<u>Original Paper</u>

Conversational Agents in Health Care: Expert Interviews to Inform the Definition, Classification, and Conceptual Framework

Laura Martinengo[1], MD, PhD; Xiaowen Lin[1], BA; Ahmad Ishqi Jabir[1,2], BSc, MSc; Tobias Kowatsch[2,3,4,5], PhD; Rifat Atun[6], MBBS, MBA, DIC; Josip Car[7,8], MSc, MD, PhD, DIC; Lorainne Tudor Car[1,8], MSc, MD, PhD

[1]Lee Kong Chian School of Medicine, Nanyang Technological University Singapore, Singapore, Singapore

[2]Future Health Technologies Programme, Campus for Research Excellence and Technological Enterprise, Singapore-ETH Centre, Singapore, Singapore

[3]Institute for Implementation Science in Health Care, University of Zurich, Zurich, Switzerland

[4]School of Medicine, University of St.Gallen, St.Gallen, Switzerland

[5]Centre for Digital Health Interventions, Department of Management, Technology, and Economics, ETH Zurich, Zurich, Switzerland

[6]Department of Global Health and Population, Harvard T.H. Chan School of Public Health, Harvard University, Cambridge, MA, United States

[7]Centre for Population Health Sciences, Lee Kong Chian School of Medicine, Nanyang Technological University Singapore, Singapore, Singapore

[8]Department of Primary Care and Public Health, School of Public Health, Imperial College London, London, United Kingdom

Corresponding Author:
Lorainne Tudor Car, MSc, MD, PhD
Lee Kong Chian School of Medicine
Nanyang Technological University Singapore
11 Mandalay Road, Level 18
Singapore, 308232
Singapore
Phone: 65 69047155
Email: <u>lorainne.tudor.car@ntu.edu.sg</u>

Abstract

Background: Conversational agents (CAs), or chatbots, are computer programs that simulate conversations with humans. The use of CAs in health care settings is recent and rapidly increasing, which often translates to poor reporting of the CA development and evaluation processes and unreliable research findings. We developed and published a conceptual framework, designing, developing, evaluating, and implementing a smartphone-delivered, rule-based conversational agent (DISCOVER), consisting of 3 iterative stages of CA design, development, and evaluation and implementation, complemented by 2 cross-cutting themes (user-centered design and data privacy and security).

Objective: This study aims to perform in-depth, semistructured interviews with multidisciplinary experts in health care CAs to share their views on the definition and classification of health care CAs and evaluate and validate the DISCOVER conceptual framework.

Methods: We conducted one-on-one semistructured interviews via Zoom (Zoom Video Communications) with 12 multidisciplinary CA experts using an interview guide based on our framework. The interviews were audio recorded, transcribed by the research team, and analyzed using thematic analysis.

Results: Following participants' input, we defined CAs as digital interfaces that use natural language to engage in a synchronous dialogue using ≥1 communication modality, such as text, voice, images, or video. CAs were classified by 13 categories: response generation method, input and output modalities, CA purpose, deployment platform, CA development modality, appearance, length of interaction, type of CA-user interaction, dialogue initiation, communication style, CA personality, human support, and type of health care intervention. Experts considered that the conceptual framework could be adapted for artificial intelligence–based CAs. However, despite recent advances in artificial intelligence, including large language models, the technology is not able to ensure safety and reliability in health care settings. Finally, aligned with participants' feedback, we present an updated iteration of the conceptual framework for health care conversational agents (CHAT) with key considerations for CA design, development, and evaluation and implementation, complemented by 3 cross-cutting themes: ethics, user involvement, and data privacy and security.

Conclusions: We present an expanded, validated CHAT and aim at guiding researchers from a variety of backgrounds and with different levels of expertise in the design, development, and evaluation and implementation of rule-based CAs in health care settings.

Introduction

Background

Conversational agents (CAs), or chatbots, are broadly defined as computer programs that simulate conversations with humans [1-3]. Although the terms CA and chatbot are often used interchangeably [1,4], they sometimes define distinct conversational systems. For example, the term chatbot generally defines text-based dialogue systems and may also be used to define dialogue systems engaging in informal conversations without a specific purpose [5,6]. CAs may communicate using a variety of input-output modalities, such as text, speech, or multimedia, and adopt diverse personalities, such as coach, peer, and expert. Given this diversity, CAs can be classified according to various dimensions, such as their purpose [1], delivery channel [1], input and output modalities [1], or the response generation model [7]. Thus, enhanced clarity about the definition and classification of health care CAs is needed to understand their scope and leverage their capabilities in health care settings, from user-initiated interventions for the self-management of chronic conditions to supporting patient-provider communication.

CAs are increasingly used in health care settings for patient education [8], triage and diagnosis [9,10], and delivery of physical and mental health interventions [11,12]. CAs may alleviate health care providers' burden by advising on the initial management of a specific complaint [13] or assisting in chronic disease management [14,15]. In addition, they can supplement providers' care in hybrid health care delivery models [16]. Most health care CAs follow a rule-based approach, offering developers full control over the conversation flow and the information provided [17]. In health care settings, rule-based CAs may reduce miscommunication and the risk of harm arising from inappropriate advice or inaccurate triaging [2,18].

Designing, Developing, Evaluating, and Implementing a Smartphone-Delivered, Rule-Based CA Conceptual Framework

The use of CAs in health care is a recent occurrence that often translates to poor reporting of the CA development and evaluation processes, which may hinder the reliability of the research findings. To offer a systematic and transparent approach to CA development and evaluation, we previously developed and published a novel conceptual framework for designing, developing, evaluating, and implementing a smartphone-delivered, rule-based conversational agent (DISCOVER) [7] (Figure 1). The framework offers a comprehensive yet simple guide for the development of rule-based CAs that aims to address all the different facets of the development of such complex interventions. The framework was developed using the methodology by Jabareen [19] and was informed by a scoping review of rule-based CAs in health care [1], a narrative review of conceptual frameworks for the development of mobile health interventions [7], and our experience of developing a rule-based CA prototype to support healthy lifestyle changes to prevent type 2 diabetes [20,21]. The development of the DISCOVER conceptual framework was described by Dhinagaran et al [7].

Figure 1. Conceptual framework for development, evaluation, and implementation of rule-based conversational agent in health care.
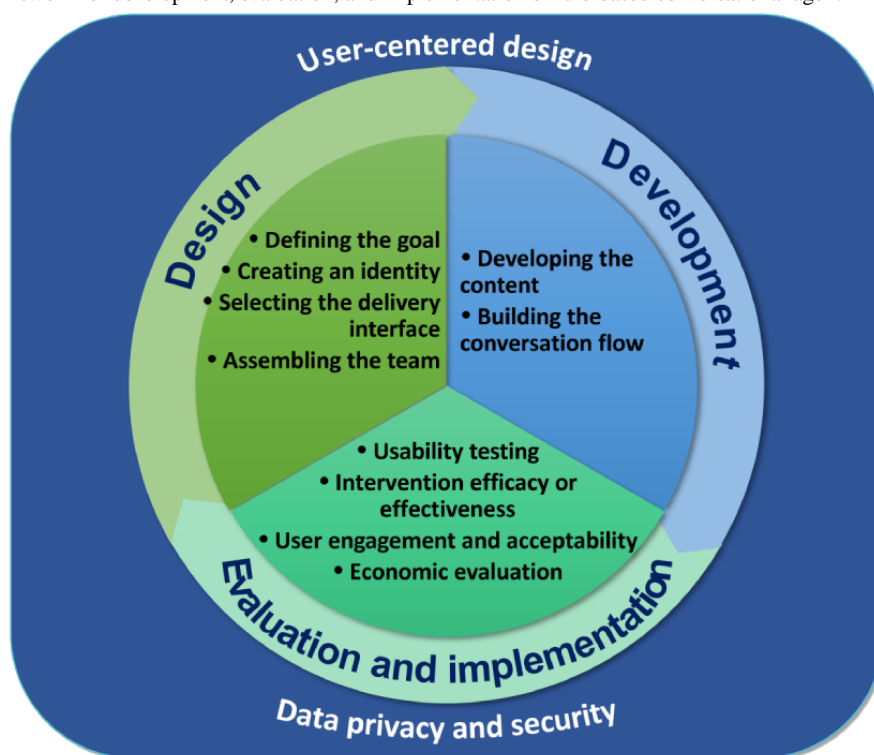
The framework consists of 3 iterative stages of the CA's design, development, and evaluation and implementation, complemented by 2 cross-cutting themes (user-centered design and data privacy and security). After development, to ensure the comprehensiveness and robustness of this framework, we validated the conceptual framework through consultation with experts in the nascent field of CA in health care [22].

Aims of the Study

We aimed to present an updated conceptual framework for developing and evaluating CAs in health care, and to suggest a revised definition and classification of health care CAs. To this end, we performed in-depth, semistructured interviews with multidisciplinary experts in CAs for health care to evaluate and validate the conceptual framework for the design, evaluation, and development and implementation of rule-based CAs.

Methods

Overview

One-on-one semistructured interviews were conducted with international experts via Zoom (Zoom Video Communications) [23]. Prospective participants were invited if they had published ≥1 peer-reviewed paper on CA interventions. Purposive sampling was used to recruit participants, who were identified through a literature search of articles and reviews on mobile health, digital health, or CA interventions and consultation with a CA expert from our team. Snowball sampling was also used to recruit additional participants. In addition, study participants were asked to provide peer recommendations for further interviews. Email invitations were sent to 50 authors from diverse fields, such as computer science, health care, and digital health, to recruit between 10 and 20 participants, as it is common practice in qualitative interviews [24] and conceptual framework

validation research [25,26]. After a positive response, follow-up emails were exchanged to fix an interview date and share further information about the study. Three days before the interview, we sent participants the informed consent form, a demographics survey, a voiceover PowerPoint presentation summarizing the conceptual framework to be discussed during the interview, and a link to join the videoconference meeting. Inclusion in the study was limited to participants who could communicate in English.

The interviews were conducted by 1 researcher (LM or AIJ) with the assistance of a second researcher (XL). LM is a pediatrician currently working full time in academic digital health research, whereas AIJ and XL hold psychology degrees with varied experience in digital health research that includes the design and development of CAs. The interviews were conducted using an interview guide with a series of illustrative questions (Multimedia Appendix 1) while acknowledging that participants' responses may add questions not previously planned. The interview guide was developed by the research team and was piloted before the start of the study. The interview was divided into 2 sections. The first section enquired about the participant's background; current role; experience with CAs; and an overview of the CA field, including the definition and classification of CAs and the advantages or disadvantages of using them in health care settings. The second section explored participants' views on the framework's overall design and individual components. The interview questions addressed the clarity and completeness of the framework, any unnecessary or missing steps, potential rearrangement of the current framework flow, and the rationale to justify each opinion. Interviews were conducted via Zoom between June and September 2022 and lasted between 60 and 90 minutes. Participants were interviewed once, except for 1 interview that was completed in a second

meeting. The interviews were audio recorded and complemented with the researcher's notes on the conversational aspects that could not be captured in the audio recording.

Ethical Considerations

This study was approved by the Institutional Review Board of Nanyang Technological University (IRB-2021-816). Informed consent was obtained from all participants before the start of the interview. Participants received a US $20 web-based retailer voucher as compensation.

Data Analysis

All interviews were transcribed verbatim by LM and XL using proprietary transcription software from the Lee Kong Chian School of Medicine Digital Learning team [27] and Microsoft Word automated transcription service [28]. Transcripts were checked for accuracy by the researchers and analyzed using thematic content analysis described by Burnard [29]. The transcribed data were analyzed sequentially, and inductive coding was used to generate the common themes and subthemes. Interview data were analyzed independently and in parallel using NVivo 12 software [30] by LM and XL. The researchers

met regularly to discuss the interview coding, and disagreements were resolved through discussion and consensus. The report followed the Standards for Reporting Qualitative Research checklist. The definition and classification of CAs were supplemented by the authors' previous and ongoing work, consisting of a series of scoping reviews on several aspects of health care CAs [1,7,31] and an analysis of definitions presented in several reviews [3,4,32-40] and a book [5].

Results

Overview

A total of 12 experts were interviewed. Most participants (9/12, 75%) had previous experience in developing rule-based CAs, whereas 25% (3/12) of the participants had developed hybrid CAs that combined rule-based decision trees with natural language processing. Table 1 presents the participants' demographic information.

The findings from the expert interviews were organized into 2 distinct sections. Multimedia Appendix 2 presents participants' quotes on the several topics discussed in the interviews.

Table 1. Participants' demographic information (N=12).

Characteristics	Values
Position, n (%)	
Professor	4 (33)
Associate or assistant professor	2 (17)
RF[a] or senior RF	3 (25)
Research associate or doctoral student	3 (25)
Workplace, n (%)	
Academic institution	9 (75)
Research institute	3 (25)
Country, n (%)	
Australia	2 (17)
China	1 (8)
Taiwan	1 (8)
Singapore	2 (17)
Switzerland	4 (33)
United States	2 (17)
Field of expertise[b], n (%)	
Computer science or artificial intelligence	2 (25)
Medical informatics, digital health, or digital mental health	8 (67)
Medicine (family medicine)	1 (8)
Psychology	2 (17)
Marketing, consumer behavior, and HCI[c]	1 (8)
Years of experience, n (%)	
<5	6 (50)
5-10	3 (25)
>10	3 (25)
Age (years), n (%)	
21-30	1 (8)
31-40	4 (33)
41-50	5 (41)
>60	2 (17)
Number of publications on CAs[d], median (range)	4.5 (2-150)
Number of publications on CAs, mean (SD)	19.6 (41.9)

[a]RF: research fellow.

[b]This section does not add up to 12, as 1 participant reported >1 field of expertise.

[c]HCI: human-computer interaction.

[d]CA: conversational agent.

Defining and Classifying CAs

All participants defined CAs as computer systems that use natural language to interact with users. In general, this definition was broad and encompassing and included voice assistants such as Siri or Alexa, which usually engage briefly with users in a "transactional" (P003) manner. At the same time, CAs often "have some kind of coherent discourse" (P003). Although some participants considered terms such as *chatbot* and *conversational agent* synonyms and used them interchangeably, most participants distinguished between them. In general, CA was regarded as a "bit broader" (P002) term that encapsulated different types of agents, including chatbots. Chatbots were seen as referring specifically to "text-based and rule-based

conversational agents" (P001). However, 1 participant distinguished CAs from chatbots and other types of dialogue systems, as CAs could engage in an empathic, personalized conversation with the user and "develop relationships" (P009) that emphasized the "social and emotional aspects of the interaction in addition to the health care tasks" (P003).

Experts classified CAs according to 13 different categories. These included commonly used ones such as CA *input modalities* (text based or speech based), *purpose* (domain specific or general purpose), or *response generation method* (rule based, artificial intelligence [AI] based, or hybrid). They also proposed other categories; for example, CAs were classified according to the *development modality* as "bespoke or off-the-shelf conversational agents" (P001) or according to the *deployment platform* as app based, often as a stand-alone CA or integrated into a website or another platform such as Facebook Messenger, Telegram, or Slack; according to the "type of communication, is it supportive, or is it just information providing" (P006). In addition, participants also categorized CAs according to their *appearance* (disembodied, embodied, or social robots), *length of interaction* (short term, medium term, and long term), *personality* (coach, peer, or expert), *type of CA-user interaction* (transactional or relational), the *inclusion of human support*, "where in the patient journey it's used" (triage, appointment management, medication adherence, others; P007), or the *domain* (health care, commerce, business, or others).

CAs in health care settings were seen as "a scalable way of delivering...personalized health services to people" (P010). Participants noted the advantages and disadvantages of using both rule-based and AI-based systems. There was widespread consensus among experts that, at present, rule-based CAs should be the norm in health care. Rule-based CAs require developers to predefine all dialogue turns, which allows greater control of the conversation flow, constituting their main advantage in health care settings. However, there are several disadvantages associated with rule-based CAs. For example, rule-based systems lack flexibility and may not address all user concerns. Moreover, CA interventions that require ongoing interactions with a user may lead to disengagement and boredom, as responses will be similar each time. Alternatively, AI-based CAs were seen as offering attractive, flexible, and innovative systems that may increase long-term user engagement. However, experts considered that current AI technology does not ensure safe and reliable conversations, particularly in health care settings. Finally, experts acknowledged that CA development is a complex, labor-intensive task, which applies to rule-based CAs and AI CAs.

A Conceptual Framework to Design, Develop, and Evaluate and Implement CAs

The Role of a Conceptual Framework for the Development of CAs in Health Care

Participants felt a conceptual framework to guide the development of rule-based CAs was useful. It provides researchers and developers with a guideline for CA development and "a benchmark for analysing it" (P007). Notably, a conceptual framework is helpful for "writing research proposals" (P003) or to "consult it to make sure I wasn't missing something" (P010). Furthermore, the framework was helpful for "developers or for companies" as they may "get caught up in the business model" and "[may not] draw from conceptual or theoretical frameworks as much as they should" (P005).

All participants agreed that the framework assessed in this study could be adapted for developing AI-based CAs, as many considerations also apply to AI-based CAs. An exception was the development of the conversation flow because AI-based CAs require the use of training data sets rather than decision trees. In addition, experts pointed out that AI-based CAs' training data should be representative of the target population, that researchers should be able to explain the algorithm's evolution, and that the CA responses should be accurate and not compromise users' safety.

Experts valued the content and layout of the existing conceptual framework. The visual presentation was considered "simple to understand and not overly complicated" (P008). Participants agreed to use a circular diagram as it represented the iterative nature of the CA development cycle. They also suggested the use of "non-standard fonts" (P008), "some more colors" (P006) to differentiate the stages, "emphasizing the arrows" (P008) to highlight the framework's iterative nature, and modifying the placement of the cross-cutting themes to avoid linking them to the specific stages of the framework. However, they suggested adding more details to each stage to make the diagram more self-explanatory. Participants had varied views about the *framework name*, appreciating that it was simple and easy to remember, although they might "not directly connect it with the framework for chatbot development" (P002).

Design of CAs

Participants agreed with the following existing concepts: defining the CA goal, determining the CA identity, identifying target users, selecting the delivery interface, and assembling a multidisciplinary team. They also added a new concept: specifying the evaluation outcomes ([Multimedia Appendix 2](#)). *Defining the CA goal* was highly relevant. Participants expressed that a clear and in-depth understanding of the health care problem is essential to "define the goal based on the problem" (P001). Still, they highlighted the relevance of defining "the central aim, why are we doing this?" (P005). Research methods used for needs assessment, such as reviews, focus group discussions, interviews, and surveys, play an important role not only in defining the goal but "for all of the content to make sure it's understandable and acceptable by [the] end audience" (P003).

The *identity of the CA* is important "in terms of who people trust and don't trust and the degree of which they think the information they're getting is trustworthy" (P009). One fundamental aspect of defining the CA identity design is to adapt the content to the local culture, including the local nuances that may make the agent more relatable to the target audience.

The *target users* refer not only to the end user of the intervention but also to the broader social circle formed by family, other members of the user's social network, and health care providers.

This is particularly important if the CA intervention will be used in the health care system or if the target population includes older adults or individuals with intensive care needs. Developers may also consider the setting where the intervention would occur.

Selecting the delivery interface is "a key choice" (P012). Researchers "need to realize that there are differences with the platform" (P010), in terms of users' acceptance of the different tools, and regarding the technical "capabilities of the delivery interface" (P008). The choice of platform may even determine the type of outcome measures to be collected during the CA evaluation.

Assembling a *multidisciplinary team* was considered essential when designing a CA-based intervention. The team may include a "linguistic professional" (P012) to ensure that the text will be adequate for the end user, content participants familiar with traditional face-to-face interventions, and digital health intervention specialists. The team composition may also depend on the nature of the research collaboration, as 1 expert pointed out, "if I'm working with [an industry stakeholder], they tell me the team" (P008). Finally, participants suggested that it is also important "to define the roles of everyone within the team because sometimes that's a bit unclear" (P007).

Experts noted that the *type of outcomes* to be measured and the time points at which these measurements may occur during the evaluation stage should be defined during the design stage. It is crucial to consider the sources of "the data that you're going to use...it's participant reported and sensor data, but also electronic medical record data, health systems data, staff data, leadership data, focus group data. So, it's broader than just participant data" (P009). Planning for sensor data collection is essential as it must be coded while developing the CA. Special care should be taken to ensure the correct spelling of data variables, as minor orthographic mistakes may render the data unavailable.

Development of CAs

Although participants agreed on the concepts included in this stage, they suggested adding additional details to the diagram to make the content more self-explanatory, particularly for inexperienced developers.

The length of the dialogues was repeatedly mentioned as a critical consideration to be discussed early in the development process. It may take different interpretations, referring to the length of each CA turn (how many lines of speech), the length of the dialogue (number of times each speaker has a turn), and the total number of sessions required to deliver the intervention. The dialogues may convey an empathic, nonjudgmental tone, as this may influence user engagement and bonding with the CA. Moreover, it is essential to consider the language that best translates the content of face-to-face interventions into the constrained length of CA dialogues.

Participants agreed that all content in the conversations should be *evidence based*, referring to the adherence to current best practices as stated in the scientific literature and clinical guidelines and the dialogue structure that should emulate "seasoned clinicians" (P009). Health care CAs must also have

adequate systems to respond to possible *medical or psychological emergencies*. For example, they may provide crisis helpline phone numbers, include advice for the initial management of common symptoms, or hand over emergency management to a health care provider.

Error management includes inadvertent adverse events and mistakes (either from the CA or a person who enters the wrong data). However, adverse events are rarely reported in CA studies, which increases the importance of collecting "information if a person deteriorates in an outcome" (P006). Moreover, 1 expert highlighted that adverse event management might be challenging if users associate the CA with their health care provider and assume that the agent would identify adverse events when they occur. Therefore, developers may include safeguards to assist users adequately if an adverse event occurs.

Personalization implies that the "chatbot can maybe store the user preferences during the conversation" (P011). Personalization is essential, and it is considered a hallmark of CAs, which may draw information from several sources to "tailor the dialogue specifically to [the user]" (P009).

Finally, when defining the *platform,* researchers should consider the size and expertise of the multidisciplinary team. Although large teams may develop the CA "from scratch" (P001), smaller teams with limited coding expertise may "use some platforms, [such as] the Alexa skills or Juji" (P001).

Evaluation and Implementation of CAs

Participants agreed with existing evaluation and implementation topics, such as usability assessment, user engagement, intervention efficacy, and effectiveness, and suggested new topics, such as technical evaluation. They also pointed out that CA evaluation is an iterative process that often starts during content development to ensure that conversational turns occur as planned and are adequate to fulfill the CA aim. In addition, during the evaluation process, researchers should consider how the test results can be used to improve the CA further.

Participants offered varying definitions of what *usability* entails. For example, usability was defined as "not just A/B testing, but qualitative studies of what users think about the agent" (P003) and referred to user experience as a dimension of usability. Concurrently, it is important to use standardized tools to measure usability, such as the System Usability Scale [41] or the Chatbot Usability Scale [42], as "researchers tend to develop their own usability test" (P002), and this may limit comparability. *User engagement* is another important consideration, particularly in longitudinal interventions.

The *technical evaluation* of CAs has not been discussed in the original framework. However, such evaluation is important to ensure that the CA is ready and functional before evaluating it in clinical trials, probably during development. It includes not only crashes and bugs but also the evaluation of the CA content credibility and data privacy and security safeguards, among others.

In general, participants agreed that, although infrequently mentioned, *economic evaluation* of the end product is essential to ensure the viability of the CA beyond clinical trials and

consider the cost-effectiveness of the intervention. Moreover, expert P002 suggested that economic evaluations could be considered in all 3 stages of the CA development process, whereas expert P003 argued that "economic evaluation is rarely done" in research settings.

Finally, 1 expert suggested modifying the framework to "having four boxes and four quadrants" as *implementation* is different from evaluation and requires "different skill sets altogether" (P009).

Cross-Cutting Themes

Participants agreed with the existing cross-cutting themes of user-centered design and data privacy and security and suggested 2 additional ones: ethics and long-term sustainability.

Including *user-centered design* was an important aspect of CA design and evaluation. However, participants clarified that user-centered design is the name of a specific design process characterized by the user's influence in the design and should not be used to conceptualize the importance of user involvement in the design process. Participants distinguished between being user centric; seeking end users' opinions through surveys, interviews, or other qualitative methods during the CA development; and directly involving users in the design and development of the CA, as seen in participatory design frameworks. However, users may express divergent views, which may be difficult to reconcile when deciding the features to be included in the intervention.

The inclusion of data *privacy and security* was welcomed by all participants, who agreed that this topic is critical and sometimes not adequately discussed in the scientific literature. Data privacy and security were perceived as complex issues with multiple connotations, including users' behavior, which could be "very paradoxical" (P008), meaning that users may assert themselves as genuinely concerned about privacy but act as if it is not important at all. Users' concerns about data management may correlate with the extent of personal data collection. They may require more detailed explanations of how researchers will use their data if the app collects large amounts of personal data. In addition, researchers need to find a "balance between getting as much data as possible... [to] make the intervention more personalized, but also [not] to cross the boundaries of what people feel comfortable with sharing" (P007). Participants also emphasized the importance of developing adequate data management plans that align with their countries' current data protection laws during the design stage. Finally, future handling of research data may be

challenging, as technology development may facilitate the reidentification of initially anonymous data.

Ethical considerations were viewed as an essential cross-cutting theme from 2 different perspectives. One expert referred to the overarching ethical principles that should guide all health care interventions, "principles like nonmaleficence, beneficence, autonomy, fidelity" (P005). In contrast, another expert mentioned that when applying for ethics approval, researchers may need to modify the project execution to suit the current ethical best practices.

Finally, expert P010 suggested assessing early in the development (and as an ongoing theme) "where [the intervention] fits in the bigger health system" to ensure that researchers develop a sustainable, cost-effective system that addresses a real and required health care need.

Discussion

Principal Findings

This study presented the views of 12 multidisciplinary CA experts on the definition, classification, and development and evaluation of CAs in health care. Experts generally distinguished CA as an overarching term that contains several types of agents, including chatbots, and they proposed 13 categories to classify CAs. Participants agreed with the overall conceptual framework for designing, developing, evaluating, and implementing health care CAs and offered suggestions to improve the framework. Experts also agreed that the framework could be adapted for the development of AI-based CAs if the technology can ensure their safety and reliability in health care settings.

Although participants offered diverse definitions, they most clearly defined CA as an encompassing term that includes all subtypes of conversational interfaces. Thus, based on the experts' descriptions of CA and our research, we propose the following definition of CA: CAs are digital interfaces that use natural language to engage in a synchronous dialogue using ≥1 communication modality, such as text, voice, images, or video. This definition includes a variety of CAs, such as transactional, single-turn voice (or virtual) assistants (eg, Siri or Alexa); text-based and often rule-based CAs or chatbots; and complex, embodied CAs able to engage in verbal and nonverbal communication with users. In addition, a subset of embodied agents, referred to as relational agents, aims to "build and maintain long-term, social-emotional relationships with their users" [43], a feature that sets them apart from other CAs (Figure 2).

Figure 2. Types of CAs according to their characteristics and functions. CA: conversational agent.
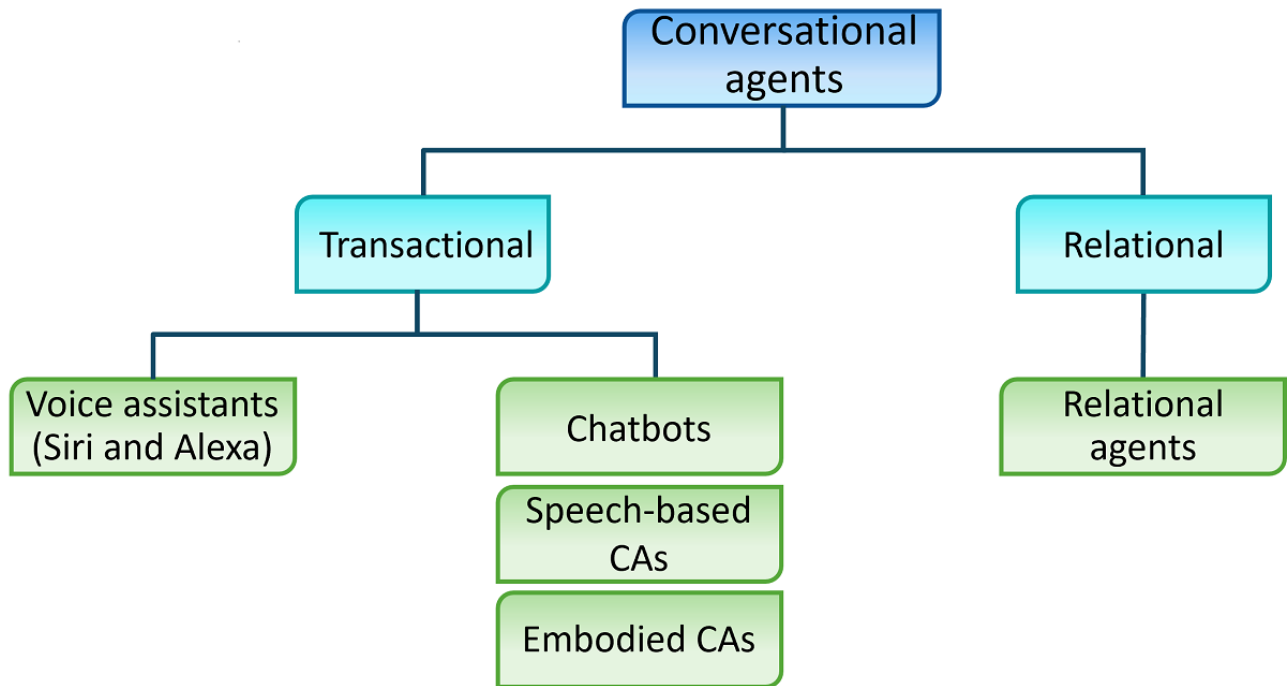
Participants suggested 13 ways to categorize CAs: domain, input modalities, purpose, response generation method, development modality, deployment platform, communication style, CA appearance, length of interaction, CA personality, type of CA-user interaction, the inclusion of human support, and where in the patient journey it is used. We compiled these categories; added a category, dialogue initiation, from the systematic review by Laranjo et al [18]; and removed the domain category as it was not specific to health care. We also expanded the CA personality category and merged the location in the patient journey to create the "type of health care intervention" category with additional information from our previous work [1]. The result was a novel classification of health care CAs,

consisting of 13 categories describing the CA appearance, communication modalities, and uses in health care settings. Denecke and May [44] recently published a technical-oriented taxonomy for health care CAs aimed at improving the reporting of the technical aspects of CA development. The authors included 18 categories grouped into 4 dimensions (agent appearance, setting, interaction, and data processing). The taxonomy included 8 categories that overlapped with our classification, including CA personality, appearance, length of interaction, response generation method, purpose, human support, input-output methods, and deployment modalities. The CA categorization is shown in Figure 3.

Figure 3. Experts' categorization of CAs. CA: conversational agent.
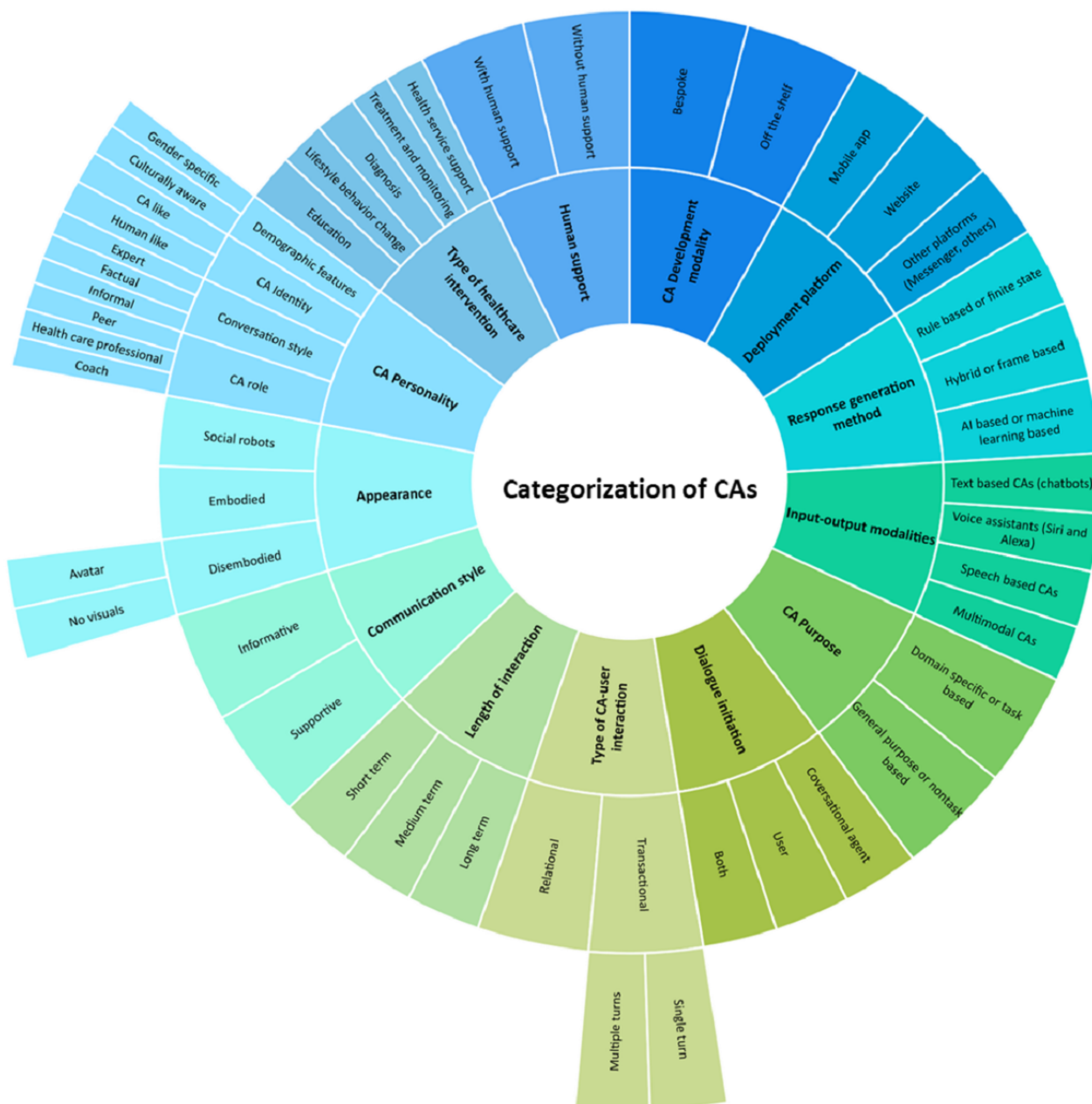
Experts' recommendations largely validated the content and structure of the previous version of the conceptual framework [7], suggesting that most elements of the original framework are congruent with participants' knowledge and experience in developing health care CAs [7]. The framework was also aligned with other frameworks guiding the design and development of digital health interventions in general [45,46]. Participants also provided valuable suggestions to improve the framework's look, content, and structure, including adding more information to the diagram to make it more self-explanatory and performing technical evaluations of the system early in the development cycle to ensure the viability of the prototype before starting costly, patient-facing tests. Furthermore, participants suggested the inclusion of ethics as a cross-cutting theme. Experts shared that the biomedical ethics principles of autonomy, nonmaleficence, beneficence, and justice [47] should guide the design of digital health care interventions. Of particular

importance are the unequal access to technology associated with inadequate digital literacy or economic disadvantage, data privacy and security breaches, and potential risks of bias and harm [47-49]. Adequate measures to reduce these risks should be considered and implemented in all stages of the CA development process.

Our conceptual framework focuses on the development of rule-based CAs. However, experts agreed that its principles could be adapted to AI-based CAs if additional guidance on topics specific to creating AI algorithms is added in the framework's development stage. Nonetheless, participants were cautious about using AI-based CAs in health care settings, given the associated risk of misunderstanding posed by systems that are unable to contextualize the conversation or understand the nuances of words and metaphors that often convey a meaning different from the textual discourse. However, the field of

conversational AI has seen significant developments over the last year, particularly with the public release of OpenAI's ChatGPT [50] in November 2022, which is a large language model that uses complex algorithms and reinforcement learning with human supervision to generate a coherent output. ChatGPT has recently been credited with passing the USMLE (United States Medical Licensing Exam), a standardized set of 3 exams required to obtain medical licensure in the United States [51]. These developments may increase the interest in adopting AI in health care settings. However, health care providers and researchers should be aware of the limitations of large language models to provide reliable and accurate responses [52], as well as issues of bias in the data [53], user safety, algorithm transparency, explainability, and liability [54-56], which are essential for ensuring the safe and reliable provision of health care.

Conceptual Framework for Health Care Conversational Agents

Participants' inputs and suggestions were adopted to develop an improved version of the conceptual framework, renamed conceptual framework for health care conversational agents (CHAT). This updated version incorporates improvements to the visual presentation and content of the framework. We also renamed the framework in response to expert comments that the previous name was not easily relatable to CAs. Visually, the structure was modified by moving the cross-cutting themes to the center to avoid linking specific cross-cutting themes to particular stages of the framework. We also emphasized the arrows illustrating the iterative process of CA development, including the term *conversational agent* within the framework, and standardizing each stage's description. Finally, following suggestions to make the framework diagram more self-explanatory, we added the development stage, and we created a checklist to supplement the information presented in Figure 4.

Figure 4. Conceptual framework for healthcare conversational agents (CHAT). CA: conversational agent; HCP: health care provider; FGD: focus group discussion; RCT: randomized controlled trial.
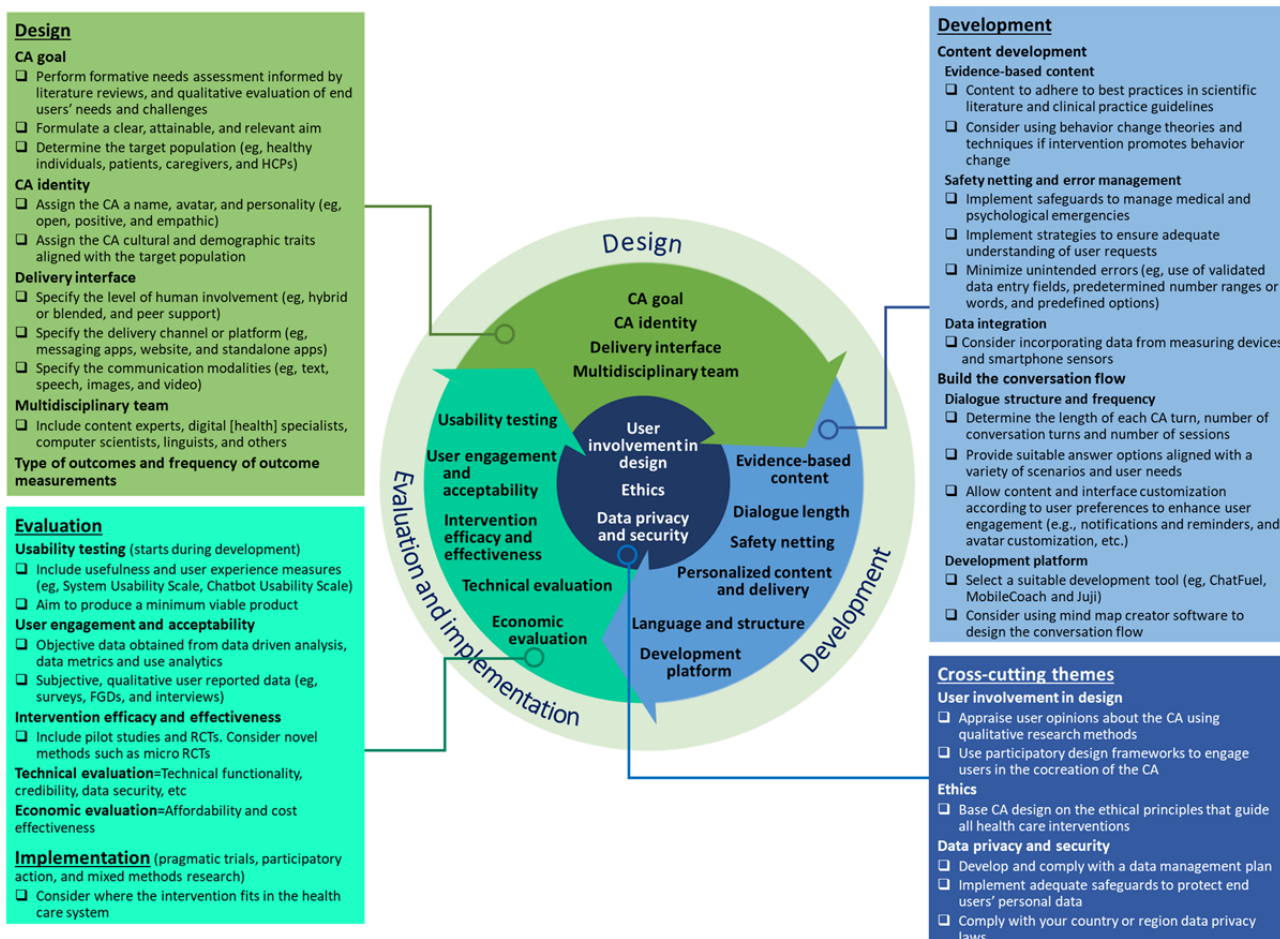
CHAT consists of 15 key topics grouped into 3 stages: the first stage, *design*, includes determining the CA goal and CA identity, selecting a delivery interface, and assembling a multidisciplinary team. The second stage, *development*, includes 2 sections: developing the content highlights the use of reliable, evidence-based sources; incorporating error management and safeguards to avoid user harm; and considering data integration with phone sensors or external devices. Building the conversation flow, considers the dialogue length, language and structure, personalization, and the CA development platform. The third stage, *evaluation and implementation* of the CA, includes the assessment of usability, user engagement,

intervention efficacy and effectiveness, as well as technical and economic evaluations. Finally, these stages are supplemented by 3 cross-cutting themes: user involvement in design, ethics, and privacy and security, which are relevant at all the framework stages.

The CHAT framework was developed to assist academic research teams of different sizes and resources in planning the design, development, and evaluation of rule-based CAs. However, the framework may also benefit developers or companies to ensure compliance with evidence-based principles.

Strengths and Limitations

This study has several strengths. First, the interviews were conducted with experts from various fields, from computer science to medicine, who shared their unique experiences working with CAs. Second, we used a comprehensive interview guide that allowed for an in-depth evaluation of the conceptual framework and the broader CA domain.

However, this study had several limitations. First, most participants were digital health or medical informatics specialists, which may have biased the study results toward more technical aspects of CA design and development. Second, most study participants worked in academic settings. Therefore, the interviews may have overlooked specific elements related to the design and development of CAs for commercial for-profit companies. Third, an early version of this conceptual framework

was used to guide the development of a healthy lifestyle CA to prevent type 2 diabetes. Further research should evaluate the use of the framework for the design of different types of CAs. Further assessment is required to evaluate the relevance of the framework steps in designing effective CA interventions. Finally, the adaptation of this framework for AI-based CAs will require more details, particularly in the development stage of the framework.

Conclusions

The use of CAs, which are complex and diverse digital health interventions in health care settings, is increasing. We invited CA experts to define, classify, and discuss the steps required to develop CAs in health care settings. According to experts, CAs are digital interfaces that use natural language to engage in a synchronous dialogue using ≥1 communication modalities such as text, voice, images, or video. CAs can be classified into 13 categories: response generation method, input and output modalities, CA purpose, deployment platform, CA development modality, appearance, length of interaction, type of CA-user interaction, dialogue initiation, communication style, CA personality, human support, and type of health care intervention. Finally, the CA development process is presented as CHAT, which consists of 3 stages of design, development, and evaluation and implementation of CAs, complemented by 3 cross-cutting themes: user involvement, data privacy and security, and ethics.

Authors' Contributions

LTC conceptualized the study; LTC and LM designed the study; LM, XL, and AIJ conducted the interviews; LM and XL transcribed the interviews and conducted the data analysis; and LM wrote the manuscript. LTC, XL, AIJ, TK, RA, and JC revised the manuscript. All authors approved the final version of the manuscript and take accountability for all aspects of the work.

Conflicts of Interest

TK is affiliated with the Centre for Digital Health Interventions, a joint initiative of the Institute for Implementation Science in Health Care, University of Zurich, the Department of Management, Technology, and Economics at ETH Zurich, and the Institute of Technology Management and School of Medicine at the University of St Gallen. The Centre for Digital Health Interventions is funded in part by CSS, a Swiss health insurer, SanusX, an Austrian health insurer, and MTIP, a Swiss digital health investor. TK is also a cofounder of Pathmate Technologies, a university spin-off company that creates and delivers digital clinical pathways. However, neither CSS, SanusX, MTIP, nor Pathmate Technologies were involved in this research. LTC is an Associate Editor for JMIR Medical Education. All other authors declare no conflict of interests.

Multimedia Appendix 1

Semistructured interviews guide.
[DOCX File , 37 KB-Multimedia Appendix 1]

Multimedia Appendix 2

Selected experts' quotes that illustrate the study findings.
[DOCX File , 25 KB-Multimedia Appendix 2]

XSL•FO

RenderX

References

1. Tudor Car L, Dhinagaran DA, Kyaw BM, Kowatsch T, Joty S, Theng YL, et al. Conversational agents in health care: scoping review and conceptual analysis. J Med Internet Res 2020 Aug 07;22(8):e17158 [FREE Full text] [doi: 10.2196/17158] [Medline: 32763886]

2. Palanica A, Flaschner P, Thommandram A, Li M, Fossat Y. Physicians' perceptions of chatbots in health care: cross-sectional web-based survey. J Med Internet Res 2019 Apr 05;21(4):e12887 [FREE Full text] [doi: 10.2196/12887] [Medline: 30950796]

3. Radziwill NM, Benton MC. Evaluating quality of chatbots and intelligent conversational agents. arXiv. Preprint posted online April 15, 2017 [FREE Full text] [doi: 10.48550/arXiv.1704.04579]

4. Vaidyam AN, Wisniewski H, Halamka JD, Kashavan MS, Torous JB. Chatbots and conversational agents in mental health: a review of the psychiatric landscape. Can J Psychiatry 2019 Jul;64(7):456-464 [FREE Full text] [doi: 10.1177/0706743719828977] [Medline: 30897957]

5. McTear M, Callejas Z, Griol D. The Conversational Interface: Talking to Smart Devices. Cham, Switzerland: Springer; 2016.

6. Lebeuf C. A taxonomy of software bots: towards a deeper understanding of software bot characteristics. University of Victoria. 2018. URL: http://hdl.handle.net/1828/10004 [accessed 2023-03-15]

7. Dhinagaran DA, Martinengo L, Ho MR, Joty S, Kowatsch T, Atun R, et al. Designing, developing, evaluating, and implementing a smartphone-delivered, rule-based conversational agent (DISCOVER): development of a conceptual framework. JMIR Mhealth Uhealth 2022 Oct 04;10(10):e38740 [FREE Full text] [doi: 10.2196/38740] [Medline: 36194462]

8. Jack BW, Bickmore T, Yinusa-Nyahkoon L, Reichert M, Julce C, Sidduri N, et al. Improving the health of young African American women in the preconception period using health information technology: a randomised controlled trial. Lancet Digit Health 2020 Sep;2(9):e475-e485 [FREE Full text] [doi: 10.1016/S2589-7500(20)30189-8] [Medline: 33328115]

9. Baker A, Perov Y, Middleton K, Baxter J, Mullarkey D, Sangar D, et al. A comparison of artificial intelligence and human doctors for the purpose of triage and diagnosis. Front Artif Intell 2020 Nov 30;3:543405 [FREE Full text] [doi: 10.3389/frai.2020.543405] [Medline: 33733203]

10. Morse KE, Ostberg NP, Jones VG, Chan AS. Use characteristics and triage acuity of a digital symptom checker in a large integrated health system: population-based descriptive study. J Med Internet Res 2020 Nov 30;22(11):e20549 [FREE Full text] [doi: 10.2196/20549] [Medline: 33170799]

11. Fitzpatrick KK, Darcy A, Vierhile M. Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (Woebot): a randomized controlled trial. JMIR Ment Health 2017 Jun 06;4(2):e19 [FREE Full text] [doi: 10.2196/mental.7785] [Medline: 28588005]

12. Gaffney H, Mansell W, Edwards R, Wright J. Manage Your Life Online (MYLO): a pilot trial of a conversational computer-based intervention for problem solving in a student sample. Behav Cogn Psychother 2014 Nov;42(6):731-746 [doi: 10.1017/S135246581300060X] [Medline: 23899405]

13. Middleton K, Butt M, Hammerla N, Hamblin S, Mehta K, Parsa A. Sorting out symptoms: design and evaluation of the'babylon check'automated triage system. arXiv. Preprint posted online June 7, 2016 [FREE Full text] [doi: 10.48550/arXiv.1606.02041]

14. Gong E, Baptista S, Russell A, Scuffham P, Riddell M, Speight J, et al. My diabetes coach, a mobile app-based interactive conversational agent to support type 2 diabetes self-management: randomized effectiveness-implementation trial. J Med Internet Res 2020 Nov 05;22(11):e20322 [FREE Full text] [doi: 10.2196/20322] [Medline: 33151154]

15. Echeazarra L, Pereira J, Saracho R. TensioBot: a Chatbot assistant for self-managed in-house blood pressure checking. J Med Syst 2021 Mar 15;45(4):54 [doi: 10.1007/s10916-021-01730-x] [Medline: 33723721]

16. Rabinowitz AR, Collier G, Vaccaro M, Wingfield R. Development of RehaBot-a conversational agent for promoting rewarding activities in users with traumatic brain injury. J Head Trauma Rehabil 2022 May;37(3):144-151 [doi: 10.1097/HTR.0000000000000770] [Medline: 35293365]

17. Mohamad Suhaili S, Salim N, Jambli MN. Service chatbots: a systematic review. Expert Syst Appl 2021 Dec;184:115461 [FREE Full text] [doi: 10.1016/j.eswa.2021.115461]

18. Laranjo L, Dunn AG, Tong HL, Kocaballi AB, Chen J, Bashir R, et al. Conversational agents in healthcare: a systematic review. J Am Med Inform Assoc 2018 Sep 01;25(9):1248-1258 [FREE Full text] [doi: 10.1093/jamia/ocy072] [Medline: 30010941]

19. Jabareen Y. Building a conceptual framework: philosophy, definitions, and procedure. Int J Qual Methods 2009 Dec;8(4):49-62 [FREE Full text] [doi: 10.1177/160940690900800406]

20. Dhinagaran DA, Sathish T, Soong A, Theng YL, Best J, Tudor Car L. Conversational agent for healthy lifestyle behavior change: web-based feasibility study. JMIR Form Res 2021 Dec 03;5(12):e27956 [FREE Full text] [doi: 10.2196/27956] [Medline: 34870611]

21. Dhinagaran DA, Sathish T, Kowatsch T, Griva K, Best JD, Tudor Car L. Public perceptions of diabetes, healthy living, and conversational agents in Singapore: needs assessment. JMIR Form Res 2021 Nov 11;5(11):e30435 [FREE Full text] [doi: 10.2196/30435] [Medline: 34762053]

XSL·FO

RenderX

22. Pfadenhauer M. At eye level: the expert interview — a talk between expert and quasi-expert. In: Bogner A, Littig B, Menz W, editors. Interviewing Experts. London, UK: Palgrave Macmillan; 2009.

23. Zoom. Zoom Video Communications. URL: https://zoom.us/ [accessed 2022-06-01]

24. Vasileiou K, Barnett J, Thorpe S, Young T. Characterising and justifying sample size sufficiency in interview-based studies: systematic analysis of qualitative health research over a 15-year period. BMC Med Res Methodol 2018 Nov 21;18(1):148 [FREE Full text] [doi: 10.1186/s12874-018-0594-7] [Medline: 30463515]

25. Martin P, Alberti C, Gottot S, Bourmaud A, de La Rochebrochard E. Expert opinions on web-based peer education interventions for youth sexual health promotion: qualitative study. J Med Internet Res 2020 Nov 24;22(11):e18650 [FREE Full text] [doi: 10.2196/18650] [Medline: 33231552]

26. Mauco KL, Scott RE, Mars M. Development of an eHealth readiness assessment framework for Botswana and other developing countries: interview study. JMIR Med Inform 2019 Aug 22;7(3):e12949 [FREE Full text] [doi: 10.2196/12949] [Medline: 31441429]

27. Transcription software. Lee Kong Chian School of Medicine Digital Learning. URL: https://sites.google.com/ntu.edu.sg/digitallearning/home/speech-to-text?authuser=0 [accessed 2022-07-01]

28. Microsoft Word automated transcription service. Microsoft. URL: https://support.microsoft.com/en-us/office/transcribe-your-recordings-7fc2efec-245e-45f0-b053-2a97531ecf57 [accessed 2023-10-10]

29. Burnard P. A method of analysing interview transcripts in qualitative research. Nurse Educ Today 1991 Dec;11(6):461-466 [doi: 10.1016/0260-6917(91)90009-y] [Medline: 1775125]

30. NVivo 12. Lumivero. URL: https://www.qsrinternational.com/nvivo-qualitative-data-analysis-software/home [accessed 2023-10-10]

31. Jabir AI, Martinengo L, Lin X, Torous J, Subramaniam M, Tudor Car L. Evaluating conversational agents for mental health: scoping review of outcomes and outcome measurement instruments. J Med Internet Res 2023 Apr 19;25:e44548 [FREE Full text] [doi: 10.2196/44548] [Medline: 37074762]

32. Lebeuf C, Storey MA, Zagalsky A. Software Bots. IEEE Softw 2018 Jan;35(1):18-23 [FREE Full text] [doi: 10.1109/ms.2017.4541027]

33. Sezgin E, Militello LK, Huang Y, Lin S. A scoping review of patient-facing, behavioral health interventions with voice assistant technology targeting self-management and healthy lifestyle behaviors. Transl Behav Med 2020 Aug 07;10(3):606-628 [doi: 10.1093/tbm/ibz141] [Medline: 32766865]

34. Dowling M, Rickwood D. Online counseling and therapy for mental health problems: a systematic review of individual synchronous interventions using chat. J Technol Hum Serv 2013;31(1):1-21 [FREE Full text] [doi: 10.1080/15228835.2012.728508]

35. Bendig E, Erb B, Schulze-Thuesing L, Baumeister H. The next generation: chatbots in clinical psychology and psychotherapy to foster mental health – a scoping review. Verhaltenstherapie 2019 Aug 20;32(Suppl. 1):64-76 [FREE Full text] [doi: 10.1159/000501812]

36. Lim SM, Shiau CW, Cheng LJ, Lau Y. Chatbot-delivered psychotherapy for adults with depressive and anxiety symptoms: a systematic review and meta-regression. Behav Ther 2022 Mar;53(2):334-347 [doi: 10.1016/j.beth.2021.09.007] [Medline: 35227408]

37. Provoost S, Lau HM, Ruwaard J, Riper H. Embodied conversational agents in clinical psychology: a scoping review. J Med Internet Res 2017 May 09;19(5):e151 [doi: 10.2196/jmir.6553] [Medline: 28487267]

38. Abdellatif A, Badran K, Costa DE, Shihab E. A comparison of natural language understanding platforms for chatbots in software engineering. IIEEE Trans Software Eng 2022 Aug 1;48(8):3087-3102 [FREE Full text] [doi: 10.1109/TSE.2021.3078384]

39. McTear MF. Spoken dialogue technology: enabling the conversational user interface. ACM Comput Surv 2002 Mar;34(1):90-169 [doi: 10.1145/505282.505285]

40. Agarwal R, Wadhwa M. Review of state-of-the-art design techniques for chatbots. SN Comput Sci 2020 Jul 29;1(5):246 [FREE Full text] [doi: 10.1007/s42979-020-00255-3]

41. Brooke J. SUS: a 'quick and dirty' usability scale. In: Jordan PW, Thomas B, McClelland IL, Weerdmeester B, editors. Usability Evaluation in Industry. London, UK: CRC Press; 1996.

42. Borsci S, Malizia A, Schmettow M, van der Velde F, Tariverdiyeva G, Balaji D, et al. The Chatbot Usability Scale: the Design and Pilot of a usability scale for interaction with AI-based conversational agents. Pers Ubiquit Comput 2021 Jul 21;26(1):95-119 [FREE Full text] [doi: 10.1007/s00779-021-01582-9]

43. Bickmore T, Schulman D, Yin L. Maintaining engagement in long-term interventions with relational agents. Appl Artif Intell 2010 Jul 01;24(6):648-666 [FREE Full text] [doi: 10.1080/08839514.2010.492259] [Medline: 21318052]

44. Denecke K, May R. Investigating conversational agents in healthcare: application of a technical-oriented taxonomy. Procedia Comput Sci 2023;219:1289-1296 [FREE Full text] [doi: 10.1016/j.procs.2023.01.413]

45. Mummah SA, Robinson TN, King AC, Gardner CD, Sutton S. IDEAS (Integrate, Design, Assess, and Share): a framework and toolkit of strategies for the development of more effective digital interventions to change health behavior. J Med Internet Res 2016 Dec 16;18(12):e317 [FREE Full text] [doi: 10.2196/jmir.5927] [Medline: 27986647]

46. Collins LM, Murphy SA, Nair VN, Strecher VJ. A strategy for optimizing and evaluating behavioral interventions. Ann Behav Med 2005 Aug;30(1):65-73 [doi: 10.1207/s15324796abm3001_8] [Medline: 16097907]

47. Jumelle AK, Ispas I. Ethical issues in digital health. In: Fricker SA, Thümmler C, Gavras A, editors. Requirements Engineering for Digital Health. Cham, Switzerland: Springer; 2015:75-93

48. Luxton DD. Ethical implications of conversational agents in global public health. Bull World Health Organ 2020 Apr 01;98(4):285-287 [FREE Full text] [doi: 10.2471/BLT.19.237636] [Medline: 32284654]

49. Brall C, Schröder-Bäck P, Maeckelberghe E. Ethical aspects of digital health from a justice point of view. Eur J Public Health 2019 Oct 01;29(Supplement_3):18-22 [FREE Full text] [doi: 10.1093/eurpub/ckz167] [Medline: 31738439]

50. ChatGPT: optimizing language models for dialogue. OpenAI. URL: https://openai.com/blog/chatgpt/ [accessed 2023-01-25]

51. Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepaño C, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. PLOS Digit Health 2023 Feb;2(2):e0000198 [FREE Full text] [doi: 10.1371/journal.pdig.0000198] [Medline: 36812645]

52. Darcy A. Why generative AI is not yet ready for mental healthcare. Woebot Health. 2023. URL: https://woebothealth.com/why-generative-ai-is-not-yet-ready-for-mental-healthcare/ [accessed 2023-04-18]

53. Mbakwe AB, Lourentzou I, Celi LA, Mechanic OJ, Dagan A. ChatGPT passing USMLE shines a spotlight on the flaws of medical education. PLOS Digit Health 2023 Feb;2(2):e0000205 [FREE Full text] [doi: 10.1371/journal.pdig.0000205] [Medline: 36812618]

54. McGreevey JD3, Hanson CW3, Koppel R. Clinical, legal, and ethical aspects of artificial intelligence-assisted conversational agents in health care. JAMA 2020 Aug 11;324(6):552-553 [doi: 10.1001/jama.2020.2724] [Medline: 32706386]

55. Ruane E, Birhane A, Ventresque A. Conversational AI: social and ethical considerations. In: Proceedings of the 27th AIAI Irish Conference on Artificial Intelligence and Cognitive Science. 2019 Presented at: AIAI '19; December 5-6, 2019; Galway, Ireland p. 104-115 URL: https://ceur-ws.org/Vol-2563/aics_12.pdf

56. Ghassemi M, Oakden-Rayner L, Beam AL. The false hope of current approaches to explainable artificial intelligence in health care. Lancet Digit Health 2021 Nov;3(11):e745-e750 [FREE Full text] [doi: 10.1016/S2589-7500(21)00208-9] [Medline: 34711379]

Abbreviations

AI: artificial intelligence
CA: conversational agent
CHAT: conceptual framework for health care conversational agents
DISCOVER: designing, developing, evaluating, and implementing a smartphone-delivered, rule-based conversational agent
USMLE: United States Medical Licensing Exam

XSL•FO
RenderX