Darer JD, Elmore JG, Farag N, et al. Inviting patients to read their doctors' notes: a quasi-experimental study and a look ahead. Ann Intern Med 2012 Oct 2;157(7):461-470 [FREE Full text] [doi: 10.7326/0003-4819-157-7-201210020-00002] [Medline: 23027317]
5. Wiljer D, Bogomilsky S, Catton P, Murray C, Stewart J, Minden M. Getting results for hematology patients through access to the electronic health record. Can Oncol Nurs J 2006;16(3):154-164. [Medline: 17523576]
6. Tang PC, Lansky D. The missing link: bridging the patient-provider health information gap. Health Aff (Millwood) 2005;24(5):1290-1295 [FREE Full text] [doi: 10.1377/hlthaff.24.5.1290] [Medline: 16162575]
7. Mossanen M, True LD, Wright JL, Vakar-Lopez F, Lavallee D, Gore JL. Surgical pathology and the patient: a systematic review evaluating the primary audience of pathology reports. Hum Pathol 2014 Nov;45(11):2192-2201. [doi: 10.1016/j.humpath.2014.07.008] [Medline: 25149550]
8. Keselman A, Slaughter L, Smith CA, Kim H, Divita G, Browne A, et al. Towards consumer-friendly PHRs: patients' experience with reviewing their health records. AMIA Annu Symp Proc 2007:399-403 [FREE Full text] [Medline: 18693866]
9. Keselman A, Smith CA. A classification of errors in lay comprehension of medical documents. J Biomed Inform 2012 Dec;45(6):1151-1163 [FREE Full text] [doi: 10.1016/j.jbi.2012.07.012] [Medline: 22925723]
10. Pyper C, Amery J, Watson M, Crook C. Patients' experiences when accessing their on-line electronic patient records in primary care. Br J Gen Pract 2004 Jan;54(498):38-43 [FREE Full text] [Medline: 14965405]
11. Mák G, Smith FH, Leaver C, Hagens S, Zelmer J. The effects of web-based patient access to laboratory results in british columbia: a patient survey on comprehension and anxiety. J Med Internet Res 2015;17(8):e191 [FREE Full text] [doi: 10.2196/jmir.4350] [Medline: 26242801]
12. Chapman K, Abraham C, Jenkins V, Fallowfield L. Lay understanding of terms used in cancer consultations. Psychooncology 2003 Sep;12(6):557-566. [doi: 10.1002/pon.673] [Medline: 12923796]
13. Lerner EB, Jehle DV, Janicke DM, Moscati RM. Medical communication: do our patients understand? Am J Emerg Med 2000 Nov;18(7):764-766. [doi: 10.1053/ajem.2000.18040] [Medline: 11103725]
14. Zeng QT, Tse T, Divita G, Keselman A, Crowell J, Browne AC, et al. Term identification methods for consumer health vocabulary development. J Med Internet Res 2007 Feb 28;9(1):e4 [FREE Full text] [doi: 10.2196/jmir.9.1.e4] [Medline: 17478413]
15. Zielstorff RD. Controlled vocabularies for consumer health. J Biomed Inform 2003;36(4-5):326-333 [FREE Full text] [Medline: 14643728]
16. Patrick TB, Monga HK, Sievert ME, Houston HJ, Longo DR. Evaluation of controlled vocabulary resources for development of a consumer entry vocabulary for diabetes. J Med Internet Res 2001;3(3):E24 [FREE Full text] [doi: 10.2196/jmir.3.3.e24] [Medline: 11720966]
17. Kutner M, Greenburg E, Jin Y, Paulsen C. NCES. The Health Literacy of America's Adults: Results From the 2003 National Assessment of Adult Literacy URL: https://nces.ed.gov/pubs2006/2006483.pdf [accessed 2017-01-24] [WebCite Cache ID 6nkJV3Mzm]
18. Boles CD, Liu Y, November-Rider D. Readability levels of dental patient education brochures. J Dent Hyg 2016 Feb;90(1):28-34. [Medline: 26896514]
19. Huang G, Fang CH, Agarwal N, Bhagat N, Eloy JA, Langer PD. Assessment of online patient education materials from major ophthalmologic associations. JAMA Ophthalmol 2015 Apr;133(4):449-454. [doi: 10.1001/jamaophthalmol.2014.6104] [Medline: 25654639]

20. Grossman SA, Piantadosi S, Covahey C. Are informed consent forms that describe clinical oncology research protocols readable by most patients and their families? J Clin Oncol 1994 Oct;12(10):2211-2215. [doi: 10.1200/jco.1994.12.10.2211] [Medline: 7931491]
21. Redish J. Readability formulas have even more limitations than Klare discusses. ACM J Comput Doc 2000 Aug 01;24(3):132-137. [doi: 10.1145/344599.344637]
22. Kim H, Goryachev S, Rosemblat G, Browne A, Keselman A, Zeng-Treitler Q. Beyond surface characteristics: a new health text-specific readability measurement. AMIA Annu Symp Proc 2007 Oct 11:418-422 [FREE Full text] [Medline: 18693870]
23. Leroy G, Miller T, Rosemblat G, Browne A. A balanced approach to health information evaluation: a vocabulary-based naïve Bayes classifier and readability formulas. J Am Soc Inf Sci 2008 Jul;59(9):1409-1419. [doi: 10.1002/asi.20837]
24. Klare GR. Assessing readability. Read Res Q 1974;10(1):62-102. [doi: 10.2307/747086]
25. Flesch R. A new readability yardstick. J Appl Psychol 1948 Jun;32(3):221-233. [Medline: 18867058]
26. McLaughlin GH. SMOG grading-a new readability formula. Journal of reading 1969;12(8):639-646.
27. Gunning R. The technique of clear writing. New York: McGraw-Hill; 1968.
28. Williamson JML, Martin AG. Analysis of patient information leaflets provided by a district general hospital by the Flesch and Flesch-Kincaid method. Int J Clin Pract 2010 Dec;64(13):1824-1831. [doi: 10.1111/j.1742-1241.2010.02408.x] [Medline: 21070533]
29. Wilson M. Readability and patient education materials used for low-income populations. Clin Nurse Spec 2009;23(1):33-40; quiz 41. [doi: 10.1097/01.NUR.0000343079.50214.31] [Medline: 19098513]
30. Woodmansey K. Readability of educational materials for endodontic patients. J Endod 2010 Oct;36(10):1703-1706. [doi: 10.1016/j.joen.2010.07.005] [Medline: 20850682]
31. Cheng C, Dunn M. Health literacy and the Internet: a study on the readability of Australian online health information. Aust N Z J Public Health 2015 Aug;39(4):309-314. [doi: 10.1111/1753-6405.12341] [Medline: 25716142]
32. Eltorai AE, Ghanian S, Adams Jr CA, Born CT, Daniels AH. Readability of patient education materials on the American association for surgery of trauma website. Arch Trauma Res 2014 Jun;3(2):e18161 [FREE Full text] [doi: 10.5812/atr.18161] [Medline: 25147778]
33. Khurana RN, Lee PP, Challa P. Readability of ocular medication inserts. J Glaucoma 2003 Feb;12(1):50-53. [Medline: 12567112]
34. Wallace LS, Keenum AJ, Roskos SE, Blake GH, Colwell ST, Weiss BD. Suitability and readability of consumer medical information accompanying prescription medication samples. Patient Educ Couns 2008 Mar;70(3):420-425. [doi: 10.1016/j.pec.2007.11.017] [Medline: 18178363]
35. Tarnowski KJ, Allen DM, Mayhall C, Kelly PA. Readability of pediatric biomedical research informed consent forms. Pediatrics 1990 Jan;85(1):58-62. [Medline: 2296494]
36. Paasche-Orlow MK, Taylor HA, Brancati FL. Readability standards for informed-consent forms as compared with actual readability. N Engl J Med 2003 Feb 20;348(8):721-726. [doi: 10.1056/NEJMsa021212] [Medline: 12594317]
37. Wu DT, Hanauer DA, Mei Q, Clark PM, An LC, Proulx J, et al. Assessing the readability of ClinicalTrials.gov. J Am Med Inform Assoc 2016 Mar;23(2):269-275. [doi: 10.1093/jamia/ocv062] [Medline: 26269536]
38. Thomas GR, Eng L, de Wolff JF, Grover SC. An evaluation of Wikipedia as a resource for patient education in nephrology. Semin Dial 2013;26(2):159-163. [doi: 10.1111/sdi.12059] [Medline: 23432369]
39. Azer SA, AlSwaidan NM, Alshwairikh LA, AlShammari JM. Accuracy and readability of cardiovascular entries on Wikipedia: are they reliable learning resources for medical students? BMJ Open 2015 Oct 06;5(10):e008187 [FREE Full text] [doi: 10.1136/bmjopen-2015-008187] [Medline: 26443650]
40. Wang L, Miller MJ, Schmitt MR, Wen FK. Assessing readability formula differences with written health information materials: application, results, and recommendations. Res Social Adm Pharm 2013;9(5):503-516. [doi: 10.1016/j.sapharm.2012.05.009] [Medline: 22835706]
41. Lee S, Lee W, Kim H, Stout PA. A comparison of objective characteristics and user perception of web sites. J Interact Advert 2004 Mar;4(2):61-75. [doi: 10.1080/15252019.2004.10722088]
42. Flesh.sourceforge. Flesh URL: http://flesh.sourceforge.net/ [accessed 2017-01-19] [WebCite Cache ID 6ndpfKhaB]
43. Silver NC, Dunlap WP. Averaging correlation coefficients: Should Fisher's z transformation be used? J Appl Psychol 1987;72(1):146-148. [doi: 10.1037/0021-9010.72.1.146]
44. Hansberry DR, Agarwal N, Gonzales SF, Baker SR. Are we effectively informing patients? A quantitative analysis of on-line patient education resources from the American Society of Neuroradiology. AJNR Am J Neuroradiol 2014 Jul;35(7):1270-1275 [FREE Full text] [doi: 10.3174/ajnr.A3854] [Medline: 24763420]
45. Hansberry DR, Agarwal N, Baker SR. Health literacy and online educational resources: an opportunity to educate patients. AJR Am J Roentgenol 2015 Jan;204(1):111-116. [doi: 10.2214/AJR.14.13086] [Medline: 25539245]
46. Vargas CR, Koolen PG, Chuang DJ, Ganor O, Lee BT. Online patient resources for breast reconstruction: an analysis of readability. Plast Reconstr Surg 2014 Sep;134(3):406-413. [doi: 10.1097/PRS.0000000000000472] [Medline: 25158700]
47. Vargas CR, Chuang DJ, Ganor O, Lee BT. Readability of online patient resources for the operative treatment of breast cancer. Surgery 2014 Aug;156(2):311-318. [doi: 10.1016/j.surg.2014.03.004] [Medline: 24953268]

48. Taki S, Campbell KJ, Russell CG, Elliott R, Laws R, Denney-Wilson E. Infant feeding websites and apps: a systematic assessment of quality and content. Interact J Med Res 2015 Sep 29;4(3):e18 [FREE Full text] [doi: 10.2196/ijmr.4323] [Medline: 26420339]
49. Piñero-López M, Modamio P, Lastra CF, Mariño EL. Readability analysis of the package leaflets for biological medicines available on the internet between 2007 and 2013: an analytical longitudinal study. J Med Internet Res 2016 May 25;18(5):e100 [FREE Full text] [doi: 10.2196/jmir.5145] [Medline: 27226241]
50. Kang T, Elhadad N, Weng C. Initial readability assessment of clinical trial eligibility criteria. AMIA Annu Symp Proc 2015;2015:687-696 [FREE Full text] [Medline: 26958204]
51. Polepalli RB, Houston T, Brandt C, Fang H, Yu H. Improving patients' electronic health record comprehension with NoteAid. Stud Health Technol Inform 2013;192:714-718. [Medline: 23920650]
52. Elhadad N. Comprehending technical texts: predicting and defining unfamiliar terms. AMIA Annu Symp Proc 2006:239-243 [FREE Full text] [Medline: 17238339]
53. Zheng J, Yu H. Key concept identification for medical information retrieval. 2015 Presented at: The Conference on Empirical Methods in Natural Language Processing; 2015; Lisbon p. 579-584.
54. Aronson AR, Lang F. An overview of MetaMap: historical perspective and recent advances. J Am Med Inform Assoc 2010;17(3):229-236 [FREE Full text] [doi: 10.1136/jamia.2009.002733] [Medline: 20442139]
55. Xu H, Stetson PD, Friedman C. A study of abbreviations in clinical notes. AMIA Annu Symp Proc 2007 Oct 11:821-825 [FREE Full text] [Medline: 18693951]
56. Kim Y, Hurdle J, Meystre SM. Using UMLS lexical resources to disambiguate abbreviations in clinical text. AMIA Annu Symp Proc 2011;2011:715-722 [FREE Full text] [Medline: 22195128]
57. Štajner S, Evans R, Orasan C, Mitkov R. What can readability measures really tell us about text complexity. In: Proceedings of the 8th International Conference on Language Resources and Evaluation. 2012 Presented at: LREC Workshop on natural language processing for improving textual accessibility; 2012; Istanbul, Turkey p. 14-22.
58. Van Oosten OP, Tanghe D, Hoste V. Towards an improved methodology for automated readability prediction. : European Language Resources Association (ELRA); 2010 Presented at: The 7th Conference on International Language Resources and Evaluation (LREC); 2010; Malta p. 775-782.
59. Badarudeen S, Sabharwal S. Assessing readability of patient education materials: current role in orthopaedics. Clin Orthop Relat Res 2010 Oct;468(10):2572-2580 [FREE Full text] [doi: 10.1007/s11999-010-1380-y] [Medline: 20496023]
60. Baldwin RS, Peleg-Bruckner Z, McClintock AH. Effects of topic interest and prior knowledge on reading comprehension. Read Res Q 1985;20(4):497. [doi: 10.2307/747856]
61. Si L, Callan J. A statistical model for scientific readability. 2001 Presented at: The tenth international conference on information knowledge management; 2001; Atlanta p. 574-576.
62. Collins-Thompson K, Callan J. A language modeling approach to predicting reading difficulty. 2004 Presented at: Proceedings of HLT-NAACL; 2004; Boston p. 193-200. [doi: 10.1002/asi.20243]
63. Schwarm SE, Ostendorf M. Reading level assessment using support vector machines and statistical language models. 2005 Presented at: The 43rd Annual Meeting on Association for Computational Linguistics; 2005; Ann Arbor p. 523-530.
64. De Felippe N, Kar F. Readability of information related to the parenting of a child with a cleft. Interact J Med Res 2015 Jul 08;4(3):e14 [FREE Full text] [doi: 10.2196/ijmr.4210] [Medline: 26155814]

Abbreviations

AMT: Amazon Mechanical Turk
EHR: electronic health records
FKGL: Flesch-Kincaid Grade Level
GFI: Gunning-Fog Index
SMOG: Simple Measure of Gobbledygook

Edited by L Buis; submitted 10.11.16; peer-reviewed by H Feldman, D Redd, K Koo; comments to author 12.12.16; revised version received 25.12.16; accepted 23.01.17; published 02.03.17.

Please cite as:
Zheng J, Yu H
Readability Formulas and User Perceptions of Electronic Health Records Difficulty: A Corpus Study
J Med Internet Res 2017;19(3):e59
URL: http://www.jmir.org/2017/3/e59/
doi:10.2196/jmir.6962
PMID:28254738

Original Paper

Exacerbations in Chronic Obstructive Pulmonary Disease: Identification and Prediction Using a Digital Health System

Syed Ahmar Shah[1,2], DPhil (Oxon); Carmelo Velardo[2], PhD; Andrew Farmer[3], DM; Lionel Tarassenko[2], DPhil (Oxon)

[1]Nuffield Department of Clinical Neurosciences, University of Oxford, Oxford, United Kingdom

[2]Institute of Biomedical Engineering, Department of Engineering Science, University of Oxford, Oxford, United Kingdom

[3]Nuffield Department of Primary Care Health Sciences, University of Oxford, Oxford, United Kingdom

Corresponding Author:
Syed Ahmar Shah, DPhil (Oxon)
Nuffield Department of Clinical Neurosciences
University of Oxford
Level 6, West Wing, John Radcliffe Hospital
Oxford,
United Kingdom
Phone: 44 1865 234764
Fax: 44 1865 273010
Email: syed.shah@ndcn.ox.ac.uk

Abstract

Background: Chronic obstructive pulmonary disease (COPD) is a progressive, chronic respiratory disease with a significant socioeconomic burden. Exacerbations, the sudden and sustained worsening of symptoms, can lead to hospitalization and reduce quality of life. Major limitations of previous telemonitoring interventions for COPD include low compliance, lack of consensus on what constitutes an exacerbation, limited numbers of patients, and short monitoring periods. We developed a telemonitoring system based on a digital health platform that was used to collect data from the 1-year EDGE (Self Management and Support Programme) COPD clinical trial aiming at daily monitoring in a heterogeneous group of patients with moderate to severe COPD.

Objective: The objectives of the study were as follows: first, to develop a systematic and reproducible approach to exacerbation identification and to track the progression of patient condition during remote monitoring; and second, to develop a robust algorithm able to predict COPD exacerbation, based on vital signs acquired from a pulse oximeter.

Methods: We used data from 110 patients, with a combined monitoring period of more than 35,000 days. We propose a finite-state machine–based approach for modeling COPD exacerbation to gain a deeper insight into COPD patient condition during home monitoring to take account of the time course of symptoms. A robust algorithm based on short-period trend analysis and logistic regression using vital signs derived from a pulse oximeter is also developed to predict exacerbations.

Results: On the basis of 27,260 sessions recorded during the clinical trial (average usage of 5.3 times per week for 12 months), there were 361 exacerbation events. There was considerable variation in the length of exacerbation events, with a mean length of 8.8 days. The mean value of oxygen saturation was lower, and both the pulse rate and respiratory rate were higher before an impending exacerbation episode, compared with stable periods. On the basis of the classifier developed in this work, prediction of COPD exacerbation episodes with 60%-80% sensitivity will result in 68%-36% specificity.

Conclusions: All 3 vital signs acquired from a pulse oximeter (pulse rate, oxygen saturation, and respiratory rate) are predictive of COPD exacerbation events, with oxygen saturation being the most predictive, followed by respiratory rate and pulse rate. Combination of these vital signs with a robust algorithm based on machine learning leads to further improvement in positive predictive accuracy.

Trial Registration: International Standard Randomized Controlled Trial Number (ISRCTN): 40367841; http://www.isrctn.com/ISRCTN40367841 (Archived by WebCite at http://www.webcitation.org/6olpMWNpc)

(J Med Internet Res 2017;19(3):e69) doi:10.2196/jmir.7207

KEYWORDS

COPD; disease exacerbation; mobile health; self-management; pulse oximetry; respiratory rate; clinical prediction rule; algorithms

Introduction

Chronic obstructive pulmonary disease (COPD) is expected to become the fourth leading cause of death by 2030 [1], imposing a high socioeconomic burden worldwide [2]. It is a chronic, progressive condition caused by airway infection due to smoking or long-term exposure to pollutants (eg, dust, fumes, poor air quality) irritating lungs. Common symptoms of COPD include coughing, chest tightness, breathlessness, fatigue, dizziness, and wheezing. Patients with COPD suffer from "exacerbations," a sustained worsening of symptoms often leading to hospitalization. In the United Kingdom, COPD exacerbations are now the leading cause of hospitalization, accounting for 15.9% of all hospital admissions [3]. There is growing evidence to suggest that increased frequency of exacerbation episodes leads to faster decline in lung function, reduced quality of life, greater numbers of hospital admissions, and increased health care cost [4]. Exacerbation is thus widely used as a major outcome in clinical studies for patients with COPD, with increasing efforts to try and reduce them. These efforts are hampered by the wide variation in the definitions of exacerbation used in clinical studies, with some based on symptoms (eg, increased breathlessness, sputum volume/color), some on events (eg, requiring hospitalization, taking antibiotics, and/or steroids), and some using a combination of both symptoms and events [4]. Symptoms are typically reported by the patients themselves using a paper-based diary and hence their recording depends on a subjective assessment. Poor adherence to protocol and data validity issues in clinical studies are often cited as major disadvantages of this approach [4]. Event-based definitions, on the other hand, are based on assessment by health care professionals (HCPs) after contact with them has been made by the patient. The disadvantage here is the low level of reporting to HCPs, with as many as 50% of exacerbations being unreported [5]. Until recently, exacerbations could only be identified from patient interviews or from analysis of records of treatment in health care databases or of symptom diary cards filled manually by patients [6].

Telemonitoring allows patients to record their symptoms in real time. Symptom data can then be stored and transmitted, with instant access possible for HCPs. As a result, there have been a number of attempts to use telemonitoring solutions to help manage patients with COPD through early detection of deterioration [7]. However, low compliance, lack of consensus on what constitutes an exacerbation, the small number of patients in most studies, and the relatively short periods of time during which monitoring takes place (less than 6 months) are limitations of COPD telemonitoring interventions [7]. In order to address these challenges, we developed the EDGE (Self Management and Support Programme) digital health system, a telemonitoring platform first evaluated during a cohort study [8] and subsequently used for 12 months, in a randomized controlled trial (RCT), by 110 patients in the intervention group [9]. The EDGE digital system allows patients to report their symptoms directly on an Internet-enabled tablet computer with a user-friendly interface, thereby eliminating most data validity issues. In the RCT, patients were expected to manage their use of relievers, antibiotics, and steroids and record it as part of the symptom diary on the tablet computer. Medication usage data linked to the symptom data were thus available, and so events could be defined based on medication data without the requirement of contact with an HCP. The system was intuitive and easy to use leading to high patient compliance and the collection of large amounts of COPD patient data [9].

Exacerbations are characterized by a number of different scenarios: for example, a patient's medication changes but the symptoms do not change; a patient's symptoms change, but the intake of medication does not change; or the patient's symptoms get worse without the patient taking any medication and there is no change for a few days before the patient eventually takes his or her medication. A systematic approach to address such a complex problem is to model it using a finite-state machine (FSM). Assuming that exacerbations can be correctly identified using such an approach, the next step is to investigate if a robust algorithm can be developed to predict exacerbation episodes. This is especially challenging because the protocol for the RCT allowed patients to self-monitor and record their symptom and medication data at a time of their choosing. There were periods, for some patients, during which they self-monitored less frequently and other periods with more frequent data but for different times of the day. Any predictive algorithm must be robust to such irregular sampling and take the variable timing of data recording into account.

In previous work, we developed an algorithm using the EDGE cohort study data, with a limited definition of exacerbation—based on self-reported medication usage, self-reported symptoms, and pulse oximetry data [10,11]. We did not differentiate between correlation and prediction, as the entire period, both before and during an exacerbation, was considered. Ideally, any predictive algorithm developed should identify periods likely to be associated with an impending exacerbation and differentiate them from periods where a patient's condition is stable (including periods of sustained high medication usage).

Methods

Overview

This section first describes the EDGE digital health system used to collect data from the 110 patients in the intervention group during the RCT. Next, the FSM model for identifying exacerbation episodes from symptom and medication usage data is introduced. This is followed by a description of the procedure for extracting "stable" and "prodromal" periods to provide the training data for an exacerbation prediction algorithm relying on robust features. Finally, the classification and validation methodologies adopted are presented.

Self Management and Support Programme System

The EDGE system is a digital health system with an Internet-enabled tablet computer at its core to support self-monitoring and self-management. The main components include a customized application developed for patients with COPD in collaboration with clinicians and patients, a Bluetooth-enabled pulse oximeter, and a secure back-end server. The application includes several self-monitoring modules,

including an interactive symptom diary with a series of questions related to COPD symptoms and to medication usage. A "session" consists of completing a symptom diary questionnaire and a brief recording with a pulse oximeter using a finger probe and typically lasts about 100 seconds [8]. The interactive diary questionnaire allows patients to self-report whether their symptoms have improved, become worse, or not changed compared with their understanding of what is "usual" for them. Chest tightness, breathlessness, sputum volume, and purulence are treated as "major" symptoms, while having a cold or sore throat, and an assessment of feeling generally run down, are treated as "minor" symptoms. Patients also report whether they are taking relievers, antibiotics, or steroids or any combination of the 3.

Finite-State Machine Model for COPD Exacerbation

According to the World Health Organization (WHO) and US National Heart, Lung, and Blood Institute Global Initiative for Chronic Obstructive Lung Disease (GOLD), an exacerbation is defined as "an event in the natural course of the disease characterized by a change in the patient's baseline dyspnea, cough, or sputum that is beyond normal day-to-day variations, is acute in onset, and may warrant a change in regular medication in a patient with underlying COPD." Although the definition is comprehensive, it raises a number of questions (eg, what are "normal day-to-day variations"?), thereby limiting its adoption as a practical definition in the context of clinical studies.

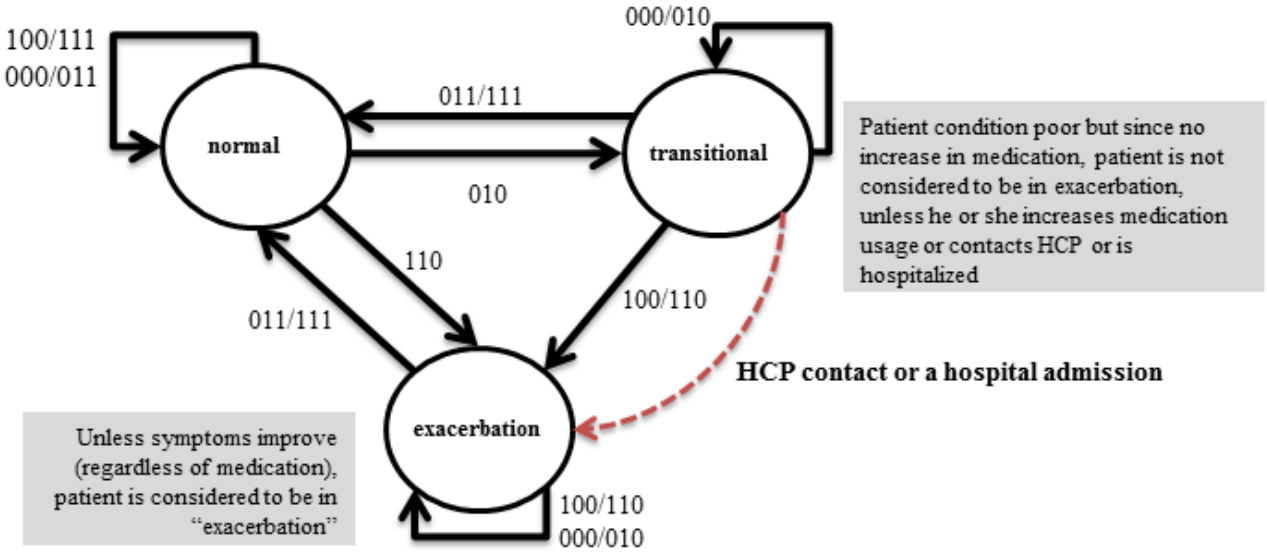
An FSM is a mathematical abstraction extensively used in the domain of logic design [12]. In FSM models, there are a finite number of "states" and a machine can transition from one state to another depending on "inputs." In this context, the "machine" will model a patient, and "inputs" are changes in symptoms and/or medication uptake. A patient will be in a particular "state" at any given time. Depending on changes in symptoms or medication usage, the patient's state can transition to another state. In addition to a "normal" state and an "exacerbation" state, we define a "transitional" state to account for those situations in which the patients are not in their normal state but their condition is not poor enough for them to be considered as having an exacerbation.

As already mentioned, there are a large number of exacerbation definitions. The following definition of exacerbation, based on clinical experience and a comprehensive review of the literature, was adopted by the clinical experts running the EDGE RCT: "An exacerbation is defined as a change in medication or an HCP contact (hospital admission or any documented contact, face-to-face contact or telephone calls) in the presence of a significant increase in symptoms."

"Change in medication" or "medication event" is defined as (1) increased use of reliever inhaler for at least 48 hours and (2) starting oral steroids and/or antibiotics. "Significant increase in symptoms" is defined as the presence of at least two symptoms, one of which should be a major symptom.

The 3 states in the FSM-based model are "normal" (patient is completely normal), "transitional" (patient's symptoms have become worse but no medication event or hospital admission or contact with HCP), and "exacerbation." Completion of a new symptom diary questionnaire generates a new set of inputs. Figure 1 shows the FSM model with its 3 states and the transitions between them depending on the 3-bit "inputs" encoding different combinations of symptoms and medication usage. The first bit encodes medication usage (0 = no medication or usual medication; 1 = medication event, as defined above). The second and third bits encode the status of the self-reported symptoms (00 = no change in symptoms; 10 = symptoms get worse; 11 = symptoms improve). It can be seen from Figure 1 that patients stay in the normal state regardless of whether they take medication or not, unless their symptoms get worse. Once a patient's symptoms get worse, he or she either transitions to a transitional state or into the exacerbation state, depending on whether there is a change in medication or medication event, as defined above. In the transitional state, a patient returns to the normal state if his or her symptoms improve (regardless of whether a medication event occurs or does not occur). Whenever the symptoms have not improved, a transition to the exacerbation state occurs instead if there is a medication event, or if the patient is hospitalized or makes contact with an HCP (regarding his or her COPD symptoms). In the exacerbation state, a patient only transitions back to the normal state if his or her symptoms improve. The definition of symptom improvement is a mirror image of the definition of "symptoms getting worse" given above: at least two symptoms, one of which must be a major symptom, must get better in comparison with the data recorded in the previous diary session.

Figure 1. A finite-state machine to model chronic obstructive pulmonary disease patient condition and its changes depending on symptoms and medication usage. HCP: health care professional.



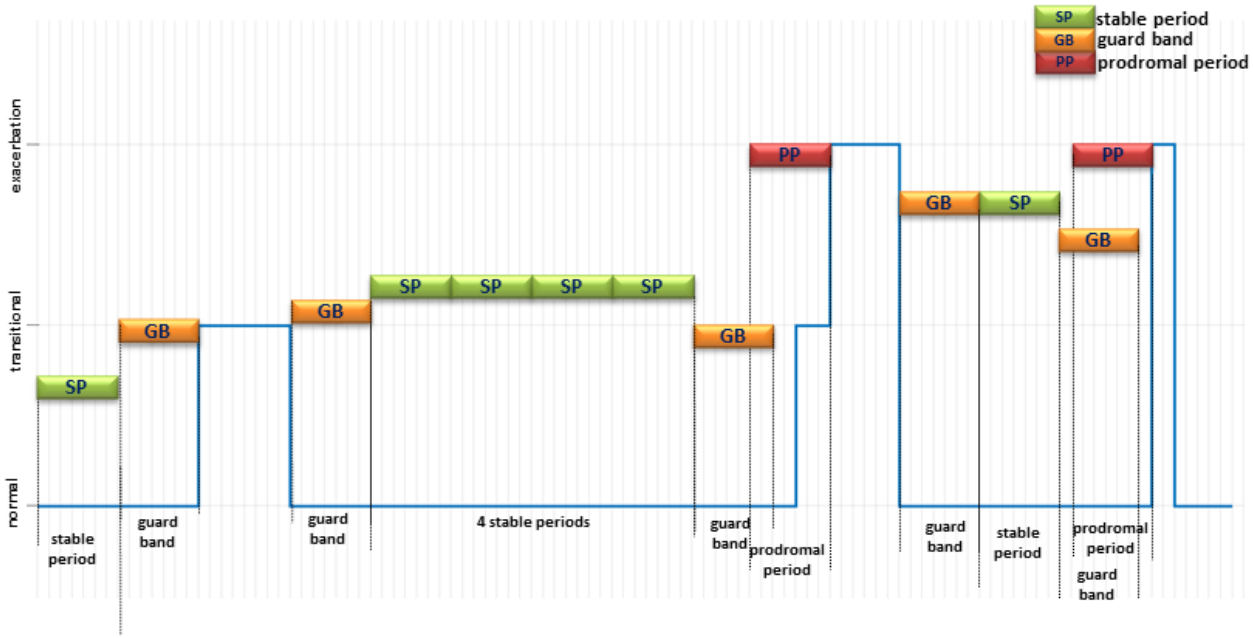
Input	Interpretation	Input	Interpretation
000	no or usual medication, no change in symptoms	100	medication increase, no change in symptoms
010	no or usual medication, symptoms get worse	110	medication increase, symptoms get worse
011	no or usual medication, symptoms improve	111	medication increase, symptoms improve

Selection of Independent Periods

In order to develop algorithms for prediction of exacerbation episodes, it is important first to identify the periods of time (in days) during which a patient's condition is stable, as well as periods of time when a patient's condition can be considered to be deteriorating before an exacerbation event. A previous study [13] has suggested that symptoms tend to worsen during the 7 days immediately before an exacerbation episode. We therefore define a time window of 7 days before an exacerbation event and label all such instances in our dataset as "prodromal" periods. Similarly, "stable" periods are defined as 7-day periods during which a patient must be in the normal state throughout. These periods have a 7-day "guard band" at the start and the end of the period. This is to ensure that, for every stable period, there are at least 7 days at the start of the stable period since a transitional or exacerbation state ended and at least 7 days at the end of the stable period before the next transitional or exacerbation state occurs.

Figure 2 illustrates the identification of "stable" and "prodromal" periods for a single subject who was in different states over the course of the trial.

Figure 2. Identification of "stable periods" and "prodromal periods" over the course of monitoring of a single patient. Notice that there are only 4 stable periods identified between the time when the patient returns to normal from the transitional state and the time when the patient goes from normal to the transitional state. There is only 1 stable period after the patient returns to normal after the first exacerbation and before the second exacerbation occurs. In this illustration, each vertical line represents a single day, and "stable," "prodromal," and "guard bands" are all 7 days long.



Vital Signs From Pulse Oximetry

Commercial pulse oximeters typically provide only oxygen saturation (SpO_2) and pulse rate measures. However, it is also possible to extract respiratory rate using the waveform—called photoplethysmogram (PPG)—recorded by a pulse oximeter with an appropriate algorithm. Briefly, respiratory rate modulates both the amplitude (primarily due to the mechanical effect of breathing) and frequency (due to respiratory sinus arrhythmia) of the PPG [14]. In the analysis presented in this paper, we focus on extracting respiratory rate from the amplitude modulation of the PPG because the phenomenon of respiratory sinus arrhythmia is reduced both with age [15] and pharmacological interventions [16]. The method used for estimating respiratory rate from the PPG waveform was based on computing the median frequency spectrum from a number of autoregressive models of the 30-second waveform segments recorded during the self-monitoring with the finger probe pulse oximeter. This novel signal processing approach was developed by us to estimate respiratory rate from PPG data in a different clinical setting [17]. Estimation of the respiratory rate from the amplitude modulation of the PPG requires a preprocessing stage in which an appropriate low-pass filter attenuates as much high-frequency content (corresponding to the cardiac frequency) as possible while preserving the low-frequency content (corresponding to the respiratory frequency). Because the work described in [17] was based on a population of pediatric patients with different values of respiratory and heart rates, the preprocessing stage was modified to take into account the lower values of respiratory and heart rates in the adult population. The cutoff frequencies of the transition band of the low-pass filter were varied according to the heart rate of the patient in a given session (ie, from 0.5 × heart rate to 1.2 × heart rate). Consequently, for each recording session from a participant in the RCT, 3 vital signs (heart rate, SpO_2, and respiratory rate) were obtained from the 30-second pulse oximetry data.

Feature Extraction

The labeling shown in Figure 2 is used to create a database of vital sign data from 7-day stable periods and 7-day prodromal periods (ahead of the exacerbation events). This database will enable us to construct a 2-class classifier, with the vital sign data as its inputs, to assign the data to either the stable or prodromal class. Classification of a 7-day period as prodromal is of course equivalent to exacerbation prediction.

It is important to capture the time course of each vital sign during these 7-day periods. A number of traditional algorithms developed for time series analysis cannot be directly applied because of the irregular number of recordings over a given period, both within and across subjects. Rather than applying approximations, for example, grouping multiple recording sessions within a day together (a rare occurrence), or imputing missing data (a much more common phenomenon), we fit a straight line, using a least squares criterion, to each set of vital sign data in each period. This corresponds to finding the minimum of $E(m,c)$ defined in equation (a) in Figure 3, where m is the gradient, x_i is the time at which the i^{th} session was recorded in a given 7-day period, y_i is the corresponding value of the vital sign, and c is the y-intercept value. For each vital sign in each 7-day period, the minimization of equation (a) will provide the gradient, m, whose magnitude will reflect whether the value of the vital sign is increasing, decreasing, or not changing, and an intercept value that is dependent on both the mean value and the gradient.

Each 7-day period is normalized in time with respect to the first self-monitoring session in that period (the first session in each 7-day period is taken to be at time 0). Figure 4 illustrates this for 2 cases, a "stable" period and a "prodromal" period. In this particular illustration, it can be seen that the patient in the stable period had a higher mean respiratory rate, which was gradually decreasing during the 7 days. The patient in the prodromal period has a lower mean respiratory rate but there is a clear increase over the 7 days. With this simple linear fit method, we capture the trend of each vital sign over the 7 days. Periods with irregular self-monitoring times or missing data can still be analyzed.

Figure 3. (a) Least Squares Fitting to determine the gradient and y-intercept in each 7-day period(b) Hypothesis function (c) Sigmoid function(d) Cost function for logistic regression.

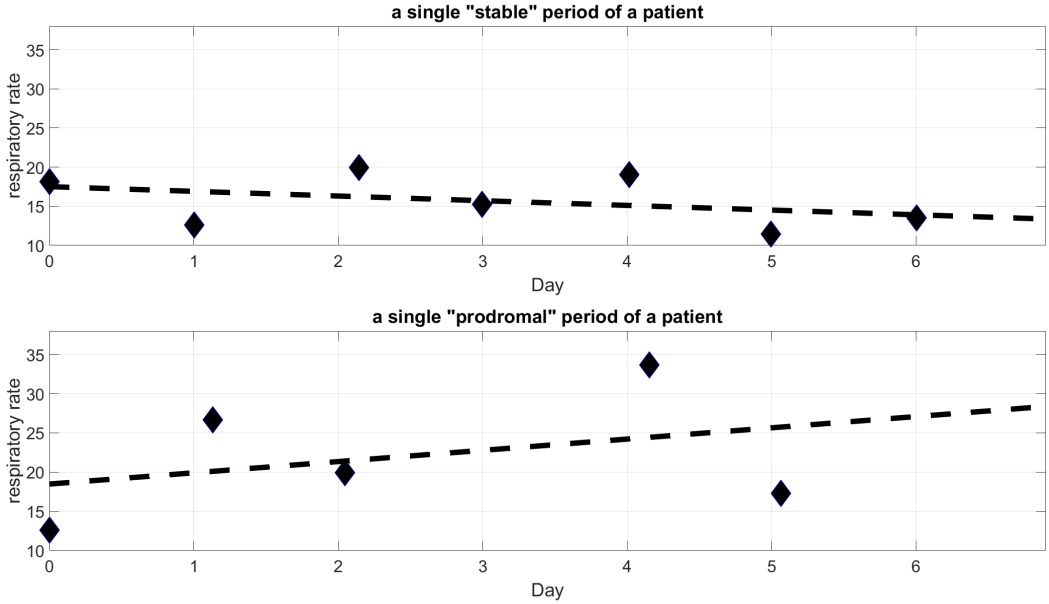
(a) $$E(m,c) = \sum_i (mx_i + c - y_i)^2$$

(b) $$h_\Theta(x) = s(\Theta_0 + \Theta_1 x_1 + \Theta_2 x_2 \ldots . \Theta_N x_N)$$

(c) $$s(y) = \frac{1}{1+e^{-y}}$$

(d) $$C(\Theta) = -\frac{1}{m}\left(\sum_{i=1}^{m} y^i log h_\Theta(x^{(i)}) + (1-y^i)\log(1-h_\Theta(x^{(i)}))\right)$$

Figure 4. Illustration of straight line fitting during a "stable" period (gradient: −0.60, y-intercept: 17.5) and a "prodromal" period (gradient: 1.43, y-intercept: 18.47).



Classification and Validation

A logistic classifier was applied to the vital sign feature vectors derived from both the stable and prodromal periods in order to investigate the discriminatory power of these features for classification (and hence exacerbation prediction). Equation (b) in Figure 3 gives the hypothesis function used in logistic regression where x_n and θ_n are the n^{th} feature and parameter, respectively, and s (y) is a sigmoid function (equation (c) in Figure 3). The output of the sigmoid function is bounded between 0 and 1 and can thus be interpreted as a probability or, in this case, the likelihood of a patient's condition deteriorating and leading to exacerbation (the prodromal period). The mean and gradient values (x_n) of each of the 3 vital signs are the inputs to the logistic classifier.

The cost function given by equation (d) in Figure 3 outputs a large value when the predicted value $h_\theta(x)$, is very different from the true class value, y, and a very small output when the predicted value is close to the true class value. An additional benefit of C (θ) as a cost function is that it is convex and thus a global minimum will always exist, which can easily be determined with a gradient descent algorithm.

Correct classification of a prodromal period was deemed to be a true positive and correct classification of a stable period was deemed to be a true negative. Incorrect classifications of prodromal periods and stable periods were taken to be false negatives and false positives, respectively. To test the generalization capability of the classification algorithm, we used 10-fold cross-validation. The data from both classes were randomly divided into 10 folds, with 9 folds used for training and 1 fold for testing. This was repeated 10 times, each time using a different fold for testing. This allowed us to obtain classification results on the whole dataset. Because the partitioning of data into 10 folds is random, the process of applying 10-fold cross-validation was repeated 1000 times, thereby allowing us to find the possible range of classifier performance. For performance evaluation, the receiver operating characteristic (ROC) curve was determined for each iteration and the mean ROC curve along with the 95% confidence interval was then computed. We then estimated the area under the curve (AUC) to compare the different classifiers.

Results

Dataset

Table 1 provides summary statistics on the number of days, sessions, and compliance of the 100 patients in the RCT intervention group with near-complete symptom diary and pulse oximetry data. On average, a patient self-monitored for 354 days, with a minimum usage of 5 days every week throughout this period.

Table 1. Overall number of days, sessions, and compliance of 100 patients in the EDGE (Self Management and Support Programme) study with both symptom diary and pulse oximetry data.

Metric	Across all patients	Per patient	
		Mean (SD)	Range
Number of days in study	35,439	354 (31)	194-403
Number of sessions	27,260	273 (68)	61-373
Number of days system used	27,074	271 (67)	60-373
Mean compliance (days/week)		5.3 (1.2)	1.8-6.9

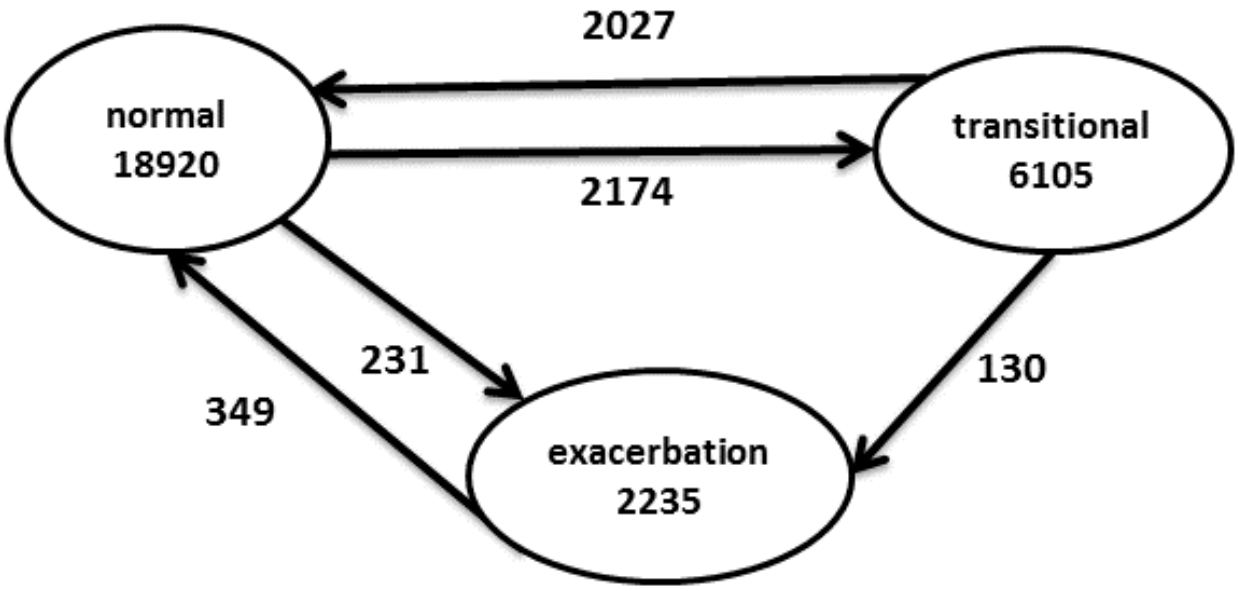
Identification of Exacerbations

There were a total of 361 exacerbation events after applying the proposed FSM model to the 27260 sessions from 100 patients. Figure 5 summarizes the total number of sessions in each state (given inside the circles), as well as the number of times that a patient transitioned from one state to the other (given next to the corresponding arrow). As expected, patients were mostly in the normal state (18,920/27,260, 69.41% of sessions completed) and were in the transitional state for 22.40% (6105/27,260) of the sessions. In 2174 instances, patients in the normal state entered the transitional state as a result of their symptoms deteriorating but without increasing their medication. In 93.24% of these cases (2027/2174), their symptoms subsequently improved and they returned to the normal state. For 97.08% of the 2027 cases (1968/2027), this happened without any medication increase; in only 2.91% of the cases (59/2027) did patients increase their medication before returning to the normal state. In the cases for which patients did not increase their medication while returning to normal state, the mean and median number of days in the transitional state were 3.2 days and 1.8 days. In the cases for which patients increased their medication while returning to normal state, the mean and median number of days in the transitional state were 4.2 days and 3.0 days.

Only 2235 sessions (8.20% of total sessions, 2235/27,260) were completed while the patient was in the exacerbation state. In 64.0% of cases (231/361), a patient's state transitioned directly from the normal to the exacerbation state, corresponding to patients increasing their medication as soon as their self-reported symptoms got worse.

All patients were assumed to be in the normal state at the start of the RCT, but only 71 patients ended up being in the normal state at the end of the study, with the remaining 29 patients ending in either the transitional state or the exacerbation state. Thus, there were 29 more transitions out of the normal state (2405) than into the normal state (2376) and also 17 and 12 extra transitions into the transitional and exacerbation states, respectively.

Figure 5. Summary of state transitions illustrating the total number of sessions in each state (inside the circles) and the number of instances of each transition between states across all patients (numbers next to the arrow between states) in the randomized controlled trial.



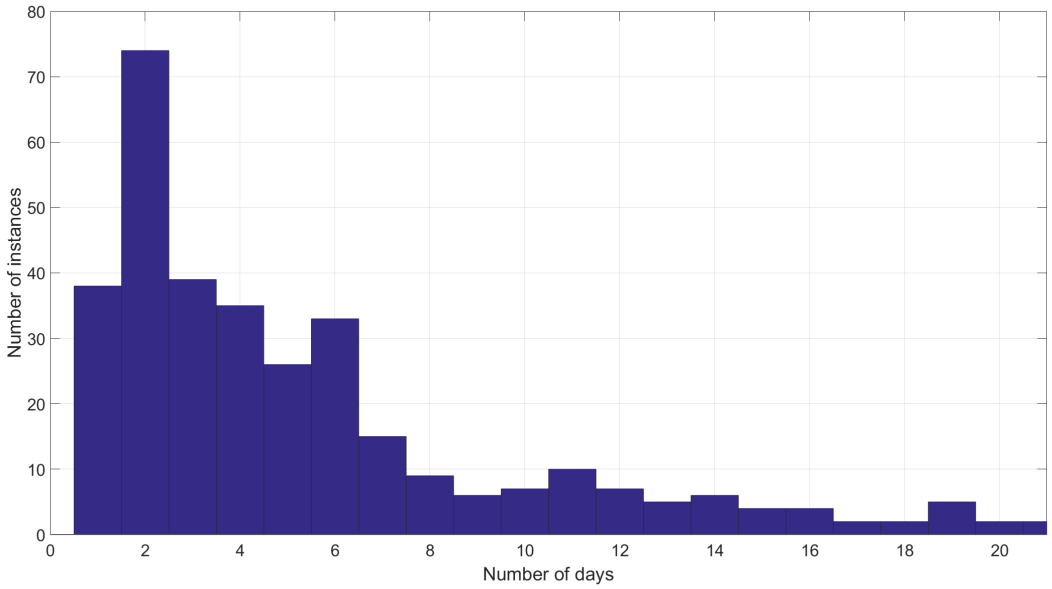
Distribution of Exacerbations

Table 2 and Figure 6 illustrate the distribution of length of exacerbation episodes. It can be seen that most of the exacerbation episodes lasted for up to 7 days (260/361, 72%). However, there was considerable variation in the length of an exacerbation episode across the population, with a mean length of 8.8 days and a median length of 4 days. On the basis of the total number of monitoring days, patients were in the exacerbation state for about 9% of the time (3180/35,439 monitoring days).

Table 2. Length of exacerbation events across all patients (91.7% of exacerbation episodes, 331/361, were within 21 days).

Metric	Number
Total number of events	361
Event length 1-7 days, n (%)	260 (72.0)
Event length 8-14 days, n (%)	50 (13.9)
Event length 15-21 days, n (%)	21 (5.8)
Mean length of exacerbation, days	8.8
Median length (interquartile range) of exacerbation, days	4 (2-9)

Figure 6. Distribution of exacerbation length (only those with episodes lasting 21 days or less are shown).



Characteristics of Vital Signs During Stable and Prodromal Periods

Table 3 summarizes the total number of periods and the corresponding number of sessions, as well as summary statistics for the various features extracted for each of these periods, both stable and prodromal periods. Even though there were 361 exacerbation events, in only 304 of those cases were there at least 7 days immediately before the exacerbation event during which the patient was either in the normal or the transitional state. The remaining cases refer to exacerbation events that occurred within 7 days of a previous exacerbation event. In 18 out of these 304 cases, there was only a single diary session completed by the patient and therefore it was not possible to extract features. This explains why there are 286 prodromal periods extracted from this dataset, from a total of 361 exacerbation events.

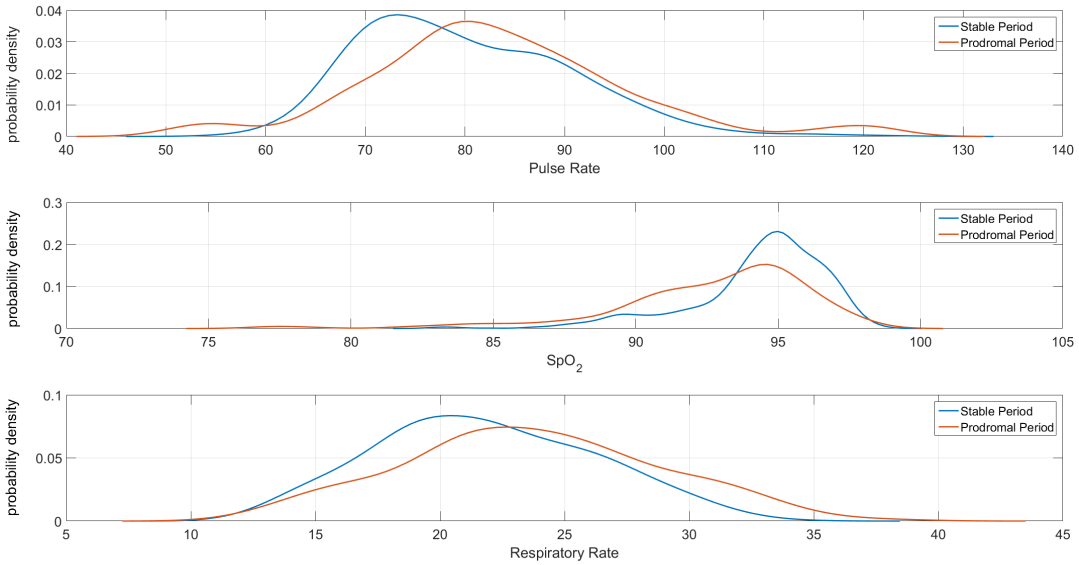
It can be seen from Table 3 that the mean values for all 3 vital signs are as expected for COPD patients with worsening symptoms: lower mean for SpO_2 and higher mean for both respiratory rate and heart rate in the prodromal periods in comparison with the stable periods. Figure 7 illustrates the distribution of the 3 vital signs in the stable and prodromal periods. The small difference in the means and the significant overlap between the distribution of the 3 vital signs possibly explain the difficulty of designing algorithms capable of predicting exacerbation episodes.

Table 3. Summary statistics of different features in the stable and prodromal periods.

Metric	"Stable" period	"Prodromal" period
Total number of 7-day periods	1005	286
Mean number of sessions in these 7-day periods	5.9	4.6
Mean value of SpO_2[a] (%)	94	93
Mean value of pulse rate (beats per minute)	80	83
Mean value of respiratory rate (breaths per minute)	22	24

[a]SpO_2: oxygen saturation.

Figure 7. Distribution of vital signs (pulse rate, oxygen saturation or SpO_2, and respiratory rate) acquired from a pulse oximeter in "stable" and "prodromal" periods.



Prediction of Exacerbations

Finally, Figure 8 illustrates the mean ROC curve, bounded by the 95% confidence interval, after applying the logistic classifier to the stable and prodromal periods as explained in the "Classification and Validation" section. Table 4 summarizes the performance results using AUC measures from the mean ROC curve (along with the 95% confidence interval) based on classification with each of the features extracted from each of the vital signs separately and in all possible combinations. In addition, specificity at specific sensitivity measures is also provided to give a better sense of the level of performance that could be expected for a given operating point (threshold) selected from the mean ROC curve. All 3 vital signs have a degree of predictivity because the mean AUC is greater than .5 for all of them. SpO_2 seems to be the most predictive vital sign, followed by respiratory rate estimated using our algorithm. In comparison with respiratory rate and SpO_2, pulse rate seems to be the least predictive in this population.

Table 4. Comparison of classifier performance using features extracted from each of the vital signs separately and in combination; the sensitivity and specificity values are those extracted from mean receiver operating characteristic curve.

Vital sign	Mean AUC[a] (95% CI)	Specificity range (%) for 60%-80% sensitivity from mean ROC[b] curve
Pulse rate (PR)	0.578 (0.578-0.578)	52-31
SpO_2[c]	0.658 (0.657-0.658)	62-38
Respiratory rate (RR)	0.605 (0.604-0.605)	53-32
PR + SpO_2	0.664 (0.664-0.664)	63-40
RR + PR	0.612 (0.612-0.612)	55-27
RR + SpO_2	0.672 (0.671-0.672)	64-36
RR + PR + SpO_2	0.682 (0.681-0.682)	68-36

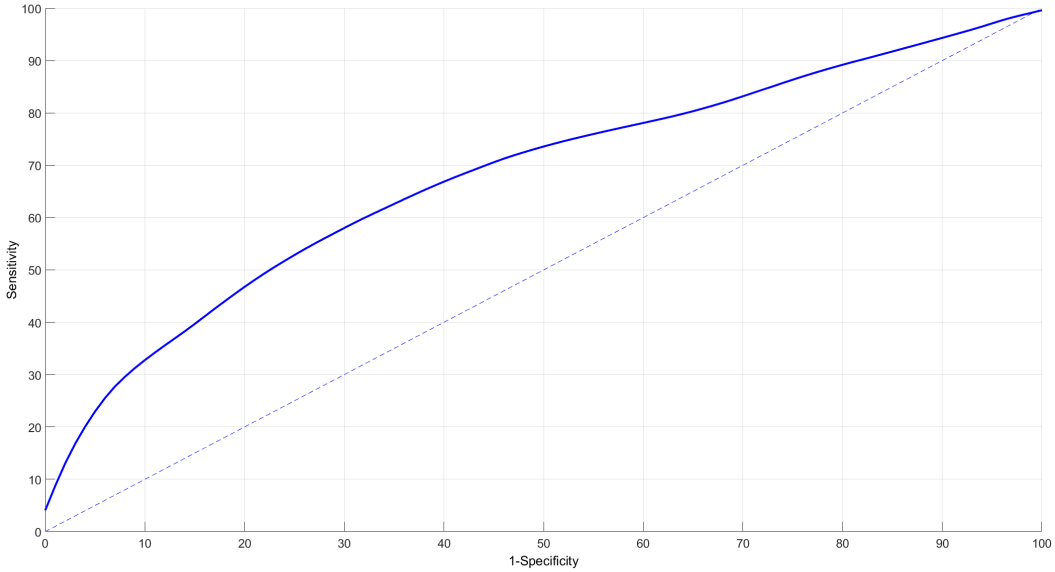
[a]AUC: area under the curve.

[b]ROC: receiver operating characteristic.

[c]SpO_2: oxygen saturation.

The amount of improvement possible as a result of combining vital signs in the classification algorithm can be seen from the two lower rows in Table 4. All 3 vital signs are derived from a single device and have a degree of correlation for physiological reasons. However, it can be seen that classifiers using respiratory rate, pulse rate, and SpO_2 as inputs give the best AUC results, suggesting that there is more predictive information captured when these vital signs are used together. These results suggest that, in order to predict 8 exacerbations out of 10 (80% sensitivity), there would be approximately 6 false alarms out of 10 (36% specificity). If instead the aim was to predict 6 exacerbations out of 10 (60% sensitivity), this would give rise to approximately 3 false alarms in every 10 (68% specificity).

Figure 8. Receiver operating characteristic showing the sensitivity and specificity (mean of 1000 iterations) when using the mean and gradient of pulse rate, SpO_2, and respiratory rate as input features to the classifiers. Note that the confidence interval is not shown because it is very narrow and the corresponding curves overlap the mean line on the scale shown.



Discussion

Principal Findings

Our digital health system enabled reliable and regular self-monitoring of both symptom and pulse oximetry data by patients with COPD for a sustained period of 12 months. This allowed the acquisition of a unique dataset, with more than 35,000 monitoring days (average of 5.3 times per week). We have also presented a systematic approach to modeling COPD exacerbations based on FSMs. This provides a deeper insight into how a COPD patient's condition progresses over time. For example, this study showed that in about 97% of the cases for which patients went from the normal state to the transitional state and then back again to the normal state, they did so without increasing their medication. This could be described as normal variation (ie, normal variation in symptoms that does not warrant a change in regular medication). We are not aware of any previous study that has been able to quantify and report this finding from such a large dataset.

A limitation of this study is that the results presented are based on the application of a specific definition of exacerbation. In addition, both the symptom diary information and the medication data used to identify exacerbation episodes are self-reported by the patients. Finally, any exacerbation event occurring while the patient is in a transitional state and contacts an HCP or is admitted to hospital is not identified in our analysis because of lack of availability of this information.

The first step in developing a predictive algorithm is to decouple periods for which a patient's condition is completely stable from those for which a patient is approaching an exacerbation event (the prodromal period). This paper described a possible approach to the development of a predictive algorithm robust to irregularly sampled data as well as missing data. The features used as inputs to the proposed algorithm were the mean and gradient of each of the vital signs (pulse rate and SpO_2) acquired by the pulse oximeter or estimated from the pulse oximetry PPG data (respiratory rate) for that 7-day period. A predictive algorithm based on physiological data rather than self-reported symptoms could be really helpful to support self-management in COPD. Pinnock et al [18] report that analysis of previous telemonitoring data suggests that it is the minority of patients who are able to log discrete episodes of increased breathlessness, cough, and sputum.

This study has demonstrated and quantified the predictive power of 3 vital signs (pulse rate, SpO_2, and respiratory rate) and also shown that they can reliably be obtained in home-monitoring with only a pulse oximeter. This coupled with the fact that only a minority of patients with COPD are able to log episodes of worsening symptoms suggests that the use of a pulse oximeter should be considered in any future study involving remote monitoring of patients with COPD. In addition to extracting respiratory rate, we also explored the possibility of gaining more information from the pulse oximeter waveform (the PPG). A patient in the prodromal period is likely to have a more irregular respiratory pattern, with a greater proportion of power in higher-frequency bands as his or her condition worsens. However, no features associated with the spectral content of the respiratory signal extracted from the PPG provided any further information once the respiratory rate was extracted. We also explored the use of the standard deviation and y-intercept of each of the vital signs in the 7-day periods as additional features, but again this did not lead to improvement in the performance of the predictive algorithm. We also investigated if the classification performance could be improved using support vector machines, decision trees, or k-nearest neighbor–based classifiers [19]. However, none of these algorithms led to any significant improvement with respect to logistic regression. This suggests that further improvements are more likely to come from the use of independent features not

explored in this study (eg, temperature, activity monitoring [20], heart rate variability [21], or automated sputum analysis [22]).

We are not aware of any previous work that has proposed FSM or similar models in order to identify COPD exacerbations and track the condition of patients with COPD during remote monitoring. Furthermore, we are not aware of any previous study that has reported respiratory rate measurements from patients with COPD during home monitoring (except one study [23], which only looked at patients requiring domiciliary oxygen, a very specific and narrow segment of the population of patients with COPD, for only 3 months). In this paper, we have reported the distribution of 3 vital signs (pulse rate, SpO_2, and respiratory rate) during both stable and prodromal periods based on measurements from 100 patients with COPD, providing very frequent measurements over a 12-month period. The majority of previous studies using telemonitoring for COPD were shorter than 12 months [7]. Of the few studies [24-27] that were at least 12 months long, only one [25] had at least 100 patients. However, in that study [25], the alerting algorithm was based on thresholds applied to the symptom score derived from a self-reported symptom diary.

Conclusions

Given the heterogeneous nature of COPD exacerbations and the large number of definitions in the literature, the FSM-based approach described in this paper can help systematize the analysis of COPD exacerbations and improve our understanding of how symptoms worsen and of the impact of medication.

We showed that all the 3 vital signs (SpO_2, respiratory rate, pulse rate) are predictive of exacerbation events, with a combination of these vital signs resulting in the best AUC result. SpO_2 was the most predictive vital sign, followed by respiratory rate, and pulse rate was the least predictive. It is important to think about the trade-off between sensitivity and specificity in this context: false alerts will result in overmedication, but the ability to predict exacerbations may avoid costly hospital admissions.

None of the other classification algorithms (support vector machines, decision trees, and k-nearest neighbor–based classifiers) led to any significant improvement compared with logistic regression. Additional improvements in COPD exacerbation prediction are likely to come instead from the use of additional physiological measures, behavioral measures, and biomarkers (eg, temperature, heart rate variability [21], activity [20], and sputum analysis [22]) as input to the prediction algorithm. The other alternative is to personalize the prediction algorithms over time and adapt the classifier to the range of vital sign values that are typical for that particular individual.

Acknowledgments

The authors thank all members of the EDGE COPD research team: Dr Oliver Gibson, Dr Andreas Triantafyllidis, Dr Gari Clifford, Dr Carl Heneghan, Heather Rutter, Dr Veronika Williams, Dr Maxine Hardinge, Dr Christy Toms, Dr Louise Jones, Stephanie Robinson, Dr Jonathan Price, Linda Heritage, and the patients who took part in the pilot study and RCT. This publication presents independent research supported by the Health Innovation Challenge Fund (HICF-1010-032), a parallel funding partnership between the Department of Health and Wellcome Trust. LT and AF receive funding from the National Institute for Health Research (NIHR) Biomedical Research Centre, and AF is an NIHR Senior Investigator. The views expressed in this publication are those of the authors and not necessarily those of the Department of Health or Wellcome Trust.

Authors' Contributions

SAS and CV led the data analysis and technical development. AF and LT led the project and secured funding for it. SAS wrote the initial draft of the paper and all coauthors contributed to and approved the final version of the manuscript.

Conflicts of Interest

None declared.

References

1. Mathers CD, Loncar D. Projections of global mortality and burden of disease from 2002 to 2030. PLoS Med 2006 Nov;3(11):e442 [FREE Full text] [Medline: 17132052]
2. Mannino DM, Buist AS. Global burden of COPD: risk factors, prevalence, and future trends. Lancet 2007 Sep 1;370(9589):765-773. [doi: 10.1016/S0140-6736(07)61380-4] [Medline: 17765526]
3. Wedzicha JA, Seemungal TA. COPD exacerbations: defining their cause and prevention. Lancet 2007 Sep 1;370(9589):786-796. [doi: 10.1016/S0140-6736(07)61382-8] [Medline: 17765528]
4. Pauwels R, Calverley P, Buist AS, Rennard S, Fukuchi Y, Stahl E, et al. COPD exacerbations: the importance of a standard definition. Respir Med 2004 Feb;98(2):99-107. [Medline: 14971871]
5. Seemungal TA, Donaldson GC, Bhowmik A, Jeffries DJ, Wedzicha JA. Time course and recovery of exacerbations in patients with chronic obstructive pulmonary disease. Am J Respir Crit Care Med 2000 May;161(5):1608-1613. [doi: 10.1164/ajrccm.161.5.9908022] [Medline: 10806163]

6. Ure J, Pinnock H, Hanley J, Kidd G, McCall SE, Tarling A, et al. Piloting tele-monitoring in COPD: a mixed methods exploration of issues in design and implementation. Prim Care Respir J 2012 Mar;21(1):57-64 [FREE Full text] [doi: 10.4104/pcrj.2011.00065] [Medline: 21785816]
7. Sanchez-Morillo D, Fernandez-Granero MA, Leon-Jimenez A. Use of predictive algorithms in-home monitoring of chronic obstructive pulmonary disease and asthma: a systematic review. Chron Respir Dis 2016 Aug;13(3):264-283. [doi: 10.1177/1479972316642365] [Medline: 27097638]
8. Hardinge M, Rutter H, Velardo C, Shah SA, Williams V, Tarassenko L, et al. Using a mobile health application to support self-management in chronic obstructive pulmonary disease: a six-month cohort study. BMC Med Inform Decis Mak 2015 Jun 18;15:46 [FREE Full text] [doi: 10.1186/s12911-015-0171-5] [Medline: 26084626]
9. Velardo C, Shah SA, Gibson O, Clifford G, Heneghan C, Rutter H, et al. Digital health system for personalised COPD long-term management. BMC Med Inform Decis Mak 2017 Feb 20;17(1):19. [doi: 10.1186/s12911-017-0414-8] [Medline: 28219430]
10. Shah SA, Velardo C, Gibson OJ, Rutter H, Farmer A, Tarassenko L. Personalized alerts for patients with COPD using pulse oximetry and symptom scores. Conf Proc IEEE Eng Med Biol Soc 2014;2014:3164-3167. [doi: 10.1109/EMBC.2014.6944294] [Medline: 25570662]
11. Velardo C, Shah SA, Gibson O, Rutter H, Farmer A, Tarassenko L. Automatic generation of personalised alert thresholds for patients with COPD. 2014 Presented at: European Signal Processing Conference (EUSIPCO); 1-5 September 2014; Lisbon, Portugal URL: http://ieeexplore.ieee.org/document/6952738/
12. Mano MM, Kime CR. Logic and computer design fundamentals. Upper Saddle River, NJ: Prentice Hall; 2000.
13. Donaldson GC, Wedzicha JA. COPD exacerbations .1: Epidemiology. Thorax 2006 Feb;61(2):164-168 [FREE Full text] [doi: 10.1136/thx.2005.041806] [Medline: 16443707]
14. Shah SA. Vital sign monitoring and data fusion for paediatric triage (DPhil Thesis). University of Oxford; 2012. URL: https://ora.ox.ac.uk/objects/uuid:80ae66e3-849b-4df1-b064-f9eb7530200d/datastreams/ATTACHMENT01 [WebCite Cache ID 6olnxjFQ2]
15. Stein PK, Kleiger RE, Rottman JN. Differing effects of age on heart rate variability in men and women. Am J Cardiol 1997 Aug 01;80(3):302-305. [Medline: 9264423]
16. Rajendra Acharya U, Paul Joseph K, Kannathal N, Lim CM, Suri JS. Heart rate variability: a review. Med Biol Eng Comput 2006 Dec;44(12):1031-1051. [doi: 10.1007/s11517-006-0119-0] [Medline: 17111118]
17. Shah SA, Fleming S, Thompson M, Tarassenko L. Respiratory rate estimation during triage of children in hospitals. J Med Eng Technol 2015;39(8):514-524. [doi: 10.3109/03091902.2015.1105316] [Medline: 26548638]
18. Pinnock H, Steed L, Jordan R. Supported self-management for COPD: making progress, but there are still challenges. Eur Respir J 2016 Jul;48(1):6-9. [doi: 10.1183/13993003.00576-2016] [Medline: 27365502]
19. Theodoridis S, Koutroumbas K. Pattern Recognition, Fourth Edition. London: Academic Press; 2009.
20. Waschki B, Spruit MA, Watz H, Albert PS, Shrikrishna D, Groenen M, et al. Physical activity monitoring in COPD: compliance and associations with clinical characteristics in a multicenter study. Respir Med 2012 Apr;106(4):522-530 [FREE Full text] [doi: 10.1016/j.rmed.2011.10.022] [Medline: 22118987]
21. Carvalho TD, Pastre CM, de Godoy MF, Fereira C, Pitta FO, de Abreu LC, et al. Fractal correlation property of heart rate variability in chronic obstructive pulmonary disease. Int J Chron Obstruct Pulmon Dis 2011 Jan 06;6:23-28 [FREE Full text] [doi: 10.2147/COPD.S15099] [Medline: 21311690]
22. Bafadhel M, McKenna S, Terry S, Mistry V, Reid C, Haldar P, et al. Acute exacerbations of chronic obstructive pulmonary disease: identification of biologic clusters and their biomarkers. Am J Respir Crit Care Med 2011 Sep 15;184(6):662-671. [doi: 10.1164/rccm.201104-0597OC] [Medline: 21680942]
23. Yañez AM, Guerrero D, Pérez de Alejo R, Garcia-Rio F, Alvarez-Sala JL, Calle-Rubio M, et al. Monitoring breathing rate at home allows early identification of COPD exacerbations. Chest 2012 Dec;142(6):1524-1529. [doi: 10.1378/chest.11-2728] [Medline: 22797131]
24. Johnston NW, Lambert K, Hussack P, Gerhardsson de Verdier M, Higenbottam T, Lewis J, et al. Detection of COPD Exacerbations and compliance with patient-reported daily symptom diaries using a smart phone-based information system [corrected]. Chest 2013 Aug;144(2):507-514. [doi: 10.1378/chest.12-2308] [Medline: 23519329]
25. Pinnock H, Hanley J, McCloughan L, Todd A, Krishan A, Lewis S, et al. Effectiveness of telemonitoring integrated into existing clinical services on hospital admission for exacerbation of chronic obstructive pulmonary disease: researcher blind, multicentre, randomised controlled trial. Br Med J 2013;347:f6070 [FREE Full text] [Medline: 24136634]
26. Mohktar MS, Redmond SJ, Antoniades NC, Rochford PD, Pretto JJ, Basilakis J, et al. Predicting the risk of exacerbation in patients with chronic obstructive pulmonary disease using home telehealth measurement data. Artif Intell Med 2015 Jan;63(1):51-59. [doi: 10.1016/j.artmed.2014.12.003] [Medline: 25704112]
27. Antoniades NC, Rochford PD, Pretto JJ, Pierce RJ, Gogler J, Steinkrug J, et al. Pilot study of remote telemonitoring in COPD. Telemed J E Health 2012 Oct;18(8):634-640. [doi: 10.1089/tmj.2011.0231] [Medline: 22957501]

Abbreviations

AUC: area under the curve
COPD: chronic obstructive pulmonary disease
EDGE: Self Management and Support Programme
FSM: finite-state machine
GOLD: Global Initiative for Chronic Obstructive Lung Disease
HCP: health care professional
PPG: photoplethysmogram
RCT: randomized controlled trial
ROC: receiver operating characteristic
WHO: World Health Organization

Edited by G Eysenbach; submitted 21.12.16; peer-reviewed by B Kim, V Barber; comments to author 01.02.17; accepted 14.02.17; published 07.03.17.

Please cite as:
Shah SA, Velardo C, Farmer A, Tarassenko L
Exacerbations in Chronic Obstructive Pulmonary Disease: Identification and Prediction Using a Digital Health System
J Med Internet Res 2017;19(3):e69
URL: http://www.jmir.org/2017/3/e69/
doi:10.2196/jmir.7207
PMID:28270380

Viewpoint

A Learning Health Care System Using Computer-Aided Diagnosis

Amos Cahan[1], MD; James J Cimino[2], MD

[1]IBM TJ Watson Research Center, Yorktown Heights, NY, United States

[2]Informatics Institute, University of Alabama at Birmingham, Birmingham, AL, United States

Corresponding Author:
Amos Cahan, MD
IBM TJ Watson Research Center
1101 Kitchawan Rd
Route 134
Yorktown Heights, NY, 10598
United States
Phone: 1 914 945 2590
Fax: 1 914 945 2590
Email: acahan@us.ibm.com

Abstract

Physicians intuitively apply pattern recognition when evaluating a patient. Rational diagnosis making requires that clinical patterns be put in the context of disease prior probability, yet physicians often exhibit flawed probabilistic reasoning. Difficulties in making a diagnosis are reflected in the high rates of deadly and costly diagnostic errors. Introduced 6 decades ago, computerized diagnosis support systems are still not widely used by internists. These systems cannot efficiently recognize patterns and are unable to consider the base rate of potential diagnoses. We review the limitations of current computer-aided diagnosis support systems. We then portray future diagnosis support systems and provide a conceptual framework for their development. We argue for capturing physician knowledge using a novel knowledge representation model of the clinical picture. This model (based on structured patient presentation patterns) holds not only symptoms and signs but also their temporal and semantic interrelations. We call for the collection of crowdsourced, automatically deidentified, structured patient patterns as means to support distributed knowledge accumulation and maintenance. In this approach, each structured patient pattern adds to a self-growing and -maintaining knowledge base, sharing the experience of physicians worldwide. Besides supporting diagnosis by relating the symptoms and signs with the final diagnosis recorded, the collective pattern map can also provide disease base-rate estimates and real-time surveillance for early detection of outbreaks. We explain how health care in resource-limited settings can benefit from using this approach and how it can be applied to provide feedback-rich medical education for both students and practitioners.

(J Med Internet Res 2017;19(3):e54) doi:10.2196/jmir.6663

KEYWORDS

diagnostic errors; diagnosis, computer-assisted; decision support systems, clinical; pattern recognition, automated; knowledge bases; knowledge management; diagnosis support systems; crowdsourcing; structured knowledge representation

Why We Need Computer-Aided Diagnosis

We Make Too Many Diagnostic Errors

Two main questions are key when evaluating a patient in the context of constructing a differential diagnosis: The first is "How representative is the presentation of the patient to a set of manifestations of a known disease?" In other words, to what degree is there a match between a set of symptoms, signs, and laboratory results and the clinical features of the disease. The second is "What is the likelihood of encountering that disease in a patient like this?" Answering this question requires knowing the base rate (ie, incidence) of the disease and accounting for any patient risk factors that may alter the patient's prior probability of having the disease.

Good clinicians are characterized by their ability to cluster findings around a single process or cause. Their intuitive clinical assessment heavily relies on pattern recognition [1]. It has been argued [2] that, to develop skilled intuition, a predictable environment and adequate opportunity to practice skills are needed. Indeed, physicians are more likely to be wrong in cases where they have encountered too few instances of a pattern to recognize it.

However, a perfect match between a patient's presentation and a typical clinical picture of a disease is no guarantee that the

patient indeed has that disease. The aphorism coined by Dr Theodore Woodward, "When you hear hoofbeats, think of horses not zebras" [3], reflects the importance of the disease's prior probability in the population to which the patient belongs. In other words, an atypical presentation of a common disease is probably more likely to be encountered than a classic presentation of a rare disease.

Although probabilistic reasoning is key to medical diagnosis, physicians, like humans in general, perform poorly in this aspect; probability overestimation and low between-physician agreement are common [4-6]. The US National Academy of Medicine's recently published *Improving Diagnosis in Health Care* report [7] points to the unacceptable number of patients harmed by diagnostic errors. With medical knowledge rapidly expanding, information lacunae are common [8]. Doctors use heuristics (mental shortcuts) to compensate for knowledge gaps, but this practice involves substantial biases [9,10], which may contribute to diagnostic errors. Thus, to minimize diagnostic errors, doctors need help.

Attempts to develop efficient computer-aided diagnosis support systems (DSSs) [11,12], including differential diagnosis generators, have been made since the 1960s (reviewed in [13,14]). Whereas narrow spectrum, rule-based systems, such as electrocardiogram interpreters, have become ubiquitous over the years, DSSs in the general or internal medicine domain have not [15]. This, despite a demonstrated positive effect on physician performance [16]. Several general medicine DSSs are available commercially (eg, DXplain [17], GIDEON [18], and Isabel [19]), but their routine clinical use remains limited. In fact, only 5 of 11 differential diagnosis generators included in a recent systematic review [20] are currently used in practice. Despite the digital revolution of the health care system in the last decade, a 2013 review of DSSs concluded that progress in the development of DSSs during this time was minimal [21]. However, recent years have seen a new class of computerized diagnosis tools aimed at patients. These resources, called "symptom checkers," suggest potential diagnoses explaining a user's set of symptoms as reported through a user interface [22]. A recent evaluation of 23 symptom checkers [23] using standardized cases found that the correct diagnosis was listed among the top 20 options in 58% (95% CI 55%-62%) and appeared first on the list in 34% (95% CI 31%-37%). Symptom checkers provided inappropriate triage advice in 20% of emergent and 45% of nonemergent cases.

Why Diagnosis Support Systems Fail

Poor specificity of DSSs is reflected by the large number of possible diagnoses suggested. Berner and coworkers [24], in a report on leading general DSSs, used "relevance," a measure closely related to specificity. They defined relevance as the average proportion of computer-generated diagnoses considered reasonable by clinical experts. Mean relevance scores were low (ranging from 0.19 to 0.37). Poor relevance rates mean that a workup plan based on DSS-suggested diagnoses would be impractical and would expose patients to undue risks [25].

Indeed, by design, DSSs focus primarily on sensitivity at the expense of specificity [26], yet in Berner and colleague's report, the overall sensitivity (the proportion of cases where the correct diagnosis was included in the computer-generated differential diagnosis list) ranged from 0.52 to 0.71. When they considered only the top 5 and 10 diagnoses on the list, rates were only about 0.25-0.35 for relevance and 0.35-0.45 for sensitivity. A follow-up study found similar results in a different set of DSSs [26]. In fact, a Google Web search has been reported to provide the correct diagnosis at similar rates [27].

Current DSSs cannot efficiently match patients and diseases on patterns, since they rely on a unidimensional projection of clinical information; typically, the system uses a vector of "findings" (symptoms, signs and laboratory results) provided by the user to generate a differential diagnosis. Some systems differentiate between acute and more prolonged processes [17,18], but none are able to cluster findings based on their course in time. Using this "bag of findings" approach makes DSSs agnostic to key clinical clues. For instance, chest pain and dyspnea appearing during physical exercise strongly suggest angina, whereas shortness of breath with subsequent chest pain may suggest pneumothorax or a pulmonary infarction. From a DSS point of view, these conditions are indistinguishable, as both have the same findings: "chest pain" and "dyspnea" (Figure 1). Indeed, some systems define higher-level concepts such as "angina chest pain" [28], but these are rigid and limited in number.

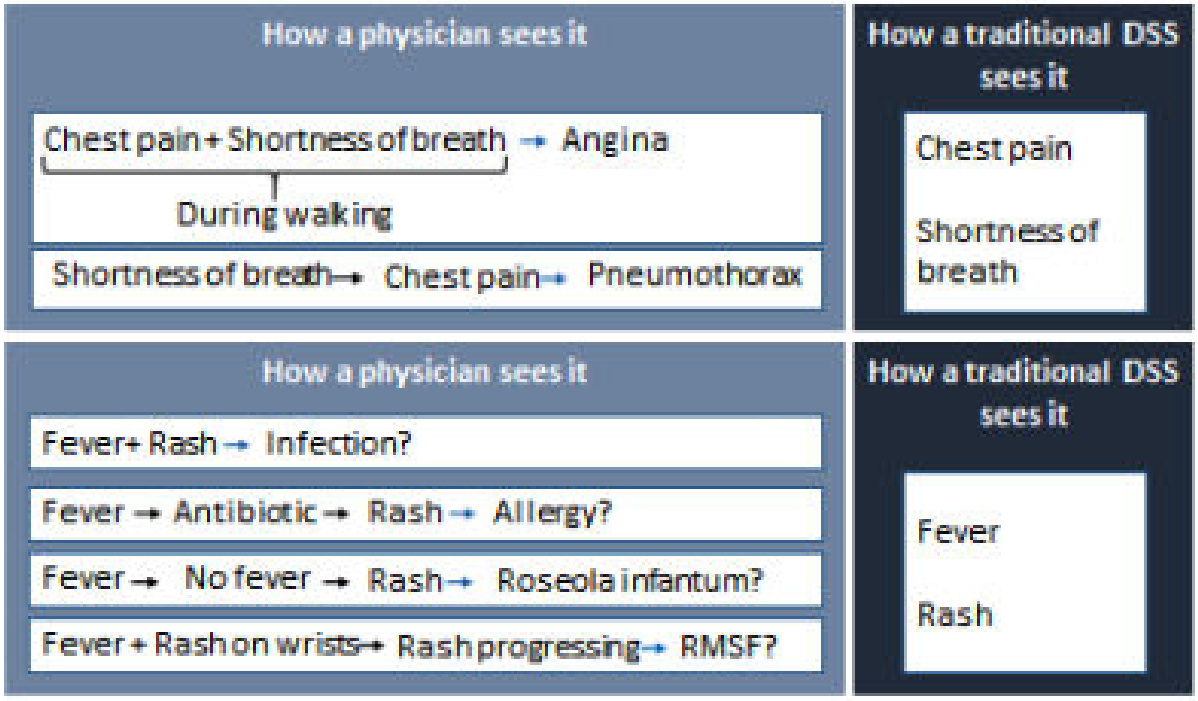
An incomplete knowledge base further limits DSS performance. INTERNIST-1 included 570 diseases [29], and DXplain's knowledge base has over 2400 diseases and 5000 findings [30]. Nevertheless, when 4 DSSs were evaluated on a set of challenging clinical cases, the correct diagnosis was absent from their knowledge base in 9% to 27% of cases [24]. Manually curating and maintaining a comprehensive knowledge base in the face of rapid knowledge growth is extremely demanding. In fact, leading informaticians have acknowledged that even working toward a complete DSS knowledge base is infeasible [15]. In a seminal paper published in 1959 [31], Ledley and Lusted proposed a physician-maintained, notched card-based "learning device." This device could be used to collect and reuse associations between symptoms and diseases, with each card representing a patient. This early attempt to capture physician knowledge as applied to a single patient in a structured, machine-interpretable format allowed, in theory, for supporting a learning DSS. An attempt to implement knowledge accumulation through actual cases was made in the 1980s by the creators of the discontinued ILIAD expert system [28]. More recently, the Web-based Human Diagnosis Project [32] was introduced, which allows clinicians to upload real or simulated case vignettes, and challenges other clinicians to solve those cases. This information is used to generate a knowledge graph associating symptoms, signs, and diagnoses. Nevertheless, to the best of our knowledge, there is no commercially available DSS that has self-learning capabilities.

Almost 60 years after Ledley and Lusted [31] had laid the foundations for probabilistic reasoning in diagnosis making (with ILIAD being an exception to some degree), DSSs still do not hold disease base rates as part of their knowledge bases and are unable to account for factors that may alter the prior probability of a disease. Some diseases are limited to certain geographic locations, while the incidence of others varies by

the time of year or by race. Readily available in the digital age, these data remain unused by current DSSs to refine their differential diagnosis.

Finally, DSSs do not align well with clinicians' work flow. A few DSSs now offer variable degrees of direct connectivity to the electronic health record (EHR) [33]. Some can extract data from the EHR using natural language processing tools, although this may adversely affect performance [34].

Figure 1. Examples of the view of a set of findings in a patient by a physician and a traditional diagnosis support system (DSS): chest pain and shortness of breath (upper panel), fever and rash (lower panel). Temporal and semantic interrelations between findings are crucial in putting findings in the right clinical context. RMSF, Rocky mountain spotted fever.



Next-Generation Diagnosis Support Systems

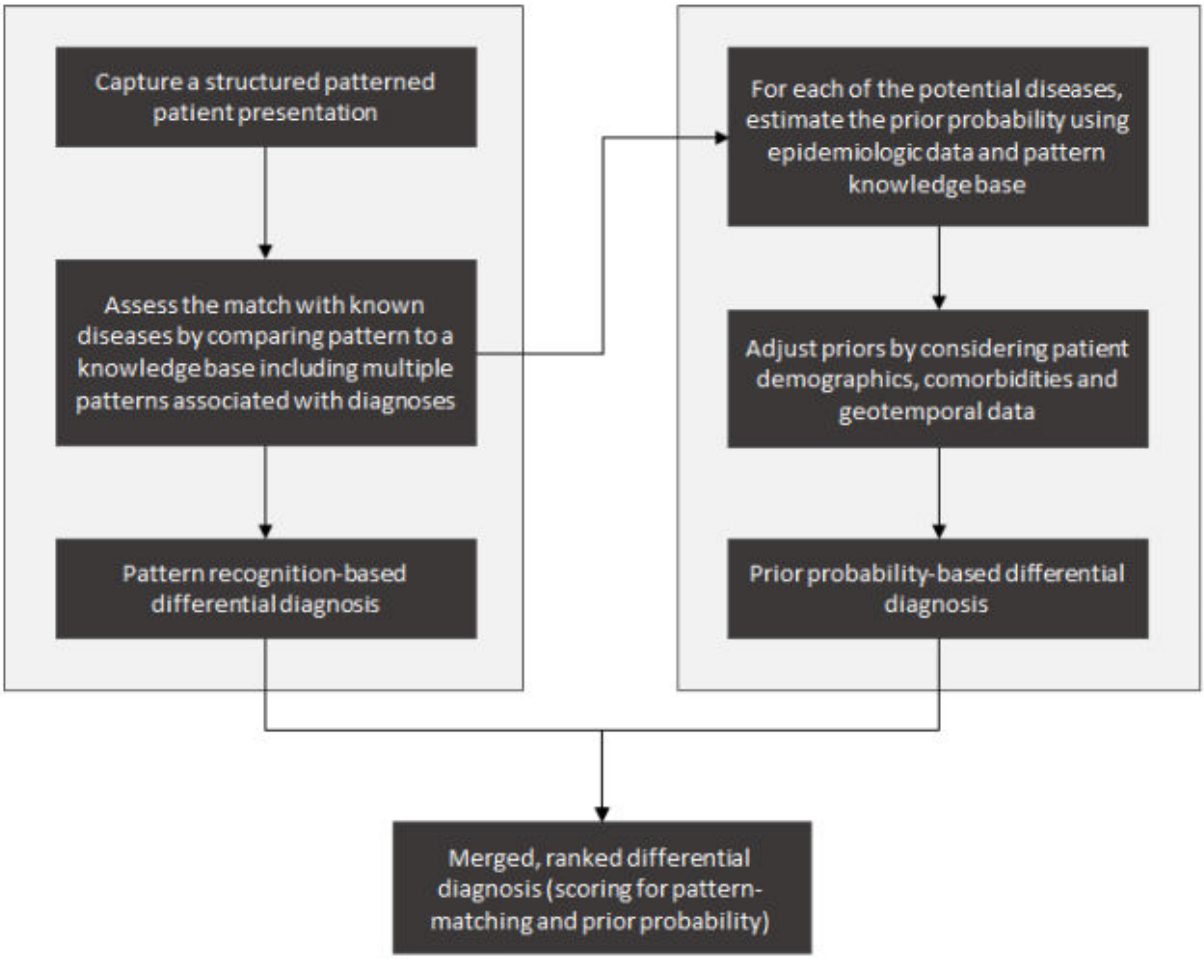
Reviewing DSSs in 1994 [13], Miller noted that

> *We may understand, in theory, how to develop systems that take into account gradations of symptoms, the degrees of uncertainty...the severity of each illness under consideration, the pathophysiologic mechanisms of disease, and/or the time courses of illness. However, it is not yet practical to build such broad-based systems for patient care.*

More recently, Weber et al note that "industries have figured out...that big data becomes transformative when disparate data sets can be linked at the individual person level" [35]. Technology is now ripe to enable the development of next-generation DSSs (NGDSSs) based on these key insights.

Here we portray NGDSSs and provide a conceptual framework for their development (Figure 2). To be effective, NGDSSs will have to (1) support pattern recognition-based diagnosis by capturing a richer clinical picture, (2) provide personalized prior-probability assessments, (3) maintain a comprehensive and current knowledge base, and (4) better align with clinicians' workflow. In the next sections, we discuss NGDSSs characteristics.

Figure 2. An integrated approach to computer-aided diagnosis. The process addresses the 2 questions that lead to likely diagnoses: Left side: How similar is the presentation of the patient to a set of manifestations of a known disease? Right side: What is the likelihood of encountering that disease in a patient like this?



Structured Patterned Patient Presentation

For hundreds of years, physicians have been documenting their thoughts in the form of free text. Computer systems work best with structured data, but free-text input prevails even in the age of EHRs. A digital form of Ledley and Lusted's learning device could provide a structured representation of a patient's symptoms, extracted using natural language processing, as they relate to a disease. However, such a representation would, again, be of the "bag of findings" kind, as natural language processing techniques cannot reliably generate structured representations of complex clinical concepts documented in the EHR notes. In particular, this is true for temporal and semantic interrelations between symptoms and signs, which are key in forming the clinical patterns recognized by physicians. Thus, new ways to provide DSSs with a structured clinical picture are needed.

We suggest that a structured, higher cognitive-level patient representation can be constructed in real time through a (graphical) machine-physician interaction. We refer to this representation as a "structured presentation pattern" or "structured pattern." A structured pattern can be thought of as a model, which can represent physician knowledge and reasoning in a machine-interpretable format. A structured pattern should ideally represent key symptoms and signs associated with a particular patient's presentation and their temporal and semantic interrelations. This allows for translation of a list of findings (symptoms and signs) into multiple distinct structured patterns according to the temporal course of the disease and other relations between findings. Through this approach, a differential diagnosis constructed by NGDSSs is likely to be more specific than one based on a list of findings.

Temporal Patterns

The creators of the pioneering INTERNIST-1 attributed its insufficient clinical reliability in part to it's being temporally naive [36]. Kohane [37], appreciating the importance of temporal patterns in clinical reasoning, concluded that "knowledge bases that fail to capture the temporal component of the course of disease omit useful diagnostic knowledge." It is hard to estimate the percentage of cases in which the course in time of the manifestations of a disease plays a substantial role in diagnosis making. Yet the fundamental categorization in clinical medicine of illnesses as acute, subacute, and chronic attests to the central role of the temporal dimension in differentiating between diseases. Still, efforts to capture, model, and represent temporal relations between clinical entities [37-40] have not yet matured to support temporally aware DSSs.

Semantic Patterns

One piece of evidence can be the cause of another, one may support or contradict the other, or one may be more reliable than another. Interpretation of a symptom or sign is ever dependent on the clinical context, which is, in a sense, a sum of all such interrelations. For example, a patient with suspected brucellosis may be unsure of having consumed potentially unpasteurized milk products in the preceding weeks. A physician auscultating the heart may hear an extra sound during diastole but have doubts as to whether this is an opening snap or a third heart sound. A patient's record may document contradicting views of the etiology of a prior illness (eg, convulsion vs transient ischemic attack). Making at least some of these semantic or contextual interrelations interpretable by NGDSSs is likely to improve their performance on pattern recognition.

Using Big Data to Create a Differential Diagnosis

Patient notes include protected health information, which is why individual medical records cannot be readily shared. Free text can only be considered to be deidentified after it has been manually reviewed. In contrast, user-generated structured patient patterns are readily automatically de-deidentifiable.

This opens the way to *real-time* sharing of high-quality deidentified clinical patient information across physicians and institutions. The collective experience of physicians worldwide can be stored in a structured knowledge base made available to support pattern recognition-based diagnosis. Similarity analysis could support this process by computing the degree of match between a patient's pattern and patterns of other patients in the knowledge base who already have a diagnosis. This can provide an answer to the first diagnostic question mentioned above, namely, "How similar is the presentation of the patient to a set of manifestations of a known disease?"

Disease prevalence by parameters such as age, sex, race, and geotemporal distribution can be extracted from various sources, including published reports, large EHR repositories, administrative claims data, social media, and environmental data (eg, weather). These sources can feed an NGDSS knowledge base. Patient demographic data automatically extracted from the EHR can personalize prior-probability estimates. Few findings typical of a particular disease are invariably present in every case of it. The probability of a certain symptom occurring, a certain sign being noted, or a particular laboratory abnormality being found in a given disease is available from published reports [41] and can be used by NGDSSs. In fact, as the NGDSS knowledge base expands, the collective body of structured patterns contributed by physicians can serve as a living epidemiologic database, providing real-time statistics on the incidence of symptoms, signs, and diseases. This will enable NGDSSs to address the second diagnostic question: "What is the likelihood of encountering that disease in a patient like this?"

A Democratized Medical Knowledge Base

DSSs partly rely on the fact that disease manifestations change relatively little over time, yet as new diseases arise (with obvious examples being human immunodeficiency virus and Zika virus infections) and new disease correlates are found (eg, genetic traits), continuous updates are necessary [13].

Using structured patterns, crowdsourcing of knowledge collection and reuse becomes possible. Crowdsourcing may be a sustainable strategy in a reality of exploding knowledge and limited resources (Figure 3).

Each time a physician adds a patient pattern (subsequently labeled with a diagnosis code assigned to that patient), the NGDSS knowledge base is enriched. An initial core body of knowledge may be manually curated by translating disease entries in a textbook into structured patterns of diseases (Figure 4, panel C provides an example [42]).

Figure 3. Generating a real-time structured representation of a patient presentation supports a computer-aided diagnostic process (blue arrows) and a learning health care system through knowledge reuse (gray arrows).

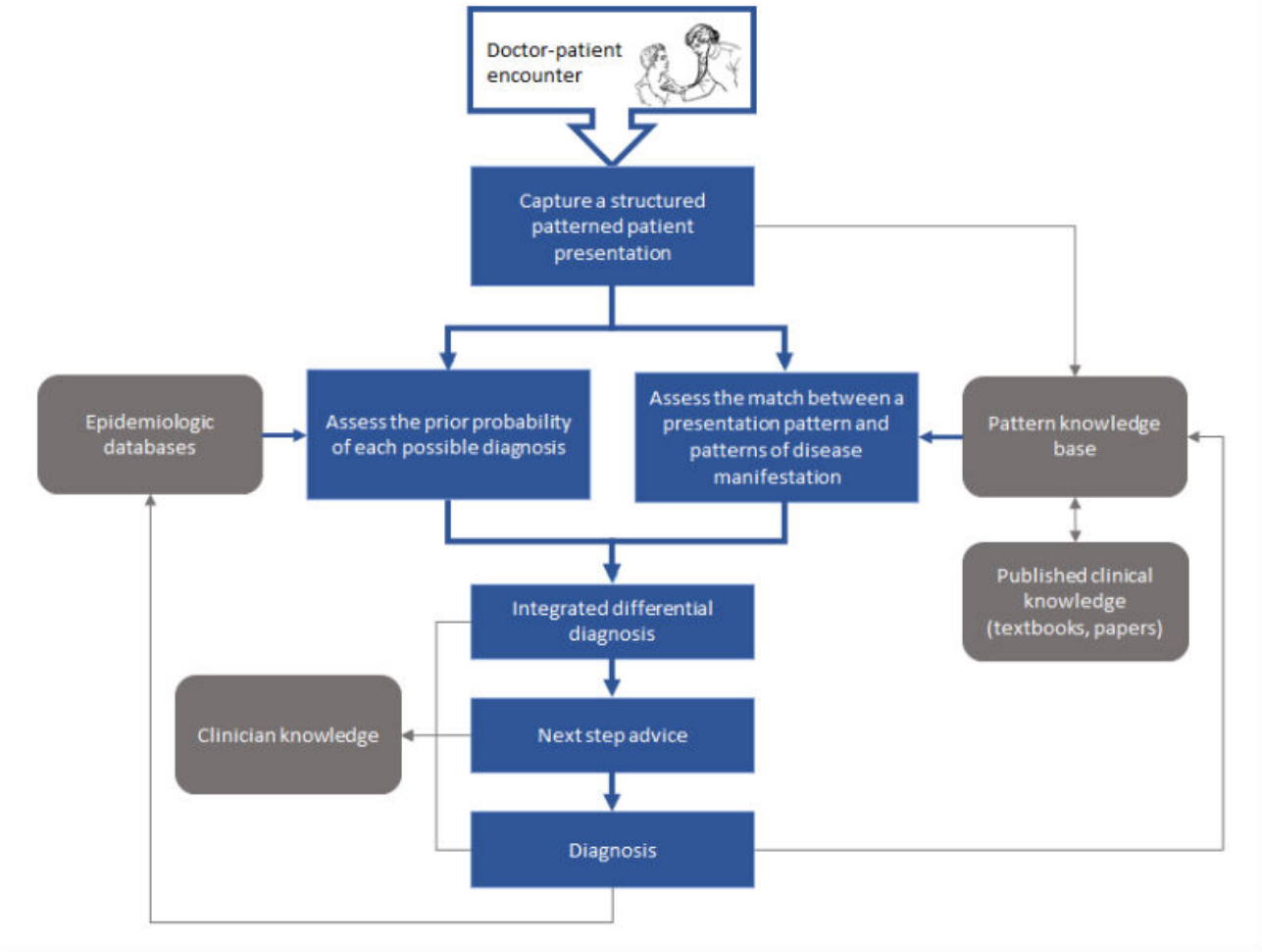
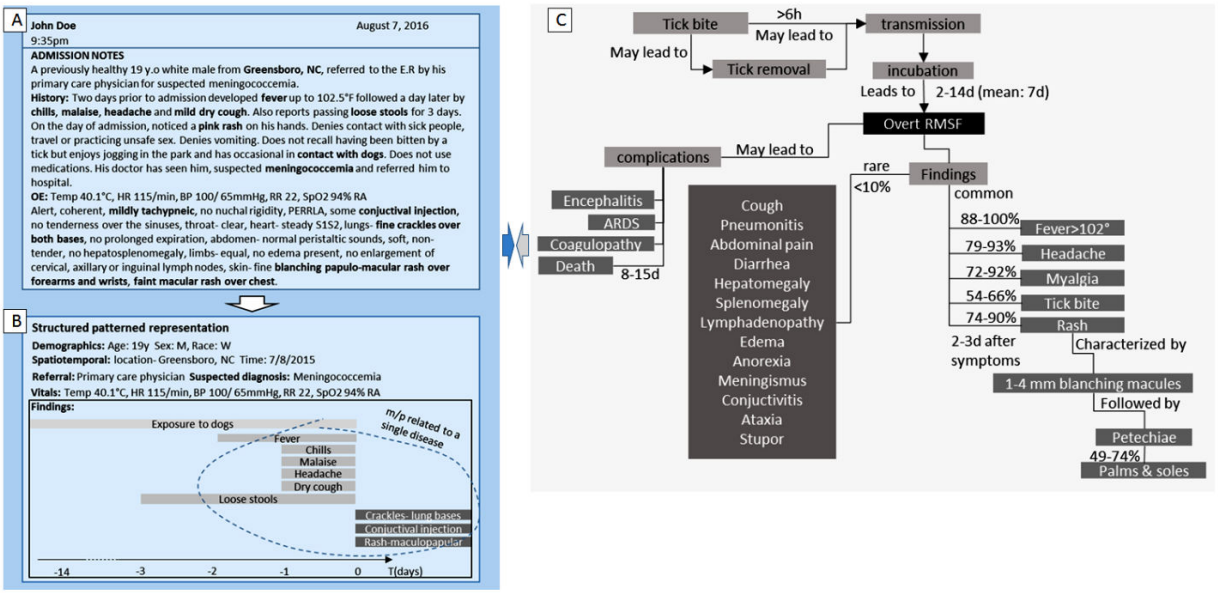


Figure 4. Structured patient and disease representation. (A) A simulated view of an electronic health record with admission notes. Key terms are highlighted automatically using a real-time natural language processing engine or marked by the user. (B) Selected terms are then manipulated by the user by means of a touch screen to create a pattern representing key temporal and semantic interrelations between terms in a structured format. This pattern is augmented by automatically extracted relevant clinical data, demographics, and other metadata. (C) Structured patterned representation of the manifestations of Rocky Mountain spotted fever derived from a review article. Applying analytics schemes for assessing patient and disease similarity in context of disease prevalence can inform the generation of a ranked differential diagnosis for the patient in question.



Prioritizing Possible Diagnoses

Considerations beyond the prior probability of potential diagnoses on the differential diagnosis list come into play when making clinical decisions on investigation and treatment. The most probable disease may be of little practical importance to the patient's outcome. On the other hand, missing the diagnosis of a severe, albeit less-likely, disease on the differential diagnosis may have grave consequences. Thus, the test and test-treat thresholds [43] may vary by disease severity. Seriousness may be reflected by measurable factors such as survival rate, complication rate, quality of life, and impact on productivity. Scoring for disease severity can be used to help in prioritizing the use of diagnostic tests and treatment modalities.

Likewise, the degree of urgency of conditions on the differential diagnosis list also has practical implications. Some conditions are considered medical emergencies (eg, stroke, malignant hypertension, or myocardial infarction) and require immediate measures to be taken by the physician, whereas in others the course and outcome are not changed by delaying treatment. In presenting information to the user, an NGDSS may indicate the need to act fast when such conditions are considered.

Distinguishing Between Conditions on the Differential Diagnosis List

NGDSSs should provide next-step advice to optimize the diagnostic workup. Listing questions that, if answered, could narrow the differential diagnosis can be useful. Performance measures of diagnostic tests, contraindications for their use, and complication rates could be incorporated in their knowledge bases [44]. This information can be applied in the case of an individual patient to simulate a posttest probability given a prior probability and test result, such as using likelihood ratios [45]. Local factors such as test availability, costs, and medical insurance coverage can be considered to adapt generic recommendations to available resources. In the face of high rates of overuse and misuse of diagnostic tests [46], NGDSSs could reduce patient harm and costs by optimizing the diagnostic workup.

Alignment With Workflow

Experience with current DSSs shows that their use is hindered by poor alignment with the clinical workflow. Most DSSs require at least some degree of redundant input of clinical information. NGDSSs must seamlessly integrate with EHR systems. Cognitive computing approaches can facilitate the interaction of physicians with NGDSSs. For example, a structured pattern may be interactively created using graphic user interfaces and touch screens (see Textbox 1 and Figure 4 for a suggested practical approach).

Textbox 1. Computer-aided diagnosis using a 3-step human-computer interactive process to capture structured patterned patient and disease representation.

1. *On-the-fly, user-supervised, automated term extraction from clinical notes:* As a physician is typing in clinical notes, a real-time list of natural language processing-extracted medical terms is generated by mapping notes to standard controlled terminology. Terms are then reviewed by the physician, who can check those that apply to the patient or select keywords from the text and quickly add them to the term list.

2. *Interactive creation of a structured "pattern" by connecting related terms*: With the advent of capacitive touch screens, dragging, pinching, and swiping have become intuitive to anyone using a mobile phone. A structured patient timeline can be generated by the physician by arranging the extracted keywords on a time axis. This process would turn a unidimensional list of terms into a 2-dimensional pattern. Much like temporal relations, semantic relations between terms could be annotated by the physician. Among those could be cause-and-effect relations, degree of certainty, or contradicting evidence. Graphic determinants such as shape, size, color, or transparency may be used to distinguish between event types or represent symptom severity. Physician time constraints will limit the granularity and richness of the patterns created. Nevertheless, even a structured patterned sketch of a patient's presentation can carry more information than a list of findings.

3. *Pattern enhancement using automatically extracted patient data*. The structured presentation sketch can serve as an anchor for structured data automatically extracted from the electronic health record to enhance the pattern. Such data types may include vital signs, laboratory results, keywords from imaging and pathology reports, and background medical information (eg, comorbidities).

A Bayesian network could be continuously trained to match a new patient pattern on a large set of existing patterns, and to rank the diagnoses to which similar patterns are attributed by their prior probability.

Limitations, Challenges and Potential Solutions

Implementation of the proposed approach for NGDSSs requires major health care stakeholders to make substantial, prolonged, and coordinated efforts. To bring NGDSSs to life, major technical and regulatory challenges will have to be met. Here, we mention some of the barriers NGDSSs face and propose ways to overcome them.

Crowdsourcing Data Quality

The medical domain is characterized by tight regulation of knowledge to assure quality. In this sense, crowdsourcing is an unorthodox approach. While offering access to much more knowledge than is possible using traditional methods, crowdsourcing carries an obvious risk of collecting unreliable information. Labeling structured patterns with the diagnosis subsequently made would be accurate in some cases; however, with misdiagnosis being not uncommon [7], some labels will undoubtedly be wrong. Like in other domains where big data is used, large numbers may provide protection against misinformation. Ongoing expert evaluation of outliers in the knowledge base can also help eliminate errors. Ways to select appropriate contributors will have to be sought. These may include proof of medical education, licensure, or affiliation with a recognized institution. Feedback from peers could be used to flag errors or unreliable contributions. Contributor evaluation could be automated, for instance, by measuring the degree of similarity of patterns generated by a user to patterns with the

same label generated by other contributors. If similarity is consistently low, the reliability of that contributor would be questionable. To build the initial, core knowledge base, textbook disease entries could be translated into a structured format, for instance, with the help of medical students as part of their training. New patterns will be compared with this knowledge base and inconsistencies found could be manually reviewed. NGDSSs could be evaluated in controlled settings, such as by presenting them with standard cases (as has been done with current DSSs), or comparing their performance with that of physicians in real-life settings. The Human Diagnosis Project [32] model, in which cases are reevaluated by clinicians, can be useful as a quality control tool.

Data Sharing

Real-time sharing of structured patterns would not be possible unless authorities and other stakeholders are convinced that patient privacy is protected. The use of structured patterns eliminates the need to manually deidentify clinical notes and may facilitate sharing. However, the use of many different formats for presenting clinical data will require efforts to align EHR data from various products. The widely used Observational Medical Outcomes Partnership common data model [47], as implemented in the Observational Health Data Sciences and Informatics project [48], is a promising approach to this challenge. A central knowledge base may shorten the turnover time to provide answers to users, but medical institutions may choose to store their data on local servers, which will need to support application programming interface allowing controlled interrogation of their data.

Complex Presentations

Medical students are encouraged to find a single disease that would explain a patient's symptoms and signs. With an ageing population, comorbidities and polypharmacy are common. Comorbid conditions and medications used to treat them may alter the manifestations of a disease; the interplay between multiple factors with potential bearing on the clinical picture may make it impossible to attribute a pattern to a single etiology. This is true for an NGDSS but also, in many cases, for clinicians as well. For example, shortness of breath in the context of a respiratory tract infection may be caused by pneumonia, but in a patient with known heart failure, decompensation with pulmonary congestion may also explain the symptoms. Data extracted from the EHR of a patient and anchored to his or her structured pattern may help gain better understanding of the clinical picture. For instance, a patient's problem list and past laboratory results may put current findings in the right context. Unfortunately, variable quality of EHR data and mixture of clinical and billing information may limit the degree to which uncertainty could be reduced. New tests enhance our knowledge but may bring instances where current medical practice is simplistic to the surface. For instance, in a recent retrospective analysis, almost 5% of patients with a molecular diagnosis had 2 to 4 diagnosis accounting for their phenotype [49]. This is a challenge for humans and NGDSSs alike; however, NGDSSs are better positioned to handle genomic data (as well as other high-volume data types) and learn genotype-phenotype associations.

Alignment With Workflow

Admittedly, at least in the foreseeable future, even the most user-friendly NGDSSs would require clinicians to invest time in acquainting themselves with their use and interacting with them in the clinical setting. This is a challenge for work-overloaded physicians, many of whom do not trust DSSs. Attempts to structure history taking (eg, [50]) have not gained popularity due to usability issues. Thus, for an NGDSS to be used, it has to minimize interaction time. Even more importantly, NGDSSs will have to reward users by providing useful insights. One way to give clinicians back the time invested in interacting with an NGDSS would be to automatically translate a user-generated structured pattern into narrative notes. Apart from generating a differential diagnosis list, presenting an extract of the evidence used by the system when reaching its conclusions, as well as links to source documents, could make the system's conclusions more understandable to physicians and increase their trust. The direct feedback an NGDSS could provide can serve as a powerful tool in developing clinical expertise and may be a strong driving force for clinicians to use it. Using NGDSSs may also save time by eliminating the need to run a Web search or look for relevant evidence in other sources (eg, UpToDate, textbooks). Nevertheless, due to the extra effort required to use NGDSSs, more obscure or difficult cases are likely to be overrepresented in their knowledge base. Ways to correct for this when estimating prevalence will have to be developed.

A Learning Health Care System

NGDSSs can realize the vision of Ledley and Lusted [31] for a learning health care system through knowledge reuse. We describe some of the potential benefits of this approach in this section (Figure 3).

Better Characterization of Diseases and Syndromes

There are probably as-yet unidentified diseases and syndromes. Some syndromes go unnoticed due to their rarity, and those could be identified through analysis of very large datasets. A structured clinical pattern can serve as an anchor for all patient-related structured data (eg, laboratory results, imaging tests) in a patient record. The result would be a rich representation of the manifestations of the patient's disease(s), in which laboratory and imaging results are put in clinical context. This can help us understand accumulating genetic, proteomic, and microbiomic information and its association with clinical disease at the individual patient level. Integrated knowledge can help break down syndromes (eg, systemic lupus erythematosus and inflammatory bowel disease) to their underlying causes. Cohorting patients with similar structured patterns could potentially support more accurate outcome prediction and more reliable detection of adverse reactions to medications and other interventions.

Improved Medical Education

Apprenticeship is a major pillar in the training of clinicians, appreciating that effective learning takes place through practice and direct, immediate feedback [51]. Whereas clinicians do receive feedback on their decisions through following up on

their patients, minimizing trial-and-error-based learning is advisable. An attending physician provides feedback to his or her residents during rounds. However, most of the medical work is done between attending visits and does not involve high-quality feedback. An efficient and reliable NGDSS could serve as a mentor to both practicing physicians and medical students. As offered by current DSSs, NGDSSs could generate patient vignettes used for problem-based learning and diagnostic performance evaluation.

Enhanced Disease Surveillance Powered By Real-Time Clinical Data

Users' Internet activity has been shown to detect disease outbreaks before regulatory agencies can detect them [52,53]. Likewise, when NGDSSs are used, real-time clinical data may enhance early outbreak detection with a higher signal to noise ratio. Early detection is key to containing an outbreak, since by the time the first cases are identified, other individuals have likely been infected and may be spreading the disease. Enhancing the ability to detect clusters of unusual cases seen in the clinical setting in real time, even before a clinical or laboratory diagnosis has been established, can enable taking control measures earlier.

Improved Health Care in Limited-Resource Settings

Populations living in limited-resource settings are typically underrepresented in published medical reports. Cultural issues, limited availability of health professionals and diagnostic tests, and other factors may influence the ways diseases are first encountered and diagnosed by clinicians in such settings. Indeed, diagnostic errors in primary care are more common in low- and medium-income countries [54]. NGDSSs could help bypass the publication bias by collecting, analyzing, and sharing locally relevant knowledge. Global knowledge collection is particularly important in the age of global travel. Returning travelers can present with diseases acquired during travel that are rarely encountered in their homeland. Comparing their presentation patterns with patterns commonly seen in the places they had visited could help local physicians overcome knowledge gaps and the availability bias.

Conclusion

Computer-aided diagnosis has for decades been the Holy Grail of medical informaticians. The extreme complexity of constructing an efficient and sustainable system is reflected by the infrequent clinical use of DSSs despite the vast efforts that have been put into developing them.

On the one hand is an expanding domain knowledge, increasingly complex patients, and a high burden of diagnostic errors. On the other, EHR systems have become ubiquitous; powerful computers enable sophisticated analytics; the Internet can connect physicians from around the globe in real time; and human-computer interaction technologies have ripened. Taken together, there is both a real need for NGDSSs and the technology to meet it. We are laying a conceptual framework for developing NGDSSs that relies on structuring clinical notes; real-time sharing of patient structured patterns; democratization of knowledge generation, maintenance, and reuse; and integration of epidemiologic data to support the complicated task of making a diagnosis.

Development of NGDSSs will be very demanding, yet we argue that their potential utility justifies the investment required to realize them. The future of computer-aided medical diagnosis lies ahead and will likely change the way medicine is practiced.

Acknowledgments

No financial support was provided for this study.

The authors thank Peter Santhanam, Yishai Feldman, and Mohammad Sandhogi-Hamedani for their input.

An abstract (poster) related to this work was presented at the Diagnostic Error in Medicine 1st European Conference in Rotterdam, the Netherlands on June 30, 2016.

Conflicts of Interest

AC is employed by IBM. JJC has no conflicts of interest to disclose.

References

1. Del Mar C, Doust J, Glasziou P. Clinical Thinking. Malden, MA: Blackwell Publishing Inc; 2006.
2. Kahneman D, Klein G. Conditions for intuitive expertise: a failure to disagree. Am Psychol 2009 Sep;64(6):515-526. [doi: 10.1037/a0016755] [Medline: 19739881]
3. Sotos JG. Zebra Cards: An Aid To Obscure Diagnosis. Mt Vernon, VA: Mt Vernon Book Systems; 2006.
4. Phelps MA, Levitt MA. Pretest probability estimates: a pitfall to the clinical utility of evidence-based medicine? Acad Emerg Med 2004 Jun;11(6):692-694 [FREE Full text] [Medline: 15175211]
5. Dolan JG, Bordley DR, Mushlin AI. An evaluation of clinicians' subjective prior probability estimates. Med Decis Making 1986;6(4):216-223. [Medline: 3773651]
6. Cahan A, Gilon D, Manor O, Paltiel O. Probabilistic reasoning and clinical decision-making: do doctors overestimate diagnostic probabilities? QJM 2003 Oct;96(10):763-769 [FREE Full text] [Medline: 14500863]
7. Balogh EP, Miller BT, Ball JR, editors. Improving Diagnosis in Health Care. Washington, DC: National Academy of Sciences; 2015.

8. McDonald CJ. Medical heuristics: the silent adjudicators of clinical practice. Ann Intern Med 1996 Jan 1;124(1 Pt 1):56-62. [Medline: 7503478]
9. Tversky A, Kahneman D. Judgment under uncertainty: heuristics and biases. Science 1974 Sep 27;185(4157):1124-1131. [doi: 10.1126/science.185.4157.1124] [Medline: 17835457]
10. Redelmeier DA, Koehler DJ, Liberman V, Tversky A. Probability judgement in medicine: discounting unspecified possibilities. Med Decis Making 1995;15(3):227-230. [Medline: 7564936]
11. Ledley RS, Lusted LB. The role of computers in medical diagnosis. Med Dok 1961 Jul;5:70-78. [Medline: 13760192]
12. Gorry GA, Barnett GO. Sequential diagnosis by computer. JAMA 1968 Sep 16;205(12):849-854. [Medline: 5695508]
13. Miller RA. Medical diagnostic decision support systems--past, present, and future: a threaded bibliography and brief commentary. J Am Med Inform Assoc 1994;1(1):8-27 [FREE Full text] [Medline: 7719792]
14. Miller RA. Computer-assisted diagnostic decision support: history, challenges, and possible paths forward. Adv Health Sci Educ Theory Pract 2009 Sep;14 Suppl 1:89-106. [doi: 10.1007/s10459-009-9186-y] [Medline: 19672686]
15. Mitchell JA, Gerdin U, Lindberg DA, Lovis C, Martin-Sanchez FJ, Miller RA, et al. 50 years of informatics research on decision support: what's next. Methods Inf Med 2011;50(6):525-535. [doi: 10.3414/ME11-06-0004] [Medline: 22146915]
16. Garg AX, Adhikari NKJ, McDonald H, Rosas-Arellano MP, Devereaux PJ, Beyene J, et al. Effects of computerized clinical decision support systems on practitioner performance and patient outcomes: a systematic review. JAMA 2005 Mar 9;293(10):1223-1238. [doi: 10.1001/jama.293.10.1223] [Medline: 15755945]
17. Barnett GO, Famiglietti KT, Kim RJ, Hoffer EP, Feldman MJ. DXplain on the Internet. Proc AMIA Symp 1998:607-611 [FREE Full text] [Medline: 9929291]
18. Berger SA. GIDEON: a computer program for diagnosis, simulation, and informatics in the fields of geographic medicine and emerging diseases. Emerg Infect Dis 2001;7(3 Suppl):550 [FREE Full text] [doi: 10.3201/eid0707.010729] [Medline: 11485667]
19. Vardell E, Moore M. Isabel, a clinical decision support system. Med Ref Serv Q 2011;30(2):158-166. [doi: 10.1080/02763869.2011.562800] [Medline: 21534115]
20. Riches N, Panagioti M, Alam R, Cheraghi-Sohi S, Campbell S, Esmail A, et al. The effectiveness of electronic differential diagnoses (DDX) generators: a systematic review and meta-analysis. PLoS One 2016;11(3):e0148991 [FREE Full text] [doi: 10.1371/journal.pone.0148991] [Medline: 26954234]
21. El-Kareh R, Hasan O, Schiff GD. Use of health information technology to reduce diagnostic errors. BMJ Qual Saf 2013 Oct;22 Suppl 2:ii40-ii51 [FREE Full text] [doi: 10.1136/bmjqs-2013-001884] [Medline: 23852973]
22. Lupton D, Jutel A. 'It's like having a physician in your pocket!' A critical analysis of self-diagnosis smartphone apps. Soc Sci Med 2015 May;133:128-135. [doi: 10.1016/j.socscimed.2015.04.004] [Medline: 25864149]
23. Semigran HL, Linder JA, Gidengil C, Mehrotra A. Evaluation of symptom checkers for self diagnosis and triage: audit study. BMJ 2015;351:h3480 [FREE Full text] [Medline: 26157077]
24. Berner ES, Webster GD, Shugerman AA, Jackson JR, Algina J, Baker AL, et al. Performance of four computer-based diagnostic systems. N Engl J Med 1994 Jun 23;330(25):1792-1796. [doi: 10.1056/NEJM199406233302506] [Medline: 8190157]
25. Umscheid CA, Hanson CW. A follow-up report card on computer-assisted diagnosis--the grade: C+. J Gen Intern Med 2012 Feb;27(2):142-144 [FREE Full text] [doi: 10.1007/s11606-011-1944-x] [Medline: 22160816]
26. Bond WF, Schwartz LM, Weaver KR, Levick D, Giuliano M, Graber ML. Differential diagnosis generators: an evaluation of currently available computer programs. J Gen Intern Med 2012 Feb;27(2):213-219 [FREE Full text] [doi: 10.1007/s11606-011-1804-8] [Medline: 21789717]
27. Tang H, Ng JKH. Googling for a diagnosis--use of Google as a diagnostic aid: internet based study. BMJ 2006 Dec 2;333(7579):1143-1145 [FREE Full text] [doi: 10.1136/bmj.39003.640567.AE] [Medline: 17098763]
28. Warner H, Haug P, Bouhaddou O, Lincoln M, Warner H, Sorenson D. ILIAD as an expert consultant to teach differential diagnosis. Proc Annu Symp Comput Appl Med Care 1988;Nov 9:371-376.
29. Miller RA, McNeil MA, Challinor SM, Masarie FE, Myers JD. The INTERNIST-1/QUICK MEDICAL REFERENCE project--status report. West J Med 1986 Dec;145(6):816-822 [FREE Full text] [Medline: 3544509]
30. DxPlain. Boston, MA: The Massachussets General Hospital Laboratory of Computer Science; 2015. URL: http://www.mghlcs.org/projects/dxplain [accessed 2016-09-18] [WebCite Cache ID 6kcFQQxqe]
31. Ledley RS, Lusted LB. Reasoning foundations of medical diagnosis; symbolic logic, probability, and value theory aid our understanding of how physicians reason. Science 1959 Jul 3;130(3366):9-21. [Medline: 13668531]
32. Human Diagnosis Project.: Human Dx, Ltd; 2016. URL: https://www.humandx.org/ [accessed 2016-09-19] [WebCite Cache ID 6kdEt5zBX]
33. Pryor TA, Gardner RM, Clayton PD, Warner HR. The HELP system. J Med Syst 1983 Apr;7(2):87-102. [Medline: 6688267]
34. Graber ML, Mathew A. Performance of a web-based clinical diagnosis support system for internists. J Gen Intern Med 2008 Jan;23 Suppl 1:37-40 [FREE Full text] [doi: 10.1007/s11606-007-0271-8] [Medline: 18095042]
35. Weber GM, Mandl KD, Kohane IS. Finding the missing link for big biomedical data. JAMA 2014 Jun 25;311(24):2479-2480. [doi: 10.1001/jama.2014.4228] [Medline: 24854141]

36. Miller RA, Pople HE, Myers JD. Internist-1, an experimental computer-based diagnostic consultant for general internal medicine. N Engl J Med 1982 Aug 19;307(8):468-476. [doi: 10.1056/NEJM198208193070803] [Medline: 7048091]
37. Kohane I. Temporal Reasoning in Medical Expert Systems: Technical Report. Cambridge, MA: Massachusetts Institute of Technology; 1987.
38. [No authors listed]. Present and future of isolation systems in Auckland. N Z Med J 1975 Dec 24;82(554):410-413. [Medline: 1062725]
39. Shahar Y, Combi C. Timing is everything. Time-oriented clinical information systems. West J Med 1998 Feb;168(2):105-113 [FREE Full text] [Medline: 9499744]
40. Denekamp Y, Gal A. A clinical problem-oriented decision support model based on extended temporal database functionalities. AMIA Annu Symp Proc 2005:936 [FREE Full text] [Medline: 16779223]
41. Williamson M, Snyder L, Wallach J. Wallach's Interpretation of Diagnostic Tests. 9th edition. Philadelphia, PA: Wolters Kluwer; 2011.
42. Thorner AR, Walker DH, Petri WA. Rocky Mountain spotted fever. Clin Infect Dis 1998 Dec;27(6):1353-9; quiz 1360. [Medline: 9868640]
43. Pauker SG, Kassirer JP. The threshold approach to clinical decision making. N Engl J Med 1980 May 15;302(20):1109-1117. [doi: 10.1056/NEJM198005153022003] [Medline: 7366635]
44. Hauser RG, Jackson BR, Shirts BH. A bayesian approach to laboratory utilization management. J Pathol Inform 2015;6:10 [FREE Full text] [doi: 10.4103/2153-3539.151921] [Medline: 25774321]
45. Halkin A, Reichman J, Schwaber M, Paltiel O, Brezis M. Likelihood ratios: getting diagnostic testing into perspective. QJM 1998 Apr;91(4):247-258 [FREE Full text] [Medline: 9666947]
46. AHIP white paper focuses on patients and diagnostic imaging. Washington, DC: AHIP Coverage; 2011 Apr 18. URL: http://03664d9.netsolhost.com/devwordpress/2011/04/18/ahip-white-paper-focuses-on-patients-and-diagnostic-imaging/ [accessed 2016-09-18] [WebCite Cache ID 6kcGEbVEn]
47. Reich C, Ryan PB, Stang PE, Rocca M. Evaluation of alternative standardized terminologies for medical conditions within a network of observational healthcare databases. J Biomed Inform 2012 Aug;45(4):689-696 [FREE Full text] [doi: 10.1016/j.jbi.2012.05.002] [Medline: 22683994]
48. Hripcsak G, Duke JD, Shah NH, Reich CG, Huser V, Schuemie MJ, et al. Observational Health Data Sciences and Informatics (OHDSI): opportunities for observational researchers. Stud Health Technol Inform 2015;216:574-578 [FREE Full text] [Medline: 26262116]
49. Posey JE, Harel T, Liu P, Rosenfeld JA, James RA, Coban Akdemir ZH, et al. Resolution of disease phenotypes resulting from multilocus genomic variation. N Engl J Med 2017 Jan 05;376(1):21-31. [doi: 10.1056/NEJMoa1516767] [Medline: 27959697]
50. Fung CH. Computerized condition-specific templates for improving care of geriatric syndromes in a primary care setting. J Gen Intern Med 2006 Sep;21(9):989-994 [FREE Full text] [doi: 10.1111/j.1525-1497.2006.00522.x] [Medline: 16918747]
51. Archer JC. State of the science in health professional education: effective feedback. Med Educ 2010 Jan;44(1):101-108. [doi: 10.1111/j.1365-2923.2009.03546.x] [Medline: 20078761]
52. McIver DJ, Brownstein JS. Wikipedia usage estimates prevalence of influenza-like illness in the United States in near real-time. PLoS Comput Biol 2014 Apr;10(4):e1003581 [FREE Full text] [doi: 10.1371/journal.pcbi.1003581] [Medline: 24743682]
53. Velardi P, Stilo G, Tozzi AE, Gesualdo F. Twitter mining for fine-grained syndromic surveillance. Artif Intell Med 2014 Jul;61(3):153-163. [doi: 10.1016/j.artmed.2014.01.002] [Medline: 24613716]
54. World Health Organization. Diagnostic errors: technical series on safer primary care. Geneva, Switzerland: WHO; 2016. URL: http://apps.who.int/iris/bitstream/10665/252410/1/9789241511636-eng.pdf?ua=1 [accessed 2017-03-02] [WebCite Cache ID 6ofYC0mp5]

Abbreviations

DSS: diagnosis support system
EHR: electronic health record
NGDSS: next-generation diagnosis support system

Edited by G Eysenbach; submitted 19.09.16; peer-reviewed by G Weber, R El-Kareh; comments to author 27.12.16; revised version received 04.01.17; accepted 12.02.17; published 08.03.17.

Please cite as:
Cahan A, Cimino JJ
A Learning Health Care System Using Computer-Aided Diagnosis
J Med Internet Res 2017;19(3):e54
URL: http://www.jmir.org/2017/3/e54/
doi:10.2196/jmir.6663
PMID:28274905

Original Paper

Use of GetCheckedOnline, a Comprehensive Web-based Testing Service for Sexually Transmitted and Blood-Borne Infections

Mark Gilbert[1,2], MD; Travis Salway[1], PhD; Devon Haag[1], MSc; Christopher K Fairley[3,4], PhD; Jason Wong[1,2], MD; Troy Grennan[1,5], MD; Zhaida Uddin[6], MA; Christopher S Buchner[7], MHA; Tom Wong[8], MD; Mel Krajden[1,9], MD; Mark Tyndall[1,2], MD; Jean Shoveller[2,10], PhD; Gina Ogilvie[1,2], MD, DrPH

[1]BC Centre for Disease Control, Vancouver, BC, Canada
[2]School of Population and Public Health, University of British Columbia, Vancouver, BC, Canada
[3]Melbourne Sexual Health Centre, Melbourne, Australia
[4]Central Clinical School, Monash University, Melbourne, Australia
[5]Division of Infectious Diseases, Faculty of Medicine, University of British Columbia, Vancouver, BC, Canada
[6]Ottawa Public Health, Ottawa, ON, Canada
[7]Fraser Health, Vancouver, BC, Canada
[8]Health Canada, Ottawa, ON, Canada
[9]Department of Pathology and Laboratory Medicine, Faculty of Medicine, University of British Columbia, Vancouver, BC, Canada
[10]British Columbia Centre for Excellence in HIV/AIDS, Vancouver, BC, Canada

Corresponding Author:
Mark Gilbert, MD
BC Centre for Disease Control
655 West 12th Ave
Vancouver, BC, V5Z 4R4
Canada
Phone: 1 604 707 5619
Fax: 1 604 707 2401
Email: mark.gilbert@bccdc.ca

Abstract

Background: The British Columbia Centre for Disease Control implemented a comprehensive Web-based testing service GetCheckedOnline (GCO) in September 2014 in Vancouver, Canada. GCO's objectives are to increase testing for sexually transmitted and blood-borne infections (STBBIs), reach high-prevalence populations facing testing barriers, and increase clinical STI service capacity. GCO was promoted through email invitations to provincial STI clinic clients, access codes to clients unable to access immediate clinic-based testing (deferred testers), and a campaign to gay, bisexual, and other men who have sex with men (MSM).

Objective: The objective of the study was to report on characteristics of GCO users, use and test outcomes (overall and by promotional strategy) during this pilot phase.

Methods: We used GCO program data, website metrics, and provincial STI clinic records to describe temporal trends, progression through the service pathway, and demographic, risk, and testing outcomes for individuals creating GCO accounts during the first 15 months of implementation.

Results: Of 868 clients creating accounts, 318 (36.6%) submitted specimens, of whom 96 (30.2%) tested more than once and 10 (3.1%) had a positive STI diagnosis. The proportion of clients submitting specimens increased steadily over the course of the pilot phase following introduction of deferred tester codes. Clients were diverse with respect to age, gender, and ethnicity, although youth and individuals of nonwhite ethnicity were underrepresented. Of the 506 clients completing risk assessments, 215 (42.5%) were MSM, 89 (17.6%) were symptomatic, 47 (9.3%) were STI contacts, 232 (45.8%) reported condomless sex, 146 (28.9%) reported ≥4 partners in the past 3 months, and 76 (15.0%) reported a recent STI. A total of 63 (12.5%) GCO clients were testing for the first time. For 868 accounts created, 337 (38.8%) were by clinic invitations (0 diagnoses), 298 (34.3%) were by deferred testers (6 diagnoses), 194 (22.4%) were by promotional campaign (3 diagnoses), and 39 (4.5%) were by other means (1 diagnosis).

Conclusions: Our evaluation suggests that GCO is an acceptable and feasible approach to engage individuals in testing. Use by first-time testers, repeated use, and STI diagnosis of individuals unable to access immediate clinic-based testing suggest GCO

may facilitate uptake of STBBI testing and earlier diagnosis. Use by MSM and individuals reporting sexual risk suggests GCO may reach populations with a higher risk of STI. Motivation to test (eg, unable to access clinical services immediately) appears a key factor underlying GCO use. These findings identify areas for refinement of the testing model, further promotion, and future research (including understanding reasons for drop-off through the service pathway and more comprehensive evaluation of effectiveness). Increased uptake and diagnosis corresponding with expansion of the service within British Columbia will permit future evaluation of this service across varying populations and settings.

(J Med Internet Res 2017;19(3):e81) doi:10.2196/jmir.7097

KEYWORDS

Internet; sexually transmitted diseases; diagnostic tests; health care delivery; health services research; intervention study

Introduction

Globally, health systems are implementing new digital applications of existing health interventions to improve health care access and health outcomes. In the field of sexual health, Web-based testing services are widely considered to overcome barriers faced by individuals seeking testing for sexually transmitted and blood-borne infections (STBBIs) [1]. By offering an opportunity to test at a local laboratory or at home without needing to see a provider or present to a clinic, these services may eliminate known barriers related to conventional testing services (although to date this has not been well studied). Such barriers include feeling ashamed or embarrassed about getting a sexually transmitted infection (STI) test, fears of negative reactions from providers at disclosure of sexual behaviors, or clinic access barriers such as limited working hours or long wait times for appointments [2,3]. Web-based testing services are highly acceptable across all ages, offer privacy and anonymity, reach individuals at higher risk of STI, and may be cost-effective [4-7]. Web-based testing models are varied and are often designed as population screening programs for STIs, usually chlamydia [8,9]. Alternatively, Web-based testing can be offered as an integrated extension of existing clinical STI services and offer testing for multiple STBBIs, although fewer such models are known to exist or have been well evaluated in the published literature [5,10,11].

Following extensive consultation, formative research, and usability testing, the British Columbia Centre for Disease Control (BCCDC) implemented a comprehensive Web-based testing service for chlamydia, gonorrhea, syphilis, human immunodeficiency virus (HIV), and hepatitis C virus (HCV), which is operated as an integrated extension of its provincial STI clinic [12]. Called *GetCheckedOnline* (GCO) [13], the service has three main objectives: (1) to improve sexual health by increasing the uptake and frequency of STBBI testing and earlier diagnosis; (2) to reach populations with a greater prevalence of infection and facing barriers to testing access, such as gay, bisexual, and other men who have sex with men (MSM), youth, and people living in rural areas; and (3) to increase the capacity of STI clinic services and allow clinical resources to be more focused on more complex STI cases (eg, through reducing wait times and asymptomatic client visits).

In this study, our primary objective was to describe the use, test outcomes, and characteristics of GCO users during the first 15 months of operation. This pilot phase involved specimen collection sites in Vancouver, British Columbia, and focused on promotion to MSM and STI clinic clients in Vancouver. As a secondary objective, we aimed to describe differences in use and test outcomes between the strategies used to promote GCO.

Methods

Service Overview

The pilot phase of GCO began in September 2014. Details of the development process and service pathway have been described in detail elsewhere [12]. In brief, use of GCO involves proceeding through 5 steps: (1) account creation, (2) start and complete a risk assessment, (3) create and print a laboratory form, (4) submit specimens, and (5) receive results.

In step 1, clients create an online account, which includes collection of basic demographic information (eg, sex, age).

In step 2, clients answer questions on sexual history (eg, partner gender, prior testing history), which are collected in order to tailor test recommendations, educational messages, and testing reminders. Clients reporting symptoms or contact with a sexual partner with a diagnosed STI are recommended to go to a clinic to receive treatment but are not barred from proceeding with testing.

In step 3, clients view test recommendations, provide consent to get tested, and print their laboratory form.

In step 4, clients then present with the laboratory form to 1 of 6 designated collection centers in Vancouver where specimens are collected.

Finally, in step 5, clients receive a notification email when results are ready with a link to the GCO website to access their results. Results are provided online if all results are negative or by phone if any result is positive. If a test is invalid, such as a problem with a specimen, clients are notified to contact the clinic.

Steps 2 to 5 above constitute a single “test episode” and repeat each time a client uses GCO. Test recommendations are consistent with standard clinical practices of the BCCDC provincial STI clinic. Chlamydia (urine), gonorrhea (urine), HIV, and syphilis tests are recommended for all clients; HCV testing is recommended for clients reporting sharing drug paraphernalia and is available on an opt-in basis for MSM.

Promotional Strategies

Access to GCO is available through email invitation or through access codes that can be entered on the GCO home page.

Between September 9, 2014, and December 31, 2015, users were offered GCO through 1 of 3 promotional strategies: (1) email invitations to clients of the provincial STI clinic at BCCDC (hereafter referred to as "clinic client invitations") beginning in September 2014; (2) access codes given to clients who presented for testing to the provincial STI clinic or 2 Vancouver STI clinics accessed by MSM but were unable to get a same-day appointment, or called for an appointment and did not want to wait ("deferred testers"), beginning in March 2015; and (3) access codes distributed through a promotional campaign emphasizing the convenience of the service to gay and bisexual men in Vancouver ("promotional campaign") [14] between April and September 2015. During the latter half of the pilot phase, as interest in the service grew, access codes were also distributed by local STI clinics and gay men's health organizations as well as to individuals contacting the BCCDC to request access ("other").

Data Collection and Analysis

Client demographics, risk assessment responses, and data on progression through each step in the service pathway were extracted from the GCO database. Specimen submission was considered our ultimate indication of service uptake, with the proportion of clients creating accounts proceeding to submit specimens at least once our primary measure of service uptake. STI diagnoses and relevant clinical and public health outcomes for GCO clients with a positive test result were extracted through a chart review of the electronic medical chart used by the provincial STI clinic for documenting all follow-up of positive STI diagnoses in Vancouver (including for GCO clients). Outcomes examined were result delivery (clinic staff contacted and provided clients with their positive result), treatment (client was treated using appropriate antibiotic regimens), and partner notification (documentation of whether sexual partners were notified or not).

For this analysis, we restricted data to clients who created accounts between September 9, 2014, and December 31, 2015. We included test episode outcomes through March 31, 2016, to allow adequate opportunity for those clients who created accounts late in the pilot enrollment period to order tests, submit specimens, and receive results. We examined progression through the 5 steps in the service pathway and described the demographic characteristics of GCO clients, responses to the risk assessment, repeated testing patterns, and customization patterns (opting in or out of recommended tests). We examined temporal trends by month for the number of accounts created, laboratory forms created, specimens submitted (counted once per test episode), and positive test results (excluding laboratory forms subsequently cancelled by clients, as well as repeated tests ordered <14 days after the initial test episode, which were assumed to be retests and part of the same testing episode). We also calculated 3-month moving averages for the percentage of clients who submitted specimens at least once. Finally, we stratified temporal trends and test episode outcomes by promotional strategy, using numbers of unique clients.

We used chi-square or *t* tests as appropriate (two-sided; $P<.05$ considered statistically significant). Analyses were completed using R version 3.1 (R Foundation for Statistical Computing). Institutional ethical review was not required, as this analysis constitutes an evaluation of a public health program and use of GCO program data in this way is permitted under its terms of use agreed to by all clients.

Results

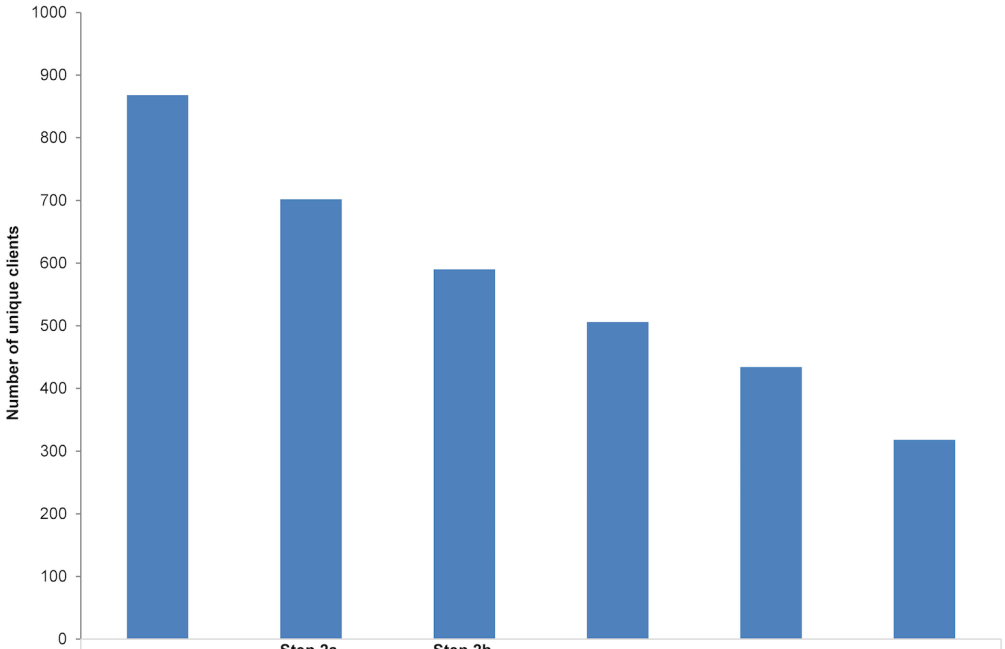
Use of GetCheckedOnline and Testing Outcomes

Between September 2014 and December 2015, a total of 868 unique clients created GCO accounts. Approximately 15%-25% of clients discontinued at each step of the GCO testing process, with a cumulative attrition between account creation and specimen submission of 63.4% (Figure 1). A total of 30.8% of clients submitted specimens more than once (96/318, range 2-6 times) for a total of 462 submitted sets of specimens and an average interval between test episodes of 119 days (range 17-517 days). Among the 462 submitting specimens, 8 (1.7%) opted out of urine chlamydia and gonorrhea tests, 27 (5.8%) opted out of HIV tests, and 23 (5.0%) opted out of syphilis tests. Among the 34 completed episodes in which drug equipment sharing was reported, 3 (8.8%) opted out of HCV tests; among 175 completed episodes among MSM, 80 (45.7%) opted in to receive an HCV test.

Of the 318 clients submitting specimens, STI was diagnosed in 10 (3.1%) clients (3 with chlamydia only, 4 with gonorrhea only, 1 with chlamydia and gonorrhea dual infection, and 2 with syphilis only). Notably, a chlamydia and a gonorrhea infection were detected separately on 2 rectal swabs collected after self-collected swabs were introduced in February 2016, just before the end of the analysis period for test episode outcomes (March 31, 2016). Of these 10 clients, 2 had never before tested for STBBI, 4 reported symptoms of STI, and 1 reported contact with a partner with an STI. All clients received their results by phone within 6 days of diagnosis. A total of 5 clients were treated at the provincial STI clinic at BCCDC, 4 clients reported receiving treatment at another clinic, and 1 client did not confirm receipt of treatment. All 10 clients reported notifying partners.

For the 308 clients with negative results, 234 (76.0%) were known to have clicked the link in their notification email to access their test results. It is possible, however, that some clients logged in directly to their GCO accounts to access their results without using the notification link.

Figure 1. Completion of test episodes, by GetCheckedOnline service pathway steps.



	Step 1. Accounts created	Step 2a. Assessments started	Step 2b. Assessments completed	Step 3a. Lab forms created	Step 3b. Lab forms printed	Step 4. Specimens submitted
Number of clients	868	702	590	506	434	318
% of previous step		80.9%	84.0%	85.8%	85.8%	73.3%
% of total (N=868)		80.9%	68.0%	58.3%	50.0%	36.6%

Characteristics of Users

GCO clients were diverse with respect to age, gender, and ethnicity (Table 1). Clients ranged in age from 16 to 79 years, and 71.3% (619/868) identified as male. Reporting white (73.9 %, 566/766) and Chinese (10.2%, 78/766 ethnicity was most common; 1.7% (13/766) identified as First Nations or Métis. Most clients resided within Vancouver and surrounding suburbs.

Table 1. Characteristics of *GetCheckedOnline* clients provided during account creation, September 2014 to December 2015.

Characteristic[a]	Total accounts created (N=868)
Age in years, minimum-maximum (median)	16-79 (32)
Age categories, years, n (%)	
16-19	9 (1.0)
20-24	105 (12.1)
25-29	217 (25.0)
30-39	291 (33.5)
40-59	215 (24.8)
60+	31 (3.6)
Gender, n (%)	
Male	619 (71.3)
Female	240 (27.6)
Transgender females (MTF[b])	1 (0.1)
Transgender males (FTM[c])	5 (0.6)
Other	3 (0.3)
Ethnicity, n (%)	
First Nations[d]	10 (1.3)
Métis	3 (0.4)
White	566 (73.9)
Chinese	78 (10.2)
South Asian	20 (2.6)
Filipino	15 (2.0)
Korean	3 (0.4)
Southeast Asian	9 (1.2)
Japanese	3 (3.9)
West Asian	5 (0.7)
Latin American	15 (2.0)
Black	6 (0.8)
Arab	3 (0.4)
Other	30 (3.9)
No response[e]	102
Region, n (%)	
City of Vancouver	534 (69.6)
Suburban, Greater Vancouver	203 (26.5)
Other	30 (3.9)
No response or invalid entry[d]	101

[a]Restricted to unique *GetCheckedOnline* clients (ie, not counting multiple test episodes).

[b]MTF: male to female.

[c]FTM: female to male.

[d]No client identified as Inuit.

[e]Excluded from calculation of column percentages.

Among the 506 unique clients completing risk assessments, based on documentation on at least one risk assessment, 215 (42.5%) were males or transgender males who reported having male sex partners (MSM), 89 (17.6%) reported symptoms, 146

(28.9%) reported 4 or more sex partners in the past 3 months, 47 (9.3%) reported being a contact to an STI, 232 (45.8%) reported condomless anal or vaginal sex in the past 3 months (20/506, 4.0%, with an HIV-positive partner), 76 (15.0%) reported a recent STI diagnosis, and 41 (8.1%) had ever shared syringes or other drug paraphernalia. Of the 506 unique clients, 63 (12.5%) reported no previous test for STI or HIV at the time of completion of the first risk assessment; for 157 (31.0%) clients it had been more than 1 year since their last test. Responses indicating a potential need for HIV postexposure prophylaxis (based on condomless sex with an HIV-positive partner in the last 72 hours) or emergency contraception (based on condomless vaginal sex without other forms of birth control in the last 5 days) were provided at least once to 13 (2.6%) and 57 (11.3%) clients, respectively. The frequencies of risk assessment variables analyzed across all 695 risk assessments completed by these 506 clients are provided in Table 2.

Table 2. Responses to clinical assessment among *GetCheckedOnline* clients creating laboratory forms, N=695 assessments (completed by 506 clients).

Variable	Response categories	Per assessment (N=695)	
		n	%
Any symptoms reported[a]	Yes	103	14.8
	No	528	76.0
	Don't know	40	5.8
	Prefer not to answer	24	3.5
Contact to an STI[b,c]	Yes	50	7.2
	No	458	65.9
	Don't know	167	24.0
	Prefer not to answer	15	2.2
	Not applicable[d]	5	0.7
Gender of sex partners[e]	Males or transgender males with male partners (MSM[f])	295	42.4
	Males or transgender males with female partners	255	36.7
	Males or transgender males with transgender partners	7	1.3
	Females or transgender females with male partners	159	22.9
	Females or transgender females with female partners	31	4.5
	Females or transgender females with transgender partners	2	1.2
	Other	1	0.1
	Prefer not to answer	18	2.6
Number of sex partners (vaginal, oral, or anal sex), last 3 months	0	22	3.2
	1	162	23.3
	2-3	294	42.3
	4-9	158	22.7
	10+	30	4.3
	Don't know	1	0.1
	Prefer not to answer	27	3.9
	Not applicable	1	0.1
Types of sex, last 3 months[e]	Vaginal	406	58.4
	Oral receiver	587	84.5
	Oral giver	537	77.3
	Anal bottom	188	27.1
	Anal top	223	32.1
	Prefer not to answer	26	3.7
Condomless anal or vaginal sex, last 3 months	No	372	53.5
	Yes	286	41.2
	Don't know	9	1.3
	Prefer not to answer	25	3.6
	Not applicable	3	0.4
Condomless anal or vaginal sex with HIV[g]-positive partner, last 3 months	No	450	64.7
	Yes	22	3.2

Variable	Response categories	Per assessment (N=695)	
		n	%
	Don't know	157	22.6
	Prefer not to answer	18	2.6
	Not applicable	48	6.9
Condomless (or condom broke during) anal or vaginal sex with partner known or thought to be HIV-positive, last 72 hours[h]	No	573	82.4
	Yes	14	2.0
	Don't know	44	6.3
	Prefer not to answer	17	2.4
	Not applicable	47	6.8
Condomless (or condom broke during) vaginal sex without using other form of birth control, last 5 days[i]	No	485	69.8
	Yes	63	9.1
	Don't know	12	1.7
	Prefer not to answer	19	2.7
	Not applicable	116	16.7
STI diagnosis, last 12 months	No	561	80.7
	Yes	103	14.8
	Don't know	15	2.2
	Prefer not to answer	16	2.3
Shared drug equipment	No	606	87.2
	Yes	57	8.2
	Don't know	5	0.7
	Prefer not to answer	13	1.9
	Not applicable	14	2.0
Last STI or HIV test	Never	64	9.2
	Last 3 months	162	23.3
	3-6 months	170	24.5
	6 months to 1 year	117	16.8
	>1 year	157	22.6
	Don't know	12	1.7
	Prefer not to answer	14	2.0

[a]Symptoms listed: painful urination, sores on or near genitals, rash on any part of the body, anal discharge, pain, blood, lesion, vaginal discharge, odor, itch, abnormal vaginal bleeding, lower abdominal pain, pain during intercourse, discharge from penis, swelling in testicles.

[b]STI: sexually transmitted infection.

[c]Sex partner who has recently tested positive for STI or told respondent he or she needs to get tested.

[d]Not applicable: client selected response.

[e]Categories not mutually exclusive, thus numbers do not sum to 100%.

[f]MSM: men who have sex with men.

[g]HIV: human immunodeficiency virus.

[h]Respondents who answered “yes” were shown information about how to access HIV postexposure prophylaxis.

[i]Respondents who answered “yes” were shown information about how to access emergency contraception.

Temporal Trends and Differences by Promotional Strategy

Account creation per month increased from the GCO program launch in September 2014 to a peak in April 2015, coinciding with the launch of the deferred tester codes and the promotional campaign (Figure 2). The percentage of clients proceeding from account creation to at least one specimen submission increased from 8.3% in the first 3 months of the pilot phase when accounts were created by clinic client invitations to 78.5% at the end of the pilot phase (P<.001) when the majority of accounts were created by deferred testers.

We observed large and statistically significant differences in progression through the steps of the service when examined across the 3 promotional strategies (Table 3). Of 868 accounts, 337 (38.8%) were created by clinic invitees, 298 (34.3%) by deferred testers, and 194 (22.4%) by promotional campaign clients; 113/130 (86.9%) of campaign clients who created a laboratory form were MSM. The percentage of clients submitting specimens was highest for deferred testers (184/298, 61.7%), followed by promotional campaign clients (58/194, 29.9%) and clinic invitees (62/337, 18.4%; P<.001). Of the 10 clients with positive results, 6 were deferred testers, 3 were promotional campaign clients, and 1 was referred by a local gay health organization. Differences in percent positivity by promotional strategy were not significant, although statistical power was limited given the small number of positive results.

Table 3. Uptake of *GetCheckedOnline* and related steps of service pathway, by promotional strategy.

Promotional strategy[a]	Accounts created n (%)	Laboratory forms created[b,c] n (%)	Specimens submitted[b,c] n (%)	Positive results[d] n (%)	Repeated testing[b,d] n (%)
Clinic client invitations	337	102 (30.3)	62 (18.4)	0 (0.0)	29 (46.8)
Deferred testers	298	249 (83.6)	184 (61.7)	6 (3.3)	39 (21.2)
Promotional campaign	194	130 (67.0)	58 (29.9)	3 (5.2)	23 (39.7)
Other[e]	39	25 (64.1)	14 (35.9)	1 (7.1)	5 (35.7)
Total	868	506 (58.3)	318 (36.6)	10 (3.1)	96 (30.2)

[a]All metrics restricted to unique clients only (in order to calculate percentages with proper denominators).

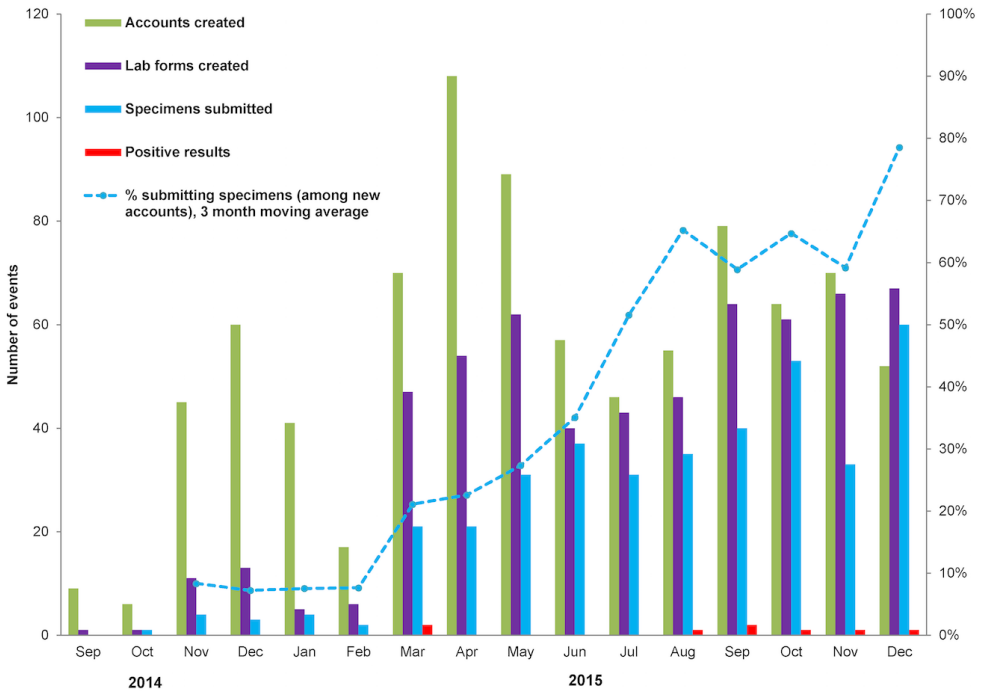
[b]P<.05 for chi-square test comparing proportions across promotional strategies.

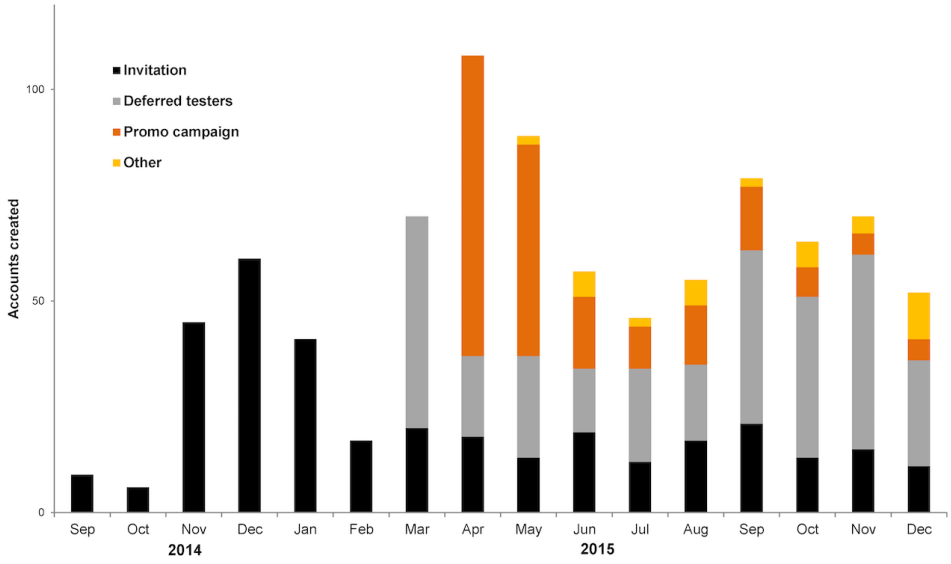
[c]Denominator is accounts created.

[d]Denominator is those who submitted specimens.

[e]Includes referrals from other primary care and community clinics, gay health organizations, and those requesting access.

Figure 2. Promotion and uptake of GetCheckedOnline by month; (top) accounts, lab forms, specimen submissions, and positive results; (bottom) accounts created by promotional strategy.





Discussion

Principal Findings

In this evaluation, we have shown GCO to be an acceptable and feasible approach to engage clients in STI testing. These preliminary findings suggest that broader implementation of GCO may have the potential to improve sexual health and increase uptake and frequency of STBBI testing, given the repeated use by clients and use by both clients who had never previously tested for STI or HIV and clients reporting it had been more than 1 year since they were last tested. Of the 10 individuals with a positive result, 6 were deferred testers who were seeking testing but unable to be seen immediately in a clinic; it is possible that these individuals accessed testing and their STI was diagnosed sooner through GCO. Furthermore, the addition of self-collected throat and rectal swabs just before the end of the analysis period led to the diagnosis of STI in 2 clients, confirming the importance of including these testing options as part of a Web-based testing service to diagnose STI that may otherwise be missed [15]. Beyond testing, clients reporting specific risk events also received tailored educational messages regarding postexposure prophylaxis and emergency contraception.

GCO appears to also be reaching individuals who may be at greater risk of infection. Of GCO clients, 42% were MSM and, based on responses to the risk assessment, a large proportion of GCO clients reported risk factors for STBBI, including condomless sex (41%), prior STI (15%, in past 12 months), high number of partners (27%, 4+ in past 3 months), and sharing of syringes or drug paraphernalia (8%). We also hypothesize that GCO may be reaching individuals who face barriers to accessing clinic-based STBBI testing. For example, clients reported a range of stigmatized gender identities and same-gender sexual activities, who may be more likely to face testing barriers through conventional testing services (eg, discomfort disclosing sexual identity or sexual behavior because of associated stigma) [16,17]. However, some populations facing testing barriers to STBBI testing are not yet well-represented among GCO clients. Individuals younger than 25 years of age comprised 13% of GCO users yet, in 2014, comprised 50% and 28% of chlamydia and gonorrhea infections in British Columbia [18]. There are also inequities based on ethnicity; almost three-quarters of GCO clients reported white ethnicity, yet 42% of respondents in the 2006 census for the Greater Vancouver region identified as a visible minority [19]. Consideration of how GCO may be promoted or adapted for youth and diverse ethnicities is needed.

During this GCO pilot phase, the provincial STI clinic was able to offer testing to 318 additional clients, of which the majority were asymptomatic. While representing a small percentage of the total clients seen by the provincial STI clinic during this time period, these findings suggest that GCO has the potential to offset growing demands on clinic capacity and that this potential will increase over time as use of the service expands. However, despite recommendations to visit a clinic, a small number of clients with symptoms or who had a partner with an STI proceeded to test through GCO. As these clients may have not received appropriate clinical management, further research is needed to understand the motivations of these clients for using GCO and whether any program modifications are needed.

Overall, 3% of clients testing through GCO had a new STI diagnosis, which may be lower than typical of many STI clinic settings. If we consider diffusion of innovations theory, innovators and early adopters of GCO may not necessarily reflect the risk profile of the future population of GCO clients and so diagnosis rates may change over time [20]. Interestingly, we note as demonstrated in Figure 2 (top) that the majority of the STI diagnoses were made after July 2015, suggesting there may be temporal trends toward increasing risk of infection among GCO users that may be explained by this theory. We postulate that a lower prevalence of STBBIs among GCO users may also be a reflection of the intervention itself, as clients experiencing symptoms or having a partner with a diagnosed STI (and thereby a higher probability of infection) are recommended not to proceed and instead to present to a clinic for testing. We also had no users with diagnosed HIV or HCV, which may reflect lower overall prevalence of these viral infections compared with bacterial STIs or that awareness or uptake of GCO has not yet penetrated higher-risk sexual networks for these infections. With accumulation of more data we hope to be able to investigate these questions and examine more carefully how prevalence may differ across different subgroups of GCO users.

Service Pathway

We observed consistent drop-off throughout the steps of the service pathway (Figure 1). This is not unexpected, having been reported for other Web-based STI testing services involving downloading laboratory requisitions (where between 10% and 33% of clients downloading requisitions submit specimens for testing) [11,21]. We postulate that many individuals who create an account and do not submit specimens are curious and learning about the service but not motivated to test at that time. However, the drop-off may represent true barriers posed by the GCO service itself (eg, need to print laboratory requisitions, dissatisfaction with the testing model). Client interviews and further analysis of GCO program data such as website metrics, time between steps, and characteristics of clients completing and not completing testing are being undertaken to determine if there may be modifiable factors related to website design or the GCO model that could facilitate progress through the service pathway (eg, optimizing the user experience, reminders of specimens not submitted).

Effect of Promotional Strategies

Motivation to test appeared to increase over the pilot phase, with an increased proportion of individuals submitting specimens over time, which is likely explained by differences between promotional strategies. The first promotional strategy used was email invitations to clinic attendees, which had the lowest specimen submission rate and no resulting diagnoses; as emails were collected from STI clinic clients, many of whom had just completed a testing encounter, motivation to use GCO was likely low. Deferred testers were likely the most motivated to get tested through GCO as they were deliberately seeking testing at the time of learning about GCO; unsurprisingly, this

group had the highest specimen submission rates. While specimens submitted by MSM had the highest positivity, only 58 specimens were submitted through the promotional campaign. As we have previously found high intentions to use GCO among Canadian MSM [22], we hypothesize that this may be related to the availability of existing low-barrier testing clinics for MSM in Vancouver, and the perceived benefits of seeing health care providers in these clinics who are competent in providing sexual health care for MSM are greater than the perceived benefits of GCO. Further evaluation of the campaign and reach of GCO among MSM in Vancouver is underway. Regardless, these findings suggest that clinic-based promotion (particularly for deferred testers) is an effective means of engaging individuals to use GCO.

Limitations

While most of the data presented in this report are administrative records of service use, some analyses are based on self-reported data collected routinely through GCO. While we generally saw high levels of completion for the risk assessment questions, ethnicity and region were not reported for 12% of individuals creating accounts, which may affect our conclusions about these variables. Many outstanding questions remain about the individual, health service, and population impacts of GCO that cannot be answered using program data, including understanding the experiences of clients choosing to use GCO, how they compare with clinic clients in terms of risk of infection and testing barriers, and the impacts on testing patterns and prevention of ongoing transmission. We are currently undertaking a comprehensive program of research that will aim to answer these questions, as well as further confirm these preliminary, suggestive findings [12].

Conclusions

A comprehensive, integrated STBBI testing service has been successfully launched in British Columbia. GCO is the first Web-based testing service to offer testing for multiple STBBIs in Canada, and our study adds to the sparse literature on the impact of online comprehensive STBBI testing as an extension of existing STI services. Our findings point to possible areas for refinement of the testing model and promotional strategies, such as optimizing the user experience, promoting to youth and visible minorities, and targeting individuals motivated to test. In February 2016, GCO was expanded to 2 other regions in British Columbia in partnership with regional health authorities, leading to a large increase in uptake and diagnoses. Ongoing evaluation of GCO will allow us to evaluate more fully whether our program objectives have been achieved, describe the implementation of this novel intervention across a range of populations and settings in British Columbia, and is critical for ensuring ongoing funding and sustainability of the program. The lessons learned from a comprehensive evaluation of GCO may also be more broadly relevant for understanding the impact of other self-care interventions and supporting diagnostic services for the public. Ultimately, services like GCO may prove a useful complement to—not replacement of—existing clinic or outreach-based services, as one intervention in a suite of necessary interventions needed to effectively meet the health care needs of a diverse population.

Acknowledgments

The authors would like to acknowledge the contributions of Mark Bondyra and Kimberly Thomson in the preparation of this manuscript and all of the staff involved in promotion, operation, and evaluation of GCO at the BCCDC, British Columbia Public Health Laboratory, and partner agencies. Funding for this study was provided by Canadian Institutes of Health Research Partnerships for Health System Improvement grants (PHE-114129; PHE-318068) awarded to JS, MG, and GO.

Authors' Contributions

MG led all aspects of this study and manuscript development. TS, DH, CKF, JS, and GO contributed to the design of the study, and the analysis was conducted by TS. All authors contributed to the interpretation of results and the writing of the manuscript. All authors have reviewed and approved submission of this manuscript.

Conflicts of Interest

MK has received contract or grant funding to his institution from Roche, Merck, Hologic, Boehringer Ingelheim, and Siemens that was unrelated to this work. No other authors have conflicts to disclose.

References

1. Wilson E, Free C, Morris TP, Kenward MG, Syred J, Baraitser P. Can internet-based sexual health services increase diagnoses of sexually transmitted infections (STI)? protocol for a randomized evaluation of an internet-based STI testing and results service. JMIR Res Protoc 2016 Jan 15;5(1):e9 [FREE Full text] [doi: 10.2196/resprot.4094] [Medline: 26772143]
2. Kaai S, Bullock S, Burchell AN, Major C. Factors that affect HIV testing and counseling services among heterosexuals in Canada and the United Kingdom: an integrated review. Patient Educ Couns 2012 Jul;88(1):4-15. [doi: 10.1016/j.pec.2011.11.011] [Medline: 22196985]
3. Shoveller J, Johnson J, Rosenberg M, Greaves L, Patrick DM, Oliffe JL, et al. Youth's experiences with STI testing in four communities in British Columbia, Canada. Sex Transm Infect 2009 Sep;85(5):397-401. [doi: 10.1136/sti.2008.035568] [Medline: 19508967]

4. McRee A, Esber A, Reiter PL. Acceptability of home-based chlamydia and gonorrhea testing among a national sample of sexual minority young adults. Perspect Sex Reprod Health 2015 Mar;47(1):3-10. [doi: 10.1363/47e2715] [Medline: 25776809]
5. Spielberg F, Levy V, Lensing S, Chattopadhyay I, Venkatasubramanian L, Acevedo N, et al. Fully integrated e-services for prevention, diagnosis, and treatment of sexually transmitted infections: results of a 4-county study in California. Am J Public Health 2014 Dec;104(12):2313-2320. [doi: 10.2105/AJPH.2014.302302] [Medline: 25320878]
6. Gaydos CA, Barnes M, Aumakhan B, Quinn N, Wright C, Agreda P, et al. Chlamydia trachomatis age-specific prevalence in women who used an internet-based self-screening program compared to women who were screened in family planning clinics. Sex Transm Dis 2011 Feb;38(2):74-78 [FREE Full text] [doi: 10.1097/OLQ.0b013e3182039d7f] [Medline: 21173720]
7. Huang W, Gaydos CA, Barnes MR, Jett-Goheen M, Blake DR. Cost-effectiveness analysis of Chlamydia trachomatis screening via internet-based self-collected swabs compared with clinic-based sample collection. Sex Transm Dis 2011 Sep;38(9):815-820 [FREE Full text] [doi: 10.1097/OLQ.0b013e31821b0f50] [Medline: 21844736]
8. Woodhall SC, Sile B, Talebi A, Nardone A, Baraitser P. Internet testing for Chlamydia trachomatis in England, 2006 to 2010. BMC Public Health 2012;12:1095 [FREE Full text] [doi: 10.1186/1471-2458-12-1095] [Medline: 23253518]
9. Greenland KE, Op de Coul EL, van Bergen JE, Brouwers EE, Fennema HJ, Götz HM, et al. Acceptability of the internet-based Chlamydia screening implementation in the Netherlands and insights into nonresponse. Sex Transm Dis 2011 Jun;38(6):467-474. [doi: 10.1097/OLQ.0b013e318204546e] [Medline: 21217416]
10. Levine D. Changemakers. 2007. STDTest.org - free, accessible, online STD/HIV testing URL: https://www.changemakers.com/disruptive/entries/stdtestorg-free-accessible-online-stdhiv-testing [WebCite Cache ID 6mSnN4cZX]
11. Koekenbier R, Fennerna H, Van LE, Davidovich U. The STI outpatient clinic online: the use and efficacy of online syphilis testing and the development of an full-scale testing application online. 2010 Oct 07 Presented at: Mednet; October 7, 2010; Leipzig, Germany.
12. Gilbert M, Haag D, Hottes TS, Bondyra M, Elliot E, Chabot C, et al. Get checked… where? the development of a comprehensive, integrated internet-based testing program for sexually transmitted and blood-borne infections in British Columbia, Canada. JMIR Res Protoc 2016 Sep 20;5(3):e186 [FREE Full text] [doi: 10.2196/resprot.6293] [Medline: 27649716]
13. BC Centre for Disease Control. GetCheckedOnline. A free, confidential way to test for sexually transmitted infections in Vancouver URL: https://getcheckedonline.com/ [accessed 2016-11-05] [WebCite Cache ID 6ofheh7Cy]
14. Health Initiative for Men. Just Makes Sense. Some things just make sense online: online HIV & STI testing without visiting a clinic URL: http://www.justmakessense.ca/ [accessed 2016-11-15] [WebCite Cache ID 6mSoDmz0N]
15. Schmidt AJ, Hickson F, Weatherburn P, Marcus U. Comparison of the performance of STI screening services for gay and bisexual men across 40 European cities: results from the European MSM internet survey. Sex Transm Infect 2013 Nov;89(7):575-582. [doi: 10.1136/sextrans-2012-050973] [Medline: 23744961]
16. Knight RE, Shoveller JA, Carson AM, Contreras-Whitney JG. Examining clinicians' experiences providing sexual health services for LGBTQ youth: considering social and structural determinants of health in clinical practice. Health Educ Res 2014 Aug;29(4):662-670 [FREE Full text] [doi: 10.1093/her/cyt116] [Medline: 24412811]
17. Ng BE, Moore D, Michelow W, Hogg R, Gustafson R, Robert W, et al. Relationship between disclosure of same-sex sexual activity to providers, HIV diagnosis and sexual health services for men who have sex with men in Vancouver, Canada. CJPH 2014 Apr 16;105(3):e186-e191. [doi: 10.17269/cjph.105.4212]
18. British Columbia Centre for Disease Control. STI in British Columbia: annual surveillance report 2014 URL: http://www.bccdc.ca/resource-gallery/Documents/Statistics%20and%20Research/Statistics%20and%20Reports/STI/STI_Annual_Report_2014-FINAL.pdf [accessed 2016-11-05] [WebCite Cache ID 6mSqp38LX]
19. Statistics Canada. Visible minority population, by census metropolitan areas (2006 Census) (Abbotsford-Mission, Vancouver, Victoria) URL: http://www.statcan.gc.ca/tables-tableaux/sum-som/l01/cst01/demo53g-eng.htm [accessed 2016-11-05] [WebCite Cache ID 6mSqyrPnN]
20. Zhang X, Yu P, Yan J, Ton AM. Using diffusion of innovation theory to understand the factors impacting patient acceptance and use of consumer e-health innovations: a case study in a primary care clinic. BMC Health Serv Res 2015 Feb 21;15:71 [FREE Full text] [doi: 10.1186/s12913-015-0726-2] [Medline: 25885110]
21. Levine DK, Scott KC, Klausner JD. Online syphilis testing--confidential and convenient. Sex Transm Dis 2005 Feb;32(2):139-141. [Medline: 15668622]
22. Gilbert M, Hottes TS, Kerr T, Taylor D, Fairley CK, Lester R, et al. Factors associated with intention to use internet-based testing for sexually transmitted infections among men who have sex with men. J Med Internet Res 2013;15(11):e254 [FREE Full text] [doi: 10.2196/jmir.2888] [Medline: 24240644]

Abbreviations

BCCDC: British Columbia Centre for Disease Control
GCO: GetCheckedOnline
HCV: hepatitis C virus

HIV: human immunodeficiency virus
MSM: men who have sex with men
STBBI: sexually transmitted and blood-borne infection
STI: sexually transmitted infection

Edited by G Eysenbach; submitted 05.12.16; peer-reviewed by A Bonnell, P Yu, X Zhang; comments to author 03.01.17; revised version received 11.01.17; accepted 28.01.17; published 20.03.17.

Please cite as:
Gilbert M, Salway T, Haag D, Fairley CK, Wong J, Grennan T, Uddin Z, Buchner CS, Wong T, Krajden M, Tyndall M, Shoveller J, Ogilvie G
Use of GetCheckedOnline, a Comprehensive Web-based Testing Service for Sexually Transmitted and Blood-Borne Infections
J Med Internet Res 2017;19(3):e81
URL: http://www.jmir.org/2017/3/e81/
doi:10.2196/jmir.7097
PMID:28320690

Original Paper

Supervised Machine Learning Algorithms Can Classify Open-Text Feedback of Doctor Performance With Human-Level Accuracy

Chris Gibbons[1,2*], PhD; Suzanne Richards[3*], PhD; Jose Maria Valderas[4*], PhD; John Campbell[4*], MD

[1]Centre for Health Services Research, University of Cambridge, Cambridge, United Kingdom

[2]The Psychometrics Centre, University of Cambridge, Cambridge, United Kingdom

[3]Leeds Institute for Health Sciences, University of Leeds, Leeds, United Kingdom

[4]Primary Care Research Group, University of Exeter, Exeter, United Kingdom

[*]all authors contributed equally

Corresponding Author:
Chris Gibbons, PhD
The Psychometrics Centre
University of Cambridge
16 Mill Lane
Cambridge, CB2 1RH
United Kingdom
Phone: 44 1223 765 203
Fax: 44 1223 765 203
Email: cg598@cam.ac.uk

Abstract

Background: Machine learning techniques may be an effective and efficient way to classify open-text reports on doctor's activity for the purposes of quality assurance, safety, and continuing professional development.

Objective: The objective of the study was to evaluate the accuracy of machine learning algorithms trained to classify open-text reports of doctor performance and to assess the potential for classifications to identify significant differences in doctors' professional performance in the United Kingdom.

Methods: We used 1636 open-text comments (34,283 words) relating to the performance of 548 doctors collected from a survey of clinicians' colleagues using the General Medical Council Colleague Questionnaire (GMC-CQ). We coded 77.75% (1272/1636) of the comments into 5 global themes (innovation, interpersonal skills, popularity, professionalism, and respect) using a qualitative framework. We trained 8 machine learning algorithms to classify comments and assessed their performance using several training samples. We evaluated doctor performance using the GMC-CQ and compared scores between doctors with different classifications using *t* tests.

Results: Individual algorithm performance was high (range *F* score=.68 to .83). Interrater agreement between the algorithms and the human coder was highest for codes relating to "popular" (recall=.97), "innovator" (recall=.98), and "respected" (recall=.87) codes and was lower for the "interpersonal" (recall=.80) and "professional" (recall=.82) codes. A 10-fold cross-validation demonstrated similar performance in each analysis. When combined together into an ensemble of multiple algorithms, mean human-computer interrater agreement was .88. Comments that were classified as "respected," "professional," and "interpersonal" related to higher doctor scores on the GMC-CQ compared with comments that were not classified ($P<.05$). Scores did not vary between doctors who were rated as popular or innovative and those who were not rated at all ($P>.05$).

Conclusions: Machine learning algorithms can classify open-text feedback of doctor performance into multiple themes derived by human raters with high performance. Colleague open-text comments that signal respect, professionalism, and being interpersonal may be key indicators of doctor's performance.

(J Med Internet Res 2017;19(3):e65) doi:10.2196/jmir.6533

KEYWORDS

machine learning; surveys and questionnaires; feedback; data mining; work performance

Introduction

Multisource "360-degree" feedback is increasingly used across business and health sectors to give workers insights into their performance and to identify areas in which improvements may be made. Such feedback often includes different reporting modalities that most commonly take the form of validated questionnaires or open-text comments. In the United Kingdom, large-scale national surveys include open-text feedback, such as the Friends and Family Test, the Inpatient Survey, and the Cancer Patient Experience Survey.

The complexity of open-text information means that, unlike the scores from validated patient-reported experiences and outcome measures, the words cannot simply be "added up" to create insight and meaning. As such, the task of making sense of such data has historically been completed manually by skilled qualitative analysts.

As the volume of text increases, qualitative data can quickly become difficult to manage and draw insights from. Coding and interpreting large bodies of qualitative information received from open-text comments collected is labor-intensive and is at risk of bias if multiple raters use subtly different coding heuristics. Where human raters systematically analyze qualitative data, there remain issues with both time and financial constraints of doing so, as well as potential challenges in ensuring intercoder consistency [1].

The term *machine learning* refers to the application of a growing number of algorithms that are able to complete diverse computational tasks, including mastering complex computer games [2], understanding the meaning of sentences [3], and successfully predicting psychological profiles from the Internet behavior [4,5].

Although machine learning appears to be eminently suitable for the task of classifying open-text data from national surveys, its potential is largely untested in the context of comments made by medical professionals about doctors' performance. Classification algorithms have been previously applied to patient comments about the experience of living beyond cancer [6], clinical incident reports [7,8], and sentiment analysis of digital footprints including Twitter and online blogs [9,10].

While algorithms have demonstrated excellent performance in diverse tasks, there is no evidence specifically relating to their ability to classify comments about doctors made by their colleagues as part of a formal evaluation. Although doctors' performance might be best assessed by fellow professionals who know them very well, positive reporting bias in open-text reports may occlude differences in performance [11,12]. The challenge therefore is to classify differences in text that is often positively worded and to use these classifications to signal differences in doctors' performance.

The objective of this study was to train and evaluate an ensemble of machine learning algorithms to accurately classify open-text reports of doctors, which are known to be positively biased, and to assess the potential for theory-based classifications in open text to signal differences in doctors' professional performance in the United Kingdom.

Methods

Sample

We collected data from all non–training-grade doctors from 11 sites in England and Wales between March 2008 and January 2011. We recruited doctors from 4 acute hospital trusts, an anesthetics department, 1 mental health trust, 4 primary care organizations, and 1 independent (non–National Health Service) health care organization. We provided all doctors with detailed information regarding the study before they consented to take part in it; they were told they could withdraw at any point without justification. Detailed description of this sample is reported elsewhere [7,8].

Doctors were asked to suggest up to 20 colleagues (half of whom were to be medically qualified) who could provide multisource "360-degree" feedback regarding their professional performance.

Multisource feedback was elicited using the General Medical Council Colleague Questionnaire (GMC-CQ), a reliable measure of doctor performance that is validated for use in the United Kingdom [13]. The GMC-CQ contains 18 items assessing diverse aspects of doctor performance and a section for entering open-text feedback.

Text Categorization

Qualitative analysts inductively coded the open-text feedback from the GMC-CQ into 5 themes relating to (1) innovation and openness to change (59/1636 comments, 3.6%); (2) interpersonal skills and caring (432/1636 comments, 26.4%); (3) popularity (131/1636 comments, 8%); (4) professionalism (701/1636 comments, 42.8%); and (5) respect or esteem in which the doctor was held (346/1636 comments, 21.1%) [12]. We refer to these categories throughout the rest of the paper as innovator, interpersonal skills, popularity, professionalism, and respect. Classification of a comment into more than one theme was possible. Of the 1636 reports, 1211 (74%) were classified as belonging to at least one of these categories. Similarly, classification of doctors into more than one category was possible, and 648 (28.8%) reports were classified into one or more of the 5 categories; as such, there were 2858 human-labeled comments in the entire corpus.

The number of comments in each category, the distribution of words, and statistical comparison of the word length are provided in Table 1. Significant analysis of variance (ANOVA; with post hoc Tukey test) results indicate that the number of words in texts that were granted the label of "innovator" was significantly greater than all other categories, whereas comments which received a label or "respected" or no label at all were significantly shorter.

Table 1. Number of comments, distribution of words, and statistical comparison for each of the 5 categories.

Categories	Reports in category	Length of report, mean (SD)	ANOVA[a] *P* value
Innovator	59	41.99 (30.84)	<.001
Interpersonal	432	23.87 (16.39)	.99
Popular	131	25.49 (16.74)	.97
Professional	701	24.46 (17.34)	.91
Respected	346	20.69 (19.13)	.03
More than 1 category	1189	21.63 (16.76)	.56
No categories	425	19.54 (13.62)	<.001

[a]ANOVA: analysis of variance; conducted with post hoc Tukey tests.

The qualitative researchers followed Holsti's approach [14]. Using rigorous data coding and verification procedures, which included double coding and independent verification within a qualitative framework [12], the resultant data were coded in such a way as to support quantitative data analysis.

Comments were generally, though not always, positive. In our sample, 91.5% (1497/1636) of all comments were positive, 5.93% (97/1636) of the comments were mixed, containing both a positive and a negative statement about the doctor, and the remaining 2.57% (42/1636) of comments were either neutral or negative (see Table 2). Prior publications relating to this dataset give further information on the process of ascertaining the polarity of comments [12].

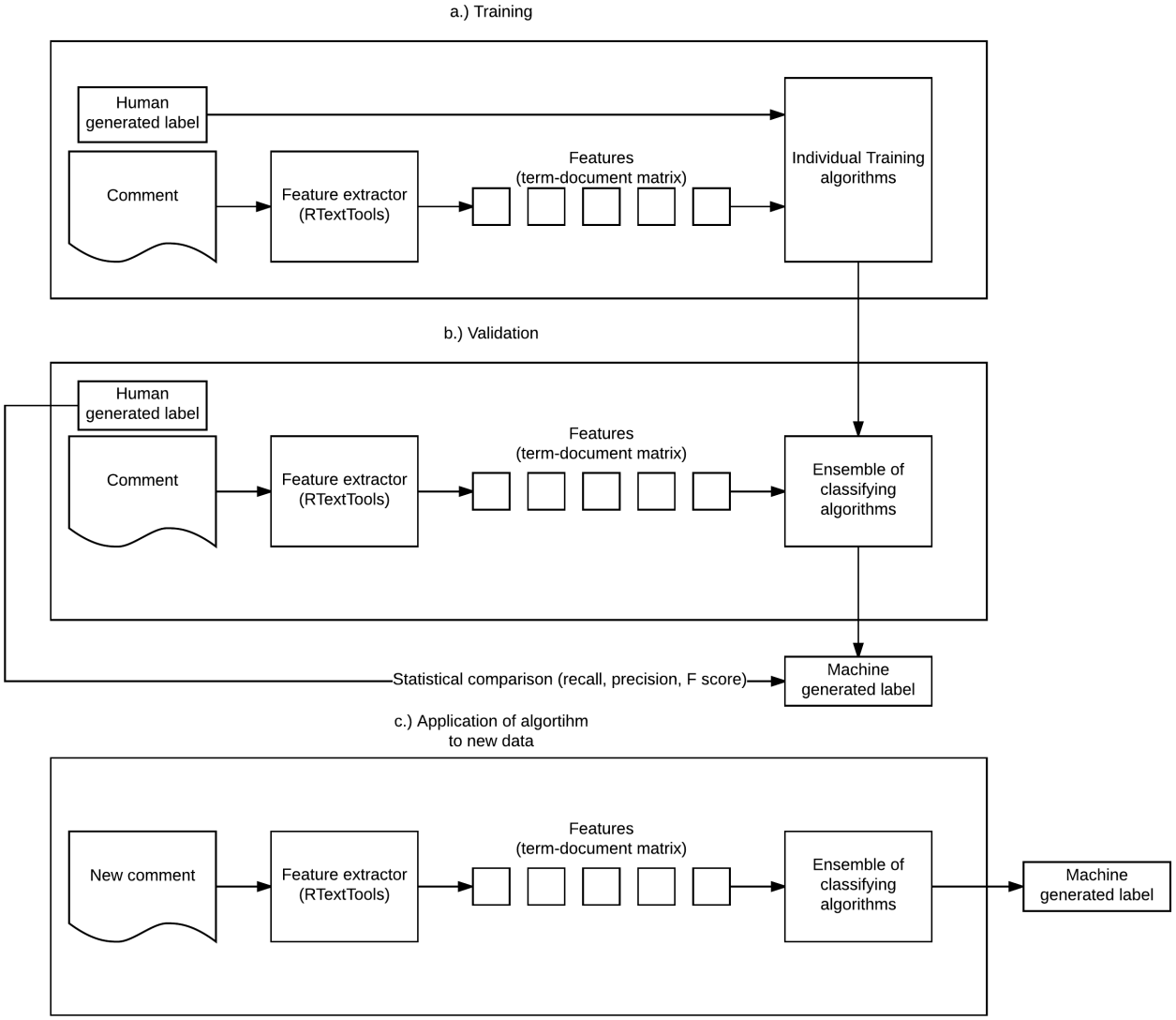
Assessment of Machine Learning Algorithm Performance

The process of training, validating, and deploying the algorithms is illustrated in Figure 1.

Table 2. Example quotes from each category

Theme	Comment
Innovator	"It is clear from the advice he gives that he is aware of [the] current good practice, is highly motivated, very practical and very much a team player. His advice, when working with consultant colleagues was respected, and he recognized where practice/primary care limitations were and yet looked for opportunities for change and improvement."
	"She has an admirable level of commitment and enthusiasm for her patients and her work. She has been instrumental in promoting change and improvement in her department. She is a great asset to the department and the hospital."
Interpersonal	"She is a very good, committed colleague always keen to improve, very liked by her patients and highly valued by all who work with her."
	"Very approachable and professional."
Popular	"Excellent well liked and easy working colleague."
	"Very popular doctor. Works to high standards."
Professional	"I find this doctor to be very efficient, caring, honest and very professional."
	"I find that he very easy and helpful to work with, he always has time for patients and staff."
Respected	"A first class colleague."
	"Pleasant and valued colleague."
Not coded by qualitative rater (given label of 0)	"Supportive colleague, excellent time management skills."
	"I think I have a good working relationship with this doctor. I have been impressed with his openness to Psychological work with his patients and his support for my work. In my opinion he gives thorough consideration to his diagnosis."

Figure 1. Flow diagram of the stages “training,” “validation,” and “application to new data.”.



Feature Selection

The first step within each stage is the identification of features within the comments. The features used in this study were identified and stored using a term-document matrix that describes the frequency of terms that appear in each of the comments. The term-document matrix uses a bag-of-words structure that counts the number of terms in each comment and does not consider the order in which the words appear. Term-document matrices are a simple way to represent text data that are computationally straightforward. The matrix comprised unweighted words and was cleaned by stemming, removing numbers, and removing sparse terms (where a certain word was only used in fewer than 0.02% of cases) [15]. Sparse-term removal reduced the number of terms from 1737 to 616. The final term-document matrix contained a lexicon of 616 unique words (columns) for 1636 comments (rows). The matrix density was 5.8%.

The term frequencies for each comment were therefore used as features that the algorithms used to classify the text. An example of a term-document matrix is provided in Table 3. For each of the 5 categories, texts with a human classification in that category were labeled with a 1 and those without were labeled with a 0.

Table 3. An example term-document matrix for 3 texts.

Texts	Terms											
	a	and	colleague	doctor	great	is	patients	respected	this	trouble-some	well	with
Text 1[a]	1	0	1	0	1	0	0	0	0	0	0	0
Text 2[b]	1	0	1	0	0	0	0	0	0	1	0	0
Text 3[c]	0	1	0	1	1	1	1	1	1	0	1	1

[a]Text 1: “A great colleague.”

[b]Text 2: “A troublesome colleague.”

[c]Text 3: “This doctor is well respected, and great with patients.”

Once the features have been extracted, they are used to “train” the algorithms to describe the relationship between the features and the classification.

In the validation stage the classifications given by the ensemble to new data are compared with the classifications made by the human qualitative analysts. If the results of the validation stage are acceptable, the algorithms can be exported and used to classify new data independently of the dataset that was used to train and validate the models.

The steps of these stages are given in greater detail in Figure 1.

Algorithms

“RTextTools” brings together other packages that contain different machine learning algorithms and provides a system by which the performance of each algorithm can be assessed both individually and as a collected ensemble of different methods that are combined to maximize performance in the training dataset. We included all of the available algorithms within RTextTools apart from the neural network, which did not converge in pilot assessments. The algorithms were support vector machine (SVM) using the radial basis function kernel with the penalty parameter of error term set to 1 and a gamma parameter set to 1/number of features [16], scaled linear discriminant analysis (SLDA) with eigenvalue threshold set to ≥1, bootstrapped boosting (bagging) with 25 bootstrap replications [17], boosting [18], random classification and regression forests with 500 trees [19], classification and regression tree [20], maximum entropy without regularization [21], and generalized linear models with L1 (lasso) penalized regularization (GLM/LASSO) [22].

A review and summary of supervised machine learning algorithms can be found elsewhere [23].

Training and Validation

Algorithms were trained with a corpus of 1000 randomly selected precoded comments (see Figure 1, part “a”) and validated on the remaining 636 comments (see Figure 1, part “b”).

We assessed algorithm performance using statistics of (1) recall (analogous to sensitivity)—what proportion of cases in a class are correctly assigned to the class; (2) precision (analogous to specificity)—how often a case that is predicted to belong to a class does belong to that class; and (3) *F* score, which is a combination of both recall and precision where 1 represents the best performance and 0 the worst performance [24]. To maximize performance, algorithms are combined into a consensus “ensemble” consisting of multiple algorithms. The consensus ensemble is a collection of algorithms that make the same prediction concerning the class of a text in the training dataset. We included the group of algorithms that had full agreement on every document in the dataset in the training sample. Classification was performed using majority voting between the algorithms in the ensemble.

When assessed as an ensemble of multiple algorithms working together, recall is evaluated alongside coverage (the proportion of cases within the dataset to which the recall value applies) [21]. The *F* value is analogous to interrater reliability and, as such, we will accept agreements ≥.80 between the algorithms and the human codes as evidence that the algorithms can complete the categorization task with acceptable accuracy.

n-Fold Validation

In addition to the standard assessment of algorithm performance using a validation dataset (636 comments), stability of algorithm performance across different data was also tested using an n-fold cross-validation. In the current analysis, a 10-fold validation was used in which 10 randomly selected samples of 1000 comments were selected from the dataset and validated using the remaining documents.

This analysis will indicate the robustness of the algorithms and their suitability for application to novel data and is preferable to split-half validation or bootstrapping [6,25].

Sample Size Accuracy Trade-Off

As well as the precision of the algorithms it is important to assess their training efficiency (the relationship between performance and size of the training dataset) so that we might best understand how to apply these techniques in practice. We compared training efficiency performance using randomly selected training sets of 1000, 750, 500, 250, 100, and 50 to accurately classify randomly selected comments from a fixed-size validation set (636 cases).

Assessment of Group Differences

To assess the ability of these categories to highlight differences in global performance, we investigated the differences in GMC-CQ scores for doctors whose comments were classified into at least 1 of 5 categories and those who were not placed in any category. We hypothesized that doctors who were placed

into one or more categories would perform better than those doctors who were not classified into any of the categories.

We conducted this analysis using both the machine-coded dataset (the entire dataset was recoded by the algorithm blinded to the codes given by the human rater and the original human-rated dataset.

Questionnaire data were scored using the graded response model [22]. All items fit the graded response model (chi-square interaction P>.01) and overall model fit was good (root mean square error of approximation =.048, comparative fit index =.97) [26,27]. The scale's marginal reliability was .76. This analysis was conducted so that interval-scaled logit scores (theta) could be extracted from the model to use in the comparative analysis. This technique has been shown to increase sensitivity to detect change on questionnaire measures of quality of life [28]. Correlation between theta values and scale raw scores was .95. Further details on the process of item response theory scoring and analysis can be found elsewhere [27,29,30].

Statistical Analysis

Computational text classification and statistical analysis were conducted within the R statistical programming environment (R Foundation) [31] using the "RTextTools" package for training the algorithms and the "base" package for conducting between-group comparisons. Figures were plotted using "ggplot2."

Ethical Approval

The study was originally considered by the Devon and Torbay NHS Research Ethics Committee but judged not to require a formal ethics submission. No subsequent ethical approval was sought for the secondary analyses on the anonymized datasets presented here.

Results

Assessment of Algorithm Performance

Table 4 presents the summary performance statistics for the algorithms and their individual F scores and the recall values for the ensemble of algorithms.

Table 4. Summary of algorithm and ensemble performance in the main analysis.

Model[a]	Metric	Innovator	Interpersonal	Popular	Professional	Respected	Average
Support vector machine	F score	.73	.69	.84	.79	.73	.76
Scaled linear discriminant analysis	F score	.77	.65	.88	.73	.77	.76
Boosting	F score	.75	.77	.81	.76	.75	.77
Bootstrap boosting	F score	.87	.85	.83	.80	.82	.83
Random forests	F score	.67	.59	.87	.78	.74	.75
Decision tree	F score	.80	.75	.88	.78	.80	.80
Generalized linear model	F score	.89	.82	.88	.81	.89	.85
Maximum entropy	F score	.70	.62	.73	.65	.70	.68
Final ensemble (3+ models with agreement for the entire dataset)	Recall with 100% agreement	.98	.80	.97	.82	.87	.89
10-Fold validation mean (range)	F score	.97 (.96-.98)	.80 (.74-.86)	.97 (.96-.98)	.79 (.75-.83)	.86 (.84-.89)	.88

[a]Training set size=1000; validation=636.

Innovator

The GLM/LASSO algorithm was the single highest-performing algorithm for correctly classifying the open-text comments into the "innovator" category. The ensemble of 3 algorithms (GLM/LASSO, bootstrapped boosting, and regression tree) has a 98% recall agreement with the human coder. The 10-fold validation indicated robust accuracy scores between .96 and .98 (mean .97). In the whole dataset, 48 comments (3.5%) were classified as an innovator by the algorithms.

Interpersonal

The bootstrapped boosting algorithm was the best-performing algorithm for categorizing open-text comments into the "interpersonal skills" category. The ensemble of 3 algorithms (boosting, GLM/LASSO, and bootstrapped boosting) demonstrated an 80% recall agreement with the human-coded dataset, the lowest performance for any of the classes. The 10-fold validation indicated similar performance in each fold and agreement values between .74 and .86 (mean .80). The algorithms classified 435 comments (28.4%) as "interpersonal."

Popular

All algorithms performed exceptionally well in classifying open-text comments into the "popular" category, with F scores greater than .80 for all but maximum entropy (F score=.73). The ensemble performance (SLDA, decision tree, and GLM/LASSO) was also excellent, with an interrater recall agreement of .97 (10-fold validation range .96-.98, mean .97). In total, 107 comments (8.3%) were placed in the "popular" category.

Professional

Similar performance was evident for many algorithms including SVM, random forests, and GLM/LASSO . Overall ensemble performance (GLM/LASSO, bootstrap boosting, and SVM) had an interrater recall of .82. The 10-fold validation suggested good agreement between the algorithm and the human analyst (mean .79, range .75-.83). The algorithms classified almost half of the comments in the whole dataset into the "professional" category (643 comments, 42.7%).

Respected

Once again, the GLM/LASSO algorithm showed the strongest single performance in the classification task for the "respected" category (*F* score=.89). The overall performance of the ensemble was very high, with .87 recall between the human coder and the 3-algorithm ensemble. The 10-fold validation demonstrated greater agreement between the human analyst and the algorithms (mean .86, range .84-.89). In the whole dataset, the ensemble classified 243 (16.6%) comments into the "respected" category.

Overall Performance

The GLM/LASSO algorithm was the strongest performing individual algorithm and the maximum entropy the worst. The overall average performance was nevertheless high (*F* score=.77). Average agreement between the human coder and the ensemble of algorithms was high (.89).

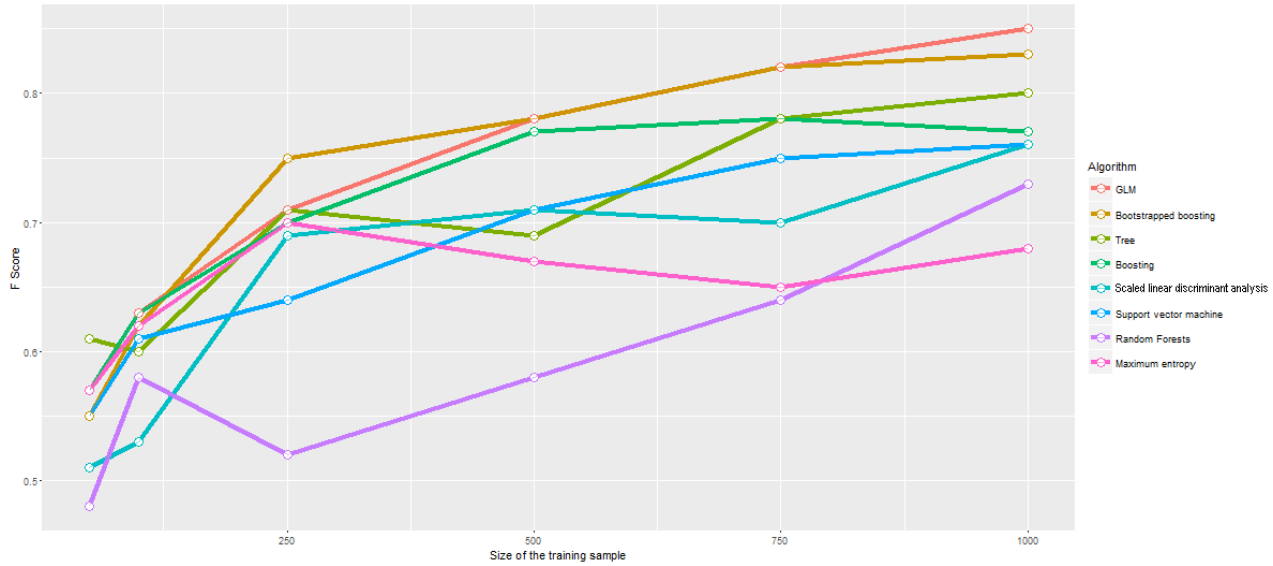
n-Fold Validation of Ensemble Accuracy

Results for the 10-fold cross-validation were very similar for the final recall values for the individual samples. The n-fold result displayed a tight distribution over the 10 samples (Table 4), indicating that the ensemble performs robustly across different samples.

Algorithm Performance With Differing Sample Sizes

As the training sample size was reduced, the algorithms continued to perform well but fell sharply when the training dataset was reduced to fewer than 250 comments. Figure 2 shows the algorithm performance with different training sample sizes.

Figure 2. Algorithm performance with differing training sample sizes. Performance decreases as expected with smaller training corpora.



Assessment of Group Differences

The *t* tests demonstrated a significant difference in the GMC-CQ scores between doctors who received comments that placed them into 1 of the 5 categories and those who did not ($t_{173.81}$=0.77, *P*=.001). Although the results were significant, there was sizeable overlap in the distributions as shown in Figure 3, indicating that open-text classification alone was not sufficient to reliably distinguish between doctors' performance. The largest difference in mean performance was between doctors who were classified as "respected" and those without a classification ($t_{629.17}$=3.75, *P*<.001). There was no difference in mean performance scores between doctors who were classified as "popular" (*P*=.44) and those who were not. Similarly, being rated as "innovative" did not signal higher performance (*P*=.99), although the low numbers in the analysis suggest a lack of power to detect an effect (n=48). Table 5 presents the results of these analyses conducted with both the machine learning classifications (Table 5, "Panel A") and the human classifications (Table 5, "Panel B"). The results are similar between the human-rated and machine-rated datasets, with stronger effect sizes being reported in the human-classified group.

Table 5. Comparison of means between doctors classified into a category and those who were unclassified.

Categories	Panel A[a]					Panel B[b]				
	Mean score (logits)	Reports in category	*t* test (vs no category rating)			Mean score (logits)	Reports in category	*t* test (vs no category rating)		
			t	df[c]	*P*			*t*	df	*P*
Innovator	0.00	48	0.00	55.74	.99	0.01	59	1.14	35.69	.26
Interpersonal	1.97	435	1.98	857.97	.04	0.07	432	2.97	346.63	<.01
Popular	−0.05	107	−0.88	176.42	.38	0.13	131	1.32	149.05	.19
Professional	−0.03	643	2.51	901.34	.01	0.1	701	3.47	286.99	<.001
Respected	0.15	243	3.75	629.17	<.001	0.44	346	5.58	300.13	<.001
More than 1 category	0.04	1081	0.77	173.81	.001	0.12	1189	3.81	239.8	<.001
No categories	−0.09	413	N/A[d]	N/A	N/A	−0.4	425	N/A	N/A	N/A

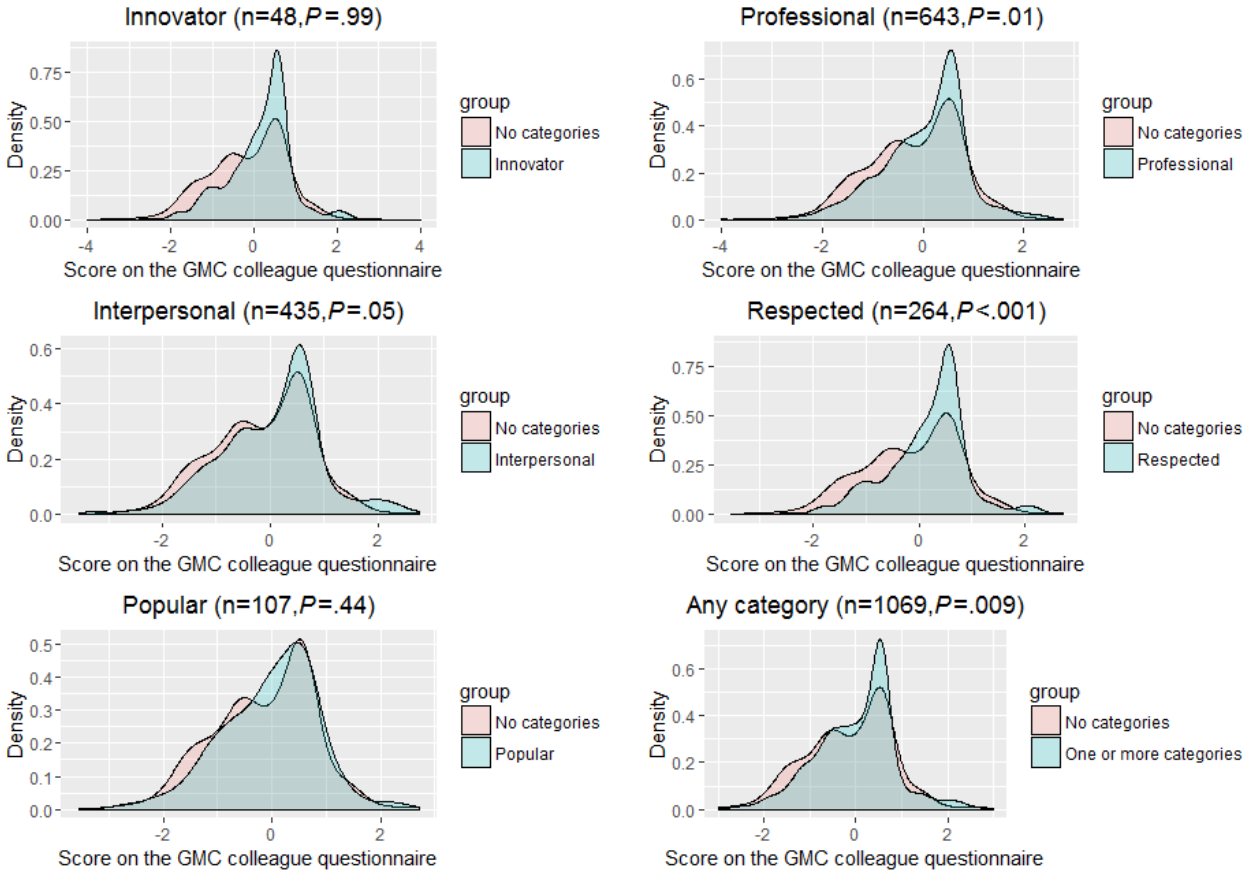
[a]Panel A: analysis using machine ensemble classifications on entire corpus.

[b]Panel B: analysis using human rater classifications on entire corpus.

[c]df: degrees of freedom.

[d]N/A: not applicable.

Figure 3. Comparison of General Medical Council Colleague Questionnaire (GMC-CQ) scores between doctors who were placed in 1 of the 5 categories versus those who were not (positive comments only). Significance (P) values for the t tests are shown to indicate the relationship between the 2 groups.



Discussion

Principal Findings

This study demonstrates the ability of machine learning algorithms to categorize qualitative data with high performance. The integration of such algorithms into data analysis tool kits for nationwide surveys may allow rich qualitative data to be analyzed without the resource burden associated with expert human ratings across an entire corpus [32].

We also demonstrate the ability of categories to highlight differences in overall doctor performance that were statistically significant. We hypothesized that doctors who were classified into 1 of the 5 categories would have higher scores on the GMC-CQ than those who were unclassified. We found partial support for this hypothesis: doctors who were classified as "respected," "professional," and "interpersonal" tended to outperform unclassified doctors, whereas no significant difference in performance was evident between doctors who were classified as "popular," "innovative," and those who were not classified into any of the 5 categories. However, the number of doctors who received a classification of "innovative" was low, which resulted in tests with low power and potential for type II error. Doctors with multiple ratings performed better than those without any ratings. Doctors who were classified as "respected" had the highest performance of each group in both the human-rated and machine-rated datasets.

These techniques have clear potential for developing actionable insights in diverse specialties: they have also been used to classify patient-derived open-text comments in national cancer surveys [6]. A key advantage of these techniques is the possibility of deploying trained algorithms to operate on data as they are being collected, allowing real-time feedback and insight from open-text data [33], which may be used to monitor performance, and possibly safety, in the future. It is important to remember that although machine learning algorithms can perform to a high level in prediction or classification tasks, some operate as "black box" and it is often difficult, or even impossible, to generate theory or convey insight into *how* the algorithms arrived at their final solutions.

We trained algorithms using precoded data and validated their performance on uncoded data, an example of "supervised" machine learning. We demonstrate strong performance using a relatively simple sparse "term-document matrix" method of identifying features in open text. The term-document matrix simply counts the instances of a word's use within a comment and does not consider the order in which the words are presented. This approach has been used in similar studies published in the medical literature [6,8].

It is possible to extract features using different, more complex, methods. Feature extraction using *n*-grams offers a means to retain some of the context in which words are used. An *n*-gram tokenizes sequences (of length *n*) of words as features, which may provide better information than the simple word-count strategy utilized in term-document matrix. Similarly, dimension reduction or clustering techniques such as latent Dirichlet allocation or singular value decomposition may have been used to reduce the sparsity within the matrix. There are no simple guidelines suggesting the optimal matrix density for use in this context and the possible benefits of clustering and dimension reduction must be counterbalanced by the caveat that these techniques can reduce the interpretability and accuracy of predictions [4].

Although performance of these techniques is demonstrably high, further research may be warranted to explore the extent to which they can improve the accuracy of classification algorithms in this context and at what cost (eg, computational burden or interpretability).

Natural language processing algorithms and their related software, driven by market forces in the technology industry, are improving at a remarkable rate and, paradoxically given the economic motivation for their development, increasingly being distributed at no cost under open-source licenses. As the field develops, we may expect these sorts of algorithms to successfully classify more complex corpora and perhaps even identify important elements within open-text comments without task-specific training data, which is known as "unsupervised" machine learning.

Limitations

This study has some limitations. The average performance of our algorithms is likely to be high in some instances given the low incidence of category. For example, algorithm performance was exceptionally high (recall=.97) for the "innovator" category, where only 4.2% of doctors were rated as innovative. In this instance, the low number of classifications in the population would have meant that a "dumb guess" that simply rated each doctor as "not innovative" would have demonstrated a 95.8% agreement rate. However, while algorithm performance was somewhat lower for categories with balanced distributions (eg, 46% of doctors were rated as professional), it was still acceptably high (recall=.82, 10-fold accuracy=.87).

Because of a somewhat small dataset, the trained ensemble was used to reclassify the whole corpora on which the algorithms were originally trained. It is likely that the performance of the algorithms will be higher when reclassifying the data they were originally trained on. The rationale for this decision was to maximize the number of classified categories and therefore maintain statistical power in the analyses. The results of these assessments were broadly the same as the results from the human-classified dataset, albeit the effect sizes were consistently smaller in the analyses using the machine-classified codes. This may be especially important as it appeared that the machine-labeled dataset was less sensitive to differences in doctor's performance than the human-labeled dataset.

A further possible limitation of the dataset was the necessity to have a small training-set to validation-set ratio (3:2) to keep sufficient number of comments in the validation sample. While this advantaged the statistical analysis of difference in performance signaling for different categories, it may have hampered the performance of certain algorithms by not providing sufficient training data.

The significant positive skew in doctors' ratings and the scarcity of comments that were outright negative meant that we were unable to conduct a sentiment analysis. We expect this to be the case in a population where most subjects are performing well, and this is probably representative of most datasets that collect open-text information on doctor performance. The content of uncategorized comments reveals a trend of doctors saying *something* positive about the colleagues if it did not relate to key elements of their medical practice (eg, "supportive colleague. Excellent time management skills"). The data used here were collected in relation to the high-stakes General

Medical Council revalidation exam, which may introduce a barrier to honest reporting of negative aspects of a doctor's practice. In addition to contextual factors, innovation surrounding the manner in which such comments are elicited, including less direct conversational techniques, may also reduce reporting biases.

Although algorithm performance was generally high for each of the individual machine learning techniques, it is apparent that the generalized linear model with lasso regularization had the highest performance for each of the classes. The precise reason for this improved performance is somewhat opaque, but the lasso regularization technique is especially suitable for classification problems using sparse matrices [34]. It is somewhat surprising that classification and regression trees outperformed the random forests; this may be attributed to the sparsity of the matrix and the low number of classifications made in some of the categories leading to high misclassification error in the random trees. Their performance may have been improved using a dimensional reduction technique such as singular value decomposition or latent Dirichlet allocation, which reduces sparsity in the matrix but which may also lead to a loss of information for other algorithms and uninterpretable results [35,36].

Similarly, it is not immediately clear as to why certain codes could be classified with greater accuracy than others. The differences in performance between classes may be explained by differences in the conceptual basis of each class; both humans and algorithms may find it easier to classify comments that reflect easily defined concepts such as being "popular" (the class for which algorithm performance was highest), rather than less well-defined concepts such as being "interpersonal" (the class for which algorithm performance was lowest) [37].

There may be an opportunity for similar techniques to be applied to patient experience data to build algorithms that can correctly classify and perhaps, using sentiment analysis, quantify open text in national-scale patient experience surveys and provide feedback that is more meaningful to both patients and practitioners. Computational analysis of open-text comments may be of greater usefulness when it is used to identify issues that were not previously envisaged.

Conclusions

This study demonstrates excellent performance for an ensemble of machine learning algorithms tasked to classify open-text comments of doctors' performance. These algorithms perform well, even where limited time and resources are available to code training datasets. We demonstrate that machine identification of qualitatively derived, theory-based open-text classifications can signpost significant differences in a doctor's performance, even when comments are exclusively positive. These findings may inform future predictive models of performance and support real-time evaluation to improve quality and safety.

Acknowledgments

We thank Karen Alexander, the National Institute for Health Research (NIHR) Adaptive Tests for Long-Term Conditions (ATLanTiC) patient and public involvement partner, for providing critical insight, comments, and editing the manuscript. Data collection and qualitative coding were funded by the UK General Medical Council as an unrestricted research award. Support for the novel work presented in this paper was given by a postdoctoral fellowship award for CG (NIHR-PDF-2014-07-028).

Conflicts of Interest

JC has been an advisor to the UK General Medical Council and has received only direct costs associated with the presentation of this work.

References

1. NHSsurveys. Inpatient Survey 2007: Analysis of the patients' free text comments URL: http://www.nhssurveys.org/Filestore/documents/Freetext_comments_Exec_Summary.pdf
2. Mnih V, Kavukcuoglu K, Silver D, Rusu AA, Veness J, Bellemare MG, et al. Human-level control through deep reinforcement learning. Nature 2015 Feb 26;518(7540):529-533. [doi: 10.1038/nature14236] [Medline: 25719670]
3. Cambria E, White B. Jumping NLP curves: a review of natural language processing research [Review Article]. IEEE Comput Intell Mag 2014 May;9(2):48-57. [doi: 10.1109/MCI.2014.2307227]
4. Kosinski M, Stillwell D, Graepel T. Private traits and attributes are predictable from digital records of human behavior. Proc Natl Acad Sci USA 2013 Apr 09;110(15):5802-5805 [FREE Full text] [doi: 10.1073/pnas.1218772110] [Medline: 23479631]
5. Youyou W, Kosinski M, Stillwell D. Computer-based personality judgments are more accurate than those made by humans. Proc Natl Acad Sci USA 2015 Jan 27;112(4):1036-1040 [FREE Full text] [doi: 10.1073/pnas.1418680112] [Medline: 25583507]
6. Wagland R, Recio-Saucedo A, Simon M, Bracher M, Hunt K, Foster C, et al. Development and testing of a text-mining approach to analyse patients' comments on their experiences of colorectal cancer care. BMJ Qual Saf 2016 Aug;25(8):604-614. [doi: 10.1136/bmjqs-2015-004063] [Medline: 26512131]
7. Ong MS, Magrabi F, Coiera E. Automated categorisation of clinical incident reports using statistical text classification. Qual Saf Health Care 2010 Dec;19(6):e55. [doi: 10.1136/qshc.2009.036657] [Medline: 20724392]

8. Ong MS, Magrabi F, Coiera E. Automated identification of extreme-risk events in clinical incident reports. J Am Med Inform Assoc 2012 Jun;19(e1):e110-e118 [FREE Full text] [doi: 10.1136/amiajnl-2011-000562] [Medline: 22237865]
9. Hawkins JB, Brownstein JS, Tuli G, Runels T, Broecker K, Nsoesie EO, et al. Measuring patient-perceived quality of care in US hospitals using Twitter. BMJ Qual Saf 2016 Jun;25(6):404-413 [FREE Full text] [doi: 10.1136/bmjqs-2015-004309] [Medline: 26464518]
10. Greaves F, Ramirez-Cano D, Millett C, Darzi A, Donaldson L. Use of sentiment analysis for capturing patient experience from free-text comments posted online. J Med Internet Res 2013 Nov 01;15(11):e239 [FREE Full text] [doi: 10.2196/jmir.2721] [Medline: 24184993]
11. Campbell JL, Roberts M, Wright C, Hill J, Greco M, Taylor M, et al. Factors associated with variability in the assessment of UK doctors' professionalism: analysis of survey results. Br Med J 2011 Oct 27;343:d6212 [FREE Full text] [Medline: 22034193]
12. Richards SH, Campbell JL, Walshaw E, Dickens A, Greco M. A multi-method analysis of free-text comments from the UK General Medical Council Colleague Questionnaires. Med Educ 2009 Aug;43(8):757-766. [doi: 10.1111/j.1365-2923.2009.03416.x] [Medline: 19659489]
13. Campbell JL, Richards SH, Dickens A, Greco M, Narayanan A, Brearley S. Assessing the professional performance of UK doctors: an evaluation of the utility of the General Medical Council patient and colleague questionnaires. Qual Saf Health Care 2008 Jun;17(3):187-193. [doi: 10.1136/qshc.2007.024679] [Medline: 18519625]
14. Holsti O. Content Analysis for the Social Sciences and Humanities. MA: Addison-Wesley; 1969.
15. Meyer D, Hornik K, Feinerer I. Text mining infrastructure in R. J Stat Softw 2008;25(5):-. [doi: 10.18637/jss.v025.i05]
16. Hornik E, Leisch F, Meyer D, Weingessel A. Researchgate. 2005. E1071: Misc Functions of the Department of Statistics (E1071), TU Wien URL: https://www.researchgate.net/publication/221678005_E1071_Misc_Functions_of_the_Department_of_Statistics_E1071_TU_Wien
17. Peters A, Hothorn T, Lausen B. ipred: Improved Predictors. R News 2002;2(2):33-36.
18. Tuszynski J. Cran.r-project. 2002. Tools: moving window statistics, GIF, Base64, ROC AUC, etc URL: https://cran.r-project.org/web/packages/caTools/caTools.pdf
19. Liaw A, Wiener M. Classification and regression by randomForest. R news 2002;2(3):18-22.
20. Ripley B. Cran.r-project. 2012. tree: Classification and regression trees URL: https://cran.r-project.org/web/packages/tree/tree.pdf
21. Jurka T, Collingwood L, Boydstun AE, Grossman E, Van Atteveldt W. RTextTools: A Supervised Learning Package for Text Classification. The R Journal 2013;5(1):6-12.
22. Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. J Stat Softw 2010;33(1):1-22 [FREE Full text] [Medline: 20808728]
23. Kotsiantis SB, Zaharakis ID, Pintelas PE. Machine learning: a review of classification and combining techniques. Artif Intell Rev 2007 Nov 10;26(3):159-190. [doi: 10.1007/s10462-007-9052-3]
24. Sokolova M, Szpakowicz S. Beyond accuracy, F-score and ROC: a family of discriminant measures for performance evaluation. Adv Artif Intell 2006:1015-1021. [doi: 10.1007/11941439_114]
25. Kohavi R. A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. IJCAI 1995;1(2):1137-1143.
26. Samejima F. Graded Response Model. In: Handbook of Modern Item Response Theory. New York, NY: Springer; 1997:95-107.
27. Reeve BB, Hays RD, Bjorner JB, Cook K, Crane PK, Teresi JA, et al. Psychometric evaluation and calibration of health-related quality of life item banks. Medical Care 2007;45(Suppl 1):S22-S31. [doi: 10.1097/01.mlr.0000250483.85507.04] [Medline: 17443115]
28. Birbeck GL, Kim S, Hays RD, Vickrey BG. Quality of life measures in epilepsy: how well can they detect change over time? Neurology 2000 May 09;54(9):1822-1827. [Medline: 10802791]
29. Bee P, Gibbons C, Callaghan P, Fraser C, Lovell K. Evaluating and quantifying user and carer involvement in mental health care planning (EQUIP): co-development of a new patient-reported outcome measure. PLoS One 2016;11(3):e0149973 [FREE Full text] [doi: 10.1371/journal.pone.0149973] [Medline: 26963252]
30. Gibbons C, Bower P, Lovell K, Valderas J, Skevington S. Electronic quality of life assessment using computer-adaptive testing. J Med Internet Res 2016 Sep 30;18(9):e240 [FREE Full text] [doi: 10.2196/jmir.6053] [Medline: 27694100]
31. R Core Team. Gbif. 2014. R: A language environment for statistical computing (Version 3.0.2) URL: http://www.gbif.org/resource/81287
32. Corner J, Wagland R, Glaser A, Richards SM. Qualitative analysis of patients' feedback from a PROMs survey of cancer patients in England. BMJ Open 2013;3(4):e002316 [FREE Full text] [doi: 10.1136/bmjopen-2012-002316] [Medline: 23578681]
33. Ainsworth J, Buchan I. Combining health data uses to ignite health system learning. Methods Inf Med 2015;54(6):479-487. [doi: 10.3414/ME15-01-0064] [Medline: 26395036]
34. Hastie T, Tibshirani R, Friedman J. Elements of statistical learning. New York: Springer-Verlag; 2001.

35. Golub GH, Reinsch C. Singular value decomposition and least squares solutions. Numer Math 1970 Apr;14(5):403-420. [doi: 10.1007/BF02163027]
36. Blei D, Ng AY, Jordan I. Latent Dirichlet Allocation. J Mach Learn Res 2003;3:993-1022.
37. Efron B, Hastie T. Computer Age Statistical Inference. 1st ed. Cambridge: Cambridge University Press; 2016:3-11.

Abbreviations

ANOVA: analysis of variance
GMC-CQ: General Medical Council Colleague Questionnaire
NIHR: National Institute for Health Research
SLDA: scaled linear discriminant analysis
SVM: support vector machine

Edited by CL Parra-Calderón; submitted 23.08.16; peer-reviewed by NP Cruz-Díaz, P Giacomelli, H Singh; comments to author 12.09.16; revised version received 30.09.16; accepted 29.11.16; published 15.03.17.

Please cite as:
Gibbons C, Richards S, Valderas JM, Campbell J
Supervised Machine Learning Algorithms Can Classify Open-Text Feedback of Doctor Performance With Human-Level Accuracy
J Med Internet Res 2017;19(3):e65
URL: http://www.jmir.org/2017/3/e65/
doi:10.2196/jmir.6533
PMID:28298265

Original Paper

Costs and Efficiency of Online and Offline Recruitment Methods: A Web-Based Cohort Study

Tina Christensen[1], MA; Anders H Riis[1], MSc; Elizabeth E Hatch[2], MS, PhD; Lauren A Wise[2], ScM, PhD; Marie G Nielsen[3], MSPH; Kenneth J Rothman[2,4], DrPH; Henrik Toft Sørensen[1], MD, PhD, DMSc; Ellen M Mikkelsen[1], MPH, PhD

[1]Aarhus University Hospital, Department of Clinical Epidemiology, Aarhus N, Denmark

[2]Boston University School of Public Health, Department of Epidemiology, Boston, MA, United States

[3]Aarhus University, Research Unit for General Practice, Department of Public Health, Aarhus C, Denmark

[4]RTI Health Solutions, Research Triangle Park, NC, United States

Corresponding Author:
Tina Christensen, MA
Aarhus University Hospital
Department of Clinical Epidemiology
Olof Palmes Allé 43-45
Aarhus N, 8200
Denmark
Phone: 45 87168225
Fax: 45 87167215
Email: tc@clin.au.dk

Abstract

Background: The Internet is widely used to conduct research studies on health issues. Many different methods are used to recruit participants for such studies, but little is known about how various recruitment methods compare in terms of efficiency and costs.

Objective: The aim of our study was to compare online and offline recruitment methods for Internet-based studies in terms of efficiency (number of recruited participants) and costs per participant.

Methods: We employed several online and offline recruitment methods to enroll 18- to 45-year-old women in an Internet-based Danish prospective cohort study on fertility. Offline methods included press releases, posters, and flyers. Online methods comprised advertisements placed on five different websites, including Facebook and Netdoktor.dk. We defined seven categories of mutually exclusive recruitment methods and used electronic tracking via unique Uniform Resource Locator (URL) and self-reported data to identify the recruitment method for each participant. For each method, we calculated the average cost per participant and efficiency, that is, the total number of recruited participants.

Results: We recruited 8252 study participants. Of these, 534 were excluded as they could not be assigned to a specific recruitment method. The final study population included 7724 participants, of whom 803 (10.4%) were recruited by offline methods, 3985 (51.6%) by online methods, 2382 (30.8%) by online methods not initiated by us, and 554 (7.2%) by other methods. Overall, the average cost per participant was €6.22 for online methods initiated by us versus €9.06 for offline methods. Costs per participant ranged from €2.74 to €105.53 for online methods and from €0 to €67.50 for offline methods. Lowest average costs per participant were for those recruited from Netdoktor.dk (€2.99) and from Facebook (€3.44).

Conclusions: In our Internet-based cohort study, online recruitment methods were superior to offline methods in terms of efficiency (total number of participants enrolled). The average cost per recruited participant was also lower for online than for offline methods, although costs varied greatly among both online and offline recruitment methods. We observed a decrease in the efficiency of some online recruitment methods over time, suggesting that it may be optimal to adopt multiple online methods.

(J Med Internet Res 2017;19(3):e58) doi:10.2196/jmir.6716

KEYWORDS

participant recruitment; Web-based study; costs per participant; Internet; advertising

Introduction

Recruiting participants for epidemiologic research is increasingly difficult as the number of projects competing for people's attention increases and response rates decline [1]. In cohort studies, participant recruitment and data collection are associated with a heavy workload and high costs [2]. Widespread access to the Internet now offers an alternative strategy to recruit participants into cohort studies and to collect data. The Internet offers technical advantages in data collection that can reduce administrative procedures and improve data quality. Examples are incorporating skip patterns that avoid displaying irrelevant questions, building in internal consistency checks, and avoiding errors that occur during manual data entry [3]. Furthermore, the Internet is an effective tool to reach populations who are otherwise challenging to enroll because they have sensitive health concerns, including urinary incontinence, sexual health disorders, or mental health problems [4-6].

Pregnancy planners also constitute a hard-to-reach population as they do not typically announce their pregnancy intentions [7-10]. Among recruitment methods that have proven to be effective in enrolling participants are study-related media publicity [6,11], online advertisements [6,7,10,12-15], printed advertisements [6], and Web-based social groups [6,10].

Some studies have reported the costs of a single online recruitment method [14,16-19], and others have reported on costs of using several online methods [20,21], or both offline and online methods [6,10,12,22,23]. Nevertheless, little is known about how recruitment methods compare in terms of efficiency (number of recruited participants) and costs per participant [20]. Here, we compare the efficiency and costs per recruited participant of online and offline recruitment methods used to enroll women in a Danish cohort study of pregnancy planners that relied on the Internet for enrollment and data collection.

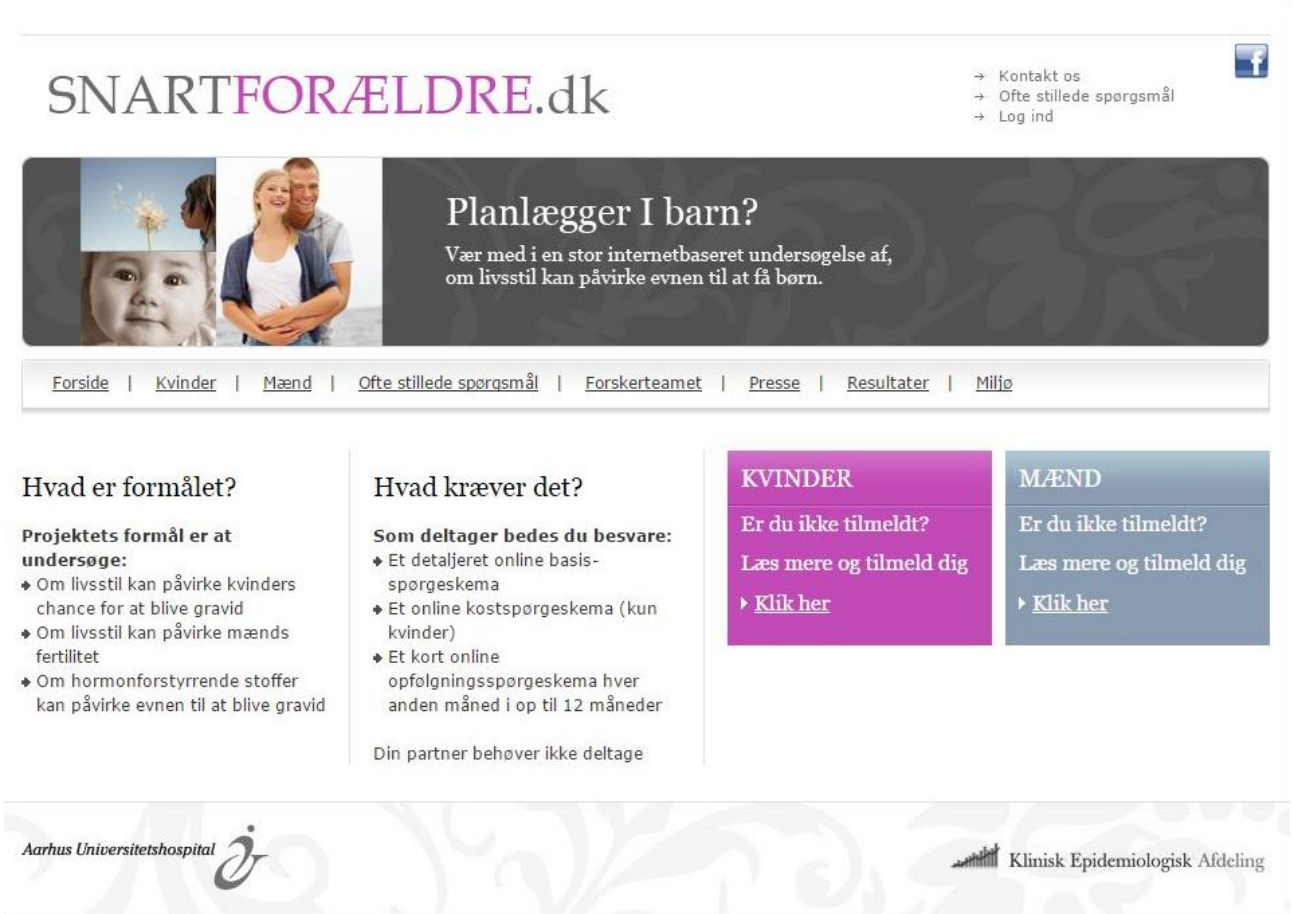
Methods

Setting

The Snart-Gravid.dk (Soon-Pregnant) and the SnartForaeldre.dk (SoonParents) studies are related prospective cohort studies on lifestyle and fertility [8,9,24,25]. In both studies, participants enroll via the Internet and all data are collected by means of Web-based questionnaires. Snart-Gravid.dk was launched in June 2007. It was succeeded in August 2011 by SnartForaeldre.dk, which incorporates a dietary questionnaire and includes male partners (Figure 1). Recruitment for SnartForaeldre.dk is ongoing [26].

The study period for this paper, which focuses on female recruitment in both studies (in the following referred to as one study), is June 2007 through December 2013.

Figure 1. The frontpage of the SnartForaeldre.dk website.



Study Population

The recruitment area for Snart-Gravid.dk and SnartForaledre.dk covers all of Denmark (total population of 5.7 million on January 1, 2016, and 58,205 births in 2015) [27]. Eligible female participants have a Danish civil registration number, are aged 18-45 years old (18-40 years in Snart-Gravid.dk), in a stable relationship with a male partner, and attempting to become pregnant. To enroll, participants have to (1) visit the study website, which contains information on eligibility criteria and informed consent and (2) complete a Web-based screening questionnaire to determine eligibility. As part of this process, participants provide their civil registration number, which is a unique, personal 10-digit number assigned to all Danish residents at birth or upon immigration. No incentives are offered to participants, who are followed until they achieve a pregnancy, start fertility treatment, or end of observation (12 cycles). In both studies, enrollment is based on completion of a comprehensive baseline questionnaire, which takes an average of 22 minutes [28].

A substudy, SnartForaldre.dk/Milieu, was introduced to pilot test the feasibility of collecting blood and urine samples to investigate associations between endocrine disruptors and fertility. Participation in the substudy required enrollment in SnartForaeldre.dk and residence in the Aalborg area (population of 205,407 in December 2013) [29].

Recruitment Methods

During the study period, we employed several online and offline recruitment methods to increase awareness about the study and to encourage enrollment. All advertisements were designed by the same graphics designer and used images, colors, and text phrases identical with or similar to the respective study website to enhance recognition. The most frequently used text was “Planning to get pregnant? Help us find out whether lifestyle influences your ability to become pregnant.”

Offline Recruitment Methods

Three offline recruitment methods were used to attract attention to our study: press releases, posters, and flyers.

A total of 6 press releases with various topics were issued; of which, 3 press releases were written and disseminated with the assistance of contracted, external journalists, 1 was handled by study staff alone, and 2 were issued with help from an internal journalist from Aarhus University. We obtained information from media surveillance companies on the numbers of printed and Web-based articles, and radio or TV features that the press releases resulted in.

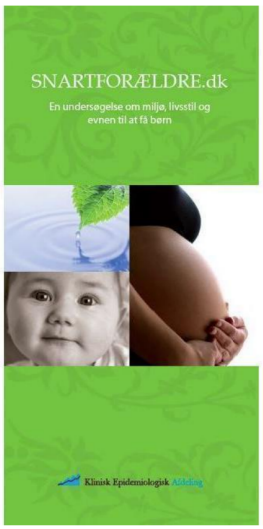
The costs of obtaining the services of external journalists and a media surveillance company were €2281.29 and €1761.91 for the first 2 press releases, respectively, and €1772.54 for the fourth press release. Issuing the other press releases involved no direct costs as internal institutional resources were used.

A small A4-size (210x297 mm) poster with information about the study was designed (Figure 2). Each copy was equipped with a block of 50 post-it notes providing study information. The poster included a quick response (QR) code providing a direct link to the study website when scanned with a smart phone. Study staff and colleagues placed 133 posters on notice boards in libraries, hospital canteens, fitness centers, grocery stores, and other public places. In total, €377.76 was paid for the design of the poster, the QR-code, and the post-it notes and for printing.

Flyers advertising SnartForaeldre.dk, and the substudy, SnartForaeldre.dk/Milieu were distributed to pharmacies, a few stores, and 54 general practitioners’ (GP) offices in the city of Aalborg (Figure 3). Costs for designing the flyer and printing 3500 copies were €604.31.

Figure 2. Poster advertising SnartForaeldre.dk.

Figure 3. Flyer advertising SnartForaeldre.dk and the substudy SnartForaeldre.dk/Milieu.



Online Recruitment Methods

Online campaigns, primarily consisting of online advertisements, were placed on 6 different websites: Netdoktor.dk, Minmave.dk, Facebook, Sundhedsguiden.dk, Baby.dk, and Aarhus University (AU). When participants clicked on an advertisement, they were automatically redirected to the study website.

Netdoktor.dk is a popular general health website. Two online advertisements of different sizes were posted 2 weeks after the launch of Snart-Gravid.dk (Figure 4). Netdoktor.dk agreed to show the advertisement until it had generated 2500 participants at a fixed price of €3.93 each. For the SnartForaeldre.dk study (Figure 5), Netdoktor.dk committed itself to recruit 10,000 women at a fixed price of €2.99 per enrolled participant. Netdoktor.dk controlled the timing of advertisements to be able to prioritize campaigns by full-price advertisers.

Facebook is a popular social media website. In total, 8 advertisements, lasting 8-10 days each, were posted on Facebook. The advertisements were targeted at potential participants, that is, female, 18-40 years old, speaking Danish, and married or in a relationship. Seven advertisements targeted women living anywhere in Denmark and one targeted only women living in the city of Aalborg.

Advertising on Facebook is managed on the Internet and the price is based on a bidding system. Depending on the number of advertisers aiming at the same target group, the system suggests a price range, within which it is likely that your advertisement will be shown. It is possible to pay per 1000 impressions, that is, the number of times an advertisement is shown on the site, or per click and to choose a daily spending limit. We paid per click and chose a limit of €13.33 per day for 8-10 days at a time. The SnartForaeldre.dk study staff also created a page on Facebook to promote the study. Postings were made by staff and thus incurred no direct costs.

Sundhedsguiden.dk is a general health website. We made an agreement including 100,000 impressions of banner advertisements at an overall fixed price of €1077 over a period of 8 months.

Minmave.dk is a website for pregnant women, that is, women who want to become pregnant and women with infants. We negotiated an agreement covering 5 one-month campaigns, each consisting of 500,000 impressions of banner advertisements, 50,000 pop up/overlay banner impressions within the category "Fertility," and short texts in 3 electronic newsletters distributed to subscribers. An overall fixed price of €2019 was paid for each of the 5 campaigns.

Baby.dk is a website aimed at parents of infants, pregnant women, and women wanting to become pregnant. We made an agreement including continuous display of a top banner advertisement and a continuous "ownership" picture on a fertility page for 3 months, as well as placement of 2 stories about our study on the website at an overall fixed price of €673.

The Aarhus University Communications Department placed 2 short postings about SnartForaeldre.dk, including links to the study website, on the AU intranet, which reaches approximately 38,000 students and on the AU Facebook page. The postings were free of charge.

Figure 4. Online advertisement for Snart-Gravid.dk posted on Netdoktor.dk.

Figure 5. Online advertisement for SnartForaeldre.dk posted on Netdoktor.dk.

Assessment of Number of Participants Recruited by Each Method

We defined seven categories of mutually exclusive recruitment methods: online advertisements, press releases, posters, flyers, SnartForaeldre homepage, other homepages, and "other." We used two data sources, electronic tracking and self-reported data, to determine the recruitment method for each participant. The number of participants enrolled as a result of online advertisements was obtained by electronically tracking participants' click on the advertisements via Uniform Resource Locator (URL) using different URL codes for each online campaign. URL tracking overruled any self-reported response by the participants. Facebook provided detailed statistics regarding the number of impressions and the number of clicks for each advertisement placed on its website. In addition, we calculated a click-through rate, that is, the number of times someone clicked on an advertisement divided by the number of times it was shown, and a conversion rate, that is, the number of people who enrolled after having clicked on the advertisement (and thus registered by its unique URL) divided by the total number of people who clicked on the advertisement.

The remaining participants were categorized according to their response to the following question on the baseline questionnaire: "How did you hear about the present study?" The response options were: "SnartForaeldre website," "Facebook," "Netdoktor," "Other websites or blogs or chatrooms," "Poster," "Flyer from my GP," "Radio," "Television or teletext," "Newspapers," "Weekly magazines," "Previously participated

in study," and "Word of mouth," as well as an open-ended "Other." It was possible to mark multiple response options. Any open-ended response that could be linked unambiguously to an online advertisement was categorized as such.

The number of women who enrolled as a result of press releases was determined by identifying those who, within a period of 2 weeks after a press release, marked at least one of the following responses: "Other homepages," "Radio," "TV or teletext," "Newspapers," "Weekly magazines," or "Other magazines," without marking any other answers. In a sensitivity analysis, we expanded the time period to 4 weeks.

In a subanalysis, we estimated the average number of participants recruited per 30 days for each online recruitment method.

Finally, we excluded women who did not respond to the question "How did you hear about the study?" or who outside a press release window marked more than one answer. The reason for these exclusions was the inability to assign respondents to one specific recruitment category.

Costs

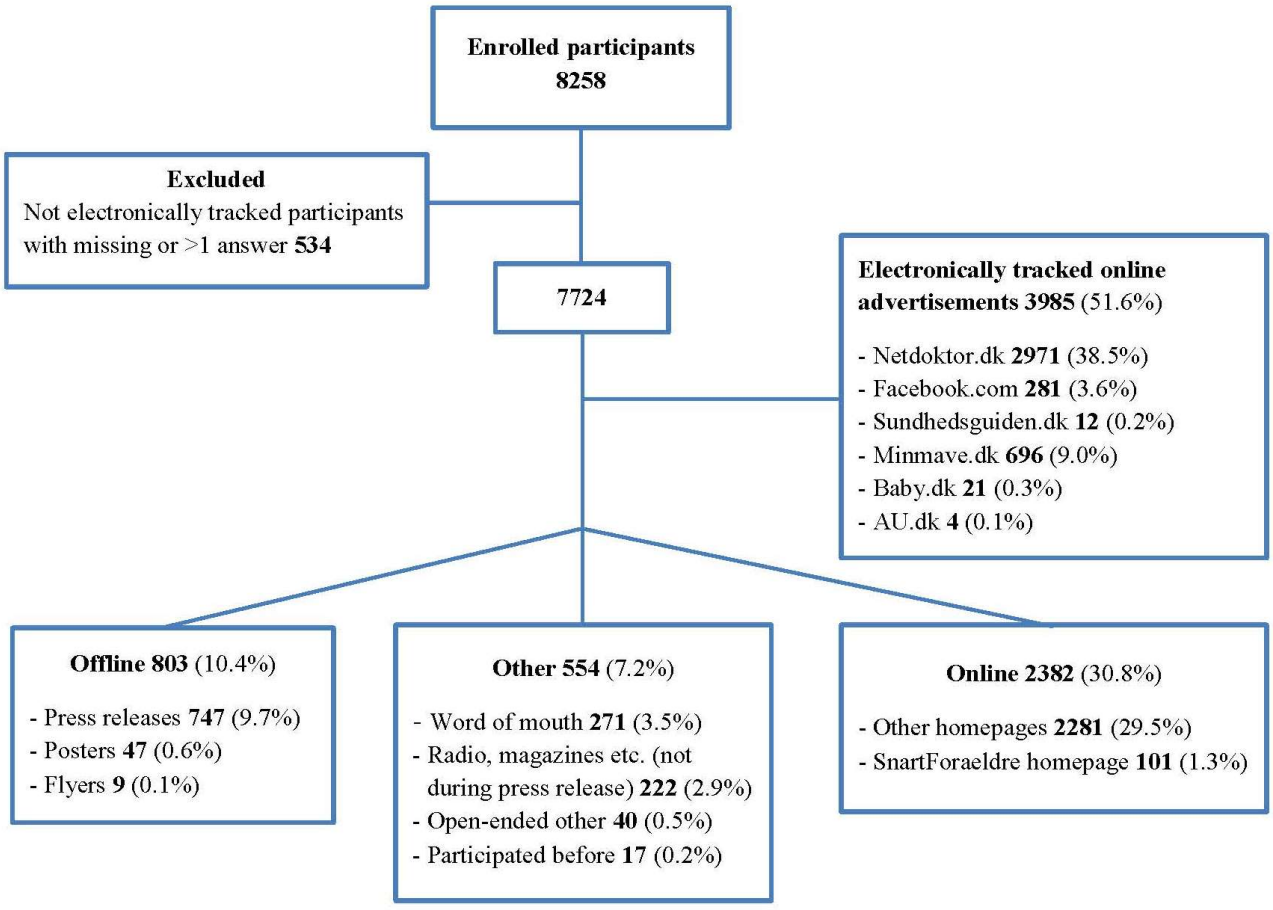
For each recruitment method, costs per enrolled participant were calculated by dividing the total direct expenses (which did not include salary expenses for study staff) by the number of participants recruited from that method. For the recruitment methods "SnartForaeldre homepage," "other homepages," and "other," costs per enrolled participant were not calculated as these methods were free of charge. All costs are reported in euros, applying exchange rates as of March 2012.

Results

Overall Recruitment

In total, 8258 participants were recruited for our Web-based cohort study during the study period (Figure 6). We excluded 534 (6.5%) participants from this analysis as they either did not answer the question on how they heard about the study (n=19) or they provided more than one answer (515 respondents, among whom 452 provided two answers, 59 provided three answers, and 4 gave four answers). Thus, the final study population consisted of 7724 participants. Of these, 3985 (51.6%) were recruited by online advertisements, 803 (10.4%) were recruited by offline methods, and 2382 (30.8%) were recruited from two other online methods, that is, "SnartForaeldre homepage" and "other homepages" (see Figure 6). Other methods generated 554 participants (7.2%), of whom 271 (48.9%) learned about the study from word-of-mouth, whereas 222 (40.1%) heard about the study through radio, TV, newspapers, or magazines outside the time window of a press release. Of the 3985 participants recruited from online advertisements, 3866 were tracked electronically and 119 were categorized based on self-reported data.

Figure 6. Number of participants by recruitment method.



Offline Recruitment Methods

For the offline methods, six press releases yielded 747 participants in total. The press releases generated between 8 and 81 articles/features and recruited between 21 and 506 participants at a cost of €0 to €24.28 per participant. Three press releases were issued without external assistance and were not assigned a price (Table 1). The sensitivity analysis in which the time window of the definition of the "press release recruitment" was changed to 4 instead of 2 weeks did not change the cost per participant for the last four press releases. However, for the press release "Launch of Snart-Gravid.dk" the average cost per participant decreased from €4.51 to €3.35 and for the second press release "Funding received to continue study," the average cost declined from €18.35 to €11.82. Distribution of 133 posters resulted in 47 participants at a cost of €8.04 each, whereas the distribution of 3500 flyers for display led to 9 participants, each costing €67.50 (Table 1). The average cost per participant recruited via offline recruitment methods initiated by study staff was €9.06.

Table 1. Costs and number of participants recruited by offline recruitment methods.

Method		Characteristics	Participants N	Average costs per participant €
Press releases				
	Launch of Snart-Gravid.dk	35 articles/features	506	4.51
	Funding received to continue study	50 articles/features	96	18.35
	Investigating impact of H1N1 influenza	8 articles/features	28	0.00
	Launch of SnartForaeldre.dk	58 articles/features	73	24.28
	Results: physical activity	84 articles/features	23	0.00
	Results: oral contraceptives	81 articles/features	21	0.00
Posters		133 placed in canteens, supermarkets, and so on	47	8.04
Flyers		3500 distributed to GP[a] offices, pharmacies, and so on	9	67.50
All offline methods			803	9.06

[a]GP: general practitioners.

Online Recruitment Methods

Online methods comprised online campaigns including banner advertisements on various websites. A total of 4 years and 7 months of frequent advertisements on Netdoktor.dk generated 2912 participants, each at a fixed price of €3.93 in Snart-Gravid.dk and €2.99 in SnartForaeldre.dk. Factoring in 59 participants who stated that they heard about the study from Netdoktor.dk, but who were not tracked electronically, the costs were €2.74 per participant in SnartForaeldre.dk (Table 2). The average number of participants recruited via Netdoktor.dk per 30 days was 73 in Snart-Gravid.dk and 27 in SnartForaeldre.dk.

Table 2. Costs and number of participants recruited by online recruitment methods.

Online advertising campaigns	Characteristics	Participants N			Average costs per participant €	
		Tracked	Tracked per 30 days	Self-reported	Tracked	Tracked and self-reported
Netdoktor.dk						
Snart-Gravid.dk	One campaign[a], 31 months	2273	73	0	3.93[a]	N/A[b]
SnartForældre.dk	One campaign[a], 24 months	639	27	59	2.99[a]	2.74
Minmave.dk	Five campaigns, 1 month each	689	138	7	14.93	14.78
	October 1, 2012	212		N/A	9.70	N/A
	January 20, 2013	158		N/A	13.02	N/A
	March 1, 2013	142		N/A	14.49	N/A
	May 14, 2013	88		N/A	23.37	N/A
	August 31, 2013	89		N/A	23.11	N/A
Sundhedsguiden.dk	One campaign, 8 months	11	1	1	115.12	105.53
Baby.dk	One campaign, 3 months	19	6	2	45.40	41.08
AU.dk	Postings on AU[c] intranet and AU Facebook page	4	N/A	0	0.00	N/A
Facebook	Seven nationwide advertisements, 8-10 days each	230	121	50	3.44	2.83
All online methods		3866	53	119	6.22	6.04

[a]Negotiated, fixed price per recruited participant.

[b]N/A: not available.

[c]AU: Aarhus University.

Five one-month campaigns on Minmave.dk resulted in 689 participants at an average cost of €14.93 each. Factoring in 7 participants who stated that they heard about the study from Minmave.dk, but who were not tracked by a URL code, the average cost was €14.78 each. However, the number of recruited participants decreased from 212 participants in the first campaign to 89 in the last campaign. Consequently, the corresponding cost per recruited participant increased from €9.70 to €23.11 in the course of the campaigns (Table 2). On average, 138 participants were recruited from Minmave.dk per 30 days.

An advertisement on Sundhedsguiden.dk yielded 11 participants at €115.12 each over 8 months, whereas a 3-month campaign on Baby.dk generated 19 participants at €45.40 each. When we included participants who reported that they knew about the study from these websites but were not tracked, the cost per participant was €105.53 for Sundhedsguiden.dk and €41.08 for Baby.dk (Table 2).

Seven nationwide advertising periods of 8-10 days on Facebook yielded 230 participants at an average cost of €3.44 each (Table 2). Factoring in 50 participants who stated that they heard about the study from Facebook, but who were not tracked electronically, the average cost per participant was €2.83. It was found that one Facebook advertisement targeted only at women living in Aalborg, that is, a small target population, resulted in 1 participant at the cost of €21.94 (Table 3). In terms of individual nationwide Facebook advertisements, the number of recruited participants varied between 22 and 84 for the first four advertisements, whereas the number was 11 or less for the last three advertisements. The average cost per participant ranged from €1.32 to €22.18. The lowest number of impressions was 16,776 and the highest was 409,129. The click-through rate varied between 0.13% and 1.85%. The conversion rate steadily decreased from 12.07% for the first Facebook advertisement to 1.61% for the last. The average number of participants recruited from Facebook per 30 days was 121.

The average cost per participant recruited via any online recruitment method initiated by study staff was €6.22 for those tracked electronically compared with an average of €9.06 for offline methods.

Table 3. Costs and number of participants tracked by Uniform Resource Locator codes in advertisements on Facebook.

Target area	Time	Impressions[a] n	Clicks[b] n	Click-through rate[c] %	Participants n	Conversion rate[d] %	Average costs per participant €
Nationwide	March 30 to April 7, 2010	409,129	696	0.17	84	12.07	1.32
Nationwide	June 29 to July 7, 2010	250,324	587	0.23	60	10.22	1.85
Nationwide	November 25 to December 2, 2010	271,176	403	0.15	42	10.42	2.64
Nationwide	September 22-29, 2011	247,412	324	0.13	22	6.79	5.04
Aalborg	October 9-16, 2011	84,565	39	0.05	1	2.56	21.94
Nationwide	May 3-12, 2012	79,624	157	0.20	6	3.82	16.65
Nationwide	October 31 to November 9, 2013	23,423	344	1.47	11	3.20	12.53
Nationwide	December 21-28, 2013	16,776	311	1.85	5	1.61	22.18
Nationwide, in total		1,297,864	2822	0.22	230	8.15	3.44

[a]The number of times an advertisement was shown.

[b]The number of people who clicked on the advertisement.

[c]The proportion of people who saw the advertisement and clicked on it.

[d]The proportion of people who clicked on the advertisement and were enrolled.

Discussion

Principal Findings

In our study population of 7724 pregnancy planners recruited into a prospective cohort study, the majority (6367 participants, 82.4%) were recruited using online recruitment methods. Offline recruitment methods and other methods accounted for 10.4% (803/7724) and 7.2% (554/7724) of participants, respectively. Overall, the average cost per participant was €6.22 for online methods initiated by study staff versus €9.06 for offline methods. Though online methods were the most efficient recruitment tools, costs ranged from €2.74 to €105.53 per participant. Costs per participant for offline methods ranged from €0 (press release issued with institutional assistance, free of charge) to €67.50. We were able to monitor efficiency (the total number of recruited participants), over time for Facebook and Minmave.dk, and found that efficiency decreased while cost per participant increased over time.

Limitations

We lacked access to information on the number and timing of impressions on Netdoktor.dk. Thus, we were unable to trace any fluctuation in efficiency over time of a prime recruitment method—the advertisements on Netdoktor.dk. We can only speculate whether the efficiency of these advertisements decreased over time as was the case with our advertisements on Facebook and the campaigns on Minmave.dk.

The average number of participants recruited per 30 days from online campaigns varied from 1 (Sundhedsguiden.dk) to 138 (Minmave.dk). Minmave.dk and Facebook yielded the highest average number of participants per 30 days. However, the number of impressions for Netdoktor.dk was unknown and it varied both across and within the other recruitment methods, limiting the comparability of the numbers.

The possible imprecision of the estimated efficiency of press releases is also a concern. The number of recruited participants yielded by this method was defined according to a given time period, which may have been too short. As indicated in our sensitivity analysis, more women may have been recruited from this method than we calculated, and thus, we may have overestimated the average cost per participant.

Another limitation is the possibility that participants who reported that they learned about the study from “Other home pages” (2281 participants, 29.5%) may have seen an online advertisement but not clicked on it. Instead they may have signed up later directly via the study website. Thus, the number of participants recruited by each online advertisement may have been underestimated.

Our exclusion of 515 women who marked more than one answer to the question “How did you hear about the study?” may also

be viewed as a study limitation, as it may have led to overestimation of the cost per participant. However, had these women been assigned to two or more categories of recruitment methods, they would have been counted twice and costs would have been underestimated.

Comparison With Prior Work

Online Methods

Our special agreement with Netdoktor.dk involved paying a fixed price of less than €4 per recruited participant and resulted in 2971 participants during the study period. Such an agreement seems unusual. However, Ramo et al [20] reported a similar arrangement with a Web-based sampling service that sent out email invitations to people signed up as volunteers for completing Web-based surveys. They paid a fixed price of US $19.24 (€14.71) per completed survey and achieved 67 in 6 months. Thus, their price was much higher than the price we negotiated with Netdoktor.dk.

Compared with other studies, the overall average cost of €6.22 per participant recruited by online methods in our study is low. Graham et al [21] reported an average cost per registrant of US $209.34 (€160.06) (ranging from US $73.76 (€56.40) to US $4166.67 (€3185.84)) in their study on the effectiveness of advertisements placed on four different Spanish-language websites to recruit Latinos for a Web-based smoking cessation program. Ramo et al [20] paid US $42.77 (€32.70) per participant recruited through an Internet advertising campaign using various social networking and lifestyle-based websites.

For our North American sister study, PRESTO, Wise et al [10] reported costs of US $43.00 to $96.15 (€32.88 to €73.52) per participant for those recruited via online advertisements placed on websites other than Facebook. This agrees with the costs of the less efficient websites used for advertising our study, which cost up to €115 per participant.

The efficiency of advertisements on Facebook varied greatly in our study and decreased over time. As the number of impressions decreased markedly throughout the study period, the advertisements were shown fewer times at the same cost. Interestingly, the click-through rate was considerably higher for the last two advertisements (1.47% and 1.85%) compared with the first six (0.05-0.23%). However, this did not translate into more participants as the conversion rate decreased throughout the study period. Still, the average cost per participant recruited from Facebook advertisements was relatively low at €3.44.

Ramo and Prochaska [13] compared 20 different advertisements placed on Facebook over a 13-month period to recruit 18- to 25-year-olds for a survey on the use of tobacco and other substances. Efficiency varied widely across the advertisements. In addition, both the number of impressions and the number of clicks increased over time, but not concomitantly. Ramo et al paid an average cost of US $4.28 (€3.27) per completed survey, similar to our costs.

In comparison with the average cost of €3.44 per recruited woman via Facebook in our study, costs ranging from €6.73 to €19.48 have been reported in studies on nutrition [14,30], general health [19], smoking cessation [31], and childbirth preferences [7]. Our sister study, PRESTO, and another study by Richiardi et al, with a topic and target group similar to ours, paid US $27.77 (€21) and €25, respectively, per woman recruited from Facebook advertisements [10,17].

In contrast to our results, Thornton et al [12] reported costs of US $1.86 (€1.42) per participant recruited by Facebook advertisements for a survey about tobacco, cannabis and alcohol use, and mental health, whereas Nelson et al [15] paid US $1.36 (€1.04) per participant who completed a Web-based survey on human papillomavirus (HPV) vaccine uptake.

Bearing in mind that the price per click varies according to the Facebook bidding system and the number of advertisers opting for the same target group, costs from the various studies using Facebook depend either on the target country or area, and may not be directly comparable. The criteria by which Facebook advertisements can be tailored to reach specific target populations also vary from one country to another. Wise et al [10] reported that they used “newlyweds” as a target criterion for the PRESTO study in the United States, whereas this criterion was not available on Facebook in Denmark during our study period. Thus, the efficiency of Facebook advertisements may be influenced by how precisely the advertisement can be targeted at potential participants.

Another factor influencing comparability is the use of incentives for enrolling. Announcing incentives before enrollment may make more people want to sign up for the study in question [12,22]. Among other studies using Facebook advertisements, all used incentives [7,10,12-15,19,30], except for one [17]. Apart from the PRESTO study by Wise et al [10], which offered incentives for completing a required number of follow-up questionnaires, incentives were offered for enrollment and thus could have affected people’s inclination to participate. Some studies have reported the use of incentives worth US $15 (€11.46) [30] and US $20 (€15.28) [15] per participant. One could argue that these costs should be included in the overall recruitment cost per participant. However, this would make it more difficult to compare costs directly.

Some studies using online enrollment methods have reported problems with duplicate entries, in particular studies offering incentives [15,20,30]. We did not offer incentives and our study participants had to provide their unique 10-digit civil registration number to enroll, thus ruling out the possibility of multiple entries. The fact that multiple entries into the study was impossible is a major strength of our study and means that the reported number of participants recruited by each method reflects the actual number of distinct eligible individuals responding to the advertisement.

It may be a concern that online recruitment may not yield a sample which is representative of the background population. However, this concern is of relevance in a cross-sectional study that aims to estimate the prevalence of a disease or a risk factor in a given population at a given time. We do not recruit participants to a cross-sectional study, but into a cohort, where we follow our participants until they achieve a pregnancy, start fertility treatment, or end of observation (12 cycles). On the basis of comparisons within the cohort, we estimate associations

between various exposures and time to pregnancy. Some differences are likely to exist between our population and all Danish women trying to conceive, but these differences are unlikely to affect the internal comparisons made within our cohort. In addition, in a recent validation study, we showed that the selection bias was not a major concern in the associations under study [25].

The focus of our study on fertility could explain why our recruitment methods proved more efficient compared with some other studies. We had the advantage of an interesting topic—the expectations associated with possible, impending motherhood. Harris et al speculated that the topic of their study on contraception and pregnancy plans motivated women to enroll. Although their study offered incentives, they found that Facebook advertisements not mentioning the incentive, but rather focusing on generating new scientific knowledge to benefit other women, were highly efficient [6].

Similar to other studies that recruit hard-to-reach populations [13,15], we found that online recruitment methods were indeed a viable method to reach our target population.

As was the case with Facebook, the efficiency of the online campaigns on Minmave.dk decreased over time. The average cost per participant thus increased from €9.70 to €23.11. A possible explanation could be saturation of the target population on Minmave.dk. Another explanation could be the increasing tendency toward the use of smaller devices, like tablets and smart phones, for accessing the Internet [18]. However, due to the length of our questionnaire and the comprehensiveness of the questions, it would be cumbersome to complete it on a mobile phone.

Offline Methods

The offline methods used in our study were less efficient than the online methods in terms of total number of enrolled participants. In a French Web-based study on nutrition, Kesse-Guyot et al [11] included offline methods to recruit participants and found that only 0.61% and 1.69% of participants learned about the study from posters and flyers, respectively. Our findings are similar, as only 0.61% (47) and 0.12% (9) of our participants reported hearing about the study from posters and flyers, respectively. Kesse-Guyot et al also reported varying effects of their mass-media campaigns, which were launched with a press release and a press conference. They experienced the highest peak in enrollment after the first campaign or press release. This was also the case in our study.

When Wise et al [10] used flyers as a recruitment method, they gained 46 participants (out of 2421 in total) at a cost of US $12 (€9.18) each, whereas postcards led to 15 participants costing US $87 (€66.52) each. Thus, their experience with offline methods is similar to ours in terms of poor efficiency and relatively high costs per participant.

In general, studies employing both offline and online methods to recruit participants for Web-based data collection have found offline methods to be less efficient [6,10,22,23]. This difference may stem from the extra participant effort required by offline methods. After seeing a poster in, for example, the gym, you would have to go online and find the study homepage to enroll. In contrast, when an advertisement is provided on the Internet, enrollment is only a few clicks away [21,23]. It should also be kept in mind that offline methods are more labor-intensive for the research team [7]. This factor was not taken into account in the costs reported in our study. In addition, compared with offline methods, online methods seem more familiar to our young target population, which increasingly conducts daily communications [31].

Conclusions

We were able to recruit large numbers of women for our Web-based prospective cohort study on fertility. We found that online recruitment methods were superior to offline methods in terms of efficiency (total number of participants enrolled). Both online and offline costs per enrolled participant showed great internal variation. However, given the higher workload associated with offline methods and the lower efficiency, online methods appear the most appealing. The study topic and the use of incentives are likely to influence the efficiency of recruitment methods and the fact that efficiency of online methods may decrease over time suggests that it may be optimal to use multiple online methods.

Acknowledgments

The Snart-Gravid.dk and SnartForaeldre.dk studies were supported by the NICHD (R21050264), the Danish Medical Research Council (271-07-0338), and the NIH (R01 HD060680-01). The sponsors had no influence on the study in any way.

Authors' Contributions

All authors contributed to the design of the study. TC wrote the first and successive drafts of the paper. AHR, EM, and TC carried out the statistical analyses. All authors contributed to the interpretation of results, reviewed and approved the final manuscript.

Conflicts of Interest

None declared.

References

1. van Gelder MM, Bretveld RW, Roeleveld N. Web-based questionnaires: the future in epidemiology? Am J Epidemiol 2010 Dec 1;172(11):1292-1298 [FREE Full text] [doi: 10.1093/aje/kwq291] [Medline: 20880962]

2. Olsen J, Melbye M, Olsen SF, Sørensen TI, Aaby P, Andersen AM, et al. The Danish National Birth Cohort--its background, structure and aim. Scand J Public Health 2001 Dec;29(4):300-307. [Medline: 11775787]
3. Ekman A, Litton J. New times, new needs; e-epidemiology. Eur J Epidemiol 2007;22(5):285-292. [doi: 10.1007/s10654-007-9119-0] [Medline: 17505896]
4. Sjöström M, Umefjord G, Stenlund H, Carlbring P, Andersson G, Samuelsson E. Internet-based treatment of stress urinary incontinence: a randomised controlled study with focus on pelvic floor muscle training. BJU Int 2013 Jan 25;112(3):362-372. [doi: 10.1111/j.1464-410X.2012.11713.x]
5. Cook C. Email interviewing: generating data with a vulnerable population. J Adv Nurs 2012 Jun;68(6):1330-1339. [doi: 10.1111/j.1365-2648.2011.05843.x] [Medline: 21999402]
6. Harris ML, Loxton D, Wigginton B, Lucke JC. Recruiting online: lessons from a longitudinal survey of contraception and pregnancy intentions of young Australian women. Am J Epidemiol 2015 May 15;181(10):737-746. [doi: 10.1093/aje/kwv006] [Medline: 25883155]
7. Arcia A. Facebook Advertisements for Inexpensive Participant Recruitment Among Women in Early Pregnancy. Health Educ Behav 2014 Jun;41(3):237-241 [FREE Full text] [doi: 10.1177/1090198113504414] [Medline: 24082026]
8. Huybrechts KF, Mikkelsen EM, Christensen T, Riis AH, Hatch EE, Wise LA, et al. A successful implementation of e-epidemiology: the Danish pregnancy planning study 'Snart-Gravid'. Eur J Epidemiol 2010 May;25(5):297-304 [FREE Full text] [doi: 10.1007/s10654-010-9431-y] [Medline: 20148289]
9. Mikkelsen EM, Hatch EE, Wise LA, Rothman KJ, Riis A, Sørensen HT. Cohort profile: the Danish Web-based Pregnancy Planning Study--'Snart-Gravid'. Int J Epidemiol 2009 Aug;38(4):938-943 [FREE Full text] [doi: 10.1093/ije/dyn191] [Medline: 18782897]
10. Wise LA, Rothman KJ, Mikkelsen EM, Stanford JB, Wesselink AK, McKinnon C, et al. Design and Conduct of an Internet-Based Preconception Cohort Study in North America: Pregnancy Study Online. Paediatr Perinat Epidemiol 2015 Jul;29(4):360-371 [FREE Full text] [doi: 10.1111/ppe.12201] [Medline: 26111445]
11. Kesse-Guyot E, Andreeva V, Castetbon K, Vernay M, Touvier M, Méjean C, et al. Participant profiles according to recruitment source in a large Web-based prospective study: experience from the Nutrinet-Santé study. J Med Internet Res 2013 Sep;15(9):e205 [FREE Full text] [doi: 10.2196/jmir.2488] [Medline: 24036068]
12. Thornton LK, Harris K, Baker A, Johnson M, Kay-Lambkin FJ. Recruiting for addiction research via Facebook. Drug Alcohol Rev 2016 Jul;35(4):494-502. [doi: 10.1111/dar.12305] [Medline: 26179224]
13. Ramo DE, Prochaska JJ. Broad reach and targeted recruitment using Facebook for an online survey of young adult substance use. J Med Internet Res 2012;14(1):e28 [FREE Full text] [doi: 10.2196/jmir.1878] [Medline: 22360969]
14. Lohse B. Facebook is an effective strategy to recruit low-income women to online nutrition education. J Nutr Educ Behav 2013;45(1):69-76. [doi: 10.1016/j.jneb.2012.06.006] [Medline: 23305805]
15. Nelson EJ, Hughes J, Oakes JM, Pankow JS, Kulasingam SL. Estimation of geographic variation in human papillomavirus vaccine uptake in men and women: an online survey using facebook recruitment. J Med Internet Res 2014;16(9):e198 [FREE Full text] [doi: 10.2196/jmir.3506] [Medline: 25231937]
16. Cooper CP, Williams KN, Carey KA, Fowler CS, Frank M, Gelb CA. Advertising campaign on a major internet search engine to promote colorectal cancer screening. BMJ 2004 May 15;328(7449):1179-1180 [FREE Full text] [doi: 10.1136/bmj.328.7449.1179] [Medline: 15142926]
17. Richiardi L, Pivetta E, Merletti F. Recruiting study participants through Facebook. Epidemiology 2012 Jan;23(1):175. [doi: 10.1097/EDE.0b013e31823b5ee4] [Medline: 22157313]
18. Kapp JM, Peters C, Oliver DP. Research recruitment using Facebook advertising: big potential, big challenges. J Cancer Educ 2013 Mar;28(1):134-137. [doi: 10.1007/s13187-012-0443-z] [Medline: 23292877]
19. Fenner Y, Garland SM, Moore EE, Jayasinghe Y, Fletcher A, Tabrizi SN, et al. Web-based recruiting for health research using a social networking site: an exploratory study. J Med Internet Res 2012;14(1):e20 [FREE Full text] [doi: 10.2196/jmir.1978] [Medline: 22297093]
20. Ramo DE, Hall SM, Prochaska JJ. Reaching young adult smokers through the internet: comparison of three recruitment mechanisms. Nicotine Tob Res 2010 Jul;12(7):768-775 [FREE Full text] [doi: 10.1093/ntr/ntq086] [Medline: 20530194]
21. Graham AL, Fang Y, Moreno JL, Streiff SL, Villegas J, Muñoz RF, et al. Online advertising to reach and recruit Latino smokers to an internet cessation program: impact and costs. J Med Internet Res 2012;14(4):e116 [FREE Full text] [doi: 10.2196/jmir.2162] [Medline: 22954502]
22. Loxton D, Powers J, Anderson AE, Townsend N, Harris ML, Tuckerman R, et al. Online and Offline Recruitment of Young Women for a Longitudinal Health Survey: Findings From the Australian Longitudinal Study on Women's Health 1989-95 Cohort. J Med Internet Res 2015;17(5):e109 [FREE Full text] [doi: 10.2196/jmir.4261] [Medline: 25940876]
23. Wiist WH, Sullivan BM, Wayment HA, Warren M. A web-based survey of the relationship between buddhist religious practices, health, and psychological characteristics: research methods and preliminary results. J Relig Health 2010 Mar;49(1):18-31. [doi: 10.1007/s10943-008-9228-4] [Medline: 19107601]
24. Wise LA, Rothman KJ, Mikkelsen EM, Sørensen HT, Riis A, Hatch EE. An internet-based prospective study of body size and time-to-pregnancy. Hum Reprod 2010 Jan;25(1):253-264 [FREE Full text] [doi: 10.1093/humrep/dep360] [Medline: 19828554]

25. Hatch EE, Hahn KA, Wise LA, Mikkelsen EM, Kumar R, Fox MP, et al. Evaluation of Selection Bias in an Internet-based Study of Pregnancy Planners. Epidemiology 2016 Jan;27(1):98-104 [FREE Full text] [doi: 10.1097/EDE.0000000000000400] [Medline: 26484423]
26. SnartForaeldre. URL: https://www.snartforaeldre.dk/ [accessed 2016-09-09] [WebCite Cache ID 6kOAOzE9E]
27. DST. Statistics Denmark, Statistical Year Book URL: http://www.dst.dk/Site/Dst/Udgivelser/GetPubFile.aspx?id=22258&sid=bef [WebCite Cache ID 6kO8MPFal]
28. Rothman KJ, Mikkelsen EM, Riis A, Sørensen HT, Wise LA, Hatch EE. Randomized trial of questionnaire length. Epidemiology 2009 Jan;20(1):154. [doi: 10.1097/EDE.0b013e31818f2e96] [Medline: 19234404]
29. Aalborg Municipality. Statistics URL: http://apps.aalborgkommune.dk/statistik/webaarbog/Folketal2014/struktur/Hele_Kommunen/indexlevel1/HeleKommunen1aarsaldersklasser.html [WebCite Cache ID 6jx9yoUjM]
30. Lohse B, Wamboldt P. Purposive facebook recruitment endows cost-effective nutrition education program evaluation. JMIR Res Protoc 2013;2(2):e27 [FREE Full text] [doi: 10.2196/resprot.2713] [Medline: 23948573]
31. Ramo DE, Rodriguez TM, Chavez K, Sommer MJ, Prochaska JJ. Facebook Recruitment of Young Adult Smokers for a Cessation Trial: Methods, Metrics, and Lessons Learned. Internet Interv 2014 Apr;1(2):58-64. [doi: 10.1016/j.invent.2014.05.001] [Medline: 25045624]

Abbreviations

AU: Aarhus University
GP: general practitioner
HPV: human papillomavirus
QR: quick response
URL: Uniform Resource Locator

Edited by G Eysenbach; submitted 29.09.16; peer-reviewed by K Harris, E Nelson; comments to author 03.11.16; revised version received 16.12.16; accepted 31.01.17; published 01.03.17.

Please cite as:
Christensen T, Riis AH, Hatch EE, Wise LA, Nielsen MG, Rothman KJ, Sørensen HT, Mikkelsen EM
Costs and Efficiency of Online and Offline Recruitment Methods: A Web-Based Cohort Study
J Med Internet Res 2017;19(3):e58
URL: http://www.jmir.org/2017/3/e58/
doi:10.2196/jmir.6716
PMID:28249833

Original Paper

Using Mobile Sensing to Test Clinical Models of Depression, Social Anxiety, State Affect, and Social Isolation Among College Students

Philip I Chow[1], PhD; Karl Fua[1], PhD; Yu Huang[2], MS; Wesley Bonelli[2]; Haoyi Xiong[3], PhD; Laura E Barnes[2], PhD; Bethany A Teachman[1], PhD

[1]Department of Psychology, University of Virginia, Charlottesville, VA, United States

[2]School of Engineering and Applied Science, University of Virginia, Charlottesville, VA, United States

[3]Department of Computer Science, Missouri University of Science and Technology, Rolla, MO, United States

Corresponding Author:
Laura E Barnes, PhD
School of Engineering and Applied Science
University of Virginia
151 Engineer's Way, Olsson Hall 101B
PO Box 400747
Charlottesville, VA, 22904-4747
United States
Phone: 1 434 924 1723
Fax: 1 434 982 2972
Email: lbarnes@virginia.edu

Abstract

Background: Research in psychology demonstrates a strong link between state affect (moment-to-moment experiences of positive or negative emotionality) and trait affect (eg, relatively enduring depression and social anxiety symptoms), and a tendency to withdraw (eg, spending time at home). However, existing work is based almost exclusively on static, self-reported descriptions of emotions and behavior that limit generalizability. Despite adoption of increasingly sophisticated research designs and technology (eg, mobile sensing using a global positioning system [GPS]), little research has integrated these seemingly disparate forms of data to improve understanding of how emotional experiences in everyday life are associated with time spent at home, and whether this is influenced by depression or social anxiety symptoms.

Objective: We hypothesized that more time spent at home would be associated with more negative and less positive affect.

Methods: We recruited 72 undergraduate participants from a southeast university in the United States. We assessed depression and social anxiety symptoms using self-report instruments at baseline. An app (Sensus) installed on participants' personal mobile phones repeatedly collected in situ self-reported state affect and GPS location data for up to 2 weeks. Time spent at home was a proxy for social isolation.

Results: We tested separate models examining the relations between state affect and time spent at home, with levels of depression and social anxiety as moderators. Models differed only in the temporal links examined. One model focused on associations between changes in affect and time spent at home within short, 4-hour time windows. The other 3 models focused on associations between mean-level affect within a day and time spent at home (1) the *same* day, (2) the *following* day, and (3) the *previous* day. Overall, we obtained many of the expected main effects (although there were some null effects), in which higher social anxiety was associated with more time or greater likelihood of spending time at home, and more negative or less positive affect was linked to longer homestay. Interactions indicated that, among individuals higher in social anxiety, higher negative affect and lower positive affect within a day was associated with greater likelihood of spending time at home the following day.

Conclusions: Results demonstrate the feasibility and utility of modeling the relationship between affect and homestay using fine-grained GPS data. Although these findings must be replicated in a larger study and with clinical samples, they suggest that integrating repeated state affect assessments in situ with continuous GPS data can increase understanding of how actual homestay is related to affect in everyday life and to symptoms of anxiety and depression.

(J Med Internet Res 2017;19(3):e62) doi:10.2196/jmir.6820

KEYWORDS

mental health; depression; social anxiety; affect; homestay; mobile health; mHealth

Introduction

Our emotional or affective experiences are a central determinant of our behavior, and, in turn, how we behave has an enormous impact on how we feel. Survey results from 2014 showed that, at the trait or more enduring symptom level, 27% of US college students reported feeling "too depressed to function" and 40% reported feeling overwhelming anxiety at least once [1]. Importantly, even subthreshold levels of depression and social anxiety can have a damaging effect on current and future academic performance, relationship quality, and self-esteem [2,3]. Depression and social anxiety are marked by high trait negative affect (ie, the propensity to experience distressing emotions) and low trait positive affect (ie, the propensity to not experience pleasurable emotions) [4]. Not surprisingly, individuals higher in trait negative affect and lower in trait positive affect tend to experience more negative and less positive emotional states (ie, transient fluctuations in mood), as found in past research on depression and social anxiety [5,6].

One key corollary of both trait and state affect is thought to be the degree to which someone seeks out or avoids social contact and engages in activities that provide reinforcement from the environment. To test this idea, researchers typically use questionnaires and ask people to retrospectively report on how positive or negative they have felt and how much social contact they have had over a predetermined period (eg, the last 2 weeks) or in general. Findings generally suggest that positive affect is associated with self-reported approach behavior, such as healthy social engagement, whereas negative affect is associated with self-reported avoidance behavior, such as social isolation [7,8]. While these findings are consistent with the prediction that engaging in activities is associated with healthier emotional functioning, this methodology is limited in its ability to inform researchers about people's actual experiences in everyday life, such as the ways in which mood and time spent at home (a measure suggestive of social isolation) are linked. One concern is that a wealth of research has documented limitations in people's ability to retrospectively report on complex behavioral patterns [9]. Yet knowing how affect, symptoms, and time spent at home are actually linked in real time is critical, both to test influential theories and to guide treatment for individuals with dysregulated trait affect, such as persons with anxiety and depressive disorders. For instance, behavioral activation treatment seeks to decrease depressive symptoms by having patients seek out activities that provide reinforcement [10]. To improve behavioral activation and related treatments that strive to reduce withdrawal and avoidance behaviors (as is standard in cognitive behavior therapy for mood and anxiety disorders) [11], we need a more fine-grained understanding of how time spent at home is related to affective experiences in everyday life. Our research can help provide this understanding by using a new approach to integrate passively sensed global positioning system (GPS) data with in situ state affect assessments.

Until recently, mental health researchers had to rely mostly on static and imprecise self-report measures to infer crucial behavioral patterns, such as avoidance and social isolation (unless they were able to invest enormous resources to do in-person observations, which are rare and typically provide only a small sample of a person's behaviors). Advances in mobile phone technology now make it possible to continuously and unobtrusively monitor where someone is without needing to ask. For example, previous research has found that passively sensed location information can predict depressive symptoms with impressive accuracy [12,13], and researchers have begun to explore passively and actively sensed indicators of stress and health behaviors in college students [14], although little work has focused on how to integrate passively sensed data with affective experiences generated from in situ repeated assessments.

We hypothesized that higher levels of negative affect and lower levels of positive affect (as reflected in trait affect scores from depression and social anxiety symptoms, as well as state affect scores from ecological momentary assessment) would be associated with increased time spent at home. We focused on time spent at home to index isolation because "homestay" is widely recognized by clinicians as an indicator of social disengagement, particularly on college campuses where opportunities for social interaction are plentiful. Moreover, previous research has demonstrated that homestay is a positive predictor of depressive symptoms [13]. Although we had a central hypothesis that time spent at home would be associated with feeling worse (more symptoms, and more negative or less positive state affect), given the novelty of integrating fine-grained location data from GPS sensors with repeated in situ sampling of affect, we did not have specific hypotheses regarding the time course of the homestay-affect links, and therefore explored different models that varied the temporal connections examined. Thus, in addition to testing whether existing theoretical relationships between (trait and state) affect and social isolation could be found in a temporally rich dataset, another contribution of this research is to better understand how best to leverage time-rich data through different models (eg, by looking at mean affect and change in affect as factors associated with concurrent homestay, as well as associations between mean affect, and prior and future homestay).

Methods

Participants

In total, 72 undergraduates between 18 and 23 years old (mean 19.8, SD 2.4, 37 female) completed the study in exchange for course credit or payment. Because this research used a novel, customized mobile phone app that was only compatible with Android platforms, we recruited participants with Android phones through email advertisements sent to a university email listserv for undergraduate students, as well as through an undergraduate study participant pool. The sample reported their race/ethnicity as 42% (30/72) white, 38% (27/72) Asian, 4% (3/72) black, 4% (3/72) Latino, and 13% (9/72) multiracial or unspecified. Due to software bugs and compatibility issues that

resulted in some participants having little to no data, we excluded 9 participants from data analysis, leaving 63 participants. The study was approved by the local university institutional review board.

Self-Report Measures

Social Anxiety

The Social Interaction Anxiety Scale (SIAS) [15] assesses distress from social interactions and was administered at a baseline laboratory session. Participants rated the degree to which they agreed with 20 statements (eg, "I have difficulty talking with other people") from 0 ("not at all characteristic of me") to 4 ("extremely characteristic of me"). The SIAS has been found to differentiate individuals with social anxiety disorder from healthy control participants [15,16]. We examined SIAS scores on a continuous scale (sample mean 29.9, SD 9.6) to obtain the full range of scores (to assess moderation) and in keeping with recommendations from US National Institute of Mental Health Research Domain Criteria: Interim Guidance (Notice Number NOT-MH-11-005) [17]. Based on documented scores among individuals with a diagnosis of social anxiety disorder (mean 40.0, SD 16.0) [15], approximately 16% of our study's sample likely scored above the mean of a diagnosed sample. Internal consistency of the SIAS in our sample was good (alpha=.83).

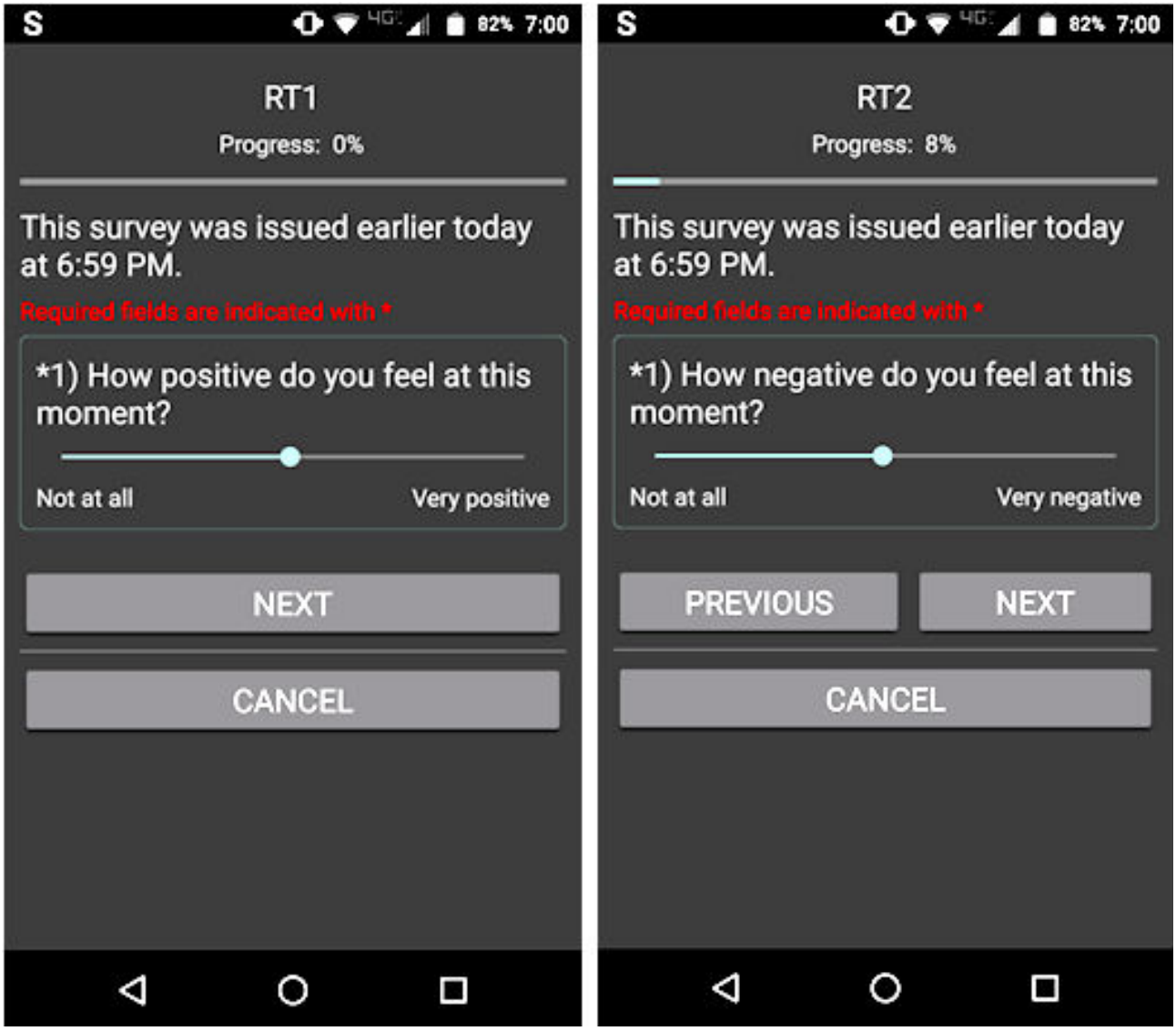
Depression

We used the 7-item depression subscale of the Depression, Anxiety and Stress Scale (DASS-21) [18] to assess depressive symptoms and administered it at a baseline laboratory session. Participants rated the degree to which statements applied to them over the past week (eg, "I felt downhearted and blue") from 0 ("did not apply to me at all") to 3 ("applied to me very much, or most of the time"). In a way similar to the SIAS analysis, we examined depression on a continuous scale (mean 3.3, SD 2.4). Scores among individuals with a diagnosis of major depressive disorder have been previously documented (mean 15.0, SD 4.6) [18]. Internal consistency of the DASS-21 depression subscale in this study's sample was excellent (alpha=.91).

State Affect (In Situ)

Each day, participants provided up to 6 separate ratings for current positive ("How positive are you feeling?) and negative ("How negative are you feeling?") affect using a visual analog scale (designed to minimize influence of one's previous ratings on the current rating) from "not at all" to "very positive" or "very negative." Each scale was coded from 0 to 100. Sliding scales were always initially presented as at the midpoint to reduce biased reporting (Figure 1).

Figure 1. Screenshot of positive (left) and negative (right) state affect rating as seen on a mobile phone screen.



Feature Extraction Using Unobtrusive Data

We indexed social isolation by the percentage of time a participant spent at home relative to other locations. While time at home does not necessarily indicate lack of social interaction, it does suggest lack of engagement with external reinforcements, and variants of homestay have been examined in previous research leveraging GPS data [13]. Identification of home location was based on the most frequent appearance of localities of each participant during nighttime (12:00 AM to 6:00 AM) every day, a method used in past research [13]. No participants had multiple home locations. On any given day, GPS data were over 90% (>173/192) complete between the hours of 10:00 AM and 6:00 PM, based on dividing the amount of data collected by the total amount of possible data. We attributed missing GPS data to users turning off their phone or shutting down the app. To handle gaps in location data, we used the last observation carried forward method.

Procedure

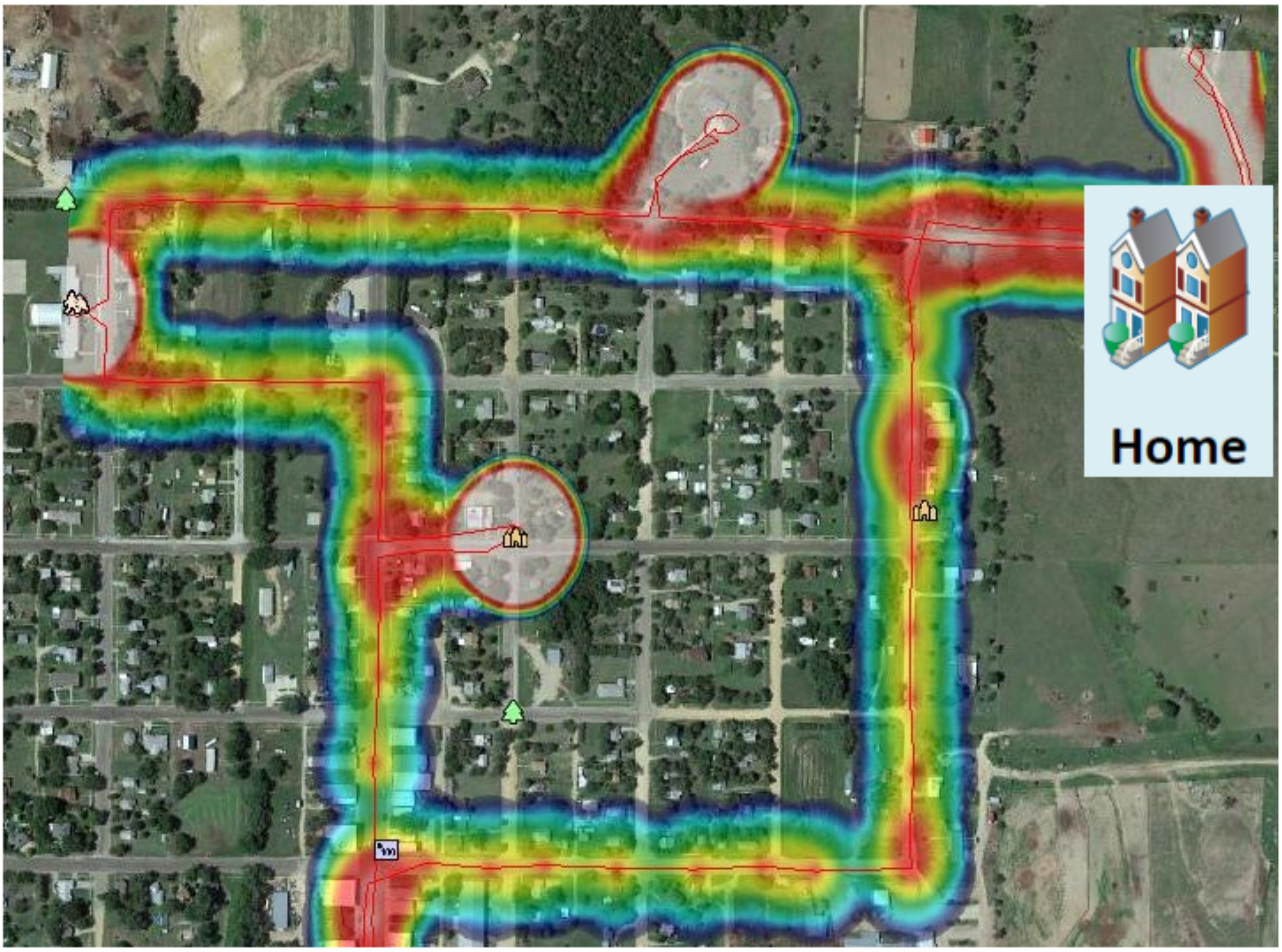
Participants were told that the study examined thoughts and feelings as people interacted with their daily environment, and were allowed to participate in the study if they owned an Android device with operating system 4.3 or higher. All data were collected in the spring semester and participants started the study on a rolling basis. There were no active participants during long scheduled breaks (eg, spring break). Participants completed 2 laboratory sessions separated by roughly 2 weeks (mean 16.41 days, SD 2.69 days). At the first laboratory session, after providing informed consent, participants completed measures of social anxiety and depression. At the end of the laboratory session, research staff assisted participants in downloading the app (Sensus [19]), a mobile crowd-sensing system, onto their personal mobile phones. The app was programmed to prompt participants up to 6 times per day between 9:00 AM and 9:00 PM, occurring at random times within six 2-hour windows; thus, prompts could occur between 2 minutes and nearly 4 hours apart. Due to software bugs, phone compatibility issues, and participants shutting down the app, on average, only 2.5 prompts per day were received. In addition to delivering self-report prompts, the app collected GPS location data continuously over the course of the study. Participants were made aware of the type of data the app would collect from their mobile phones. The app was active throughout the study period

until participants came back for the second laboratory session, when they were fully debriefed. Other measures assessing psychosocial factors not directly related to this research were administered at 2 laboratory sessions, as were additional questions during the ecological momentary assessment portion of the study. A full list of measures and questions can be obtained by contacting the lead author (PIC).

Data Preprocessing

GPS location information (Figure 2) was collected every 150 seconds and uploaded to a secure Amazon Simple Storage Service server (Amazon S3; Amazon Web Services, Inc, Seattle, WA, USA). Raw GPS data were first clustered according to spatial and temporal locality [20,21]. For this study, we used 30 minutes as the temporal threshold and 30 meters as the spatial threshold. We removed GPS points where the participant stayed less than 5 minutes (eg, roads used for travel) [12,20].

Figure 2. Example of global positioning system (GPS) location data overlaid on a satellite image. The colors indicate the amount of time spent at various locations (more red indicating more time spent at a particular location, with the red line indicating a path connecting various locations).



Data Analysis

In all statistical models, state affect served as the independent variable and time spent at home was the dependent variable. Social anxiety and depression symptoms (using the SIAS and the depression subscale of the DASS-21) were entered as the moderators. We examined 4 different models. The first model examined associations between change in state affect (we examined positive and negative affect separately) across a time window lasting up to 4 hours and time spent at home during that same time window. Change in state affect was computed as the difference between self-reported affect from the start to the end of a time window (computed separately for positive and negative affect). We based the decision to use a window length of up to 4 hours on any 2 random time prompts being timed to go off at a maximum of 4 hours from one another (note that only about 12%, 329/2741, of the total random time surveys were rendered unanalyzable because they were too far apart in timing from another survey). The next 3 models examined the associations between mean-level positive and negative state affect and ratio of time spent at home over the course of a typical workday (10:00 AM to 6:00 PM). This was done for models in which state affect was associated with (1) time spent at home the *same* day, (2) time spent at home the *following* day, and (3) time spent at home the *previous* day (models testing prior day homestay have a predictor that follows the outcome, although we wish to stress that all models tested are correlational).

Due to skewed distribution of the time-spent-at-home variable, we computed 2 sets of analyses using generalized mixed-effects models. In 1 set of analyses, using mixed-effects regression, we entered time spent at home as a continuous variable. Time spent at home scores were log transformed to address right skew. For these analyses, we computed time spent at home as

a ratio of the percentage of time an individual spent at home during a set window (ie, within a 4-hour period or over the course of a day), relative to the *average* percentage of time that individual spent at home over the *entire* study period. Thus, the ratio provides an indicator of whether an individual spent more or less time at home over a predefined window relative to the amount of time they typically spent at home during the study. Another set of analyses, using mixed-effects logistic regression, examined the *likelihood* that an individual was at home at some point during that window (in this case, we treated time spent at home as binary, where 1=spent some time at home during predetermined window, and 0=did not spend any time at home during that window).

We conducted all analyses using generalized mixed-effects models and fitted them using the lme4 package in R 3.3.2 (R Foundation) [22]. We computed effect sizes representing the amount of variance explained by the fixed effects in our generalized mixed-effects models using the MuMIn package in R [23]. For analyses with mixed-effects logistic regression, in which time spent at home was a dichotomous variable, we report the unstandardized betas from our models [24]. We used generalized mixed-effects models because they can account for changes over time and missing data more effectively than repeated-measures analyses of variance. In all models, we entered subject and day as separate random intercepts to account for differences in mean responses between participants and between days. For analyses examining time windows within a day, to control for differences in the length of time windows, we entered time window length as a random slope in the analyses.

An example of the model with time spent at home as the criterion variable, state negative affect and depression as predictor variables, and subject as the random intercept is the equation $T_{si} = \beta_0 + \beta_1(NA) + \beta_2(Dep) + \beta_3(NA \times Dep) + S_{0s} + e_{si}$, where $S_{0s} \sim N(0, \tau^2_{00})$ and $e_{si} \sim N(0, \sigma^2)$, and where T is time spent at home, NA is state negative affect, Dep is trait depression, and S_{0s} is the random intercept.

Due to the amount of time spent at home during a given time window being evaluated relative to each participants' personal average time spent at home (ie, a within-subject ratio) for the continuous measure, and the difficulties with interpreting prediction of prior versus same versus next day affect when the predictor was only measured at baseline, we computed a separate set of between-subjects analyses to examine the main effects of depression and social anxiety predicting both time spent at home (without accounting for each participants' personal average time spent at home) and likelihood of time spent at home. This was done using mixed-effects regression models in which time spent at home (as a continuous or dichotomous variable) was the response variable, depression and social anxiety were predictor variables, and subject and day were random intercepts. We computed separate models for 4-hour time windows and between 10:00 AM and 6:00 PM.

Results

Main Effects of Depression and Social Anxiety on Time Spent at Home

As expected, when examining time spent at home using 4-hour windows, there was a significant main effect for social anxiety (standardized beta=.05, *P*=.007), such that a higher level of social anxiety was associated with spending more time at home during 4-hour time windows. Depression was not associated with spending more time at home during 4-hour time windows (standardized beta=.02, *P*=.20). When examining amount of time spent at home between the hours of 10:00 AM and 6:00 PM, as expected, there once again was a significant main effect for social anxiety (standardized beta=.05, *P*=.02), such that a higher level of social anxiety was associated with spending more time at home between 10:00 AM and 6:00 PM. Similarly, those higher in depression tended to spend more time at home between 10:00 AM and 6:00 PM, although this main effect did not quite reach significance (standardized beta=.03, *P*=.06).

When examining the likelihood of time spent at home using 4-hour time windows, as expected, there was a significant main effect for social anxiety (standardized beta=.37, *P*=.02), such that a higher level of social anxiety was associated with a greater likelihood of spending time at home during 4-hour time windows. The main effect of depression (standardized beta=.14, *P*=.35) on likelihood of time spent at home was not significant. When examining time spent at home between the hours of 10:00 AM and 6:00 PM, the main effect of social anxiety was not significant (standardized beta=–.13, *P*=.61). Unexpectedly, there was a significant main effect for depression (standardized beta=–.60, *P*=.001), such that higher depression was associated with a lower likelihood of being at home.

Associations Between Change in Affect and Time Spent at Home in the Same Day Using Short Windows

We then examined the associations between changes in positive and negative affect and time spent at home within the same (up to 4-hour) time windows. For analyses in which time spent at home was a continuous variable, as expected, there was a significant main effect for change in negative affect ($F_{1,230}$=8.61, *P*=.004, R^2=.03), such that a greater increase in negative affect (between the start and end of the time window) was associated with more time spent at home during that window. Similarly, there was a significant main effect for change in positive affect ($F_{1,119}$=9.71, *P*=.002, R^2=.03), such that a greater increase in positive affect was associated with less time spent at home during that window. There were no significant interactions between state affect and either depression or social anxiety (all *P*>.10).

For analyses in which time spent at home was a dichotomous variable, there were no significant main effects for positive affect, negative affect, or interactions between state affect and either depression or social anxiety (all *P*>.10).

Associations Between Affect and Time Spent at Home Within the Same Day

We then examined associations between mean-level positive and negative affect and time spent at home the same day. For analyses in which time spent at home was a continuous variable, as expected, there was a significant main effect for negative affect ($F_{1,105}$=5.52, P=.02, R^2=.02), such that higher negative affect was related to more time spent at home on the same day. There was no significant main effect for positive affect, or interactions between state affect and either depression or social anxiety (all P>.10).

For analyses in which time spent at home was a dichotomous variable, in line with our hypotheses, there was a significant main effect for positive affect (unstandardized beta=–.03, P=.01), such that higher positive affect was associated with a lower likelihood of being at home. There was no significant main effect for negative affect, or interactions between state affect and either depression or social anxiety (all P>.10).

Associations Between Affect and Time Spent at Home the Previous Day

For models in which time spent at home was a continuous variable, there was a significant expected main effect for average positive affect ($F_{1,163}$=5.16, P=.02, R^2=.03), such that greater positive affect was associated with less time spent at home the day before. There was no significant main effect for negative affect, or significant interactions between state affect and either depression or social anxiety (all P>.10).

For analyses in which time spent at home was a dichotomous variable, for those higher in depression, a lower level of positive affect was associated with a higher likelihood of time spent at home the previous day, although this interaction did not reach significance (unstandardized beta=–.03, P=.09). There were no significant main effects for state affect, or interactions between state affect and social anxiety (all P>.10).

Associations Between Affect and Time Spent at Home the Following Day

For analyses in which time spent at home was a continuous variable, counter to hypotheses, there was a significant main effect for average positive affect ($F_{1,152}$=7.02, P=.009, R^2=.03), such that higher positive affect was associated with more time spent at home the next day. There was no significant main effect for negative affect, or interactions between state affect and either depression or social anxiety (all P>.10).

For analyses in which time spent at home was a dichotomous variable, there were no significant main effects for negative affect or positive affect (P>.10). Notably, there was a series of interactions in line with the general hypothesis. Specifically, a significant interaction between state negative affect and social anxiety being associated with time spent at home the next day (unstandardized beta=.05, P=.001) indicated that, among individuals with higher social anxiety, greater negative affect over the course of a day was associated with a greater likelihood of spending time at home the next day (Figure 3). The finding was similar for those higher in depression with a high level of negative affect, although this interaction did not reach significance (unstandardized beta=.02, P=.07). Further, there was a significant interaction between positive affect and social anxiety (unstandardized beta=–.04, P=.007), such that among individuals with greater social anxiety, lower mean-level positive affect over the course of a day was associated with a greater likelihood of spending time at home the next day.

Table 1 shows all of the results for analyses involving the association between changes in affect and time spent at home during 4-hour windows, moderated by depression and social anxiety. It also shows results for analyses involving mean affect and time spent at home on (1) the same day, (2) the previous day, and (3) the following day, moderated by depression and social anxiety, for both the amount of time spent at home and the likelihood of time spent at home. Because of the number of statistical models examined, we conducted a binomial test to determine whether the number of significant effects occurred due to random chance. We chose a binomial test because traditional alpha corrections for type I error rates (eg, Bonferroni) tend to be too conservative, reduce statistical power, and increase the risk of type II errors [25]. Assuming a 5% chance of a type I error for each possible effect in a model (effects were counted as either significant or not), the probability that the 8 (out of a possible 48) significant effects occurred due to chance is <0.001.

Table 1. Main effects of state affect, as well as interactions between state affect and either social anxiety or depression, in relation to the amount and the likelihood of time spent at home.

Affect or interactions	Change in affect in 4-hour window		Mean affect the same day		Mean affect the previous day		Mean affect the following day	
	Beta	*P* value	Beta	*P* value	Beta	*P* value	Beta	*P* value
Standardized betas for amount of time spent at home (as a continuous variable)								
NA[a]	.09	<.01	.08	.02	.03	.39	–.05	.12
NA × SIAS[b]	–.001	.98	–.03	.31	–.05	.19	–.03	.37
NA × DASS[c]	–.01	.81	–.40	.19	–.56	.11	–.01	.85
PA[d]	–.09	<.01	.05	.17	–.09	.02	.09	<.01
PA × SIAS	.01	.66	.04	.23	.03	.40	.004	.90
PA × DASS	.01	.85	.03	.27	.05	.20	–.01	.67
Unstandardized betas for likelihood of time spent at home (as a dichotomous variable)								
NA	–.02	.71	.01	.24	–.02	.27	–.01	.48
NA × SIAS	.06	.15	–.01	.29	.01	.51	.05	<.01
NA × DASS	.06	.25	–.01	.42	.01	.68	.02	.07
PA	–.03	.49	–.03	.01	–.02	.37	–.01	.49
PA × SIAS	–.03	.47	–.01	.30	–.02	.22	–.04	<.01
PA × DASS	–.02	.76	–.001	.92	–.03	.09	–.01	.28

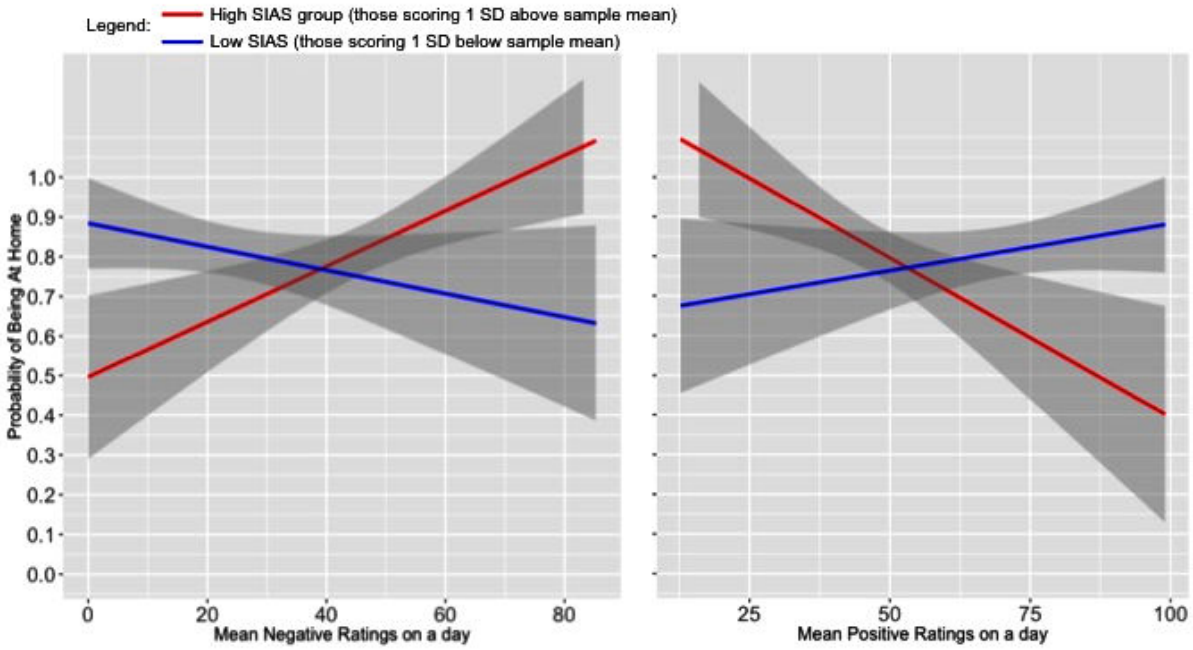
[a]NA: negative affect.

[b]SIAS: Social Interaction Anxiety Scale scores.

[c]DASS: Depression, Anxiety and Stress Scale-depression subscale.

[d]PA: positive affect.

Figure 3. Interactions of mean-level negative (left panel) and positive (right panel) affect with probability of being at home the next day, for those high (1 SD above the mean, in red) and low (1 SD below the mean, in blue) in social anxiety. The Social Interaction Anxiety Scale (SIAS) score was entered as a continuous variable in all models, although to illustrate the interaction effects, only the effects of those high and low in SIAS are plotted.



Discussion

Principal Findings

This study indicates that it is possible to integrate GPS data, a commonly available source of data from mobile phones, with repeated in situ sampling of (positive and negative) affect to enhance understanding of the relationship between affect and time spent at home, and its interaction with depression and social anxiety. By focusing on just one location-based metric, time spent at home, we were able to explore multiple models that differed in terms of their time windows and temporal links. Consistent with existing theory and past research on affect and social isolation, there were significant main effects for social anxiety and state positive or negative affect across many models, suggesting that there are multiple ways of modeling temporally rich links between affect and homestay depending on the research question. For example, researchers interested in studying the relationship between *changes* in positive affect and social isolation may wish to examine shorter time windows within a day.

Existing work in psychology examining the relations between affect and social isolation has mostly relied on data drawn from self-report measures that can have limited ecological validity. Thus, despite sound theoretical reasons to expect affect to predict social isolation, to date, to our knowledge, virtually no research has investigated how this theoretical relationship is expressed in situ in people's actual lives. A major contribution of our research is finding that temporal relationships between affect and time spent at home, a marker suggestive of social isolation, can be modeled using fine-grained data from people's daily lives, through a combination of ecological momentary assessment and passive sensing of location. Thus, research that leverages mobile sensors has the potential to not only test, but also improve, theoretical models from psychology. For example, these methods enable novel tests of the impact of contextual variables on the links between affect and social isolation. It is likely that the degree to which (trait or state) affect is linked to the amount of time someone spends at home (and thus, social isolation), and vice versa, depends on a host of important factors, such as how far someone is from home, who they are surrounded by, biological states such as hunger and fatigue, and even whether data are collected on a weekend or a weekday [26]. As precision of context-aware sensing capabilities continues to be developed, we expect researchers to develop increasingly complex models of behavior that can better account for the richness and variety of everyday life.

We found many of the anticipated main effects of state and trait affect, in that the observed significant effects were almost all in the expected direction (although there were also null findings). There were fewer significant interactions between state affect and symptoms of social anxiety and depression. Interestingly, the reliable interaction findings were with higher negative and lower positive state affect interacting with higher social anxiety, being associated with greater likelihood of spending time at home the following day, suggesting that perhaps the state and trait affect act as a "double whammy." Specifically, these findings are consistent with a wealth of research demonstrating that, in general, how (good or bad) people feel influences their desire to seek out or avoid others or engage in activities outside the home (in our case, socially anxious individuals tending to spend more time at home). We further found that spending more time at home is associated with higher negative affect that day. Not only that, it may be that for socially anxious individuals, a high level of distress or low level of pleasure on a day is associated with retreating to social isolation the next day, creating a double whammy effect. This makes sense considering that individuals high in anxiety have difficulty disengaging their attention from negative stimuli [27], and it may be hard for socially anxious people to stop thinking about the culmination of distressing (or lack of pleasurable) experiences that occurred the previous day, making them want to isolate themselves from others. Future research may wish to examine this possibility more closely by frequent in situ sampling of both emotions and thoughts, as well as passive sensing of social interaction through a wireless technology such as Bluetooth.

There were some null findings, as well as two significant findings, that were contrary to our expectations. In particular, higher positive affect was associated with more time spent at home the next day, and higher depression was associated with a lower likelihood of being at home between 10:00 AM and 6:00 PM. Rather than trying to come up with post hoc explanations for these findings, we wish to emphasize that this research is preliminary and involved running multiple sets of analyses (because we were interested in looking across different methods of operationalizing the homestay-affect link), which inevitably increased the chance for spurious findings. Thus, although we did not observe some expected main effects and interactions, the overall pattern of findings suggests a relationship between feeling worse (ie, being high in social anxiety, or having high state negative affect or low positive affect) and spending time at home, although replication is needed in larger samples.

Limitations

These results should be interpreted in light of several limitations. Although we tested different models of affect predicting time spent at home, the data are correlational and no causal relationships should be inferred. Future research that wishes to make causal claims may want to manipulate affect in a subset of participants to determine effects on social isolation and to manipulate time spent at home to observe changes in affect. Further, although we assessed depression and social anxiety using well-validated self-report measures, we did not use a diagnosed sample or administer structured clinical interviews. It is also possible that other factors outside of depression and social anxiety influenced our findings, such as alcohol use, which is commonly used as a way to decrease depressive and anxious symptoms. Although we sought to examine links between trait and state affect, the measures we used do not directly align with one another (eg, depressive symptoms are not synonymous with negative affect). Future research should try to replicate our findings using a community sample, given that undergraduate students may have similar daily routines (eg, due to common class schedules) that influence time spent at home. Given the limitations of statistical power, we were unable

to examine the potential impact of sex and race, two factors that should be more closely examined in future research involving larger samples.

In addition, while we used ecological momentary assessment to repeatedly sample affect throughout the study period, we did not obtain physiological correlates of affect such as heart rate and skin conductance. One advantage to examining physiological markers of affect is the ability to unobtrusively and continuously monitor fluctuations in states [28]. It will also be important to assess variability across different metrics of isolation, given that time spent at home is only one metric that can be used to approximate social engagement. For example, it may be possible to obtain a more reliable and precise measure of social isolation by combining time spent at home with Bluetooth sensing of individuals nearby or voice recognition software. This would allow researchers to obtain a more fine-grained understanding of whether and how long social isolation occurs at home. Relatedly, asking users to report on their activities at a given location, or inferring activities from other passively sensed data, would allow for enhanced understanding of how people spend their time at various locations and how this might influence emotions and depressive or anxious symptoms, although it would add a considerable measurement burden. Although we based our inference of home location on previous research, future work may wish to query participants about their work hours (eg, if they work a night shift) or activities between 12:00 AM and 6:00 AM, as these factors may influence ability to infer home location. Furthermore, we experienced some technical difficulties such as software bugs and compatibility issues that prevented us from collecting as much data as we would have liked. However, because these problems occurred at random, we still expect that our data can be used to make inferences about people's everyday experiences.

Conclusions

Even taking into consideration these limitations, the ability to integrate fine-grained location data with self-reported affect in situ has tremendous potential to help explicate how short-term, real-time emotional experiences are related to important behavioral patterns in both healthy emotional functioning and in depression and social anxiety. Improving our ability to assess and model variations in affect and GPS patterns may enhance detection of mental disorders through early recognition of signature patterns or change in patterns indicating an increase in isolation, as well as inform treatment planning and assessment of outcomes. Given the ubiquity of mobile phones in our society, understanding how to leverage and integrate seemingly disparate forms of actively and passively sensed data has strong potential to address the growing needs for mental health monitoring and treatment.

Acknowledgments

This research was supported by the Hobby Postdoctoral and Predoctoral Fellowships in Computational Science.

Conflicts of Interest

None declared.

References

1. American College Health Association. American College Health Association National College Health Assessment II: spring 2014 Reference Group executive summary. Hanover, MD: American College Health Association; 2014. URL: http://www.acha-ncha.org/docs/ACHA-NCHA-II_ReferenceGroup_ExecutiveSummary_Spring2014.pdf [accessed 2017-02-22] [WebCite Cache ID 6oTUD3Krv]
2. Fergusson DM, Horwood LJ, Ridder EM, Beautrais AL. Subthreshold depression in adolescence and mental health outcomes in adulthood. Arch Gen Psychiatry 2005 Jan;62(1):66-72. [doi: 10.1001/archpsyc.62.1.66] [Medline: 15630074]
3. Filho AS, Hetem LA, Ferrari MC, Trzesniak C, Martín-Santos R, Borduqui T, et al. Social anxiety disorder: what are we losing with the current diagnostic criteria? Acta Psychiatr Scand 2010 Mar;121(3):216-226. [doi: 10.1111/j.1600-0447.2009.01459.x] [Medline: 19694635]
4. Hettema JM, Neale MC, Myers JM, Prescott CA, Kendler KS. A population-based twin study of the relationship between neuroticism and internalizing disorders. Am J Psychiatry 2006 May;163(5):857-864. [doi: 10.1176/ajp.2006.163.5.857] [Medline: 16648327]
5. Kashdan TB, Steger MF. Expanding the topography of social anxiety. An experience-sampling assessment of positive emotions, positive events, and emotion suppression. Psychol Sci 2006 Feb;17(2):120-128. [doi: 10.1111/j.1467-9280.2006.01674.x] [Medline: 16466419]
6. Nutt D, Demyttenaere K, Janka Z, Aarre T, Bourin M, Canonico PL, et al. The other face of depression, reduced positive affect: the role of catecholamines in causation and cure. J Psychopharmacol 2007 Jul;21(5):461-471. [doi: 10.1177/0269881106069938] [Medline: 17050654]
7. Carver CS, White TL. Behavioral inhibition, behavioral activation, and affective responses to impending reward and punishment: the BIS/BAS Scales. J Personality Social Psychol 1994;67(2):319-333. [doi: 10.1037/0022-3514.67.2.319]
8. Kok BE, Coffey KA, Cohn MA, Catalino LI, Vacharkulksemsuk T, Algoe SB, et al. How positive emotions build physical health: perceived positive social connections account for the upward spiral between positive emotions and vagal tone. Psychological Science 2013 May 06;24(7):1123-1132. [doi: 10.1177/0956797612470827]

9. Van de Mortel TF. Faking it: social desirability response bias in self-report research.. Aust J Adv Nurs Jun 2008;25(4):40-48 [FREE Full text]
10. Kaiser R, Hubley S, Dimidjian S. Behavioural activation theory. In: Wells A, Fisher P, editors. Treating Depression: MCT, CBT and Third Wave Therapies. Hoboken, NJ: Wiley; 2015:221-241.
11. Barlow D, editor. Clinical Handbook of Psychological Disorders. New York, NY: Guilford Press; 2014.
12. Huang Y, Xiong H, Leach K, Zhang Y, Chow P, Fua K, et al. Assessing social anxiety using GPS trajectories and point-of-interest data. 2016 Presented at: 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing; September 12-16, 2016; Heidelberg, Germany p. 898-903. [doi: 10.1145/2971648.2971761]
13. Saeb S, Zhang M, Karr CJ, Schueller SM, Corden ME, Kording KP, et al. Mobile phone sensor correlates of depressive symptom severity in daily-life behavior: an exploratory study. J Med Internet Res 2015;17(7):e175 [FREE Full text] [doi: 10.2196/jmir.4273] [Medline: 26180009]
14. Wang R, Chen F, Chen Z, Li T, Harari G, Tignor S, et al. StudentLife: assessing mental health, academic performance and behavioral trends of college students using smartphones. 2014 Sep Presented at: ACM International Joint Conference on Pervasive and Ubiquitous Computing; September 13-17, 2014; Seattle, WA, USA p. 3-14. [doi: 10.1145/2632048.2632054]
15. Mattick RP, Clarke J. Development and validation of measures of social phobia scrutiny fear and social interaction anxiety. Behav Res Ther 1998 Apr;36(4):455-470. [doi: 10.1016/S0005-7967(97)10031-6]
16. Heimberg RG, Mueller GP, Holt CS, Hope DA, Liebowitz MR. Assessment of anxiety in social interaction and being observed by others: the Social Interaction Anxiety Scale and the Social Phobia Scale. Behav Ther 1992 Dec 01;23(1):53-73. [doi: 10.1016/S0005-7894(05)80308-9]
17. Insel T, Cuthbert B, Garvey M, Heinssen R, Pine DS, Quinn K, et al. Research domain criteria (RDoC): toward a new classification framework for research on mental disorders. Am J Psychiatry 2010 Jul;167(7):748-751. [doi: 10.1176/appi.ajp.2010.09091379] [Medline: 20595427]
18. Lovibond PF, Lovibond SH. The structure of negative emotional states: comparison of the Depression Anxiety Stress Scales (DASS) with the Beck Depression and Anxiety Inventories. Behav Res Ther 1995 Mar;33(3):335-343. [Medline: 7726811]
19. Xiong H, Huang Y, Barnes L, Gerber M. Sensus: a cross-platform, general-purpose system for mobile crowdsensing in human-subject studies. 2016 Sep Presented at: ACM International Joint Conference on Pervasive and Ubiquitous Computing; September 12-16, 2016; Heidelberg, Germany p. 415-426. [doi: 10.1145/2971648.2971711]
20. Kang J, Welbourne W, Stewart B, Borriello G. Extracting places from traces of locations. 2004 Presented at: 2nd ACM international workshop on Wireless mobile applicationsservices on WLAN hotspots; October 1, 2004; Philadelphia, PA, USA p. 110-118. [doi: 10.1145/1024733.1024748]
21. Hasan S, Zhan X, Ukkusuri S. Understanding urban human activitymobility patterns using large-scale location-based data from online social media. 2013 Presented at: 2nd ACM SIGKDD International Workshop on Urban Computing; August 11, 2013; Chicago, IL, USA. [doi: 10.1145/2505821.2505823]
22. Bates D, Maechler M, Bolker B, Walker S. lme4: linear mixed-effects models using Eigen and S4: R package version. 2014 Jun 23. URL: http://keziamanlove.com/wp-content/uploads/2015/04/StatsInRTutorial.pdf [accessed 2017-02-27] [WebCite Cache ID 6obHx75uP]
23. Barton K. MuMIn: multi model inference: model selection and model averaging based on information criteria. 2016 Jan 07. URL: https://cran.r-project.org/web/packages/MuMIn/MuMIn.pdf [WebCite Cache ID 6obI0n74Y]
24. Williams R. Using heterogeneous choice models to compare logit and probit coefficients across groups. Sociol Methods Res 2009 May;37(4):531-559. [doi: 10.1177/0049124109335735]
25. Perneger TV. What's wrong with Bonferroni adjustments. BMJ 1998 Apr 18;316(7139):1236-1238 [FREE Full text] [Medline: 9553006]
26. Saeb S, Lattie EG, Schueller SM, Kording KP, Mohr DC. The relationship between mobile phone location sensor data and depressive symptom severity. PeerJ 2016 Sep 29;4:e2537 [FREE Full text] [doi: 10.7717/peerj.2537]
27. Fox E, Russo R, Dutton K. Attentional bias for threat: evidence for delayed disengagement from emotional faces. Cogn Emot 2002 May 01;16(3):355-379 [FREE Full text] [doi: 10.1080/02699930143000527] [Medline: 18273395]
28. Chow PI, Bonelli W, Huang Y, Fua K, Teachman BA, Barnes LE. DEMONS: an integrated framework for examining associations between physiology and self-reported affect tied to depressive symptoms. 2016 Presented at: 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing; September 12-16, 2016; Heidelberg, Germany p. 1139-1143. [doi: 10.1145/2968219.2968300]

Abbreviations

DASS-21: Depression, Anxiety and Stress Scale
GPS: global positioning system
SIAS: Social Interaction Anxiety Scale

Edited by R Calvo; submitted 14.10.16; peer-reviewed by S Saeb, M Burns, S Schueller, F Shen; comments to author 17.11.16; revised version received 21.01.17; accepted 08.02.17; published 03.03.17.

Please cite as:
Chow PI, Fua K, Huang Y, Bonelli W, Xiong H, Barnes LE, Teachman BA
Using Mobile Sensing to Test Clinical Models of Depression, Social Anxiety, State Affect, and Social Isolation Among College Students
J Med Internet Res 2017;19(3):e62
URL: http://www.jmir.org/2017/3/e62/
doi:10.2196/jmir.6820
PMID:28258049

Original Paper

Microsoft Kinect-based Continuous Performance Test: An Objective Attention Deficit Hyperactivity Disorder Assessment

David Delgado-Gomez[1], PhD; Inmaculada Peñuelas-Calvo[2], MD; Antonio Eduardo Masó-Besga[3], MS; Silvia Vallejo-Oñate[2], MD; Itziar Baltasar Tello[2], MD; Elsa Arrua Duarte[2], MD; María Constanza Vera Varela[2], MD; Juan Carballo[4], MD, PhD; Enrique Baca-García[2,5,6,7,8,9,10,11], MD, PhD

[1]Universidad Carlos III, Department of Statistics, Madrid, Spain
[2]Hospital Universitario Fundación Jiménez Díaz, Child and Adolescents Service, Madrid, Spain
[3]Goal Systems Co, Madrid, Spain
[4]Hospital Universitario Gregorio Marañón, Child and Adolescents Service, Madrid, Spain
[5]Columbia University, New York, NY, United States
[6]IIS-Jimenez Diaz Foundation, Department of Psychiatry, Madrid, Spain
[7]Universidad Autónoma, Madrid, Spain
[8]Hospital Universitario Rey Juan Carlos, Móstoles, Madrid, Spain
[9]Hospital Universitario Infanta Elena, Valdemoro, Madrid, Spain
[10]Hospital General de Villalba, Villalba, Madrid, Spain
[11]CIBERSAM, Madrid, Spain

Corresponding Author:
Inmaculada Peñuelas-Calvo, MD
Hospital Universitario Fundación Jiménez Díaz
Child and Adolescents Service
Avenida Reyes Católicos, 2
Madrid,
Spain
Phone: 34 915417267
Fax: 34 91 542 35 36
Email: inmaculada.penuelas@quironsalud.es

Abstract

Background: One of the major challenges in mental medical care is finding out new instruments for an accurate and objective evaluation of the attention deficit hyperactivity disorder (ADHD). Early ADHD identification, severity assessment, and prompt treatment are essential to avoid the negative effects associated with this mental condition.

Objective: The aim of our study was to develop a novel ADHD assessment instrument based on Microsoft Kinect, which identifies ADHD cardinal symptoms in order to provide a more accurate evaluation.

Methods: A group of 30 children, aged 8-12 years (10.3 [SD 1.4]; male 70% [21/30]), who were referred to the Child and Adolescent Psychiatry Unit of the Department of Psychiatry at Fundación Jiménez Díaz Hospital (Madrid, Spain), were included in this study. Children were required to meet the Diagnostic and Statistical Manual of Mental Disorders (DSM-5) criteria of ADHD diagnosis. One of the parents or guardians of the children filled the Spanish version of the Strengths and Weaknesses of ADHD Symptoms and Normal Behavior (SWAN) rating scale used in clinical practice. Each child conducted a Kinect-based continuous performance test (CPT) in which the reaction time (RT), the commission errors, and the time required to complete the reaction (CT) were calculated. The correlations of the 3 predictors, obtained using Kinect methodology, with respect to the scores of the SWAN scale were calculated.

Results: The RT achieved a correlation of -.11, -.29, and -.37 with respect to the inattention, hyperactivity, and impulsivity factors of the SWAN scale. The correlations of the commission error with respect to these 3 factors were -.03, .01, and .24, respectively.

Conclusions: Our findings show a relation between the Microsoft Kinect-based version of the CPT and ADHD symptomatology assessed through parental report. Results point out the importance of future research on the development of objective measures for the diagnosis of ADHD among children and adolescents.

(J Med Internet Res 2017;19(3):e79) doi:10.2196/jmir.6985

KEYWORDS

kinect; attention deficit hyperactivity disorder; continuous performance test; impulsivity; hyperactivity

Introduction

Attention deficit hyperactivity disorder (ADHD) is the most common childhood neurodevelopmental disorder with an estimated prevalence of around 6% [1]. ADHD is characterized by a pattern of hyperactivity-impulsivity, and inattention. According to the fifth edition of the Diagnostic and Statistical Manual of Mental Disorders (DSM-5), these symptoms usually appear before the age of 12 years and, in order to be considered ADHD, must be present in 2 or more settings such as the school or at home [2]. They can be associated with worse academic performance [3] and cognitive difficulties in the future [4]. Moreover, ADHD increases the risk of developing other psychiatric or psychological disorders [5,6].

Early ADHD diagnosis is of paramount importance in order to minimize the negative cognitive, emotional, academic, and social effects associated with this condition and to restore the child's functioning. However, obtaining an accurate ADHD diagnosis is complicated, as most of the psychiatric and psychological criteria are based on clinical evidence. Questionnaires and scales are probably the most common assessment instruments. The Strengths and Weaknesses of ADHD symptoms and Normal behavior (SWAN) rating scale [7], the diagnosis criteria of the DSM-5 [2] or the ADHD rating scale [8] are some of the different tools for evaluating this medical condition. However, both questionnaires and scales present several inconveniences. For instance, the veracity and accuracy of the answers is not granted [9]. Frazier et al [10] showed that individuals might be tempted to mimic ADHD symptoms to justify a failure at school or work. Receiving a disability status to obtain certain benefits or getting access to stimulant drugs are other two common causes for people to conscientiously exaggerate their symptoms [11]. Conversely, Middeldorp et al [12] have exposed the parental psychological difficulties for internalizing ADHD symptoms, which may influence the answers of the questionnaires. In addition, questionnaires and scales are unsuitable for repeated use due to learning issues [13]. Finally, it has been pointed out that scales suffer from cultural biases [14].

In order to alleviate these weaknesses, new instruments have been developed for assessing ADHD. Currently, the continuous performance test (CPT) is one of the most popular tools for evaluating ADHD [15]. In this test, the examinee is required to press the spacebar in a keyboard as fast as possible every time certain characters appear on a computer screen. If a forbidden character, referred to as an X-stimulus appears on the screen, the examinee must inhibit the reaction. Several studies have demonstrated that CPT has the ability to differentiate children with and without ADHD [16]. CPT measures attention problems with an overall index from reaction time (RT), RT variability, and omissions [17]. In addition, impulsivity is measured by RT and commissions [17]. Notwithstanding its success, some researchers [18] have questioned its utility as a mechanism for differential diagnosis of ADHD. In particular, they found that whereas children with reading disorders have high CPT scores, those with ADHD did not show significantly different scores from clinical controls.

In order to increase the accuracy of the CPT, it has been proposed to additionally collect predictors of the participant's bodily movements during the execution of the test. One common approach consists of the use of accelerometer-based devices (actigraphy and inertial measurement units, IMUs) [19,20]. However, accelerometer-based devices appear to require a prolonged interval of acquired data in order to make an accurate prediction; moreover, they only record movement in 1 or 2 locations on the body. In addition, the intrusiveness of the method is a disadvantage. Studies combining actigraphy with CPT suggest a better accuracy of ADHD diagnosis compared with the use of a standard CPT alone [21].

A second group of techniques is based on the use of infrared motion trackers (McLean Motion Analysis Test and QbTest), small reflectors attached to the child's body for the duration of the test to aid ADHD assessment [22,23]. Studies based on these tracking systems have reported that combined CPT and infrared motion analysis differentiates ADHD children from normal controls [24]. However, these systems are time-locked and although they are able to record the path of movement, they are not yet able to integrate the movement data that are collected by different sensors [25]. As the previously described method, this technique also presents an intrusiveness problem.

One of the devices that are attracting more attention is the Microsoft Kinect. It consists of a standard RGB camera, an infrared sensor, an infrared projector, and a set of 4 microphones. Since its appearance in 2010, several studies have attempted its use for cognitive assessment. For instance, Qiu and Helbig found a statistically significant relation between the corporal posture captured with the Kinect camera and the mental load required to complete 4 different tasks [26]. They observed that the participants' head was close to the screen and the body trunk was tilted forward when the task was more cognitively demanding. In a different work, Yu et al suggested the possibility of detecting children's abnormal behavior using Kinect [27]. Likewise, Stanley proposed to estimate attention levels combining body posture and head orientation [28], aiming at improving the ADHD diagnosis procedure. In a recent work, Li et al [25] used the Kinect for monitoring the intensity of children's movement during a Go or No-Go task in order to attain a better characterization of ADHD. They proposed an

initial measure of total movement intensity, indicated by the number of displaced pixels in the silhouette of the individual in 2 consecutive frames. They apply a Fourier transform to the original time-domain signal to produce a Movement Intensity Distribution which itself is decomposed in 15 alternative nonoverlapping 1 Hz frequency bands. Their results show that each of these bands is a predictor capable of discriminating ADHD children from healthy controls. Unlike the previous studies which were limited to discriminating between two groups (ADHD and controls), these authors found that some of the frequency bands were correlated with the score obtained in the Clinical Global Impression Scale and the ADHD-Rating Scale IV. This result is interesting because none of the traditional measures obtained with the Go or No-Go task, for example, RT or commissions, was correlated with the scores of these psychometrical scales.

In this paper, we proposed a new method for evaluating the severity of ADHD. The method builds in a previous study by Delgado-Gomez et al [29], where the CPT and the Kinect device are combined in order to assess the impulsivity of 22 university students. Results of that study showed that by replacing the keyboard input with body movements, it was possible to obtain a more precise impulsivity assessment. In particular, using a Kinect device, it was found that if the examinee reacts to the stimuli by raising the dominant hand instead of pressing the spacebar, the impulsivity assessment is more accurate than with the sole use of the keyboard because the device is capable of detecting inhibited reactions. We hypothesized that this methodology will provide a more objective evaluation of ADHD to help clinicians in their diagnosis. In particular, the reaction time and number of commissions are good indicators of the severity of the ADHD.

Methods

Subjects

Thirty children, aged 8-12 years (10.3 [SD 1.4]; male 70% [21/30]), who were referred to the Child and Adolescent Psychiatry Unit of the Department of Psychiatry at Fundación Jiménez Díaz Hospital (Madrid, Spain), were included in this study. All participants met the DSM-5 criteria for ADHD [2]. Individuals with ADHD may present both inattention and hyperactivity or impulsivity or only one pattern may be predominant. These 3 traits of ADHD are commonly referred to combined-type, inattentive-type, and hyperactive or impulsive-type (50% [15/30] of the patients were diagnosed as inattentive subtype; 43% [13/30] as hyperactive or impulsive subtype; and 7% [2/30] as combined subtype).

SWAN Scale

The SWAN scale [7] is the diagnosis instrument used in this study to compare the performance of the proposed method with respect to the current clinical diagnose. It is composed of 18 items based on the DSM-5 criteria for ADHD diagnosis which measure positive attention and impulse regulation behaviors in the normal population. The scale is made up of 3 factors. The first factor is associated to inattention and comprises the first 9 items. The following 6 items characterized the hyperactivity factor. The last factor, which comprises the last 3 items, measures impulsivity. In the standard form, each item is scored from −3 to +3 (below average to above average), where 0 is “normal” and based on the population average. This work uses a Web-based mental state tracking e-tool [30] where each item is scored from 0 to 100.

Kinect CPT

In this study, we use a version of the Microsoft Kinect-based CPT proposed by Delgado-Gomez et al [29]. Originally, the Kinect-based CPT has duration of approximately 15 min. During this time, 360 letters appear sequentially on a screen with a time separation of 1, 2, or 4 s. Each time that a non–X-stimulus appears, the participant has to raise his dominant hand as soon as possible and return it to the rest position. When an X-stimulus appears, the examinee must inhibit any reaction.

The version used in this study includes some modifications to avoid the inconveniences reported by the authors. Namely, the duration of the original Kinect-based CPT resulted exhausting for the participants. Moreover, the gap of 1 s between some stimuli was too short for the examinee to be able to complete the action before the appearance of the next character. In order to tackle these inconveniences, the version used in this study takes only 3 min during which 60 letters are presented. Moreover, the time separation between stimuli is modified to 2, 3, or 5 s. A subset of 12 characters is X-stimuli and the remaining 48 are non–X-stimuli.

The RT for the non–X-stimuli, measured as the time elapse between the appearance of the stimulus and the moment in which the participant starts to raise the dominant hand, is calculated. A commissions’ index was calculated for the X-stimuli according to Delgado-Gomez et al. Concisely, it is computed as the ratio of the length of the hand displacement during the X-stimulus and the maximum displacement after the appearance of the previous and posterior stimuli. In addition to these 2 measures, the time required for completing the action of rising and returning the hand to the rest position is also calculated for the non–X-stimuli. Each participant is characterized with the median of the 48 RTs and 48 CTs and the median of the 12 commission indices. The proposed technique is illustrated in the Multimedia Appendix 1.

Ethics Procedures

Parents or guardians were required to sign an informed consent after been explained the project in detail. Participants provided assent. The consent and assent forms and the study protocol were reviewed and approved by the Institutional Review Board of Fundación Jiménez Díaz Hospital. All procedures performed were conducted in accordance with the ethical standards of the institutional and national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards.

During the experiment, a trained psychiatrist accompanied each of the 10 patients while they conducted the Kinect-based CPT. While each child was performing the Kinect-based test, the corresponding parent or guardian filled the SWAN scale.

Results

Before presenting the main results of the experiment, we present the complete performances of 2 examinees in Figures 1 and 2. Figure 1 shows the reaction to each of the 60 stimuli of a participant with a low score in the SWAN scale for impulsivity and inattention. It can be appreciated that there is only 1 commission, 1 inhibited commission (stimuli 50 and 2, respectively), and 1 partial omission (stimulus 49). It can also be seen that in most cases the RT was higher than 500 ms. Figure 2 shows the results of a child with high score in the SWAN impulsivity and inattention scale. We can see that, unlike the previous case, this patient incurred a much larger number of commissions (eg, stimuli 15, 35, and 53): 2 inhibited commissions (8 and 31) and 2 omissions (37 and 52). The occurrence of multiple reactions could also be noted (eg, stimuli 25, 41, and 44). There was 1 case where the reactor started before the stimulus had appeared (stimuli 43-44). Regarding the RT, in nearly 46% of cases, the reactions of this patient were less than 500 ms.

Table 1 shows the statistical summary of our findings. The first row shows the correlations of the RT median with respect to the inattention, hyperactivity, and impulsivity scores provided by the SWAN scale. It can be seen that impulsivity shows a statistically significant negative correlation, clearly indicating that individuals with high impulsivity tend to have low reaction times. A similar result can be observed for hyperactivity, suggesting that hyperactive patients also tend to show low reaction times. There was no evidence suggesting a strong correlation between RT and inattention.

Regarding the commission, the second row in Table 1 shows the correlation of the median of the commission indices with respect to each of the 3 factors of the SWAN scale. It can be observed that the values are positive for the impulsivity factor; and nearly 0 for hyperactivity and inattention. The results suggest that impulsive patients have more difficulties in inhibiting their reactions.

Table 1. Correlation of the median of the RT and commissions with respect to the 3 factors of the SWAN scale (inattention, hyperactivity, and impulsivity).

Explanatory variables	Inattention	Hyperactivity	Impulsivity
RT[a]	−.11 (*P*=.53)	−.29 (*P*=.11)	−.37 (*P*=.04)
Commission	−.03 (*P*=.83)	.01 (*P*=.92)	.24 (*P*=.20)

[a]RT: reaction time.

Figure 1. Performance of a participant with low scores in the impulsivity and inattention factors of the SWAN scale during the test. SWAN: Strengths and Weaknesses of ADHD Symptoms and Normal Behavior.

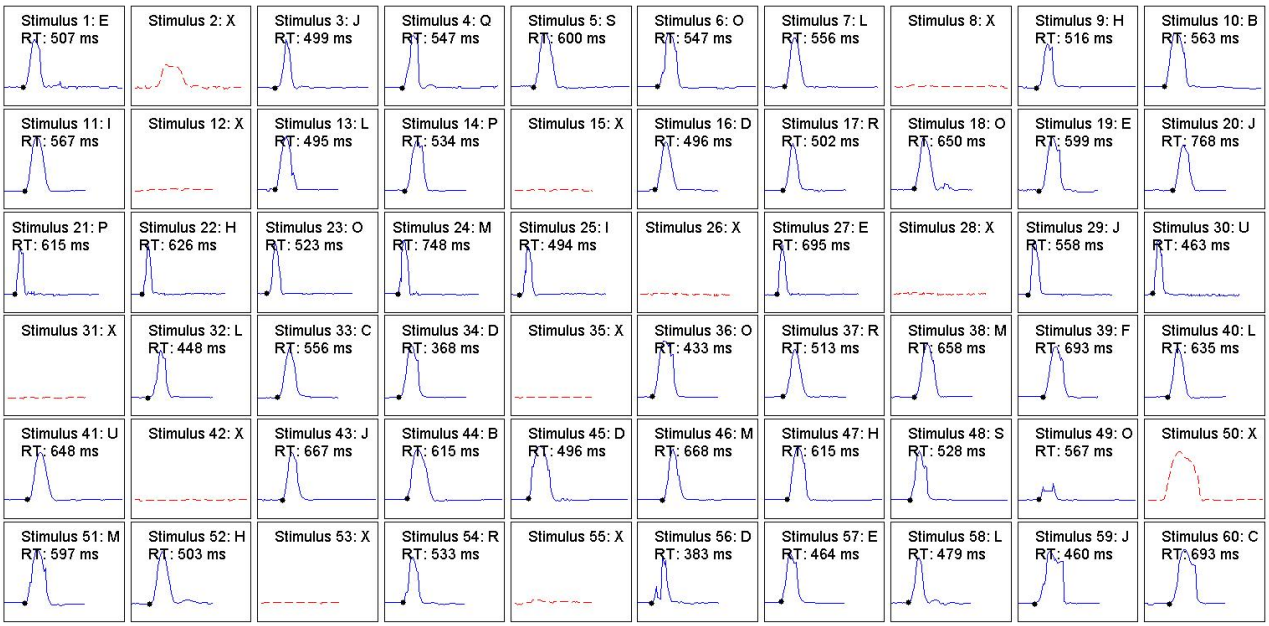
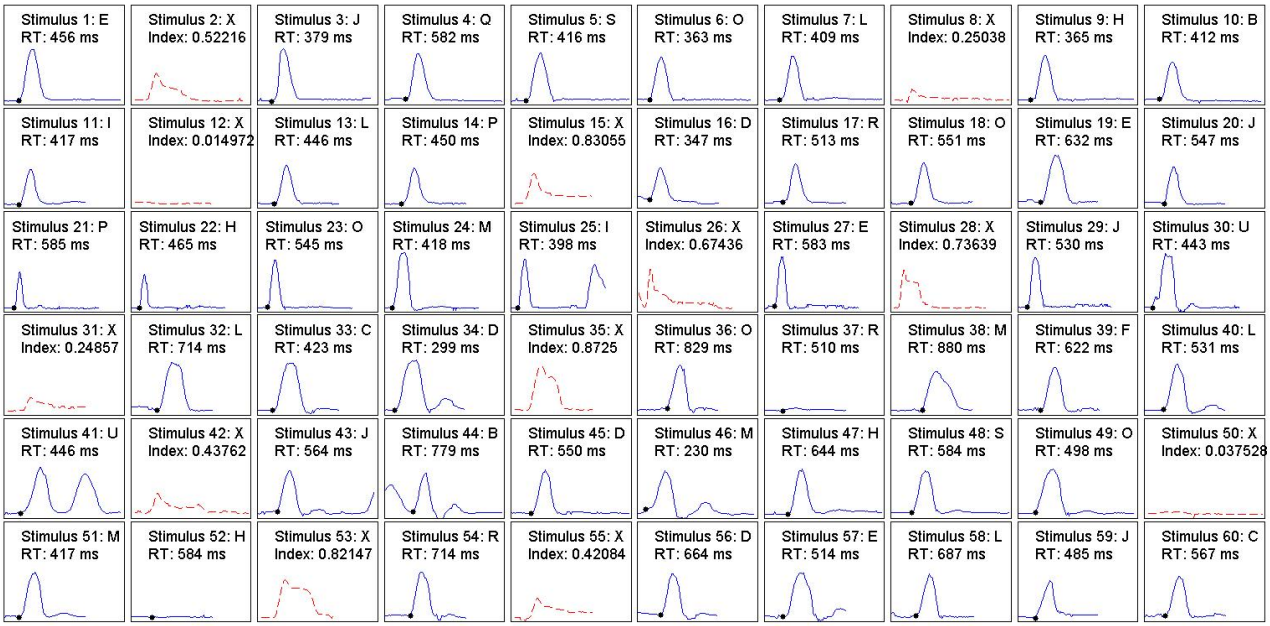


Figure 2. Performance of a participant with high scores in the impulsivity and inattention factors of the SWAN scale during the test. SWAN: Strengths and Weaknesses of ADHD Symptoms and Normal Behavior.



Discussion

In this paper, we propose an innovative approach to assess the severity of ADHD. With this aim, we have developed a completely Kinect-based version of the CPT that allows tracking the complete movement of the examinees' dominant hand, instead of only the pulse of the spacebar in the keyboard, during the performance of the test. With this tool, the examiner can not only record commissions, but also detect inhibited movements and obtain a more accurate measure of the reaction times. It is important to mention that the proposed technique differs from other available visually aided approaches, in that they still rely on the keyboard pulsation, whereas in our methodology the individual's hand takes the place of the spacebar.

In particular, our technique computes the RT and the commission indices for each participant. In order to assess their accuracy, these 2 predictors are compared with each of the 3 factors (inattention, hyperactivity, and impulsivity) of the SWAN scale. The results show that the proposed approach is capable of assessing hyperactivity and impulsivity degrees of individuals diagnosed with ADHD. Regarding the commissions, notwithstanding that the correlation with impulsivity is larger than .2, this value is not statistically significant. A possible explanation for this, as suggested by the much larger values observed in a previous work conducted with healthy population, is that the increase in the inter-stimuli time allows the patient to take a more relaxed approach to the test as then mental load is reduced. Future work will consider reducing these times. These results are in concordance with the ones reported in the literature: in particular, (1) there is a positive correlation between impulsivity and the number of commissions [31]; and (2) it is not possible to establish a correlation between inattentiveness and the number of commissions [17].

Our results have also allowed us to establish a correlation between the RT and impulsivity or hyperactivity. This result is also consistent with the available literature. Moreover, this is an interesting result as Li et al [25] did not find a relation between the traditional CPT reaction time and the severity of ADHD as measured by the ADHD-RS IV scale.

Although our results have shown the proposed approach as a promising research line, its main limitation is the small sample size, consequence of the difficulty of enrolling participants with specific characteristics (children with the age of 8 to 12 years diagnosed with ADHD). Therefore, it is the intention of the authors to replicate the experiment with a larger number of individuals, where the different ADHD subgroups are well represented.

Despite the reduced sample size, our results point out our proposed methodology as an alternative for overcoming the limitations of the current ADHD assessment instruments based mainly on scales and questionnaires. From the clinical prospective, it can be an important aid in differentiating ADHD subtypes. It could also be useful for evaluating the impact of parent's (potentially biased) perception of the child's symptoms on the scales results [12]. Moreover, even though this work has focused on the movement of the dominant hand, the potential of the proposed technique is much greater. The analysis of other body parts, which are not directly involved in the reaction to stimulus such as body leaning, legs movement, or body configuration, may provide other discriminative predictors that could improve ADHD characterization and provide more objective diagnoses. In addition, although the current work has focused on the body, face analysis can also be incorporated in future studies.

Acknowledgments

We would like to thank the Instituto de Salud Carlos III, cofounded by Fondo Europeo de Desarrollo Regional - FEDER (PI13/02200 grant).

Conflicts of Interest

None declared.

Multimedia Appendix 1

Video demonstrating the working of the proposed technique.

[MP4 File (MP4 Video), 21MB - jmir_v19i3e79_app1.mp4]

References

1. Polanczyk G, de Lima MS, Horta BL, Biederman J, Rohde LA. The worldwide prevalence of ADHD: a systematic review and metaregression analysis. Am J Psychiatry 2007 Jun;164(6):942-948. [doi: 10.1176/ajp.2007.164.6.942] [Medline: 17541055]
2. Diagnostic and statistical manual of mental disorders 5th ed. Washington DC: American Psychiatric Association; 2013.
3. Washbrook E, Propper C, Sayal K. Pre-school hyperactivity/attention problems and educational outcomes in adolescence: prospective longitudinal study. Br J Psychiatry 2013 Sep;203(3):265-271 [FREE Full text] [doi: 10.1192/bjp.bp.112.123562] [Medline: 23969481]
4. Polderman TJ, Boomsma DI, Bartels M, Verhulst FC, Huizink AC. A systematic review of prospective studies on attention problems and academic achievement. Acta Psychiatr Scand 2010 Oct;122(4):271-284. [doi: 10.1111/j.1600-0447.2010.01568.x] [Medline: 20491715]
5. Biederman J, Faraone SV. Attention-deficit hyperactivity disorder. Lancet 2005;366(9481):237-248. [doi: 10.1016/S0140-6736(05)66915-2] [Medline: 16023516]
6. Taylor E, Sandberg G, Thorley G, Giles S. The epidemiology of childhood hyperactivity. Psychol Med 1992;22(4):162-170. [doi: 10.1017/S0033291700038691]
7. Swanson JM, Schuck S, Porter MM, Carlson C, Hartman CA, Sergeant JA, et al. Categorical and dimensional definitions and evaluations of symptoms of ADHD: history of the SNAP and the SWAN rating scales. Int J Educ Psychol Assess 2012 Apr;10(1):51-70 [FREE Full text] [Medline: 26504617]
8. Zhang S, Faries DE, Vowles M, Michelson D. ADHD Rating Scale IV: psychometric properties from a multinational study as a clinician-administered instrument. Int J Methods Psychiatr Res 2005;14(4):186-201. [Medline: 16395872]
9. Marshall P, Schroeder R, O'Brien J, Fischer R, Ries A, Blesi B, et al. Effectiveness of symptom validity measures in identifying cognitive and behavioral symptom exaggeration in adult attention deficit hyperactivity disorder. Clin Neuropsychol 2010 Oct;24(7):1204-1237. [doi: 10.1080/13854046.2010.514290] [Medline: 20845231]
10. Frazier TW, Frazier AR, Busch RM, Kerwood MA, Demaree HA. Detection of simulated ADHD and reading disorder using symptom validity measures. Arch Clin Neuropsychol 2008 Sep;23(5):501-509 [FREE Full text] [doi: 10.1016/j.acn.2008.04.001] [Medline: 18534813]
11. Harrison AG. Adults faking ADHD: you must be kidding!. ADHD Rep 2006 Aug;14(4):1-7. [doi: 10.1521/adhd.2006.14.4.1]
12. Middeldorp CM, Wesseldijk LW, Hudziak JJ, Verhulst FC, Lindauer RJ, Dieleman GC. Parents of children with psychopathology: psychiatric problems and the association with their child's problems. Eur Child Adolesc Psychiatry 2016 Aug;25(8):919-927 [FREE Full text] [doi: 10.1007/s00787-015-0813-2] [Medline: 26757722]
13. Chamberlain SR, Sahakian BJ. The neuropsychiatry of impulsivity. Curr Opin Psychiatry 2007 May;20(3):255-261. [doi: 10.1097/YCO.0b013e3280ba4989] [Medline: 17415079]
14. Flaskerud JH. Cultural bias and Likert-type scales revisited. Issues Ment Health Nurs 2012 Feb;33(2):130-132. [doi: 10.3109/01612840.2011.600510] [Medline: 22273348]
15. Riccio CA, Reynolds CR, Lowe PA. Clinical applications of continuous performance tests: Measuring attention and impulsive responding in children and adults. New York: Wiley; 2001.
16. Willcutt EG, Doyle AE, Nigg JT, Faraone SV, Pennington BF. Validity of the executive function theory of attention-deficit/hyperactivity disorder: a meta-analytic review. Biol Psychiatry 2005 Jun 01;57(11):1336-1346. [doi: 10.1016/j.biopsych.2005.02.006] [Medline: 15950006]
17. Conners M, Staff V, Connelly S, Campbell M, MacLean M, Banes J. Multi-Health Syst Inc. 2000. Conners continuous performance Test II (CPT II v. 5) URL: http://www.hogrefe.se/Global/Exempelrapporter/cpt-ii-v5-progress%20SAMPLE.pdf [accessed 2017-03-09] [WebCite Cache ID 6opht8htI]
18. McGee RA, Clark SE, Symons DK. Does the Conners' continuous performance test aid in ADHD diagnosis? J Abnorm Child Psychol 2000 Oct;28(5):415-424. [Medline: 11100916]

19. Tsujii N, Okada A, Kaku R, Kuriki N, Hanada K, Shirakawa O. Differentiation between attention-deficit/hyperactivity disorder and pervasive developmental disorders with hyperactivity on objective activity levels using actigraphs. Psychiatry Clin Neurosci 2009 Jun;63(3):336-343 [FREE Full text] [doi: 10.1111/j.1440-1819.2009.01948.x] [Medline: 19566765]
20. Licht CA, Tryon WW. Are children diagnosed with the combined form of ADHD pervasively hyperactive? Behav Modif 2009 Sep;33(5):655-681. [doi: 10.1177/0145445509344167] [Medline: 19864320]
21. Gilbert H, Qin L, Li D, Zhang X, Johnstone SJ. Aiding the diagnosis of AD/HD in childhood: using actigraphy and a continuous performance test to objectively quantify symptoms. Res Dev Disabil 2016 Dec;59:35-42. [doi: 10.1016/j.ridd.2016.07.013] [Medline: 27497372]
22. Lis S, Baer N, Stein-en-Nosse C, Gallhofer B, Sammer G, Kirsch P. Objective measurement of motor activity during cognitive performance in adults with attention-deficit/hyperactivity disorder. Acta Psychiatr Scand 2010 Oct;122(4):285-294. [doi: 10.1111/j.1600-0447.2010.01549.x] [Medline: 20199487]
23. Rapport MD, Bolden J, Kofler MJ, Sarver DE, Raiker JS, Alderson RM. Hyperactivity in boys with attention-deficit/hyperactivity disorder (ADHD): a ubiquitous core symptom or manifestation of working memory deficits? J Abnorm Child Psychol 2009 May;37(4):521-534. [doi: 10.1007/s10802-008-9287-8] [Medline: 19083090]
24. Reh V, Schmidt M, Rief W, Christiansen H. Preliminary evidence for altered motion tracking-based hyperactivity in ADHD siblings. Behav Brain Funct 2014 Mar 13;10:7 [FREE Full text] [doi: 10.1186/1744-9081-10-7] [Medline: 24624931]
25. Li F, Zheng Y, Smith SD, Shic F, Moore CC, Zheng X, et al. A preliminary study of movement intensity during a Go/No-Go task and its association with ADHD outcomes and symptom severity. Child Adolesc Psychiatry Ment Health 2016;10:47 [FREE Full text] [doi: 10.1186/s13034-016-0135-2] [Medline: 27999615]
26. Qiu J, Helbig R. Body posture as an indicator of workload in mental work. Hum Factors 2012 Aug;54(4):626-635. [doi: 10.1177/0018720812437275] [Medline: 22908685]
27. Yu X, Wu L, Liu Q, Zhou H. Children tantrum behaviour analysis based on Kinect sensor. 2011 Presented at: Third Chinese Conference on Intelligent Visual Surveillance (IVS); 2011; China.
28. Stanley D. Scholarworks.rit. Rochester: Rochester Institute of Technology; 2013. Measuring attention using Microsoft Kinect URL: http://scholarworks.rit.edu/theses/4768/ [accessed 2017-03-09] [WebCite Cache ID 6opiACwzs]
29. Delgado-Gómez D, Carmona-Vázquez C, Bayona S, Ardoy-Cuadros J, Aguado D, Baca-García E, et al. Improving impulsivity assessment using movement recognition: a pilot study. Behav Res Methods 2016 Dec;48(4):1575-1579. [doi: 10.3758/s13428-015-0668-y] [Medline: 26487050]
30. Berrouiguet S, Courtet P, Perez-Rodriguez MM, Oquendo M, Baca-Garcia E. The Memind project: a new web-based mental health tracker designed for clinical management and research. Eur Psychiatry 2015;30(1):28-31. [doi: 10.1016/S0924-9338(15)31968-4]
31. Epstein JN, Erkanli A, Conners CK, Klaric J, Costello JE, Angold A. Relations between Continuous Performance Test performance measures and ADHD behaviors. J Abnorm Child Psychol 2003 Oct;31(5):543-554. [Medline: 14561061]

Abbreviations

ADHD: attention deficit hyperactivity disorder
CPT: continuous performance test
RCT: reaction time
SWAN: Strengths and Weaknesses of ADHD Symptoms and Normal Behavior

Edited by R Calvo; submitted 13.11.16; peer-reviewed by D Baron, WB Lee; comments to author 15.12.16; revised version received 27.01.17; accepted 06.02.17; published 20.03.17.

Please cite as:
Delgado-Gomez D, Peñuelas-Calvo I, Masó-Besga AE, Vallejo-Oñate S, Baltasar Tello I, Arrua Duarte E, Vera Varela MC, Carballo J, Baca-García E
Microsoft Kinect-based Continuous Performance Test: An Objective Attention Deficit Hyperactivity Disorder Assessment
J Med Internet Res 2017;19(3):e79
URL: http://www.jmir.org/2017/3/e79/
doi:10.2196/jmir.6985
PMID:28320691

Original Paper

Longitudinal Changes in Psychological States in Online Health Community Members: Understanding the Long-Term Effects of Participating in an Online Depression Community

Albert Park[1], PhD; Mike Conway[1], PhD

Department of Biomedical Informatics, School of Medicine, University of Utah, Salt Lake City, UT, United States

Corresponding Author:
Albert Park, PhD
Department of Biomedical Informatics
School of Medicine
University of Utah
421 Wakara Way #140
Salt Lake City, UT, 84108-3514
United States
Phone: 1 206 743 7843
Fax: 1 801 581 4297
Email: alpark1216@gmail.com

Abstract

Background: Major depression is a serious challenge at both the individual and population levels. Although online health communities have shown the potential to reduce the symptoms of depression, *emotional contagion theory* suggests that negative emotion can spread within a community, and prolonged interactions with other depressed individuals has potential to worsen the symptoms of depression.

Objective: The goals of our study were to investigate longitudinal changes in psychological states that are manifested through linguistic changes in depression community members who are interacting with other depressed individuals.

Methods: We examined emotion-related language usages using the Linguistic Inquiry and Word Count (LIWC) program for each member of a depression community from Reddit. To measure the changes, we applied linear least-squares regression to the LIWC scores against the interaction sequence for each member. We measured the differences in linguistic changes against three online health communities focusing on positive emotion, diabetes, and irritable bowel syndrome.

Results: On average, members of an online depression community showed improvement in 9 of 10 prespecified linguistic dimensions: “positive emotion,” “negative emotion,” “anxiety,” “anger,” “sadness,” “first person singular,” “negation,” “swear words,” and “death.” Moreover, these members improved either significantly or at least as much as members of other online health communities.

Conclusions: We provide new insights into the impact of prolonged participation in an online depression community and highlight the positive emotion change in members. The findings of this study should be interpreted with caution, because participating in an online depression community is not the sole factor for improvement or worsening of depressive symptoms. Still, the consistent statistical results including comparative analyses with different communities could indicate that the emotion-related language usage of depression community members are improving either significantly or at least as much as members of other online communities. On the basis of these findings, we contribute practical suggestions for designing online depression communities to enhance psychosocial benefit gains for members. We consider these results to be an important step toward a better understanding of the impact of prolonged participation in an online depression community, in addition to providing insights into the long-term psychosocial well-being of members.

(J Med Internet Res 2017;19(3):e71) doi:10.2196/jmir.6826

KEYWORDS

mental health; depression; consumer health information; informatics; information science; social support; psychosocial support system; community networks; self-help groups; communications media

Introduction

Social media, including online health communities, has become a popular platform for individuals to exchange social support and connect with others [1]. Research on the benefits of online health communities highlights psychosocial benefits—such as reduced depression [2,3], anxiety [3,4], stress [3], and negative mood [5]—as well as the development of a higher level of coping [6] and patient empowerment [2,6,7]. However, participants in these studies suffered from mental disorders (eg, depression, anxiety) as secondary symptoms and were contributing to health-related forums focusing on, for example, cancer. Although psychosocial benefits of online health communities have been widely studied, it is unclear whether the members of mental health communities would receive similar benefits. Thus, we studied the longitudinal changes in psychological states in members of a mental health community.

Online mental health communities can be a useful medium for individuals suffering from mental disorders [8], due to social stigma and discrimination surrounding such disorders [9-11]. However, in contrast to the substantial work focusing on the benefits of online health communities, *emotional contagion theory* [12] suggests that prolonged interactions with depressed individuals and their negative emotion can worsen the symptoms of depression. *Emotional contagion theory* is a type of emotional influence that describes the spread of one person's emotion to other people during social encounters.

Scientific evidence in support of *emotional contagion theory* in online health communities can be illustrated in two concise assertions. First, individuals' inner emotions have been shown to manifest in their choice of words in writing [13], including textual conversation within online health communities [14]. Second, both negative [15] and positive emotions [16] have been shown to spread even through text-based computer-mediated communication (eg, email, chat) as well as through a textual alert system in social media [17].

In the aforementioned study of spreading positive or negative emotion via a textual alert system in social media [17], findings suggest that individuals' emotions had spread throughout their social community. Moreover, both short-lived moods evinced in response to outside stimuli and long-term emotional states (ie, general emotional state) of community members can affect the emotions of the community. For example, in an empirical study, members of mental health communities were shown to have significant increases in anxiety, anger, and negative emotions following reports of several celebrity suicides [18]; and both positive and negative long-term emotional states slowly spread across socially connected individuals in a social network [19]. Due to this effect, researchers have suggested that interaction with other depressed individuals is a potential risk factor for depression [20,21].

Negative sentiment has been shown to be highly prevalent and more common than positive sentiment in mental health communities, including a depression community [22]. Thus, members of depression communities are likely to be exposed to other members' negative emotions. However, as members of online health communities, they are also likely to receive emotional support (ie, positive emotion) [23,24]. Thus, the extent to which one of the effects (ie, spreading of negative emotion or positive emotional support) has a greater impact on long-term participants in an online depression community is of interest. Moreover, understanding the impact of long-term participation in online mental health communities can suggest practical implications related to using, managing, designing, and researching online mental health communities that can improve the experience of online mental health community members who seek essential social support during times of difficulty. Despite the importance in clinical, practical, and public policy implications for mental health, to our knowledge, the investigation on the change of emotional states via linguistic change in depression community members has not been the focus of previous research on online mental health communities.

We aim to fill this gap in the literature with this study and answer two research questions (RQ):

RQ1. How does prolonged participation in an online depression community affect the emotional states of its members?

RQ2. How does this change in emotional states compare with other online health communities? As a control, we selected online health communities focusing on (1) positive emotion, (2) health conditions recognized for their needs for both informational and emotional support, and (3) chronic but non–life-threatening health issues that are not often related to mental disorders.

Methods

Community Platform

The data for this study consist of posts (ie, a submission that starts a conversation) and associated comments (ie, a submission that replies to posts or other comments) from several topical focused subreddits (ie, subcommunities). All subreddits found in this study are hosted on the popular social media platform, Reddit. Reddit is a social media platform that had 83 billion page views from more than 88,000 active subcommunities (subreddits) in 2015. Reddit content is accessible on the open Web. This can be contrasted with services like Facebook or specifically health-focused online community like PatientsLikeMe, where data are typically not available on the open Web, but requires user registration to access content. Members of Reddit made more than 73 million individual posts with over 725 million associated comments in the same year [25]. Reddit is highly active and the community allows posters to create submissions using throwaway and unidentifiable accounts. These features are suitable for stigmatized conditions and their discussions that may not be appropriate for identifiable accounts. Due to these two reasons, we examined posts and comments from Reddit for this study.

Subreddit Selection

The community r/depression, to our knowledge, is the largest and most active depression focused subreddit. It has been active for 7 years with 145,821 subscribers [26] in September 2016. Thus, we selected r/depression as the main community of interest for investigating the effects of prolonged participation

in an online depression community to the emotional states of members (RQ1).

To understand the significance of the change of emotional states of r/depression members, we compared r/depression against 3 other subreddits (RQ2). We first selected r/happy, a subreddit focused on sharing happy thoughts and stories. r/happy was created to focus on happy thoughts and, to our knowledge, is the largest and most active positive emotion–focused subreddit that has been active for 8 years with 94,076 subscribers [27] in September 2016. Members who are participating in the positive emotion–focused subreddit should encounter less negative emotion than the r/depression members. Thus, such a community can provide an insight into the impact of prolonged participation in an online health community without much exposure to negative emotions.

Much research on online health communities focused on two types of social support emotional and informational support [24,28-30]. Both support types are important aspects of online health communities. For instance, symptom management is an important aspect for patients with diabetes. Thus, sharing informational support, such as practical information to control one's diabetic symptoms [29] is a common practice in online diabetes communities. Also in online diabetes communities, sharing emotional support such as motivation and accountability [31] is a common practice to help members continuously manage their illnesses. To better understand the changes of emotional states in r/depression, we compared against r/diabetes, a community with abundant informational and emotional support. To our knowledge, r/diabetes is the largest and most active diabetes community on Reddit. The subreddit has been active for 8 years with 15,623 subscribers in September 2016 [32].

To better understand the effects of prolonged participation in online depression communities, we wanted to compare the depression community against online health communities that are not directly related to mental disorders in order to ensure that the changes of emotional states do not result from the secondary symptoms (ie, mental illnesses) that are associated with the primary illness. However, a confounding factor is that serious illnesses are frequently accompanied with mental illnesses, due to the distress of living with—or being diagnosed with—a serious condition. Although it may be impossible to select an online health community without any exchange of emotional support, one study investigating an online Irritable Bowel Syndrome (IBS) community characterized informational support including symptom interpretation and illness management as the main type of social support exchanged in the community [33]. Thus, we selected r/ibs, a subreddit focused on IBS, as a community focused on a chronic but non–life-threatening health issue that is not often related to mental disorders. In September 2016, r/ibs has been active for 6 years with 5251 subscribers [34] and is the largest and most active IBS community in Reddit, to our knowledge.

Data

First, we use a dataset [35] released by a Reddit member who collected the data from October 2007 to May 2015. The dataset has been used in a previous study [22]. Second, we extracted posts and associated comments from r/depression, r/happy, r/diabetes, and r/ibs. Third, we removed deleted posts or comments that are labeled as "[deleted]" from the dataset. The contents of posts or comments were deleted by members or the communities, thus we excluded them in our analyses. Fourth, to meet our study aims, we restricted our analysis to members (ie, unique user IDs) with 4 or more submissions (ie, posts and comments). In different studies [28,36], this threshold was used to determine lurkers, who are not yet a regularly contributing member of a community, thus we used the same threshold to determine regular members. Table 1 shows the overall characteristics of datasets that we examined.

Table 1. Characteristics of the analyzed datasets from r/depression, r/happy, r/diabetes, and r/ibs

Community	r/depression	r/happy	r/diabetes	r/ibs
Date of data	Dec 2008-May 2015	Jan 2008-May 2015	Jan 2009-May 2015	Aug 2012-May 2015
No. of Submissions	571,864	95,329	167,917	15,978
No. of Members	36,746	8130	4539	923

The work described in this paper was granted an exemption from review by the University of Utah's Institutional Review Board (IRB), ethics committee, [IRB 00076188] under Exemption 2 as defined in United States Federal Regulations 45 CFR 46.101(b).

Linguistic Analysis

To understand the change of emotional states, we examined emotion-related language using the Linguistic Inquiry and Word Count (LIWC) program [37]. LIWC analyzes text on a word-by-word basis, counts the predefined linguistic dimension words and then returns each linguistic dimension score as a proportion of the total number of words under analysis. The validity of LIWC's performance has been established in a large number of studies, including studies measuring perceived positive and negative emotions in online mental health communities [18], online health communities [30], text-based online communication [15,16], and social media [17,38].

LIWC consists of 4 major dimensions: "psychological processes," "linguistic processes," "personal concerns," and "spoken categories." The most notable linguistic dimension related to emotion is "psychological processes," which includes "positive emotion" and "negative emotion" as subdimensions that are hosted in "affective processes." "Negative emotion" has 3 subdimensions: "anxiety," "anger," and "sadness." Both "positive emotion" and "negative emotion" dimensions have been used in a series of studies to measure perceived positive and negative emotion in text-based communication [15-18,30,38,39], which Reddit uses to facilitate communication.

From "linguistic processes," "negation," and "swear words" have been used to predict depression in social media [38,39]. "Negation" has also been used to measure agreeableness in text-based online communication [15,16]. One of the most widely used dimensions for predicting depression is the "first person singular" [15,16,18,30,38,39]. The frequent use of "first person singular" indicates a heightened sense of self and is known to correlate with depression [40].

The subdimension, "death" from "personal concerns" has been used to measure suicidal thoughts and discussion in Reddit [18] and "assent" from "spoken categories" was used to measure agreeableness in text-based online communication [15,16] and to predict depression [39]. We made these linguistic dimension selections before the linguistic analysis without any information on the linguistic dimension frequency counts. LIWC was applied to individual posts or comments for each subreddit. A summary of the linguistic dimensions and dimensions' example words for the current study is shown in Table 2.

Table 2. Summary of Linguistic Inquiry and Word Count (LIWC) linguistic dimensions used for this study and example vocabulary.

Linguistic dimensions		Example vocabulary
Psychological processes		
	Positive emotion	happy, love, sweet
	Negative emotion	agony, hurt, nasty
	Anxiety	afraid, fearful, nervous
	Anger	abuse, annoyed, hate
	Sadness	crying, grief, sad
Linguistic processes		
	First person singular	I, me, mine
	Negation	no, not, never
	Swear words	ass, damn, fuck
Personal concerns		
	Death	death, die, kill
Spoken categories		
	Assent	agree, okay, yes

RQ1: Change of Emotional States Analysis and RQ2: Community Comparison Analysis

To measure the change of emotional states of r/depression members (RQ1), we first calculated the scores of prespecified LIWC dimensions. Then, we organized each LIWC score according to the posting time per-member basis for the community. To measure the changes of LIWC scores, we applied linear least-squares regression to the LIWC scores against the interaction sequence (ie, determined by the time of posting) for each member. We elected to perform linear the least-squares regression against the interaction sequence rather than time, because we were interested in the change caused by each interaction rather than time. For each dimension, we calculated mean, median, and 95% confidence interval (CI) of the community to reflect the overall linguistic changes in r/depression members. Then, we converted linguistic change to indicate improvement. Decrease of negative dimensions (ie, "negative emotion," "anxiety," "anger," "sadness," "first person singular," "negation," "swear words," "death") indicates improvement, although decrease of positive dimensions (ie, "positive emotion," "assent") indicates worsening. To be consistent with figures (Figures 1 and 2), we show improvement for each dimension. Then, we present an overview of improvement for each linguistic dimension along with the statistical significance against the null hypothesis (ie, prolonged participation in an online depression community has no effect on the emotional states of members).

To compare linguistic changes in communities (RQ2), we first carried out the procedure for measuring the change of emotional states for the other three communities. Then, we applied Mann-Whitney *U* tests [41] with the prespecified false discovery rate procedure [42] to *P* values. The comparison of 10 linguistic dimensions raises the need to control for Type I errors (ie, false positives); thus, we applied the false discovery rate procedure. Given the large sample of members, we followed suggestions of a previous study [43] and reported 95% CI and effect sizes (r) for the community comparisons using rank-biserial correlation [44]. We applied nonparametric tests because our data were not normally distributed.

Figure 1. An overview of improvement of each linguistic dimension for r/depression members.

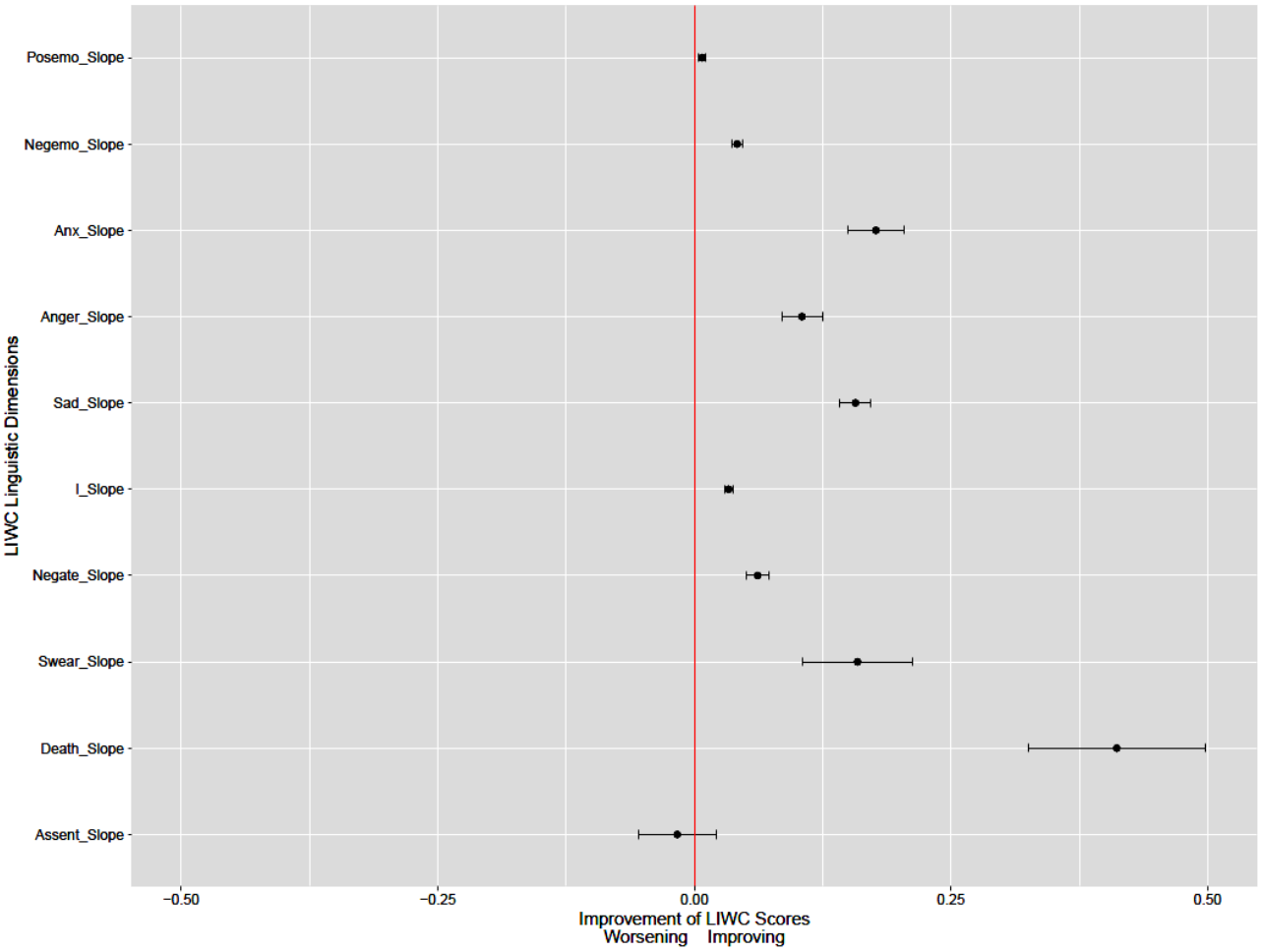
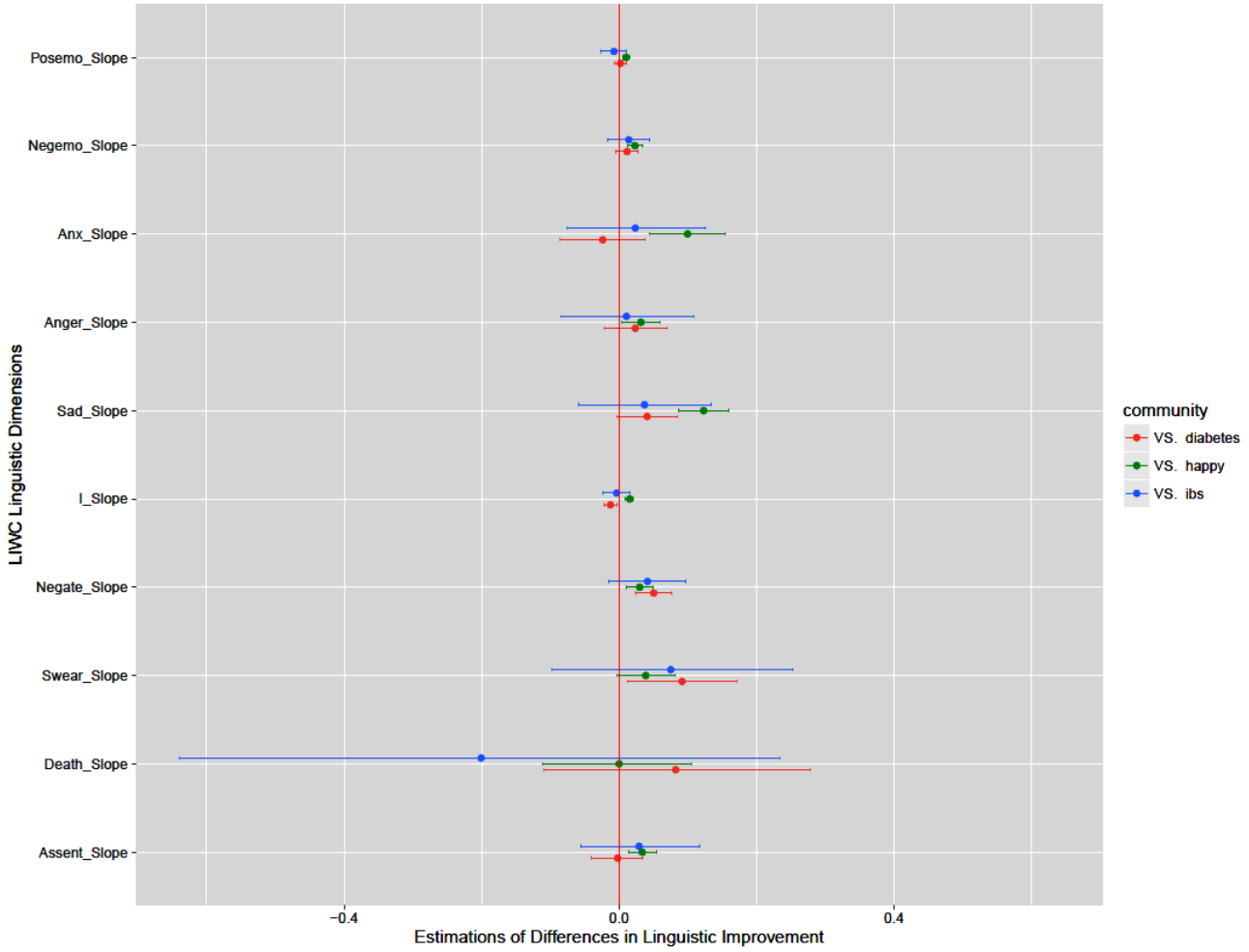


Figure 2. An overview of difference in improvement compared with r/depression.



Results

RQ 1. How Does Prolonged Participation in an Online Depression Community Affect the Emotional States of its Members?

We captured the average changes of emotional states that are manifest in linguistic changes for r/depression members. Table 3 summarizes the linguistic changes in r/depression members. On average, r/depression members showed improvement in all linguistic dimensions with the exception of “assent.”

Table 3. Linguistic changes in r/depression members. Variables are reported as the mean, median, and 95% CIs of word usage change of each dimension for all members.

Linguistic dimensions		r/depression mean, median (CI) n=36,746
Psychological processes		
	Positive emotion	0.01, 0.01 (0.004 to 0.01)
	Negative emotion	-0.04, -0.03 (-0.05 to -0.04)
	Anxiety	-0.18, -0.11 (-0.20 to -0.15)
	Anger	-0.10, -0.06 (-0.12 to -0.09)
	Sadness	-0.16, -0.13 (-0.17 to -0.14)
Linguistic processes		
	First person singular	-0.03, -0.03 (-0.04 to -0.03)
	Negation	-0.06, -0.05 (-0.07 to -0.05)
	Swear words	-0.16, 0.09 (-0.20 to -0.12)
Personal concerns		
	Death	-0.41, -0.21 (-0.47 to -0.35)
Spoken categories		
	Assent	-0.02, 0.05 (-0.05 to 0.02)

Then, we converted linguistic changes to indicate improvement and showed an overview of improvement of each linguistic dimension (Figure 1). On the x-axis, less than 0 indicates worsening (ie, decreasing use of positive words or increasing use of negative words), whereas more than 0 indicates an improvement (ie, increasing use of positive words or decreasing use of negative words). The dot indicates the mean of improvement values and the range shows 95% CI. If the range includes 0.0 (eg, "assent"), it suggests that the null hypothesis of prolonged participation in an online depression community has no effect on the emotional states of members cannot be rejected. Our analyses suggested that with the exception of "assent" dimension, members of r/depression showed significant improvement in linguistic changes. No significant change was found in the "assent" dimension.

RQ 2. How Does This Change in Emotional States Compare With Other Online Communities?

To compare the linguistic changes evident in r/depression to r/happy, r/diabetes, and r/ibs, we first calculated the mean, median, and CIs of word usage change for each linguistic dimension for each subreddit (Table 4).

Table 4. Linguistic changes in members from r/happy, r/ibs, and r/diabetes. Variables are reported as the mean, median, and 95% CIs of word usage change of each dimension for all members.

Linguistic dimensions		r/happy mean, median (CI) n=8130	r/diabetes mean, median (CI) n=4539	r/ibs mean, median (CI) n=923
Psychological processes				
	Positive emotion	0.005, 0.004 (0.0005-0.01)	0.01, 0.01 (-0.01 to 0.02)	0.02, 0.02 (-0.005 to 0.05)
	Negative emotion	-0.01, -0.01 (-0.03 to 0.001)	-0.02, -0.02 (-0.04 to 0.003)	-0.005, -0.02 (-0.05 to 0.03)
	Anxiety	0.002, -0.003 (-0.06 to 0.06)	-0.26, -0.15 (-0.35 to -0.17)	-0.18, -0.08 (-0.31 to -0.04)
	Anger	-0.02, -0.02 (-0.06 to 0.02)	-0.11, -0.04 (-0.18 to -0.03)	-0.13, -0.06 (-0.27 to 0.01)
	Sadness	0.01, -0.003 (-0.04 to 0.07)	-0.18, -0.10 (-0.25 to -0.10)	-0.15, -0.11 (-0.33 to 0.04)
Linguistic processes				
	First person singular	-0.01, -0.01 (-0.02 to -0.01)	-0.09, -0.04 (-0.11 to -0.07)	-0.03, -0.03 (-0.06 to -0.008)
	Negation	-0.02, -0.01 (-0.04 to 0.01)	0.02, 0.002 (-0.02 to 0.05)	-0.008, -0.01 (-0.08 to 0.06)
	Swear words	-0.03, -0.02 (-0.09 to 0.03)	-0.01, 0.005 (-0.20 to 0.18)	-0.06, -0.01 (0.44 to 0.33)
Personal concerns				
	Death	-0.52, -0.17 (-0.69 to -0.36)	-0.93, -0.22 (-1.35 to -0.51)	-0.98, -0.56 (-1.73 to -0.23)
Spoken categories				
	Assent	0.02, 0.01 (-0.01 to 0.05)	-0.03, 0.05 (-0.10 to 0.05)	0.04, 0.02 (-0.09 to 0.18)

On average, members of these 3 subreddits showed significant improvement in linguistic changes with increasing number of interactions. Four exceptions are “anxiety” and “sadness” from r/happy and “negation” and “assent” from r/diabetes. Then, we compared the 3 subreddits to r/depression (Table 5).

Table 5. Community comparison of average linguistic changes against r/depression.

Linguistic dimensions		versus r/happy	versus r/diabetes	versus r/ibs
Psychological processes				
	Positive emotion	r/depression	r/depression	r/ibs
		U[a]=151428562	U[a] =83095368.5	U[a] =16588110
		P=3.9e-05	*P*=.63	*P*=.43
		P_{adj}[b]=9.8e-05	P_{adj}[b] =.70	P_{adj}[b] =.70
		CI[c]=0.005-0.015	CI[c] =-0.01 to 0.01	CI[c] =-0.03 to 0.01
		r=0.04	*r*=0.09	*r*=0.02
	Negative emotion	r/depression	r/depression	r/depression
		U[a] =118117817	U[a] =77894854	U[a] =16023844.5
		P=2.7e-05	*P*=.16	*P*=.37
		P_{adj}[b] =8.9 e-05	P_{adj}[b] =.31	P_{adj}[b] =.70
		CI[c] =-0.03 to -0.01	CI[c] =-0.03 to 0.004	CI[c] =-0.05 to 0.02
		r=0.21	*r*=0.14	*r*=0.05
	Anxiety	r/depression	r/depression	r/depression
		U[a] =42155753	U[a] = 49336254	U[a] =11317589.5
		P=3.1e-04	*P*=.44	*P*=.62
		P_{adj}[b] =6.2e-04	P_{adj}[b] =.55	P_{adj}[b] =.78
		CI[c] =-0.15 to -0.05	CI[c] =-0.04 to 0.09	CI[c] =-0.13 to 0.08
		r=0.72	*r*=0.47	*r*=0.33
	Anger	r/depression	r/depression	r/depression
		U[a] =72410492.5	U[a] =54287645.5	U[a] =11059329.5
		P=.02	*P*=.30	*P*=.82
		P_{adj}[b] =.03	P_{adj}[b] =.50	P_{adj}[b] =.82
		CI[c] =-0.06 to -0.004	CI[c] =-0.07 to 0.02	CI[c] =-0.11 to 0.08
		r=0.52	*r*=0.40	*r*=0.35
	Sadness	r/depression	r/depression	r/depression
		U[a] =54288775	U[a] =59564322.5	U[a] =11862037.5
		P=1.8e-11	*P*=.07	*P*=.45
		P_{adj}[b] =1.8e-10	P_{adj}[b] =.17	P_{adj}[b] =.70
		CI[c] =-0.16 to -0.09	CI[c] =-0.08 to 0.002	CI[c] =-0.13 to 0.06
		r=0.64	*r*=0.34	*r*=0.3
Linguistic processes				
	First person singular	r/depression	r/diabetes[d]	r/ibs
		U[a] =137694516.5	U[a] =84591685.5	U[a] =16945136
		P=5.6e-07	*P*=.01	*P*=.72
		P_{adj}[b] =2.8e-06	P_{adj}[b] =.03	P_{adj}[b] =.80
		CI[c] =-0.02 to -0.009	CI[c] =0.004 to 0.02	CI[c] =-0.02 to 0.02
		r=0.08	*r*=0.11	*r*=0.01

Linguistic dimensions		versus r/happy	versus r/diabetes	versus r/ibs
	Negation	r/depression U[a] =96382143.5 *P*=.003 P_{adj}[b] =4.7e-03 CI[c] =-0.05 to 0.01 *r*=0.35	r/depression[e] U[a] =71632553 *P*=.0002 P_{adj}[b] =.002 CI[c] =-0.08 to -0.02 *r*=0.21	r/depression U[a] =14855011 *P*=.14 P_{adj}[b] =.70 CI[c] =-0.10 to 0.01 *r*=0.12
	Swear words	r/depression U[a] =40086953 *P*=.08 P_{adj}[b] =.09 CI[c] =-0.08 to 0.003 *r*=0.73	r/depression U[a] =30918210.5 *P*=.02 P_{adj}[b] =.07 CI[c] =-0.17 to -0.01 *r*=0.66	r/ibs U[a] =5505668 *P*=.40 P_{adj}[b] =.70 CI[c] =-0.25 to 0.10 *r*=0.67
Personal concerns				
	Death	r/happy U[a] =14379783 *P*=.95 P_{adj}[b] =.95 CI[c] =-0.11 to 0.11 *r*=0.9	r/depression U[a] =15714429 *P*=.40 P_{adj}[b] =.55 CI[c] =-0.28 to 0.11 *r*=0.83	r/ibs U[a] =2476626 *P*=.37 P_{adj}[b] =.70 CI[c] =-0.23 to 0.64 *r*=0.86
Spoken categories				
	Assent	r/depression U[a] =70286713.5 *P*=.001 P_{adj}[b] =.002 CI[c] =0.01-0.05 *r*=0.56	r/depression U[a] =45944192 *P*=.89 P_{adj}[b] =.89 CI[c] = -0.04 to 0.03 *r*=0.49	r/depression U[a] =8873209.5 *P*=.49 P_{adj}[b] =.70 CI[c] =-0.05 to 0.12 *r*=0.49

[a]U: Mann-Whitney *U* test values.

[b]P_{adj}: adjusted *P* values.

[c]95% CI values.

Subreddit comparisons indicate that members of r/depression improved either significantly or at least as much as (ie, no significant difference) members of other online health communities. In “positive emotion,” “negative emotion,” “anxiety,” “anger,” “sadness,” “first person singular,” “negation,” and “assent,” r/depression showed significantly more improvement than r/happy. “Swear words” and “death” showed no significant difference.

The difference in improvements between r/depression and r/diabetes were not significant, with 2 exceptions. The subreddit r/diabetes improved significantly as compared with r/depression in “first person singular,” although r/depression improved significantly as compared with r/diabetes in “negation.” There were no significant differences between r/ depression and r/ibs for all dimensions.

At this point, we converted linguistic changes to indicate improvement compared with r/depression and showed an overview of the differences with respect to each linguistic dimension (Figure 2). The dot indicates the sample estimate and the range shows 95% CI. If the range includes 0.0, it indicates that the difference between the means is not significant.

Discussion

Principal Findings on Change of Emotional States for Members of an Online Depression Community

This study has been deeply informed by recent literature focusing on the psychosocial benefits of online health communities [1-7]. We have also utilized *emotional contagion theory* [12] and empirical studies on *emotional contagion theory* [15-17,19,20] to develop this study. We address gaps in the previous literature by shedding light on the question of whether members of an online depression community experience positive or negative emotional changes after prolonged participation.

The linguistic changes have been shown to occur when emotional states are shifting [38]. In the online depression community we analyzed, we found that members of the

depression community generally showed an increased use of positive words and decreased use of negative words as the number of interactions increased. Despite prolonged interactions with other depressed individuals, r/depression members' emotional states were found to have become more positive. For r/depression members, on average all linguistic dimensions were improving with the exception of "assent." The change in usage of "assent" was found to be insignificant; and as "assent" belongs to "spoken categories," we believe many vocabularies in "assent" were used in a communicative sense, such as answering questions, a typical type of communication in online health communities [45].

Consistent with previous studies [1-7], we found a similar trend—more use of positive emotion words and less use of negative emotion words—in 3 other Reddit communities. These Reddit communities are different in size (ie, number of members), length of active years, and level of activity, while focusing on different types of health issues. Moreover, we chose these different communities on the basis of the literature that suggested a different composition of social support in these communities. Yet, in all 4 communities, we found members generally used more positive emotion words and less negative emotion words as the number of interactions increased.

The comparison of communities indicates that members of r/depression improved either significantly or at least as much as (ie, no significant difference) members of other online health communities. One unusual finding was the lack of improvement in r/happy, especially when compared with other online health communities. We believe because r/happy members were encouraged to share only positive thoughts and stories from the start, they had limited capacity for positive linguistic changes. However, further investigation is needed for a better understanding of this phenomenon.

Practical Implications for Online Community Use, Research, and Design

Many online communities utilize the efforts of moderators and even community members to monitor their communities. However, detecting changes of emotional states in members can be difficult with such manual efforts, given that the knowledge of members' (1) previous emotional states and (2) style of writing are required, as well as (3) linguistic changes can be subtle. An adaptation of our linguistic analysis method can be a basis for detecting subtle linguistic changes in individuals in massive scale networks.

Moderators of online health communities have mentally demanding tasks [46]. An automated detection of linguistic changes could reduce the mental burden of moderators and improve the quality of moderation. For example, such a system could alert moderators and members of any undesirable linguistic changes and allow them to provide timely support. Furthermore, a similar system can assist individual members and the whole community. For individual members, a similar system could be used to raise self-awareness of their changes of linguistic or emotional state. As for communities, a similar system could alert when a community-wide emotional shift occurs. Alerting of a negative shift can allow timely intervention for the whole community, whereas alerting of a positive shift provides opportunities for researchers and community managers to study the positive change. As studies on *emotional contagion theory* [15-17,19,20] demonstrated, mental health communities carry a potential for a community-wide spread of positive and negative emotions.

User Privacy

Although Reddit data (like Twitter) are public, and research using public social media is typically granted exemption from review by IRBs in the US context, ethical considerations—particularly with respect to privacy—remain [47-49]. In this paper, in order to protect user anonymity, we did not present user identifiable information (eg, direct quotations, usernames).

Limitations and Future Directions

One of the limitations in our study is that we had a dataset with 4 communities within the Reddit platform. Although Reddit is a widely used platform, it is more frequently used by young males [50,51]. Therefore, results may not be generalizable to other communities. Also, selection bias of our subjects exists. Members who choose to participate in r/depression are not necessarily representative of the depression population. Similarly, we do not have any evidence that members of r/depression are clinically diagnosed with depression. Thus, we cannot conclude that participating in an online depression community is beneficial for all depression patients. However, r/depression is worth studying given the size of r/depression, the prevalence of depressive disorder and the increasing popularity of online mental health communities.

Although beyond the scope of this study, an investigation with respect to changes of emotion in users of different systems such as Internet-based interventions for depression [52] or text-based crisis counselor system [53] is important for individuals suffering from depression. It is unlikely that any of these avenues could reach the entire population of depressed individuals [9]; thus, understanding the extent to which emotional change occurs within demographic segment in each of these different avenues is important.

Another limitation is the performance of LIWC in measuring emotional states. Although LIWC has been validated and used in many studies [15-18,30,38-40] including a study using Reddit data [18], it is important to note LIWC's limitations, such as missing potentially important misspellings and slang, which are common in online communities [54]. The findings of this study should be interpreted with caution, especially when leveraging the findings in the clinical context. Furthermore, participation in an online depression community is not the sole factor for improvement or worsening of depressive symptoms. It is also unclear why members stay or leave in this particular online depression community and how their engagement affects the results. Although a previous study suggested that emotional support (ie, positive emotion) helps members stay with an online health community [30], we cannot conclude that participating in online health communities directly caused an improvement in emotional states of members without directly asking the participants. However, this is a general limitation of any secondary analyses. Still, the consistent statistical results could

indicate that depressive symptoms of members of an online depression community were improving. Future work could bolster the current study's findings by using mixed methods such as surveys and interviews and ask the members of online depression communities about the changes in their depressive symptoms and measure depression using standardized methods like the patient health questionnaire or primary care evaluation of mental disorders [55].

We considered word usages as independent observations; nevertheless, correlations could exist in the word usages among members depending on the topic of conversation. For instance, in a conversation about death, members who are in the conversation are more likely to use words related to death regardless of their emotional states. Conversational topics in online communities, however, are known to change [56]. Thus, examining the effects of emotional states in changing topics and their associated sentiments could support our findings. Moreover, individuals could participate in multiple subreddits and use multiple user IDs, both of which violate our independent variable assumption. We also acknowledge that our large sample size could have inflated the statistical significance levels. Thus, we reported 95% CIs and effect sizes as well as compared with other online health communities to aid the interpretation of results.

Finally, our dataset did not capture passive activities like lurking. Although it has been suggested that lurkers experience fewer benefits than active participants [3], it is unclear if lurkers have less emotional changes and how their changes of emotion can affect the community when they become active participants. Understanding emotional changes in lurkers could deepen our understanding in future work.

Conclusions

We provide new insights into the impact of prolonged participation in an online depression community and highlight the positive emotion change in members. Our findings suggest that 3 other Reddit communities focusing on positive emotion, diabetes, and IBS also have a similar positive emotion change in members. The comparison of communities indicates that the emotion-related language usage of depression community members improved either significantly or at least as much as (ie, no significant difference) members of other online communities. On the basis of these findings, we contribute practical suggestions for designing online depression communities to enhance psychosocial gains for members. We consider these results to be an important step toward developing a better understanding of the impact of prolonged participation in an online depression community on emotional health, in addition to providing insights into the long-term psychosocial well-being of members.

Acknowledgments

We restricted our analysis to publicly available discussion content. The study was exempted from review by the University of Utah's Institutional Review Board (Ethics Committee) [IRB 00076188]. AP's contribution to this research was supported by the National Library of Medicine of the National Institutes of Health under training grant T15 LM007124. MC's contribution to this research was supported by the National Library of Medicine of the National Institutes of Health under award numbers R00LM011393 & K99LM011393. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. The authors thank Mr Gregory Stoddard at the University of Utah for his advice and comments on statistical methods during the preparation of this article.

Conflicts of Interest

None declared.

References

1. Fox S, Duggan M. Pew Research Center. 2013 Jan 15. Health online 2013: 35% of U.S. adults have gone online to figure out a medical condition URL: http://www.pewinternet.org/2013/01/15/health-online-2013/; [WebCite Cache ID 6Y5K9a3JH]
2. van Uden-Kraan CF, Drossaert CH, Taal E, Seydel ER, van de Laar MA. Participation in online patient support groups endorses patients' empowerment. Patient Educ Couns 2009 Jan;74(1):61-69. [doi: 10.1016/j.pec.2008.07.044] [Medline: 18778909]
3. Setoyama Y, Yamazaki Y, Namayama K. Benefits of peer support in online Japanese breast cancer communities: differences between lurkers and posters. J Med Internet Res 2011 Dec 29;13(4):e122 [FREE Full text] [doi: 10.2196/jmir.1696] [Medline: 22204869]
4. Høybye MT, Dalton S, Deltour I, Bidstrup P, Frederiksen K, Johansen C. Effect of Internet peer-support groups on psychosocial adjustment to cancer: a randomised study. Br J Cancer 2010 Apr 27;102(9):1348-1354 [FREE Full text] [doi: 10.1038/sj.bjc.6605646] [Medline: 20424614]
5. Shaw B, Hawkins R, McTavish F. Effects of insightful disclosure within computer mediated support groups on women with breast cancer. Health Commun Internet 2006;19(2):133-142 [FREE Full text] [doi: 10.1207/s15327027hc1902_5]
6. Mo PK, Coulson NS. Developing a model for online support group use, empowering processes and psychosocial outcomes for individuals living with HIV/AIDS. Psychol Health 2012;27(4):445-459. [doi: 10.1080/08870446.2011.592981] [Medline: 21854088]

7. Wentzer HS, Bygholm A. Narratives of empowerment and compliance: studies of communication in online patient support groups. Int J Med Inform 2013 Dec;82(12):e386-e394. [doi: 10.1016/j.ijmedinf.2013.01.008] [Medline: 23523082]
8. Conway M, O'Connor D. Social media, big data, and mental health: current advances and ethical implications. Curr Opin Psychol 2016 Jun;9:77-82. [doi: 10.1016/j.copsyc.2016.01.004] [Medline: 27042689]
9. Corrigan P. How stigma interferes with mental health care. Am Psychol Internet 2004;59(7):614-625 [FREE Full text] [doi: 10.1037/0003-066X.59.7.614]
10. De Choudhury M, De S. Mental health discourse on Reddit: self-disclosure, social support, and anonymity. 2014 Jun 01 Presented at: Proceedings of ICWSM, AAAI; 2014; Ann Arbor, Michigan, USA p. 71-80.
11. Barney LJ, Griffiths KM, Banfield MA. Explicit and implicit information needs of people with depression: a qualitative investigation of problems reported on an online depression support forum. BMC Psychiatry 2011 May 18;11:88 [FREE Full text] [doi: 10.1186/1471-244X-11-88] [Medline: 21592335]
12. Hatfield E, Cacioppo J, Rapson R. Emotional Contagion. Curr Dir Psychol Sci 1993 Jun;2(3):96-100 [FREE Full text] [doi: 10.1111/1467-8721.ep10770953]
13. Pennebaker JW, Mehl MR, Niederhoffer KG. Psychological aspects of natural language. use: our words, our selves. Annu Rev Psychol 2003;54:547-577. [doi: 10.1146/annurev.psych.54.101601.145041] [Medline: 12185209]
14. De Choudhury M, Kiciman E, Dredze M, Coppersmith G, Kumar M. Discovering Shifts to Suicidal Ideation from Mental Health Content in Social Media. New York, New York, USA: ACM Press; 2016 May 07 Presented at: 2016 CHI Conference on Human Factors in Computing Systems; May 07 - 12, 2016; Santa Clara, California, USA p. 2098-2110 URL: http://dl.acm.org/citation.cfm?id=2858207 [doi: 10.1145/2858036.2858207]
15. Hancock J, Gee K, Ciaccio K, Lin J. I'm sad you're sad: emotional contagion in CMC. New York, New York, USA: ACM Press; 2008 Nov 08 Presented at: 2008 ACM conference on Computer supported cooperative work; November 08 - 12, 2008; San Diego, CA, USA p. 295-298 URL: http://dl.acm.org/citation.cfm?id=1460611 [doi: 10.1145/1460563.1460611]
16. Hancock J, Landrigan C, Silver C. Expressing emotion in text-based communication. New York, New York, USA: ACM Press; 2007 Apr 28 Presented at: SIGCHI Conference on Human Factors in Computing Systems; April 28 - May 03, 2007; San Jose, California, USA p. 929-932 URL: http://portal.acm.org/citation.cfm?doid=1240624.1240764 [doi: 10.1145/1240624.1240764]
17. Kramer AD, Guillory JE, Hancock JT. Experimental evidence of massive-scale emotional contagion through social networks. Proc Natl Acad Sci U S A 2014 Jun 17;111(24):8788-8790 [FREE Full text] [doi: 10.1073/pnas.1320040111] [Medline: 24889601]
18. Kumar M, Dredze M, Coppersmith G, De Choudhury M. Detecting Changes in Suicide Content Manifested in Social Media Following Celebrity Suicides. New York, NY, USA: ACM Press; 2015 Sep 01 Presented at: 26th ACM Conference on Hypertext & Social Media; September 01 - 04, 2015; Guzelyurt, Northern Cyprus p. 85-94 URL: http://dl.acm.org/citation.cfm?id=2791026&CFID=905650047&CFTOKEN=12908080 [doi: 10.1145/2700171.2791026]
19. Hill AL, Rand DG, Nowak MA, Christakis NA. Emotions as infectious diseases in a large social network: the SISa model. Proc Biol Sci 2010 Dec 22;277(1701):3827-3835 [FREE Full text] [doi: 10.1098/rspb.2010.1217] [Medline: 20610424]
20. Rosenquist JN, Fowler JH, Christakis NA. Social network determinants of depression. Mol Psychiatry 2011 Mar;16(3):273-281 [FREE Full text] [doi: 10.1038/mp.2010.13] [Medline: 20231839]
21. Bastiampillai T, Allison S, Chan S. Is depression contagious? The importance of social networks and the implications of contagion theory. Aust N Z J Psychiatry 2013 Apr;47(4):299-303. [doi: 10.1177/0004867412471437] [Medline: 23568155]
22. Gkotsis G, Oellrich A, Hubbard T, Dobson R, Liakata M, Velupillai S. The language of mental health problems in social media. : Association for Computational Linguistics; 2016 Presented at: Proceedings of the 3rd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality; June 16, 2016; San Diego, California p. 63-73 URL: https://www.researchgate.net/profile/George_Gkotsis/publication/303365632_The_language_of_mental_health_problems_in_social_media/links/573e2d0408ae9f741b3009ff.pdf
23. Griffiths KM, Calear AL, Banfield M. Systematic review on Internet Support Groups (ISGs) and depression (1): do ISGs reduce depressive symptoms? J Med Internet Res 2009 Sep 30;11(3):e40 [FREE Full text] [doi: 10.2196/jmir.1270] [Medline: 19793719]
24. Hwang KO, Ottenbacher AJ, Green AP, Cannon-Diehl MR, Richardson O, Bernstam EV, et al. Social support in an Internet weight loss community. Int J Med Inform 2010 Jan;79(1):5-13 [FREE Full text] [doi: 10.1016/j.ijmedinf.2009.10.003] [Medline: 19945338]
25. Reddit. 2015. Reddit in 2015 Internet URL: http://www.redditblog.com/2015/12/reddit-in-2015.html [accessed 2016-01-12] [WebCite Cache ID 6eTHN0TFD]
26. Reddit. 2016. r/depression in Sept 2016 URL: https://www.reddit.com/r/depression/search?q=depression [accessed 2016-09-21] [WebCite Cache ID 6kg9g2fCa]
27. Reddit. 2016. r/happy in Sept 2016 URL: https://www.reddit.com/search?q=happy [accessed 2016-09-25] [WebCite Cache ID 6kne12XlJ]
28. Park A, Hartzler AL, Huh J, McDonald DW, Pratt W. Homophily of vocabulary usage: beneficial effects of vocabulary similarity on online health communities participation. AMIA Annu Symp Proc 2015;2015:1024-1033 [FREE Full text] [Medline: 26958240]

29. Huh J, Ackerman M. Collaborative help in chronic disease management: supporting individualized problems. New York, NY, USA: ACM Press; 2012 Feb 11 Presented at: ACM 2012 conference on Computer Supported Cooperative Work; February 11 - 15, 2012; Seattle, Washington, USA p. 853-862 URL: http://dl.acm.org/citation.cfm?id=2145331 [doi: 10.1145/2145204.2145331]
30. Wang Y, Kraut R, Levine J. To stay or leave? The relationship of emotional and informational support to commitment in online health support groups. New York, NY, USA: ACM Press; 2012 Feb 11 Presented at: ACM 2012 conference on Computer Supported Cooperative Work; February 11 - 15, 2012; Seattle, Washington, USA p. 833-842 URL: http://dl.acm.org/citation.cfm?id=2145329&CFID=905650047&CFTOKEN=12908080 [doi: 10.1145/2145204.2145329]
31. Newman M, Lauterbach D, Munson S, Resnick P, Morris M. It's not that i don't have problems, i'm just not putting them on facebook: challenges and opportunities in using online social networks for health. New York, New York, USA: ACM Press; 2011 Mar 19 Presented at: ACM 2011 conference on Computer supported cooperative work; March 19 - 23, 2011; Hangzhou, China p. 341-350 URL: http://dl.acm.org/citation.cfm?id=1958824.1958876 [doi: 10.1145/1958824.1958876]
32. Reddit. 2016. r/diabetes in Sept 2016 URL: https://www.reddit.com/search?q=diabetes [accessed 2016-09-30] [WebCite Cache ID 6kv53UDO1]
33. Coulson NS. Receiving social support online: an analysis of a computer-mediated support group for individuals living with irritable bowel syndrome. Cyberpsychol Behav 2005 Dec;8(6):580-584. [doi: 10.1089/cpb.2005.8.580] [Medline: 16332169]
34. Reddit. 2016. r/ibs in Sept 2016 URL: https://www.reddit.com/r/ibs/search?q=ibs [accessed 2016-09-30] [WebCite Cache ID 6kv4Vg7IY]
35. Reddit_Member. Reddit. I have every publicly available Reddit comment for research URL: https://www.reddit.com/r/datasets/comments/3bxlg7/i_have_every_publicly_available_reddit_comment/; [WebCite Cache ID 6kgAuNxDE]
36. Nonnecke B, Preece J. Lurker demographics: counting the silent. New York, New York, USA: ACM Press; 2000 Presented at: SIGCHI conference on Human Factors in Computing Systems; April 01 - 06, 2000; The Hague, The Netherlands p. 73-80 URL: http://dl.acm.org/citation.cfm?id=332409 [doi: 10.1145/332040.332409]
37. Pennebaker J, Booth J, Francis ME. LIWC. Austin, Texas; 2007. LIWC2007: Linguistic inquiry and word count [software program for text analysis] URL: http://liwc.wpengine.com/ [accessed 2017-02-27] [WebCite Cache ID 6obXqA7Hw]
38. De Choudhury M, Gamon M, Counts S, Horvitz E. Predicting depression via social media. 2013 Jul 08 Presented at: 7th International AAAI Conference on Weblogs and Social Media (ICWSM); July 8–11, 2013; Boston, USA URL: http://www.aaai.org/ocs/index.php/ICWSM/ICWSM13/paper/viewFile/6124/6351
39. De Choudhury M, Counts S, Horvitz E, Hoff A. Characterizing and predicting postpartum depression from shared facebook data. New York, New York, USA: ACM Press; 2014 Feb 15 Presented at: 17th ACM conference on Computer supported cooperative work & social computing - CSCW '14; February 15 - 19, 2014; Baltimore, Maryland, USA p. 626-638 URL: http://dl.acm.org/citation.cfm?doid=2531602.2531675 [doi: 10.1145/2531602.2531675]
40. Rude S, Gortner E, Pennebaker J. Language use of depressed and depression-vulnerable college students. Cogn Emot 2004 Dec;18(8):1121-1133 [FREE Full text] [doi: 10.1080/02699930441000030]
41. Hollander M, Wolfe D, Chicken E. Nonparametric statistical methods 3rd Edition. Hoboken, New Jersey, USA: Wiley; 173.
42. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Series B Stat Methodol 1995;57(1):289-300 [FREE Full text] [doi: 10.2307/2346101]
43. Lin M, Lucas H, Shmueli G. Too big to fail: large samples and the p-value problem. Inf Syst Res 2013 Dec;24(4):906-917. [doi: 10.1287/isre.2013.0480]
44. Cureton E. Rank-biserial correlation. Psychometrika 1956 Sep;21(3):287-290 [FREE Full text] [doi: 10.1007/BF02289138]
45. Arguello J, Butler B, Joyce E, Kraut R, Ling K, Rose C, et al. Talk to me: foundations for successful individual-group interactions in online communities. New York, NY, USA: ACM Press; 2006 Presented at: SIGCHI conference on Human Factors in computing systems - CHI '06; April 22 - 27, 2006; Montréal, Québec, Canada p. 959-968 URL: http://dl.acm.org/citation.cfm?id=1124916 [doi: 10.1145/1124772.1124916]
46. Calvo R, Hussain M, Milne D, Nordbo K, Hickie I, Danckwerts P. Augmenting Online Mental Health Support Services. In: Integrating Technology in Positive Psychology Practice. Hershey, Pennsylvania, USA: IGI Global; 2016:82-103.
47. Conway M. Ethical issues in using Twitter for public health surveillance and research: developing a taxonomy of ethical concepts from the research literature. J Med Internet Res 2014 Dec 22;16(12):e290 [FREE Full text] [doi: 10.2196/jmir.3617]
48. Vayena E, Salathé M, Madoff LC, Brownstein JS. Ethical challenges of big data in public health. PLoS Comput Biol 2015 Feb;11(2):e1003904 [FREE Full text] [doi: 10.1371/journal.pcbi.1003904] [Medline: 25664461]
49. Mikal J, Hurst S, Conway M. Ethical issues in using Twitter for population-level depression monitoring: a qualitative study. BMC Med Ethics 2016 Apr 14;17:22 [FREE Full text] [doi: 10.1186/s12910-016-0105-5] [Medline: 27080238]
50. Duggan M, Smith A. Pew Research Center. 2013. 6% of online adults are Reddit users URL: http://www.pewinternet.org/2013/07/03/6-of-online-adults-are-reddit-users/ [accessed 2016-02-11] [WebCite Cache ID 6fEEqzeJp]
51. Bogers T, Wernersen R. How `social' are social news sites? exploring the motivations for using Reddit.com. : iSchools; 2014 Mar 01 Presented at: iConference 2014 Proceedings; 2014; Berlin, Germany p. 329-344 URL: http://hdl.handle.net/2142/47295 [doi: 10.9776/14108]

52. Griffiths K, Farrer L, Christensen H. The efficacy of internet interventions for depression and anxiety disorders: a review of randomised controlled trials. Med J Aust 2010 Jun 07;192(11 Suppl):S4-11. [Medline: 20528707]
53. Dinakar K, Chen J, Lieberman H, Picard R, Filbin R. Mixed-Initiative Real-Time Topic Modeling & Visualization for Crisis Counseling. New York, New York, USA: ACM Press; 2015 Presented at: 20th International Conference on Intelligent User Interfaces - IUI '15; March 29 - April 01, 2015; Atlanta, Georgia, USA p. 417-426 URL: http://dl.acm.org/citation.cfm?id=2678025.2701395 [doi: 10.1145/2678025.2701395]
54. Park A, Hartzler AL, Huh J, McDonald DW, Pratt W. Automatically detecting failures in natural language processing tools for online community text. J Med Internet Res 2015 Aug 31;17(8):e212 [FREE Full text] [doi: 10.2196/jmir.4612] [Medline: 26323337]
55. Spitzer RL, Kroenke K, Williams JB. Validation and utility of a self-report version of PRIME-MD: the PHQ primary care study. Primary Care Evaluation of Mental Disorders. Patient Health Questionnaire. J Am Med Assoc 1999 Nov 10;282(18):1737-1744. [Medline: 10568646]
56. Park A, Hartzler AL, Huh J, Hsieh G, McDonald DW, Pratt W. "How did we get here?": topic drift in online health discussions. J Med Internet Res 2016 Nov 02;18(11):e284 [FREE Full text] [doi: 10.2196/jmir.6297] [Medline: 27806924]

Abbreviations

IBS: irritable bowel syndrome
LIWC: Linguistic Inquiry and Word Count
CI: Confidence Interval

Edited by R Calvo; submitted 15.10.16; peer-reviewed by D Milne, S Hussain, A Rosen; comments to author 22.11.16; revised version received 03.01.17; accepted 08.02.17; published 20.03.17.

Please cite as:
Park A, Conway M
Longitudinal Changes in Psychological States in Online Health Community Members: Understanding the Long-Term Effects of Participating in an Online Depression Community
J Med Internet Res 2017;19(3):e71
URL: http://www.jmir.org/2017/3/e71/
doi:10.2196/jmir.6826
PMID:28320692

Original Paper

Mobile Device–Based Electronic Data Capture System Used in a Clinical Randomized Controlled Trial: Advantages and Challenges

Jing Zhang[1*], MSc; Lei Sun[1*], MSc; Yu Liu[2], PhD; Hongyi Wang[3], MD, PhD; Ningling Sun[3], MD, PhD; Puhong Zhang[1], PhD

[1]The George Institute for Global Health at Peking University Health Science Center, Beijing, China

[2]Beihang University, Beijing, China

[3]The People's Hospital, Peking University, Beijing, China

[*]these authors contributed equally

Corresponding Author:
Puhong Zhang, PhD
The George Institute for Global Health at Peking University Health Science Center
Level 18, Tower B, Horizon Tower, No. 6 Zhichun Rd Haidian District | Beijing, 100088 P.R. China
Beijing,
China
Phone: 86 1082800577 ext 512
Fax: 86 1082800177
Email: zpuhong@georgeinstitute.org.cn

Abstract

Background: Electronic data capture (EDC) systems have been widely used in clinical research, but mobile device–based electronic data capture (mEDC) system has not been well evaluated.

Objective: The aim of our study was to evaluate the feasibility, advantages, and challenges of mEDC in data collection, project management, and telemonitoring in a randomized controlled trial (RCT).

Methods: We developed an mEDC to support an RCT called "Telmisartan and Hydrochlorothiazide Antihypertensive Treatment (THAT)" study, which was a multicenter, double-blinded, RCT, with the purpose of comparing the efficacy of telmisartan and hydrochlorothiazide (HCTZ) monotherapy in high-sodium-intake patients with mild to moderate hypertension during a 60 days follow-up. Semistructured interviews were conducted during and after the trial to evaluate the feasibility, advantage, and challenge of mEDC. Nvivo version 9.0 (QSR International) was used to analyze records of interviews, and a thematic framework method was used to obtain outcomes.

Results: The mEDC was successfully used to support the data collection and project management in all the 14 study hospitals. A total of 1333 patients were recruited with support of mEDC, of whom 1037 successfully completed all 4 visits. Across all visits, the average time needed for 141 questions per patient was 53 min, which were acceptable to both doctors and patients. All the interviewees, including 24 doctors, 53 patients, 1 clinical research associate (CRA), 1 project manager (PM), and 1 data manager (DM), expressed their satisfaction to nearly all the functions of the innovative mEDC in randomization, data collection, project management, quality control, and remote monitoring in real time. The average satisfaction score was 9.2 (scale, 0-10). The biggest challenge came from the stability of the mobile or Wi-Fi signal although it was not a problem in THAT study.

Conclusions: The innovative mEDC has many merits and is well acceptable in supporting data collection and project management in a timely manner in clinical trial.

(J Med Internet Res 2017;19(3):e66) doi:10.2196/jmir.6978

KEYWORDS

mEDC; electronic data capture; mobile data capture; mhealth; randomized controlled trial; clinical research

Introduction

An adequate central data management is crucial and indispensable in order to manage data capture, data integration, data storage, and data transfer in clinical trials [1]. However, paper-based data collection has considerable challenges of data management and requires additional time for double data entry, cleaning, and analysis [2]. In several cases, a conventional central data management comprising Web applications for data capture and central databases provides a suitable solution, but these are often limited by lack of reach and widespread applicability [3]. For example, since such tools are made available in an internal secured network, data capture can only be conducted from personal computers within this network. Moreover, data collection from subjects often takes place in rooms where a designated personal computer is not accessible. More importantly, most studies relying on computer-based data capture tools are not required to submit captured data in real time during collection.

In the past decade, certain mobile device–based tools such as personal digital assistants (PDAs) [4,5] and mobile phones [6,7] have been used to collect and manage data. Such tools show great performance and qualities including high-quality data, more effective training to the staff, user friendliness, and cost effectiveness. More importantly, in recent decades, access to the Internet using wireless and mobile communication technologies has tremendously increased and has been adapted for use in field research settings, particularly in medical research, because of the ease of transferring data in real time and convenience owing to the portability of mobile devices. Low- and middle-income countries such as China and India lack the infrastructure to accommodate adequate fixed-line Internet access in rural areas, and cellular networks allow access to telecommunications in such regions with limited Internet access [8]. China has the largest and fastest growing mobile Internet population, with 668 million people using the Internet (89% via mobile phones) as of June 2015 [8], which supports effective implementation of studies using mobile devices in China.

Nevertheless, mobile technology has been barely used to support implementation of pharmaceutical randomized controlled trials (RCTs). Compared with traditional computer-based clinical trial management systems, mobile device–based methods for clinical trial reporting have distinct features that can ensure data transfer and exchange, telemonitoring, and project management in real time.

From 2014 to 2015, we conducted a multicenter, randomized, double-blinded, parallel-controlled trial with the aim of comparing the efficacy of monotherapy with two types of medications in high-sodium-intake patients with mild to moderate hypertension (*T* elmisartan 40 mg/day and *H* ydrochlorothiazide [HCTZ] 25 mg/day *A* ntihypertensive *T* reatment study or the THAT study). During implementation of the THAT study, we developed an innovative, mobile device–based electronic data capture (mEDC) system to support data capture, monitoring, and project management. In addition, we performed a designated qualitative evaluation based on experiences of various types of users to determine whether mEDC can effectively facilitate data collection, project implementation and management, and real-time telemonitoring.

Methods

Design and Features of mEDC

Eligible participants for the THAT study included adults aged ≥18 years with mild to moderate hypertension not using antihypertensive or blood pressure (BP)–altering medications for at least one week. Overall, 1333 participants were recruited at clinics from 14 rural county hospitals, randomly divided into the telmisartan and HCTZ treatment groups, and followed up on the 15th, 30th, and 60th day after enrollment. Dummy telmisartan and dummy HCTZ were used to ensure double blinding. The primary outcomes included BP decrease, BP control rates, fasting blood glucose (FBG), hypokalemia, and adverse events (AEs).

The mEDC system was developed to help users (doctors, clinical research associates [CRAs], project managers [PMs], and data managers [DMs]) in data collection, telemonitoring, and project management. The mEDC, consisting of two primary components (an app installed in mobile phones and a server-based clinical trial database) was designed in accordance with the guidelines laid down by the International Conference on Harmonization Technical Requirements for Registration of Pharmaceuticals for Human Use Good Clinical Practice (ICH-GCP) [9]. We prospectively wrote design specifications to describe structure of the mEDC, and a vendor independently developed this app following our specifications. We also developed a test plan based on these specifications to justify validation of the mEDC system. In order to ensure regulatory compliance, we referred to three additional guidelines focusing on EDC systems for RCTs while designing mEDC: Good Clinical Data Management Practices proposed by the Society for Clinical Data Management (SCDM) [10], Guidance for Industry Computerized Systems Used in Clinical Investigations represented by the Food and Drug Administration [11], and Guidance for Industry Part 11 of Title 21 of the Code of Federal Regulations on Electronic Records and Electronic Signatures (21 CFR 11) [12]. Before implementing the primary study, we invited two independent research staff to validate the mEDC system. They entered a batch of data (approximately 50 simulated patients) into the system to test not only the clinical database validation, including data entry screen testing, data checking routines (eg, range and format), testing of data verification functions, and data transferring (remote data entry) but also trial-specific validation of variables such as name, label, type, and randomization of subjects [10]. Table 1 summarizes the major supportive functions and features of mEDC.

Table 1. Functions and features of mobile device–based electronic data capture (mEDC) and relevant requirements as per electronic data capture (EDC) guidelines.

No.	Functions and features of mEDC[a,b]	Requirements per EDC[a,c] guidelines and comments
1	**Site equipment and administration**	
	Logistic management	Should be maintained either onsite or be remotely accessible through electronic files (IV-B: SOP[d] [11])
	Role allocation	Access must be limited to authorized individuals (IV-D-1: Limited access [11])
2	**Patient recruitment and data collection**	
	Screening Data collection at baseline Centralized randomization Data collection during follow-up BP[e] measurement Record of adverse events	These 6 parts follow "Guidance for Industry Computerized Systems Used in Clinical Investigations" [11]: each step will be used to create, modify, maintain, archive, retrieve, or transmit source data (IV-A: Study Protocol [11]), ensure the system's date and time stamps are accurate (IV-D-3: Date/Time stamps [11]), alert the user if data are out of acceptable range, and should not automatically enter data (IV-F-1: Direct Entry Data [11])
3	**Patient referral and reference data**	
	Biological sample collection Medicine prescription	The same as No. 2 (same as above)
4	**Remote validation of data**	To assure that data are reliable, complete, and accurate (Data entry and data processing [10]); keeping track of all changes made to data in electronic records (IV-D-2: Audit trails [11]); audit trails can be particularly appropriate when users are expected to create, modify, or delete regulated records during normal operation (III-C-2: Audit trails [12])
5	**Data storage and management**	A copy of the data should be maintained at another location, typically at the clinical site (IV-C: Source documentation and records retention [11], data entry and data processing [10], measuring data quality [12]); should be accessible only by using their own password (IV-D-1: Limited access [11]) Procedures and controls should be put in place to prevent altering, browsing, querying, or reporting of data via external software applications (IV-E: External security safeguard [11]) Provided to FDA[f] and should fully describe and explain how source data were obtained and managed and how electronic records were used to capture data; have dependable system documentation (IV-F-2: Retrieving data [11]) Create and preserve electronic records, sufficient backup, and recovery procedures (IV-F-4: System controls [11])
6	**Institutional Review Board operations**	
	Clicking a photograph of informed consent forms to ensure completion of IRB[g] operations	No requirements for IRB operations in the guidelines
7	**Other functions and features**	
	Quality control for BP measurements	No requirements for quality control of BP measurements in the guidelines
	Quality control for consistence of medicine codes	Ensuring processes are defined to integrate laboratory and other non-CFR[h] data with the data from the eCRF[i] (Electronic Data Capture Principles [10])
	Project progress and status reports	Automate generation of reports on metrics and project status to facilitate project or site or patient management (Electronic Data Capture Principles [10])

[a]Mobile phone Redmi Note (5.5 inch, 1GB RAM, 32GB ROM, 4G Dual SIM, Android system v4.2) was used as the mobile device. Moreover, mEDC can be installed and used in mobile phones from most other brands that are based on the Android system.

[b]mEDC: mobile device-based electronic data capture.

[c]EDC: electronic data capture.

[d]SOP: standard operating procedure.

[e]BP: blood pressure.

[f]FDA: Food and Drug Administration.

[g]IRB: Institutional Review Board.

[h]CFR: Code of Federal Regulation.

[i]eCRF: electronic case report form.

Site Equipment, Training, and Administration

At each study site, doctors were required to take some photographs using the camera of their mobile phones to verify that they had completed the role allocation and internal training for study preparation. Delivery of all site-specific materials such as study medicines, tubes for biological samples, and other materials, both to and from each site, was confirmed by the doctors, and the materials were photographed for record. The PMs remotely checked these photographs, which were required to be uploaded in the mEDC app, and sites would initiate the process after receiving confirmation from PMs. The study protocol and standard operating procedures (SOPs) were stored in the cellphones, allowing convenient access of these instructions.

Patient Recruitment and Visits

Once patients signed informed consent forms, doctors collected necessary information using mEDC, following which the server immediately returned the results to verify whether or not the patients met the inclusion criteria. After baseline information had been collected, unique and structured randomization codes were allocated to all patients. Simultaneously, mEDC allocated study medicines and test tubes bearing the same codes as assigned to the patients. Each patient was assigned a specific time slot for site visits. mEDC helped doctors to arrange visits through 2 automatic pop-up options: sending a standard short message and calling the patient immediately. Patients who had opted to receive a short message received a message with the name of the doctor indicating when he or she should visit the doctor, if breakfast and medicines should be taken, and if the leftover medicine should be brought along. Patients who had difficulty in receiving or reading messages were contacted over telephone on their primary or alternate contact numbers that had been collected at the first visit. During visits, all data transactions between mEDC at the site and the central study server were automatically stamped with identification regarding who provided the information, when it was provided, from where, and how.

Centralized Randomization

THAT study was a double-blinded study with stratified block randomization. Each box of study medicines that contained a genuine medicine and the other dummy medicine was allocated a specific code according to the randomization list in advance. The randomization list was generated previously with gender and systolic blood pressure (SBP ≥160 mmHg) as stratified variables and a random block size of 4. A range of medicine boxes with sequential codes of randomization were delivered to each site. Site names and the matched codes list were uploaded to the mEDC server so that doctors could identify the specific medicines according to the code returned by mEDC during randomization. Biological sample tubes and other study materials with the same codes were delivered to the same site. This guaranteed that a patient had received study medicines and materials with their unique code. Doctors could trigger an unblinding procedure for patients in emergency via mEDC during follow-ups.

Quality Control for Key Procedures

BP measurement was the primary outcome of the THAT study. Doctors received strict training and were provided with uniform digital sphygmomanometers. In practice, the entire procedure was automatically controlled by mEDC. Before measurements, mEDC required doctors to click photographs to record the posture of patients at the time of measurement. During measurements, it instructed doctors to measure BP for 3 times at 2-min intervals.

Another key point for quality control was the consistence of medicine codes, that is, the codes that mEDC allocated to patients had to be the same as the codes that were on the medication boxes. Doctors were required to click photographs to record the codes on the medication boxes, and CRAs checked to ensure this consistency.

Patient Referral

At the end of the visit, doctors instructed patients to collect a blood or urine sample, and a message would pop-up on the mobile phone of the nurses responsible for blood or urine sample collection as follows: “Patient xxx with code yyy will arrive for sample collection. Please make sure to use the tubes with the same code.” A similar message would be sent to the study medication distributors to support study medication distribution. The entire process was designed to ensure consistency of patients’ codes across all procedures.

Remote Validation of Data

Similar to traditional computer-based EDC systems, most variables in the THAT study were checked for missing values, range of variable values, outliers, and irrationalities during data entry. However, in contrast to the traditional EDC systems, data entry and validation in the mEDC was performed at the time point of patients’ site visit rather than at a later time point. In addition, by relying on camera technique, mEDC could ask doctors to provide evidence by clicking photographs for key information. Photographic evidence was collected for consistency of patient identification on different materials at all 4 visits, BP measurements, and other procedures. For example, doctors were required to not only enter BP measurements but also to click a photograph for interfaces of digital sphygmomanometers and upload them on the server. The consistency between these recordings would be remotely checked by a CRA. In addition, CRAs remotely supervised the recruiting progress, data validity, missing data, and risk of loss to follow-up through mEDC and in a timely manner. Logistical checking could also be done based on the database temporarily downloaded from the server. We did not develop Web-end to track modification for each variable but used a specially designed database to record any modifications. The mEDC DM provided frequent reports on modifications by running a program based on the database.

Data Storage and Management

Owing to the need for tracking patients during referral and sending follow-up reminders, certain private data such as names and telephone numbers were recorded in mEDC. During data or project management, personal information of participants was shielded from DMs and CRAs, and the deidentified data could only be transferred to researchers and statisticians for data analysis. No hardcopy CRFs were used in the THAT study. Once submitted, collected data was then transferred to the server (located in Beijing) without any information stored in mobile phones. Instant photocopies of key information or records such as BP readings, codes of study medications, and biomedical samples were also uploaded to the server. All the data stored in the server was quarterly backed up to a designated local laptop. Secure Hash Algorithm 512 (SHA-512) [13], a more efficient and secure algorithm, was implemented to ensure that all data were stored in the server and not in local mobile phones. Only authorized doctors could log in to mEDC and access the information collected by them. During the process of data transfer, cryptographic net key was initiated when an investigator logged in. To ensure safe data exchange, the "HTTPS" was utilized in the linkage between the terminal device and the server, which formed a safe circle to prevent stealing or misuse of data.

Institutional Review Board Operations

As part of the study SOP, doctors were required to click a photograph of the page with signatures and submit it to the server, which could help CRAs verify patients' recruitment because signatures of patients and doctors were essential as per the recommendations of 21 CFR 11 [12].

Project Progress and Status Reports

Doctors could use the recruitment number of patients to confirm whether they had completed their follow-up or not, and mEDC would send reminders to doctors regarding the same. On the other hand, CRAs and PMs could easily track the progress at each site along with statistical summaries of each site, and they would receive notifications when any AE occurred.

Operation of mEDC

Certain interfaces are presented in Figure 1.

Every step of study implementation and the actions mEDC supported are presented in Figure 2.

Figure 1. Certain interfaces of mobile device–based electronic data capture (mEDC).Fig 1-1 Root directory for investigators and CRA to manage materials, visits, and quit register; Fig 1-2 Interface for investigators to input blood pressure and upload photo evidence of electronic sphygmomanometer screens with 2 minutes interval for each of three sequential blood pressure measurements;Fig 1-3 Current reports of recruitment and follow-ups for site investigators;Fig 1-4 Current reports of all sites for CRA and principle investigators.

Figure 2. Functions and operations of mobile device–based electronic data capture (mEDC).

Qualitative Evaluation

An intensified sampling method was used to select participants in different roles involved in the THAT study to evaluate the feasibility of mEDC, including 1-2 doctors at each hospital, 1 CRA, 1 PM, and 1 DM.

Semistructured qualitative interviews were conducted by 2 trained interviewers during and after the clinical trial. The structured questions included (1) individual general information and infrastructure conditions, (2) time taken for data collection at each visit, and (3) user satisfaction scores. Satisfaction scores ranged on a scale of 0-10, with a score of 10 indicating "most satisfied," and a score of 0 indicating "most dissatisfied." In-depth personal interviews were conducted to explore individuals' understanding and feelings of mEDC, which covered (1) evaluation of practical use of mEDC, (2) willingness of using such an app in the future, and (3) experience of security in data storage and transfer. Simultaneously, interviewers objectively assessed how users maneuvered the app during project implementation. Following description of interview guidelines, all participants were interviewed face-to-face, and their responses were recorded. In addition, 53 patients were interviewed through telephone calls to enquire about their experience when visited by doctors through mEDC.

Nvivo 9.0 was used for data extraction and classification after transcribing and restructuring all the records into *.doc* format. A thematic framework method was used to analyze the interview transcripts [14]. Two independent groups separately conducted data extraction and classification, and any discrepancies between the two groups were resolved by consensus. Because a non-English interview guide was used, back-translation [15] was used to ensure that the original information provided by the interviewees was accurate and reliable.

Ethical Approval

This study received approval from the Peking University at Medical Health Science Center Ethics Committee (ref: IRB00001052-14039). Signed informed consent was obtained from all participants before the qualitative interviews.

Results

Telecommunication Signal Coverage and Characteristics of Participants

Table 2 summarizes the demographic and other characteristics of the participants. All the study hospitals had access to 3G or 4G and/or WiFi signal.

Table 2. Demographic and other characteristics of all participants.

Characteristics	Doctors (n=24)	Patients (n=53)	CRA[a] (n=1)	PM[b] (n=1)	DM[c] (n=1)
Age (years), mean (SD[d])	40.6 (4.2)	53.5 (5.2)	26	34	37
Male (n)	10	23	0	0	1
Years of work experience, mean (SD)	11.2 (5.1)	-	2	10	10
Prior clinical trial experience, n (%)	7 (29.2)	-	1 (100.0)		1 (100.0)
Possession rate of any cellphone, n (%)	24 (100.0)	40 (76.0)	1 (100.0)		1 (100.0)
Possession rate of mobile phone, n (%)	24 (100.0)	13 (24.0)	1 (100.0)		1 (100.0)

[a]CRA: clinical research associate.

[b]PM: project manager.

[c]DM: data manager.

[d]SD: standard deviation.

Trial Implementation, BP Measurements, and Time Consumption for Data Collection

Overall, 1333 out of a total of 2130 participants were recruited in 14 Hebei county hospitals from October to December 2014; 670 participants were randomly allocated to the telmisartan group and 663 to the HCTZ group. At the end of the study, 1037 participants had completed three follow-up visits in 2 months. The average SBP/DBP (diastolic blood pressure) reduction in telmisartan group and HCTZ group was 12.8/7.2 and 11.5/5.3 mmHg, respectively. As an example, the details of 3 BP measurements at baseline are described in Table 3.

Table 3. Three BP measurements at baseline (n=1188).

BP	First measurement, mean (SD)	Second measurement, mean (SD)	Third measurement, mean (SD)	The average of later two measurements, mean (SD)
SBP[a]	160.1 (10.6)	155.3 (10.1)	153.1 (11.6)	154.1 (11.1)
DBP[b]	99.3 (8.9)	94.5 (8.7)	91.3 (9.3)	92.8 (8.8)

[a]SBP: systolic blood pressure.

[b]DBP: diastolic blood pressure.

The entire questionnaire had 141 questions including 54 questions for baseline visit and 29 questions for each of visits 2-4, excluding the 15 questions in the severe AEs form. As estimated by the 24 doctors, the average time used for data collection directly through mEDC was 53.0 (SD 5.3) min for all the 4 visits. All doctors considered the total time taken was reasonable for them, and most patients could bear the standby period. No patient had any complaints related to the procedure when interviewed through mEDC.

> *I think I can afford the average time spent on visiting one patient. The study did not affect my routine work. Although patients usually attend visits here in the morning, I spent only around 10 minutes per patient to complete each visit.* [Doctor, female, 46 years]

> *I feel I experienced a satisfactory process. Following reminders, my doctor told me how to conduct every step during the visit. The time spent on completing each visit was considerably short.* [Patient, female, 51 years]

Experience of Doctors

All doctors mentioned that mEDC was very convenient and could help them complete the visits smoothly. They were willing to use mEDC to implement the THAT study, although the current mEDC app still needs to be upgraded, such as addition of a function of patient indexing. Moreover, they expressed that they would be interested in participating in other clinical trials with mobile phones in the future. Table 4 summarizes the satisfaction scores.

Table 4. Satisfaction scores with mobile device–based electronic data capture (mEDC) from 24 doctors.

No.	Procedures	No. of participants (nonresponse[a]/total[b])	Satisfaction score (median, range)
1	**Procedure of implementation**		
	Inclusion and exclusion	0/18	9 (8-10)
	Informed consent	0/18	9 (8-10)
	Data collection at each visit	0/18	9 (8-10)
	Randomization	0/18	9 (8-10)
	Collection of biological samples	0/18	9 (7-10)
	Medicine delivery	0/18	9 (7-10)
	Appointment for next visit	1/18	8 (6-9)
2	**Project management**		
	Accounts management	0/18	9 (8-10)
	Logistics	1/18	9 (8-10)
	Patient indexing	0/18	8 (6-9)
3	**Quality control**		
	Reminder at each step	0/18	9 (8-10)
	Key point control	0/18	9 (8-10)
4	**Others**		
	Training for using mEDC[c]	2/18	9 (8-10)
	Wireless Internet	0/18	8 (6-9)
	Overall average score		9 (6-10)

[a]Nonresponse: number of interviewees who had no response.

[b]Total: total number of participants who attended the qualitative evaluation.

[c]mEDC: mobile device-based electronic data capture.

> *Although the study involved several steps, such as visit, blood pressure measurement, biological sample collection, medicine description, and visit appointment, the smartphone assisted me to complete every task without any trouble, and I only needed to follow the tips provided by mEDC. I can maneuver the system easily, even though I am not familiar with the functions of my smartphone.* [Doctor, male, 44 years]

> *Although I do not have prior experience in a clinical trial, I think that mEDC helped me collect the data, manage my patients, remind me of what the next step was, and when I should follow up my patients. It made me efficient and confident to execute my job. If possible, I am keen to be part of other studies with smartphones in the future. I hope that for those patients who completed their visits, an indexing function should be available in mEDC, which can help them find their doctor (me) easily.* [Doctor, male, 40 years]

Experience of CRAs and PMs

CRAs and PMs expressed that this was their first experience of using mEDC to monitor or manage a pharmaceutical clinical trial. They all believed that mEDC offered some creative functions that were not available in traditional computer-based clinical trials, such as material dispatching and tracking, remote monitoring and validation in real time, and supervising the progress of the trial. mEDC is a promising tool for use in future pharmaceutical clinical trials.

> *I could monitor quality remotely through mEDC, viewing the data instantly when doctors uploaded their data. As soon as I found that any figure or photograph was possibly wrong, like blood pressure measurements, I would contact them as soon as possible. I also noticed that compared with traditional clinical trials, doctors did not easily procrastinate, since mEDC required the doctors to correct mistakes in a timely fashion, and this feature can help reduce my workload.* [CRA, female, 26 years]

> *I believe that it is a very good tool for pharmaceutical clinical trials. The data uploading in real time, the smooth operation, and the user-friendly interface are highly impressive features. Although this was my first experience using such a cellphone-based system, I believe that it can be promoted in the field of randomized controlled trials in the future. Compared with traditional trials, especially for those trials in which the data source is not from health records, like*

> *the THAT study, this mEDC could help us manage the project and control quality in real time. The most impressive detail is that the function of the 2-minute interval between two blood pressure measurements can systematically ensure that doctors follow the SOP, so as to reduce potential bias.* [Project manager, female, 34 years]

Experience of DMs

DMs also believed that mEDC was safe and using it could help reduce workload during the process of data cleaning.

> *With almost 10 years of experience focusing on clinical data management, I consider that the whole structure of mEDC is very good and reasonable, especially the safety. For example, "HTTPS" can make sure that the process of data transfer is safe. Moreover, the server is cryptographic, and only people who know the password can have access to it. Spontaneously, real-time uploading enables data not only to be stored in phones but also to be stored on the server, which can avoid information loss to an extremely high extent. More importantly, error correction of mEDC in real time considerably reduces workload at the time of data cleaning. However, a limitation is that there is no modification track in mEDC interfaces, which would help doctors recall work experience. I hope to see this function in the next version of mEDC.* [Data manager, male, 37 years]

Discussion

Principal Findings

Mobile device-based technology or app is generally used in several domains [3], including for data collection and reporting [6]; however, at present, project management in pharmaceutical clinical trials scarcely involves mobile devices. The THAT study was successfully completed under management of mEDC, suggesting that data collection and management using mEDC was technologically feasible using a mobile phone in this study. The positive feedback of the users confirmed the feasibility of mEDC, which also establishes substantial confidence about the utilization of mEDC in future pharmaceutical clinical trials.

The significant and specific features of mEDC are data collection, monitoring, and project management in real time. Although computer-based trials proposed that they could recruit patients and collect data in real time by using Internet [16], in most cases, doctors were more likely to rely on transferring data by means of paper-to-computer when they captured clinical data, particularly in trials wherein data sources involved health records. mEDC made the aforementioned process more straightforward owing to portability of mobile phones and was more beneficial in entering data directly in real time. Under such situations, PMs can remotely monitor the progress of the project in real time. Theoretically, mEDC can reduce the time required for data capture because doctors use mobile phones to collect information directly instead of recording such information in paper-based CRFs in the first place. Although we did not compare the effectiveness between these two approaches, evidence available from several previous studies suggests that using mobile phones for data collection, instead of with pen-and-paper, eliminated data recording and entry errors, had similar interrater reliability, and took an equal amount of time per interview but with no second entry [17].

The design for quality control and data validation through mEDC was unprecedented in this time of computer-based clinical trials, particularly for validation of the primary outcome (BP measurement), which could assure that the BP of every patient was accurately measured. The photographic evidence of patients' seated posture and reminders for 2-min intervals between every two measurements could control doctors to follow the SOP of BP measurement, whereas the screenshots of the digital sphygmomanometer after each BP measurement could help CRAs check the consistency between BP values and screenshots, which can be difficult to achieve in computer-based trials. This type of design can help CRAs remotely conduct site monitoring, reduce extremely inaccurate information, improve data quality, and reduce the workload of DMs at the time of data cleaning. At the same time, Table 3 shows that among three average BP measurements recorded at baseline, the first BP measurement was higher than the other two, the second one was moderate, whereas the third one was the lowest; however, the latter two measurements exhibit a stable trend, which is in concordance with findings from a previous study [18]. This, in a way, validates the reliability of data transferred using mEDC.

From the perspective of doctors, mEDC could be used correctly and smoothly in county hospitals because the user-friendly operation system and succinct screen could help them complete a high-quality clinical trial through a built-in automated procedure of the mobile phone, regardless of prior clinical trial experience. Time consumption analysis and qualitative interviews also indicated that mEDC was easy to operate, and it did not impose any additional workload on doctors. This portable data capture tool not only benefitted doctors but also benefitted CRAs with convenience of data validation.

With regard to security, like SHA-512, which guarantees safety of data transfer [13], HTTPS assured secure connection of the mobile phone with the server and prevented it from being attacked by hackers. The central server could encrypt all transmissions and restrict each individual to appropriate access to data and operations on data.

Limitations

This study had certain limitations. The biggest concern was the signal shielding and stability of the cellular network. Although mobile network is more popular than fix-line Internet in China [3], the signal could probably be impacted due to shielding by the environment, such as in indoor areas of certain buildings with reinforcement concrete frames. The THAT study was a phase IV clinical trial, which is relatively simpler than phase I-III clinical trials; thus, the mEDC algorithm would probably need to be designed in a more complex fashion if it is to be utilized in other phases of a clinical trial. Moreover, whether available mobile phones can process algorithms that are more complicated remains unknown. The current version of mEDC was not perfect because of limited design time. For example, queries had not been designed in mEDC; therefore, any track

changes were not shown on the screen of the mobile phone. Although the track changes could be recorded in the server, doctors could not identify which data were modified; this resulted in some inconveniences. BP measurements were to be manually filled in mEDC but could not be automatically transferred via Bluetooth. Furthermore, doctors were not authorized to record electronic signatures during IRB operations but rather had to record photographs of the informed consent form pages with signatures for data retention. Nonetheless, this limitation will be addressed and fixed in an upcoming updated version of this app.

Conclusions

The mobile device-based data capture and project management system, mEDC, could help doctors complete a phase IV pharmaceutical clinical trial and was feasible for management of this trial. Moreover, doctors expressed their willingness to use this tool for study implementation. The validity, reliability, real-time feature, and user friendliness of mEDC are beneficial not only for doctors without clinical trial experience but also for CRAs and PMs. Taken together, there is a possibility for mEDC to be used in other pharmaceutical clinical trials in the future.

Acknowledgments

We thank all the investigators from the following 14 county-level hospitals in high-salt-intake areas in Hebei Province of China: Laishui County People's Hospital, Gaobeidian Hospital, Shunping County People's Hospital, Li County People's Hospital, Gaoyang County People's Hospital, Anguo County People's Hospital, Tang County People's Hospital, Mancheng County People's Hospital, Boye County People's Hospital, Rongcheng County People's Hospital, Wangdu County People's Hospital, Baigou Hospital, Dacheng County People's Hospital, and Wenan County People's Hospital.

We also thank Wenlong Zhao, from Beihang University, for helping us develop the mEDC app and maintain the server to keep it running smoothly.

We appreciate the Ministry of Science and Technology of the People's Republic of China for providing funding for this study.

Conflicts of Interest

None declared.

References

1. Meyer J, Ostrzinski S, Fredrich D, Havemann C, Krafczyk J, Hoffmann W. Efficient data management in a large-scale epidemiology research project. Comput Methods Programs Biomed 2012 Sep;107(3):425-435. [doi: 10.1016/j.cmpb.2010.12.016] [Medline: 21256617]
2. Meyer J, Fredrich D, Piegsa J, Habes M, van den Berg N, Hoffmann W. A mobile and asynchronous electronic data capture system for epidemiologic studies. Comput Methods Programs Biomed 2013 Jun;110(3):369-379. [doi: 10.1016/j.cmpb.2012.10.015] [Medline: 23195493]
3. Peiris D, Praveen D, Johnson C, Mogulluru K. Use of mHealth systems and tools for non-communicable diseases in low- and middle-income countries: a systematic review. J Cardiovasc Transl Res 2014 Nov;7(8):677-691. [doi: 10.1007/s12265-014-9581-5] [Medline: 25209729]
4. Onono MA, Carraher N, Cohen RC, Bukusi EA, Turan JM. Use of personal digital assistants for data collection in a multi-site AIDS stigma study in rural south Nyanza, Kenya. Afr Health Sci 2011 Sep;11(3):464-473 [FREE Full text] [Medline: 22275941]
5. Ojwang' JK, Lee VC, Waruru A, Ssempijja V, Ng'ang'a JG, Wakhutu BE, et al. Using information and communications technology in a national population-based survey: the Kenya AIDS Indicator Survey 2012. J Acquir Immune Defic Syndr 2014 May 01;66(Suppl 1):S123-S129 [FREE Full text] [doi: 10.1097/QAI.0000000000000116] [Medline: 24732816]
6. King C, Hall J, Banda M, Beard J, Bird J, Kazembe P, et al. Electronic data capture in a rural African setting: evaluating experiences with different systems in Malawi. Global Health Action 2017 Jan 19;7(1):25878 [FREE Full text] [doi: 10.3402/gha.v7.25878] [Medline: 25363364]
7. Jandee K, Kaewkungwal J, Khamsiriwatchara A, Lawpoolsri S, Wongwit W, Wansatid P. Effectiveness of using mobile phone image capture for collecting secondary data: a case study on immunization history data among children in remote areas of Thailand. JMIR mhealth uhealth 2015;3(3):e75. [doi: 10.2196/mhealth.4183] [Medline: 26194880]
8. Asia Pacific Internet Research Alliance. APIRA. 2015. Internet network development in China URL: http://www.apira.org/news.php?id=1734 [accessed 2017-02-26] [WebCite Cache ID 6oa4cwtz3]
9. Pieterse H, Diamant Z. Good clinical practice in clinical interventional studies. Eur Clin Respir J 2014;1:1-7 [FREE Full text] [doi: 10.3402/ecrj.v1.26422] [Medline: 26557234]
10. Cornu C, Donche A, Coffre C, Le GA, Rym B, Vaugier I, le Comité de Pilotage du groupe de travail Harmonisation des procédures de réseau des CIC. [ECRIN Standard Requirements for Good Clinical Practices-compliant Data Management in Multinational Clinical Trials]. Therapie 2015 Aug 03:2367-2373. [doi: 10.2515/therapie/2015042] [Medline: 26238126]

11. FDA. FDA.gov. Guidance for Industry Computerized Systems Used in Clinical Investigations; 2007 URL: https://www.fda.gov/OHRMS/DOCKETS/98fr/04d-0440-gdl0002.pdf [accessed 2017-02-26] [WebCite Cache ID 6oa4kGz3e]
12. Bansal A, Chamberlain R, Karr S, Kwasa S, McLaughlin B, Nguyen B, et al. A 21 CFR Part 11 compliant graphically based electronic system for clinical research documentation. J Med Syst 2012 Jun;36(3):1661-1672. [doi: 10.1007/s10916-010-9627-4] [Medline: 21107888]
13. Dobraunig CE, Mendel F. Analysis of SHA-512/224 and SHA-512/256. In: Advances in Cryptology – ASIACRYPT 2015. ASIACRYPT 2014. Lecture Notes in Computer Science, vol 9453. Berlin: Springer; 2015:978-973.
14. Pope C, Ziebland S, Mays N. Qualitative research in health care. Analysing qualitative data. Br Med J 2000 Jan 08;320(7227):114-116 [FREE Full text] [Medline: 10625273]
15. Beaton DE, Bombardier C, Guillemin F, Ferraz MB. Guidelines for the process of cross-cultural adaptation of self-report measures. Spine (Phila Pa 1976) 2000 Dec 15;25(24):3186-3191. [Medline: 11124735]
16. Marks RG, Conlon M, Ruberg SJ. Paradigm shifts in clinical trials enabled by information technology. Stat Med 2001;20(17-18):2683-2696. [Medline: 11523076]
17. Zhang S, Wu Q, van Velthoven MH, Chen L, Car J, Rudan I, et al. Smartphone versus pen-and-paper data collection of infant feeding practices in rural China. J Med Internet Res 2012 Sep 18;14(5):e119 [FREE Full text] [doi: 10.2196/jmir.2183] [Medline: 22989894]
18. Lui H, Zhang PH. Fluctuation of blood pressure in repeated measurements during a short period of time. Chinese General Practice 2011;14:1429-1430.

Abbreviations

AEs: adverse events
BP: blood pressure
21 CFR 11: Guidance for Industry Part 11 of Title 21 of the Code of Federal Regulations on Electronic Records and Electronic Signatures
CRAs: clinical research associates
DMs: data managers
eCRF: electronic case report form
EDC: electronic data capture
FBG: fasting blood glucose
FDA: Food and Drug Administration
HCTZ: hydrochlorothiazide
ICH-GCP: International Conference on Harmonization Technical Requirements for Registration of Pharmaceuticals for Human Use Good Clinical Practice
IRB: Institutional Review Board
mEDC: mobile device–based electronic data capture
PMs: project managers
SBP: systolic blood pressure
SCDM: Society for Clinical Data Management
SD: standard deviation
SHA-512: Secure Hash Algorithm 512
SOPs: standard operating procedures

Edited by G Eysenbach; submitted 09.11.16; peer-reviewed by N Unciano, J Buttolph; comments to author 08.12.16; revised version received 19.01.17; accepted 10.02.17; published 08.03.17.

Please cite as:
Zhang J, Sun L, Liu Y, Wang H, Sun N, Zhang P
Mobile Device–Based Electronic Data Capture System Used in a Clinical Randomized Controlled Trial: Advantages and Challenges
J Med Internet Res 2017;19(3):e66
URL: http://www.jmir.org/2017/3/e66/
doi:10.2196/jmir.6978
PMID:28274907

Publisher:
JMIR Publications
130 Queens Quay East.
Toronto, ON, M5A 3Y5
Phone: (+1) 416-583-2040
Email: support@jmir.org