

## Review

# Design and Evaluation Challenges of Conversational Agents in Health Care and Well-being: Selective Review Study

Ahmet Baki Kocaballi<sup>1,2\*</sup>, MSc, PhD; Emre Sezgin<sup>3,4\*</sup>, PhD; Leigh Clark<sup>5\*</sup>, PhD; John M Carroll<sup>6</sup>, PhD; Yungui Huang<sup>4</sup>, MBA, PhD; Jina Huh-Yoo<sup>7</sup>, MHCI, PhD; Junhan Kim<sup>8</sup>, MEng; Rafal Kocielnik<sup>9</sup>, PhD; Yi-Chieh Lee<sup>10</sup>, PhD; Lena Mamykina<sup>11</sup>, PhD; Elliot G Mitchell<sup>11,12</sup>, PhD; Robert J Moore<sup>13</sup>, PhD; Prasanth Murali<sup>14</sup>, MS; Elizabeth D Mynatt<sup>14</sup>, PhD; Sun Young Park<sup>8,15</sup>, PhD; Alessandro Pasta<sup>16,17</sup>, MSc, PhD; Deborah Richards<sup>18</sup>, BBus, MAppSc, PhD; Lucas M Silva<sup>19</sup>, MSc; Diva Smriti<sup>7</sup>, MSc; Brendan Spillane<sup>20</sup>, PhD; Zhan Zhang<sup>21</sup>, PhD; Tamara Zubatiy<sup>22</sup>

<sup>1</sup>School of Computer Science, University of Technology Sydney, Sydney, New South Wales, Australia

<sup>2</sup>Centre for Health Informatics, Macquarie University, Sydney, New South Wales, Australia

<sup>3</sup>Center for Biobehavioral Health, The Abigail Wexner Research Institute, Nationwide Children's Hospital, Columbus, OH, United States

<sup>4</sup>IT Research & Innovation, The Abigail Wexner Research Institute, Nationwide Children's Hospital, Columbus, OH, United States

<sup>5</sup>Bold Insight UK, London, United Kingdom

<sup>6</sup>College of Information Sciences and Technology, Pennsylvania State University, University Park, PA, United States

<sup>7</sup>College of Computing and Informatics, Drexel University, Philadelphia, PA, United States

<sup>8</sup>School of Information, University of Michigan, Ann Arbor, MI, United States

<sup>9</sup>California Institute of Technology, Pasadena, CA, United States

<sup>10</sup>Department of Computer Science, National University of Singapore, Singapore City, Singapore

<sup>11</sup>Department of Biomedical Informatics, Columbia University, New York City, NY, United States

<sup>12</sup>Steele Institute for Health Innovation, Geisinger, Danville, PA, United States

<sup>13</sup>IBM Research - Almaden, San Jose, CA, United States

<sup>14</sup>Khoury College of Computer Sciences, Northeastern University, Boston, MA, United States

<sup>15</sup>School of Art and Design, University of Michigan, Ann Arbor, MI, United States

<sup>16</sup>Cognitive Systems Section, Department of Applied Mathematics and Computer Science, Technical University of Denmark, Kongens Lyngby, Denmark

<sup>17</sup>Demant A/S, Smørum, Denmark

<sup>18</sup>School of Computing, Macquarie University, Sydney, New South Wales, Australia

<sup>19</sup>Department of Informatics, University of California, Irvine, Irvine, CA, United States

<sup>20</sup>ADAPT Centre, University College Dublin, Dublin, Ireland

<sup>21</sup>School of Computer Science and Information Systems, Pace University, New York City, NY, United States

<sup>22</sup>Georgia Institute of Technology, Atlanta, GA, United States

\* these authors contributed equally

## Corresponding Author:

Ahmet Baki Kocaballi, MSc, PhD  
School of Computer Science  
University of Technology Sydney  
Room CB11.7.103  
81 Broadway  
Sydney, New South Wales, 2007  
Australia  
Phone: 61 295141103  
Email: [baki.kocaballi@uts.edu.au](mailto:baki.kocaballi@uts.edu.au)

## Abstract

**Background:** Health care and well-being are 2 main interconnected application areas of conversational agents (CAs). There is a significant increase in research, development, and commercial implementations in this area. In parallel to the increasing interest, new challenges in designing and evaluating CAs have emerged.

**Objective:** This study aims to identify key design, development, and evaluation challenges of CAs in health care and well-being research. The focus is on the very recent projects with their emerging challenges.

**Methods:** A review study was conducted with 17 invited studies, most of which were presented at the ACM (Association for Computing Machinery) CHI 2020 conference workshop on CAs for health and well-being. Eligibility criteria required the studies to involve a CA applied to a health or well-being project (ongoing or recently finished). The participating studies were asked to report on their projects' design and evaluation challenges. We used thematic analysis to review the studies.

**Results:** The findings include a range of topics from primary care to caring for older adults to health coaching. We identified 4 major themes: (1) Domain Information and Integration, (2) User-System Interaction and Partnership, (3) Evaluation, and (4) Conversational Competence.

**Conclusions:** CAs proved their worth during the pandemic as health screening tools, and are expected to stay to further support various health care domains, especially personal health care. Growth in investment in CAs also shows the value as a personal assistant. Our study shows that while some challenges are shared with other CA application areas, safety and privacy remain the major challenges in the health care and well-being domains. An increased level of collaboration across different institutions and entities may be a promising direction to address some of the major challenges that otherwise would be too complex to be addressed by the projects with their limited scope and budget.

(*J Med Internet Res* 2022;24(11):e38525) doi: [10.2196/38525](https://doi.org/10.2196/38525)

## KEYWORDS

conversational interfaces; conversational agents; dialog systems; health care; well-being

## Introduction

Conversational agents (CAs) are applications that facilitate human-computer interaction through natural language. Automatic speech recognition (ASR) and natural language processing (NLP) models help to interpret human language and produce appropriate responses [1]. CAs (also widely known as chatbots, virtual assistants, dialog systems, or voice assistants) are used in several domains such as e-commerce, scheduling services, and question-answer systems [2]. The user-system interaction could be over text (eg, SMS text messaging over an app or web service), voice (eg, interactive voice response via phone calls, voice assistants via a smartphone or smart speaker), or multimodal (eg, visual, text, and audio feedback and interaction via a smartphone, smart speaker, or any other smart and internet of things devices) [3].

CAs have been variously used and studied in health care for supporting behavioral health and healthy living [3,4]; health information seeking [5-9]; appointment, medication, symptom tracking and chronic condition management [10-12]; and facilitating COVID-19 screening and information sharing [13,14]. Mobile phone ownership enables and increases the potential applications, availability, and access to CAs in practice. Mobile phone ownership is around 15 billion worldwide as of 2021 [15], and 97% of US adults own a mobile device [16]. Current studies and randomized trials showed that CAs could be effectively used in health care delivery and improving health outcomes, such as improving mental health [17], maternal health [18], and healthy behaviors [19]. In addition, there is an increasing investment in chatbots in the health care industry; some examples include Woebot, Babylon, and ADA Health [20].

Despite the increasing interest in using CAs in supporting health care and well-being, there are many challenges in the development, deployment, and use of CAs. Recent review studies have highlighted some of the challenges including NLP

[21,22], patient safety [1,23,24], integration with other technologies [22], information dissemination [25,26], medico-legal issues [27], and ethics [28]. In response to that, recent workshops have explored the challenges and opportunities of conversational user interfaces [29-31] and the design and evaluation of CAs in health care [32]. Here we contribute to this developing literature by reporting a self-assessment of 17 such projects.

In May 2020, a workshop entitled “Conversational Agents for Health and Well-being” was held at the *ACM (Association for Computing Machinery) Conference on Human Factors in Computing Systems (CHI 2020)* [32]. The aim was to understand the most current challenges recent research projects face and devise potential directions for future research to address those challenges. The workshop included completed or ongoing projects from 30 participants in 5 countries, covering various topics from supporting older adults to mental health and coaching to supporting everyday health. Following the workshop, participants were invited to collectively report the design and evaluation challenges of CAs in health care to provide researchers, designers, and health care professionals practical perspectives on these challenges. This paper aims to present the challenges of designing and evaluating CAs derived from recent health care projects conducted in the last 2 years.

## Methods

We followed a selective review study design focusing on the challenges of recent studies on CAs in health and well-being. Coauthors were invited to report their original CA research in health care and well-being, outlining major challenges in the design and evaluation of the CA being used in their research. Coauthors were asked to report on design and evaluation challenges they faced in their project. In their written report, each coauthor or author group (1) described their research, (2) explained challenges (limited to 3 major design challenges and 3 major evaluation challenges), (3) explained how they

addressed the challenges or how they plan to address, and (4) support their findings and suggestions with prior literature. Only the information provided through the written report was analyzed as a case study. Each case study went through an open peer review process among authors and was revised. Finalized cases were analyzed by 3 coauthors (ABK, ES, and LC). We

used thematic analysis to identify, assess, and analyze the patterns in the cases [33]. The following steps were used in the analysis process: (1) familiarizing with the data, (2) generating initial codes, (3) searching for themes, (4) reviewing and refining themes, (5) defining and naming themes, and (6) reporting the findings (Textbox 1).

**Textbox 1.** Steps used in the analysis process.

1. Familiarizing with data	<ul style="list-style-type: none"><li>To gain familiarity with the data and understand the depth and breadth of the content, coauthors (ABK, ES, and LC) read and re-read the case studies.</li></ul>
2. Generating initial code	<ul style="list-style-type: none"><li>Following an open-coding approach (without predefined codes, developed, and modified during the coding process), coauthors (ABK, ES, and LC) created initial codes independently. They reviewed the codes iteratively. The codes were compared, and group decisions and consensus created the finalized codes. Coauthors used Google Sheets to create codebooks.</li></ul>
3. Searching for themes	<ul style="list-style-type: none"><li>The codes were sorted at first to understand the frequency of occurrence. They were reviewed to find patterns and grouped into the themes collectively by the coauthors (ABK, ES, and LC). Each theme was labeled to guide the grouping and reviewed by the coauthors (ABK, ES, and LC) iteratively. Similar to the coding, coauthors reached a consensus to finalize themes.</li></ul>
4. Reviewing and refining themes	<ul style="list-style-type: none"><li>All themes were reviewed. The relationships between codes and themes were discussed by the coauthors (ABK, ES, and LC). Some of the themes were combined that were found to be related; for example, obtaining domain information and training data were combined into Domain Information and Training. Themes were reviewed and finalized for consistency regarding their content, by consensus.</li></ul>
5. Defining and naming themes	<ul style="list-style-type: none"><li>Definitions of the themes were created regarding codes, subthemes, and corresponding cases. In case of disagreements, coauthors reviewed themes to ensure consensus in theme content, definition, and labeling.</li></ul>
6. Reporting the findings	<ul style="list-style-type: none"><li>Thematic analysis results were reported through a chart with themes, subthemes, definitions, associated cases, and frequency of occurrences.</li></ul>

## Results

### Overview

The review included 17 studies covering many domains from primary care to caring for older adults to health coaching. A

summary of the projects is presented in Table 1, with a unique project number for each project to be referenced in the presentation of results. The thematic analysis identified 4 major themes: (1) Domain Information and Integration, (2) User-System Interaction and Partnership, (3) Evaluation, and (4) Conversational Competence (Table 2).

**Table 1.** A summary of participating projects, including titles, domain, CA<sup>a</sup> purpose, and CA input and output modalities.

Project number	Project title	Health/well-being domain	CA purpose	CA input/output	Project status
1	Digital Scribe: A Wizard of Oz Study	Primary care	Work with general practitioners to document patient information in a consultation	Written and spoken/written, spoken, and visual	Ongoing
2	Speech Diversity and Speech Interfaces – Considering an Inclusive Future Through Stammering [34]	Accessibility and inclusivity/speech diversity	All speech-based CAs (nonspecific projects)	Spoken/written, spoken, and visual	Ongoing
3	ADELE: An Artificial Conversational Care Agent for the Elderly [35-37]	Care of the elderly in their homes	Provide health and well-being care, advice, and monitoring	Written/written	Completed
4	Talk the Talk: How Human Conversational Agents Build Trust [38-41]	People with visual impairments	Support navigation and other activities by having the person with visual impairment share their smartphone camera feed. Based on the camera feed and conversational interaction, plus use of online tools such as Google Maps, a remotely located sighted person provides guidance to the person with visual impairment.	Spoken/spoken	Ongoing
5	Empowering Older Adults With Mild Cognitive Impairment and Their Caregivers Using Conversational Agents	Older adults with MCI <sup>b</sup>	Empower the person with MCI as well as their caregiver; amplify caregiver	Spoken/spoken and visual	Ongoing
6	Adaptive Conversational Agents for the Health Care Enterprise	Health care enterprise, payer	Help users answer analytical questions about health care enterprises	Written and spoken/written, spoken, and visual	Completed
7	Encouraging Users' Self-disclosure With a Chatbot Mediator [42-44]	Mental health care	Collect truthful self-disclosure and deliver guidance	Written/written	Ongoing
8	Motivational Interviewing Conversational Agent (MI-CA) [45,46]	Family eating habits	Deliver a counseling method named motivational interviewing in an automated manner to help parents eat healthier along with their children	Spoken/spoken	Completed
9	Conversational Agent for African Americans With Chronic Illnesses [47,48]	Chronic illness	Deliver health information on COVID-19 to African Americans with chronic illnesses	Written/written	Completed
10	Exploring Voice Assistants in Multimodal Food Journaling [49,50]	Food journaling	Entries in food journal	Spoken/written and spoken	Completed
11	t2.coach: A Chatbot Health Coach for Diabetes Self-management [51,52]	Health coaching	Health coaching and goal setting for type 2 diabetes self-management	Written/written and visual	Ongoing
12	Designing Audiologist Bots Fusing Soundscapes and User Feedback	Hearing health care	Recommending personalized hearing aid settings by gathering user feedback in real-world environments	Written and spoken/written, spoken, and visual	Ongoing
13	Utilization of Self-Diagnosis Health Chatbots in the Wild: A Case Study [53]	Self-diagnosis	Offering medical advice (eg, diagnostic suggestions) to patients based on their input (eg, symptoms)	Written and spoken/written	Completed
14	eADVICE: Providing Specialist Treatment Advice to Patients on Waiting List [54]	Children and their families referred to incontinence and sleep clinic awaiting appointment	Discuss treatments to encourage adherence	Written/written and spoken	Ongoing

Project number	Project title	Health/well-being domain	CA purpose	CA input/output	Project status
15	Symptom and Health Events Tracking at Home for Children With Special Health Care Needs (CSHCN) Using Conversational Agents [55,56]	Documentation and care coordination support for children with special health care needs	Assisting caregivers and patients with tracking and communicating symptoms and health events outside of clinical settings to reduce documentation burden and facilitate care coordination	Written and spoken/written and spoken	Completed
16	Palliative Care With Spiritual Support by Conversational Agents	Elderly care, spiritual support, human-agent/robot interaction	Development of a CA that provides end of life planning and spiritual counseling	Written/spoken and visual	Ongoing
17	HarborBot: A Chatbot for Social Needs Screening [57,58]	Social needs screening in public hospital emergency departments	Collect high-quality social needs data from vulnerable populations while increasing engagement	Written/written and spoken	Completed

<sup>a</sup>CA: conversational agent.

<sup>b</sup>MCI: mild cognitive impairment.

**Table 2.** A summary of themes and subthemes and the studies reporting them.

Themes and subthemes	Reported by
<b>Domain Information and Integration</b>	
Domain Information and Training	P2 <sup>a</sup> , P6, P10, P11, P13, P14, P17
Integration and Infrastructure	P11, P14, P15
<b>User-System Interaction and Partnership</b>	
Personalization	P3, P5, P6, P8, P9, P11, P12, P13, P14
Relationship Building	P3, P4, P7, P12, P13, P16, P17
Safety and Privacy	P3, P5, P7, P8, P9, P10, P13
User Engagement	P4, P7, P9, P10, P13, P14
<b>Evaluation</b>	
Methodological Limitations	P1, P8, P11, P12
Experimental Limitations	P1, P3, P5, P12, P13, P14, P17
Lack of Guidance on Evaluation	P1, P2, P9, P13, P15
<b>Conversational Competence</b>	
Topic Detection and ASR <sup>b</sup>	P1, P2, P3, P8, P10, P12, P15
Discoverability and Conversational Interaction Model	P10, P11, P1, P2, P17
Accessibility and Inclusivity	P2, P5, P8, P9, P14, P16, P17

<sup>a</sup>P: project.

<sup>b</sup>ASR: automatic speech recognition.

## Domain Information and Integration

### Overview and Subthemes

Health care CAs often operate within particular domains of health care that require the integration of domain-specific information and language. This theme is concerned with the challenges of obtaining the required problem domain information, collecting data to train CAs, and integrating CAs with the existing systems and infrastructures. For example, a CA to be designed for helping general practitioners in the primary care domain needs to be trained by a large number of doctor-patient conversations. Obtaining this kind of information

is challenging and resource intensive. In addition, the expert medical knowledge needs to be translated into conversational format. Similarly, CAs cannot be developed as isolated applications: they need to be integrated with the existing systems and infrastructure.

### Domain Information and Training

Two projects (P6 and P14) reported on the difficulties of obtaining domain information, in these cases, for health care enterprise and medical adherence communications. These difficulties are associated with the lack of time of domain experts (P6), the domain information being not in a



conversational format (P6) or being distributed across many subdomains (P6), and the knowledge acquisition bottleneck (P14). While P6 recruited subject matter experts to review conversation flows and response frames to format health care data for conversation-like interactions, P14 implemented incremental acquisition from the expert or user to alleviate the problem of obtaining domain-specific information.

### ***Integration and Infrastructure***

Three studies reported some challenges associated with integrating CAs into existing infrastructures. One project (P15) using CAs to support health care tracking of children with special needs at home explained the security and privacy challenges of integrating CA-collected information into clinical systems and workflows. Using interoperability standards (eg, Fast Healthcare Interoperability Resources) to share caregiver notes with health care professionals was suggested as one of the fundamental approaches for integration (P15). Many CAs are developed using underlying development platforms such as Amazon Alexa Skills or Google Dialogflow. P14 highlighted the difficulty in adapting to the rapidly evolving platforms that can potentially make the CAs designed according to the previous platform functionality stop working. Finally, P11 discussed the tension between the ubiquity and richness of the underlying platforms. For example, while SMS-based text messaging is ubiquitous, it is purely text based and does not offer some of the multimodal interaction options of other messaging platforms such as Facebook Messenger or WhatsApp (eg, 1-click suggested response buttons or carousel menus). P11 used SMS text message-based messaging to support a higher degree of accessibility and ubiquity; however, to expand the richness of the user experience they included more immersive education content with multimedia messages including infographics to elaborate on each health goal option.

## **User-System Interaction and Partnership**

### ***Overview and Subthemes***

This was one of the major themes that captured several challenges related to the characteristics and qualities of interaction between users and CAs, and the ways in which users and CAs work together. The subthemes included Personalization of CAs, Relationship Building Between CAs and Users, Safety and Privacy, and User Engagement. Some common challenges were supporting users' trust in CAs, enabling CAs to show empathy, and ensuring users' privacy.

### ***Personalization***

Personalization emerged as an essential design feature as part of the projects. Personalization was reported as a challenge in 10 projects. The personalization challenges were providing appropriate responses based on users' context (P8 and P3), tailoring conversations for different user groups with a broad age range and different health literacy levels (P14, P6, P9, P11, and P12), and minimizing the question overload in surveys (P13). While 1 project reported on the difficulties in evaluating the effects on adaptive features (P6), another project presented the challenges associated with designing for dyads (P5), which include a patient and a caregiver. P5 explained that dyads using the same device pose several challenges, including dyads'

difficulties managing technology, dyads' overall wide range of technology literacy, and how to support both members of the dyad through the functionality of the CA.

### ***Relationship Building***

Humans and CAs may have different forms of relationships ranging from very short-term relationships typically characterized by one-off task-based conversational exchanges to long-term relationships in which CAs have longer conversational interactions across different topics over a longer period. Relationship building—how a human-CA relationship is established and maintained—was reported as a challenge in 13 projects. Trust (P3, P4, P7, P13, and P16), empathy (P4, P17, and P16), self-disclosure (P7), and transparency (P12) were presented as important dimensions of relationship building. One project focusing on designing a CA for people with visual impairment (P4) explains the importance of CAs to show empathy in establishing and strengthening trust between the sighted Aira agent and the client with visual impairment. According to P4, to incorporate empathy into conversational interaction, an agent must represent not only the situation itself (eg, what objects are present, their spatial relationships, movement vectors) but also the other agent's experience and interpretation of the situation. A project designing a CA for older adults' care (P3) describes the difficulties of building trust due to concerns about security, legal issues, the sharing and storing of personal and sensitive information, and privacy and ethical concerns, among others [59]. Issues with any of these and other concerns have the potential to damage the relationship critically or terminally between the CA and the patient. To address this challenge, P3 incorporated trust-building and repair strategies from the onset. Previous research has shown that reliability is likely the preeminent factor in building trust between the patient and the artificial care agent [60,61]. By contrast, other factors such as competence, benevolence, and integrity are also likely to significantly influence [62]. Transferring users' self-disclosure was reported as a challenge in human-in-the-loop artificial intelligence systems where CAs mediate between a user and a domain expert (P7).

### ***Safety and Privacy***

Safety of user data and privacy of personal information shared with CAs have influenced the user decision and perception toward CAs. Six projects reported users' safety (P3 and P13) and privacy (P5, P7, P8, and P10) as challenges. P10 explained that while some users were worried that using the voice assistant would disturb others, other users felt discomfort or concerned with privacy when other people could hear them track their food. To tackle this issue, some suggested solutions in P10 included adopting other devices or modalities for input (eg, taking picture or text input using phone or web), as well as implementing food "template" features for quick commands (eg, saying "Alexa, journal number 1" or "Google, journal same breakfast as yesterday"). In the cases of CAs as mediators between 2 users (P7) or multiple users (P5), privacy challenges may become more critical. As many health care CAs deal with safety-critical user information and decision making, P3 and P13 noted the need for more strict and standardized evaluation measures for CAs.

## User Engagement

Establishing user engagement strategies for CAs is fundamental to improving user experience and sustained use. However, 6 projects reported enabling and maintaining user engagement as a challenge (P4, P7, P9, P10, P13, and P14). P9 explained the importance of culturally sensitive CAs to support increased trust and adoption in the context of helping African Americans with chronic diseases and the difficulties of understanding and incorporating cultural aspects into CAs. P10's focus on creating a multimodal food journal mentioned their participants' problem of remembering or discovering the voice commands to track their food, and the need to have a better mapping of commands to multiple utterance styles and advanced intent recognition. P13, a project using a self-diagnosis chatbot, reported that their users tended to drop out of the consultation with their chatbot, especially during the early stages, and pointed out the importance of examining and evaluating the mechanisms and approaches that can increase the uptake and utilization of health chatbots. P14 incorporated various strategies into the chatbot to support higher treatment adherence for pediatric patients awaiting a specialist appointment. These included developing a working alliance; having face-to-face communication; using everyday conversational language; and empathic language strategies such as choices, consequences, and nonjudgmental affirmations.

## Evaluation

### Overview and Subthemes

The Evaluation theme encapsulates 3 limitations that authors have encountered during their experiments with CAs: (1) methodological limitations that show challenges in evaluating interactions and performance of CAs; (2) experimental limitations related to the challenges with data collection and analysis and study environment; and (3) lack of guidance in evaluation describing challenges to navigate assessment of CAs without guidance or prior evidence.

### Methodological Limitations

By nature, CA interactions are designed to communicate with humans and are dependent on intellectually crafted bidirectional conversations that are hard to generate or replicate. Early efforts toward testing the CAs included scripted conversations to measure the performance of CAs objectively. Yet, scripted conversations fall short in mimicking the actual number of conversations and iterations that may occur in the real world (P1). In addition, a limited number of interactions create a barrier to evaluating the performance effectively. Creating complex scripts or testing with actual patients are some of the solutions proposed (P1).

To provide standard measures to evaluate the performance of CAs, user-CA interactions can be tested in a controlled environment using scripts, such as scenarios, role-playing, or Wizard of Oz testing. However, these scripted or simulated interactions might be limited to providing organically flowing conversations (P1) and accurate assessments due to the simulated nature of system functionalities and user scenarios (P8). Such evaluation yields the results only in a controlled environment, impacting the end user's judgment (P8). The

observations on user-CA interactions may explain a CA's efficacy to a degree yet not its effectiveness, which can be observed in the real-world environment. Using actual end users with different health and well-being needs in real-world settings and observing them longitudinally may improve the evidence of CA interactions (P1, P11, and P14).

## Experimental Limitations

### Effects of Training Materials/Unplanned Events

Training data are core requirements in developing and assessing CAs performance. However, it could be hard to gauge the effects of training materials (P5). Developers must track the features used in training and compare them with the actual CA usage. Similarly, unplanned or unforeseen events can affect the outcomes of CA interaction evaluations. The COVID-19 pandemic shifted the norms toward remote management of experiments, causing miscommunications and inefficiencies in simple user training and troubleshooting (P5).

### Accessing Vulnerable Populations

It is important to design for diverse populations, including vulnerable populations, and people with various social ills, such as homelessness, poverty, and hunger [58]. However, recruiting and engaging them with limited resources and connections with community partners can be difficult. In terms of improving CA interaction with such populations and enhancing access to the CA, inputs from heterogeneous groups and different users are needed through alternative platforms and technologies (P12 and P14). For instance, CA use evaluation in an emergency department with low-literacy users may require trained personnel to guide users and understand their experiences (P17).

### Challenges of Testing in Real-World Settings

In research, test settings and iterations often occur in controlled environments, without real-world interactions to objectively assess CAs. This results in decontextualization of testing and yields limited results. Some significant difficulties are involved in performing CA testing in real-world, authentic settings. For instance, P1, in a study with CA-supported automated documentation, reported that without a real-world electronic health record interaction, many work routines cannot be tested. Even if a CA gets integrated into electronic health record, technically, logistically, and legally it is hard to roll out. However, such efforts are necessary for understanding all end users' perceptions (nurses, doctors, patients) and the complexity of medical workflow (P17). P3 suggested a staged evaluation approach through a research platform, which will allow evaluations in mock care settings. There are still very few user-CA evaluations in real-world settings. Further efforts are required to promote real-world testing (in uncontrolled authentic environments). P13 explained that without an in-depth understanding of contextual elements in a problem domain, it is challenging for CA designers and developers to figure out how to improve the user experience and how to overcome the challenges in the actual use of health care CAs.

## ***Lack of Guidance in Evaluation***

### **Lack of Evaluation Data for Special Population Groups**

As with vulnerable populations, there is a lack of evaluation data for special population groups. This issue may lead to having no real knowledge of user experiences with CAs for people with diverse speech patterns (P2). Similarly, there is a lack of evaluation material designed (P13) specifically for marginalized or minor user groups, such as African Americans with low technology literacy (P9). Development of population-specific evaluation methods and promotion of participatory design and interactive sessions are necessary (P2 and P9).

### **Lack of Evaluation Guidelines and Metrics**

The lack of evaluation guidelines for CAs leads to creating robust frameworks for effectively and uniformly measuring impact and outcomes. In addition, the CA-based interventions' safety, efficacy, and effectiveness are lacking due to no clear guidance provided in the literature yet (P13). Metrics to measure the effects of interaction, engagement, and measures for health outcomes are necessary. One solution could be longitudinally observing interaction to identify key indicators to be used as success metrics (P15).

Theoretical approaches to evaluating the CAs may not be aligned with the standard measures. In P11, behavior change techniques showed a mismatch with usability metrics, potentially showing an inverse relationship between user engagement (quality of communication in user experience and CA usage patterns) and behavior change techniques (eg, goal setting). Combining theories and evaluation approaches might be necessary to create multifaceted evaluation metrics and triangulate use patterns.

### **Difficulties in Multimodal Testing**

CA could be potentially provided in multiple platforms and different modalities (eg, text-based chatbot, voice assistant with avatar). Evaluating individual modalities in a multimodal system is necessary, yet it is hard to evaluate the graphical and conversational interfaces separately. However, it might still be possible to design visual elements and layouts in a minimal way to reduce their effects on users' perception and assessment of the CA interface (P1).

## **Conversational Competence**

### ***Overview and Subthemes***

Several projects discussed challenges related to the conversational competence of CAs and the impacts these might have on people's interactions. Competence here refers to accurately understanding user input and responding appropriately, whether the conversation is an appropriate metaphor for design and how user interactions can be best facilitated, and how CA interactions can be made more accessible and inclusive.

### ***Topic Detection and Automated Speech Recognition***

Detecting the topic being discussed can be difficult in several CA scenarios. The unscripted nature of the social talk, often undertaken as part of caregiving interactions, makes it difficult to follow the discussed topics (P3). Even more scripted interactions such as primary care consultations may be nonlinear

and fragmented, creating further difficulties in detecting the current topic being discussed (P1). Proposed solutions include advanced topic detection methods (P3) and collaboratively built data sets to support these methods (P1). Using specific phrases to highlight topic shifts during interactions (P1) and personalized systems (P3) may also improve CAs.

Topic detection in CA in health care settings also faces challenges related to specific contexts and users. For example, tracking health care outcomes for different patients and treatments requires an understanding of specific medical terms (P15). Using services such as Amazon Comprehend Medical [63] alongside manual interventions could help expand appropriate vocabulary to improve CA comprehension. Similarly, P12 describes using CAs for people with hearing aids that may require them to use an agreed set of terms or additional supervised training to recognize "audiological intents."

More generally, there is an ongoing challenge of understanding when someone has finished an utterance while communicating with a CA (endpoint detection). This may require additional research for people with diverse speech patterns such as stammering (P2), to collect the necessary audio data and understand their interactions.

In addition to detecting topics in interaction, CAs in health care interactions face difficulties with ASR. This is also common in CA interactions outside of health care [64]. Audio may be impacted by recording quality ambient noise (eg, other devices such as televisions or other people talking), which could be combated with the use of directional microphones (P1 and P15) [21]. Again, in more nuanced interactions such as using CAs for people with hearing aids, optimal CA responses may require the processing of environmental information such as loudness and signal-to-noise ratio (P12). Inserting such information into the dialog may improve CA performance for this type of interaction.

In addition to audio quality, the language a user produces can also be a limiting factor for CAs. Active error correction by a CA may help in the misunderstanding and nonrecognition of specific terms (P10). For privacy-first interactions, processing of speech data may be performed locally on the device rather than requiring any interaction with servers, though this may reduce the performance of the CA's speech processing capabilities (P8).

### ***Discoverability and Conversational Interaction Model***

A key challenge in CA design is making the set of possible actions or commands discoverable for users [65]. Understanding and remembering how to interact with CAs in different contexts can create difficulties for people using these systems (P10 and P11). Making CAs open ended and lightweight may allow people to explore systems' capabilities and adapt them for their purposes (P10). For people with low levels of technology literacy, CA-initiated dialogs can be implemented at consistent intervals to counteract the lack of discoverability of user-initiated features (P11).

CAs rely on turn-taking-based conversational communication and interaction models; however, they may not always be appropriate or require additional scaffolding. General



practitioners taking notes using a CA, for example, may need to interrupt interactions with their patients. By contrast, screen-based technologies for the same task can support multimodal, more continuous interactions, and less intrusive data entry (P1). Continuing the multimodal use of information entry for general practitioners may improve or resolve interruptions. Furthermore, monitoring utterances in user-CA interactions may not be as smooth as human-human interactions. Consequently, enabling a CA to detect when someone has finished speaking is critical to more seamless interactions (P2). For diverse speech patterns, such as stammering, we may be able to draw on advice for interacting with various demographics, though crucially, we must understand the nuances of these interactions and design CAs with inclusivity in mind. For scenarios such as filling in forms and surveys with CAs, audio may be an inefficient modality and create design tension with sequential questions and wait times (P17). Optimizing delay usage to limit unnecessary wait times, using shorter phrases, and allowing people to opt-out of audio in multimodal CAs can be used to work against these limitations.

### ***Accessibility and Inclusivity***

Several projects discuss the need to make CAs accessible to a broad set of demographics or to focus on improving systems to make them inclusive for specific types of people. P14 identifies the need to provide equitable and easy access to CAs that improve health and well-being. They propose allowing access through a web browser and downloadable software, as well as offering technical support and downloadable fact sheets to work with or replace their CA. Off-the-shelf CAs may present difficulties for people in lower-resource areas (eg, limited internet access, financial constraints), which could be improved with systems that use more offline resources (P8), though this may present its own performance challenges that need to be addressed.

Levels of education, experience, and technological literacy can also impact users' interactions with CAs (P5, P16, and P17). These obstacles may be overcome by providing multiple modalities to compensate for people's preferences and abilities (P5 and P17) and engaging in participatory design with target demographics (P16). Evaluating such systems may require redeveloping protocols to include people with lower literacy and ensure they are able to understand the questions being posed to them (P17). Specific communities may also require additional thought in designing and evaluating CAs and building on the available sparse research literature (P2 and P9). Working with people who stammer, for example, has seen little research with CA interactions and requires a fundamental understanding of the barriers to successful interactions and how these can be overcome (P2). P9 highlights the need to consider African American communities and how their perceptions towards CA and interactions with them cannot be assumed to be the same as other communities. Consequently, novel methods and evaluation techniques may need to be developed with diverse populations in mind.

## ***Discussion***

### **Principal Findings**

This review highlights numerous challenges of CA interactions in the health care and well-being fields, including 4 major themes on Domain Information and Integration, Conversational Competence, User-System Interaction and Partnership, and Evaluation. Many challenges reported in this review echo those discussed in related CA work. ASR errors are long-standing concerns with speech-based CAs [66] that can lead users to alter their speech patterns to increase comprehension [67]. Difficulties in fostering user engagement have been highlighted when CAs cease to perform their intended utility [68]. Existing research has addressed this issue, for example, by explaining the cause of interaction errors to users or allowing users to halt interactions when they desire [69]. Accessibility concerns discussed in this review also map to those discussed in prior work on both text-based [70] and speech-based CAs [71]. Research continues to identify how CA accessibility can be improved [72] and design recommendations such as the Web Content Accessibility Guidelines [73] have been suggested as a means of addressing these. Prior studies also reported challenges associated with involving and engaging patients in their home environments [74], limitations in short-duration laboratory studies and challenges in longitudinal assessments [75], and difficulties with multicomponent system evaluation [76].

In addition to the similar challenges identified in prior CA literature, this review identifies some challenges that are specific or more critical to health care and well-being domains, including empathy, safety, recruitment of vulnerable populations, and challenges in testing authentic settings. Prior work has brought attention to open challenges across the broader CA field [77] and highlights the need to create responsible CAs that focus on fairness, transparency, and ethics. This is critical for CAs discussed in this review, particularly given the sensitive nature of interactions and the underlying data collected from these. The fragmented nature of speech-based CA work, in particular, has recently been addressed [29,78]. This prior work notes a scarcity of robust evaluation metrics and research involving real-world testing. Our review shows that there are similar problems for application of CAs in health care and well-being. However, the nature of these domains means there can be additional obstacles when entering research phases that require strict considerations of medical ethics, laws, and standardized practices. Improving consistency in CA research and implementation may benefit from reaching out to other disciplines (eg, cognitive sciences, linguistics) [78], combining existing theories and evaluation approaches (see the "Lack of Guidance in Evaluation" section), and examining guidelines from both academia and industry [79]. While broad challenges around CAs must be considered within health care and well-being contexts, this review also identifies more nuanced challenges for these domains. Future work should consider how these different challenges overlap, depending on both the expected CA scenarios and user demographics.

## Data Collection in Health Care Domains

Obtaining domain-specific information and collecting data are typically challenging. There are 2 major reasons: first, health care professionals who can provide critical domain information are very time-poor and their contribution to the projects is limited; and second, health care data are extremely sensitive, and there are a lot of privacy and safety concerns. As recruiting a full-time health care professional is not financially feasible, an incremental acquisition method might be useful (P14). Data privacy and safety are significant concerns for health care professionals, and consumers are more hesitant to share information compared with other application areas (P1). For example, the Digital Scribe project needed to collect thousands of doctor-patient conversations in primary care settings to train their NLP algorithm [21]. However, managing the audio recording system, informing patients and getting their consent, and privacy and safety concerns of not only patients but also doctors made the data collection process extremely challenging. In this case, multi-institution collaborations are needed to organize and manage data collection to reduce the frictions in the process such as employing a dedicated technical person to provide support, more automated and easy-to-control audio recording system, and streamlined patient consent gathering. Such data collection challenges are not specific to primary care as many health care settings have similar situations and requirements. Therefore, incentivizing collaborations and data sharing between institutions and creating ethical frameworks to facilitate data sharing are needed, with some already emerging in the CA community [80]. In addition, proven data safety protocols and certifications can be made available to ensure safety and increase user trust. Integrating CAs into existing systems and enabling data sharing across systems were other challenges reported. Fast Healthcare Interoperability Resources was suggested as a promising interoperability standard (P15).

## Developing Empathy in CA Interactions

Concerning the theme User-System Interaction and Partnership, empathy, safety, and privacy emerged as important challenges. Especially for CAs used in mental health applications, being able to show empathy is a valuable characteristic. Challenges related to empathy can be grouped into 2 categories: challenges of detecting the users' current emotional sensitivity to a topic and crafting an appropriate response to the user according to their situation and preferences. For example, P4, using their CA interface Aira, used client profiles (typical interaction contexts, preferred measurement units) to reference and incorporate these details in the interaction to signal empathy. P17 aimed their CAs to show empathy through more minimal means. This contrasts with other successful but more costly implementations of embodied CAs [81,82] in general medical contexts. P17 designed a series of neutral and empathetic reactions to user answers as well as other social utterances to augment the question administration dialog. They included phrases that help the user anticipate a sensitive topic before it is introduced (eg, "The next questions are about your personal safety and may be tough to answer."), provide acknowledgments for answering neutral questions (eg, "Okay, I'm getting a better idea of where

you are at," "Got it"), and empathetic reactions to sensitive questions (eg, "That must be stressful, I'm sorry to hear that."). They also pointed out some users' different preferences, such as a lower level of socialization in chat than others.

## Establishing Safety and Privacy in CA Applications

Safety and privacy concerns are more critical in health and well-being applications. A systematic review found that patient safety concerns were rarely addressed in the majority of the papers reviewed [1]. Similarly, Bickmore et al [24] found that CAs with unconstrained natural language input pose serious safety risks. Prior studies also found that commonly available voice assistants fail to answer appropriately to safety-critical user prompts [4,6]. Safety risks may occur due to misrecognized prompts, the inability to detect the severity of users' prompts (when there are no ASR errors), or gaps in clinical reasoning. A 5-stage framework for evaluating different aspects of symptom checkers might be adapted to be used for other CA applications [83]. Privacy of personal user data is also of utmost importance. Personal health data are considered one of the most sensitive types of information. Therefore, protecting the privacy of users' data becomes even more critical in health care CAs. Data privacy is particularly more relevant to CAs as many CAs process user prompts through third-party services in the cloud. Thus, there is a valid concern about the security of such information. Local NLP engines might be a solution to this (P8).

## Limitations

This review included a limited number of studies in the emerging domain of health care and well-being CAs, which were conducted within a specific period and reported during a conference workshop. The main purpose is to provide a snapshot of some of the major challenges faced by the recent projects in this area rather than providing a comprehensive overview of the challenges. Although the reported challenges provide a useful overview of some significant challenges, they should not be considered the complete set of challenges in this domain. They represent the challenges faced by some recent projects conducted by the researchers actively working in this area across the globe.

## Conclusion

This paper examined 17 recent studies of CAs in health care and well-being to identify design and evaluation challenges. While many challenges, including accessibility, personalization, and empathy provision, are shared with other application areas of CAs, safety and privacy remain the major challenges that are more critical in the health care domain.

CAs proved their worth during the pandemic as health screening tools, and are here to stay to further assist personal health care. Growth in investment in CAs also shows the value as a personal assistant. An increased level of collaboration across different institutions and entities may be a promising direction to address some of the major challenges that otherwise would be too complex to be addressed by the projects with their limited scope and budget.

## Authors' Contributions

ABK contributed to the study design. ABK, ES, LC, JMC, JH-Y, YH, JK, RK, Y-CL, LM, EGM, RJM, PM, EDM, SYP, AP, DR, LMS, DS, BS, ZZ, and TZ contributed to data reporting. ABK, ES, and LC performed thematic analysis. ABK, ES, and LC proposed the first draft. ABK, ES, LC, JMC, JH-Y, YH, JK, RK, Y-CL, LM, EGM, RJM, PM, EDM, SYP, AP, DR, LMS, DS, BS, ZZ, and TZ performed revisions and finalized subsequent drafts.

## Conflicts of Interest

None declared.

## References

- 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](https://doi.org/10.1093/jamia/ocy072)] [Medline: [30010941](https://pubmed.ncbi.nlm.nih.gov/30010941/)]
- Comes S, Schatsky D, Chauhan R. Conversational AI. Deloitte Insights. 2021. URL: <https://www.deloitte.com/us/en/insights/focus/signals-for-strategists/the-future-of-conversational-ai.html> [accessed 2022-10-19]
- 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](https://doi.org/10.1093/tbm/ibz141)] [Medline: [32766865](https://pubmed.ncbi.nlm.nih.gov/32766865/)]
- Kocaballi AB, Quiroz JC, Rezazadegan D, Berkovsky S, Magrabi F, Coiera E, et al. Responses of Conversational Agents to Health and Lifestyle Prompts: Investigation of Appropriateness and Presentation Structures. *J Med Internet Res* 2020 Feb 09;22(2):e15823 [FREE Full text] [doi: [10.2196/15823](https://doi.org/10.2196/15823)] [Medline: [32039810](https://pubmed.ncbi.nlm.nih.gov/32039810/)]
- Yang S, Lee J, Sezgin E, Bridge J, Lin S. Clinical Advice by Voice Assistants on Postpartum Depression: Cross-Sectional Investigation Using Apple Siri, Amazon Alexa, Google Assistant, and Microsoft Cortana. *JMIR Mhealth Uhealth* 2021 Jan 11;9(1):e24045 [FREE Full text] [doi: [10.2196/24045](https://doi.org/10.2196/24045)] [Medline: [33427680](https://pubmed.ncbi.nlm.nih.gov/33427680/)]
- Miner AS, Milstein A, Schueller S, Hegde R, Mangurian C, Linos E. Smartphone-Based Conversational Agents and Responses to Questions About Mental Health, Interpersonal Violence, and Physical Health. *JAMA Intern Med* 2016 May 01;176(5):619-625 [FREE Full text] [doi: [10.1001/jamainternmed.2016.0400](https://doi.org/10.1001/jamainternmed.2016.0400)] [Medline: [26974260](https://pubmed.ncbi.nlm.nih.gov/26974260/)]
- Alagha EC, Helbing RR. Evaluating the quality of voice assistants' responses to consumer health questions about vaccines: an exploratory comparison of Alexa, Google Assistant and Siri. *BMJ Health Care Inform* 2019 Nov;26(1):e100075 [FREE Full text] [doi: [10.1136/bmjhci-2019-100075](https://doi.org/10.1136/bmjhci-2019-100075)] [Medline: [31767629](https://pubmed.ncbi.nlm.nih.gov/31767629/)]
- Nobles AL, Leas EC, Caputi TL, Zhu S, Strathdee SA, Ayers JW. Responses to addiction help-seeking from Alexa, Siri, Google Assistant, Cortana, and Bixby intelligent virtual assistants. *NPJ Digit Med* 2020;3:11 [FREE Full text] [doi: [10.1038/s41746-019-0215-9](https://doi.org/10.1038/s41746-019-0215-9)] [Medline: [32025572](https://pubmed.ncbi.nlm.nih.gov/32025572/)]
- Wilson N, MacDonald EJ, Mansoor OD, Morgan J. In bed with Siri and Google Assistant: a comparison of sexual health advice. *BMJ* 2017 Dec 13;359:j5635. [doi: [10.1136/bmj.j5635](https://doi.org/10.1136/bmj.j5635)] [Medline: [29237603](https://pubmed.ncbi.nlm.nih.gov/29237603/)]
- Zand A, Sharma A, Stokes Z, Reynolds C, Montilla A, Sauk J, et al. An Exploration Into the Use of a Chatbot for Patients With Inflammatory Bowel Diseases: Retrospective Cohort Study. *J Med Internet Res* 2020 May 26;22(5):e15589 [FREE Full text] [doi: [10.2196/15589](https://doi.org/10.2196/15589)] [Medline: [32452808](https://pubmed.ncbi.nlm.nih.gov/32452808/)]
- Bérubé C, Kovacs, Fleisch, Kowatsch. Reliability of Commercial Voice Assistants' Responses to Health-Related Questions in Noncommunicable Disease Management: Factorial Experiment Assessing Response Rate and Source of Information. *J Med Internet Res* 2021 Dec 20;23(12):e32161 [FREE Full text] [doi: [10.2196/32161](https://doi.org/10.2196/32161)] [Medline: [34932003](https://pubmed.ncbi.nlm.nih.gov/34932003/)]
- Hong G, Folcarelli A, Less J, Wang C, Erbas N, Lin S. Voice Assistants and Cancer Screening: A Comparison of Alexa, Siri, Google Assistant, and Cortana. *Ann Fam Med* 2021;19(5):447-449 [FREE Full text] [doi: [10.1370/afm.2713](https://doi.org/10.1370/afm.2713)] [Medline: [34546951](https://pubmed.ncbi.nlm.nih.gov/34546951/)]
- Centers for Disease Control and Prevention (CDC). Symptoms of COVID-19. CDC. 2022. URL: <https://www.cdc.gov/coronavirus/2019-ncov/symptoms-testing/symptoms.html> [accessed 2022-10-19]
- Apple releases new COVID-19 app and website based on CDC guidance Internet. Apple Inc. 2020. URL: <https://www.apple.com/newsroom/2020/03/apple-releases-new-covid-19-app-and-website-based-on-CDC-guidance/> [accessed 2022-10-19]
- Number of mobile devices worldwide 2020-2025. Statista. 2020. URL: <https://www.statista.com/statistics/245501/multiple-mobile-device-ownership-worldwide/> [accessed 2022-10-19]
- Demographics of mobile device ownership and adoption in the United States. Pew Research Center. 2021. URL: <https://www.pewresearch.org/internet/fact-sheet/mobile/> [accessed 2022-10-19]
- Abd-Alrazaq AA, Rababeh A, Alajlani M, Bewick BM, Househ M. Effectiveness and Safety of Using Chatbots to Improve Mental Health: Systematic Review and Meta-Analysis. *J Med Internet Res* 2020 Jul 13;22(7):e16021 [FREE Full text] [doi: [10.2196/16021](https://doi.org/10.2196/16021)] [Medline: [32673216](https://pubmed.ncbi.nlm.nih.gov/32673216/)]
- Maeda E, Miyata A, Boivin J, Nomura K, Kumazawa Y, Shirasawa H, et al. Promoting fertility awareness and preconception health using a chatbot: a randomized controlled trial. *Reprod Biomed Online* 2020 Dec;41(6):1133-1143. [doi: [10.1016/j.rbmo.2020.09.006](https://doi.org/10.1016/j.rbmo.2020.09.006)] [Medline: [33039321](https://pubmed.ncbi.nlm.nih.gov/33039321/)]



19. Piao M, Ryu H, Lee H, Kim J. Use of the Healthy Lifestyle Coaching Chatbot App to Promote Stair-Climbing Habits Among Office Workers: Exploratory Randomized Controlled Trial. *JMIR Mhealth Uhealth* 2020 May 19;8(5):e15085 [FREE Full text] [doi: [10.2196/15085](https://doi.org/10.2196/15085)] [Medline: [32427114](https://pubmed.ncbi.nlm.nih.gov/32427114/)]
20. Health chatbots are here, heavily funded, and we'll have to get used to them. *Crunchbase News*. 2021. URL: <https://news.crunchbase.com/news/health-chatbots-startups-vc-funding/> [accessed 2022-10-19]
21. Quiroz JC, Laranjo L, Kocaballi AB, Berkovsky S, Rezazadegan D, Coiera E. Challenges of developing a digital scribe to reduce clinical documentation burden. *NPJ Digit Med* 2019;2:114 [FREE Full text] [doi: [10.1038/s41746-019-0190-1](https://doi.org/10.1038/s41746-019-0190-1)] [Medline: [31799422](https://pubmed.ncbi.nlm.nih.gov/31799422/)]
22. Montenegro JLZ, da Costa CA, da Rosa Righi R. Survey of conversational agents in health. *Expert Systems with Applications* 2019 Sep;129:56-67 [FREE Full text] [doi: [10.1016/j.eswa.2019.03.054](https://doi.org/10.1016/j.eswa.2019.03.054)]
23. Tudor Car L, Dhinakaran DA, Kyaw BM, Kowatsch T, Joty S, Theng Y, 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](https://doi.org/10.2196/17158)] [Medline: [32763886](https://pubmed.ncbi.nlm.nih.gov/32763886/)]
24. Bickmore TW, Trinh H, Olafsson S, O'Leary TK, Asadi R, Rickles NM, et al. Patient and Consumer Safety Risks When Using Conversational Assistants for Medical Information: An Observational Study of Siri, Alexa, and Google Assistant. *J Med Internet Res* 2018 Sep 04;20(9):e11510 [FREE Full text] [doi: [10.2196/11510](https://doi.org/10.2196/11510)] [Medline: [30181110](https://pubmed.ncbi.nlm.nih.gov/30181110/)]
25. Miner AS, Laranjo L, Kocaballi AB. Chatbots in the fight against the COVID-19 pandemic. *NPJ Digit Med* 2020;3:65 [FREE Full text] [doi: [10.1038/s41746-020-0280-0](https://doi.org/10.1038/s41746-020-0280-0)] [Medline: [32377576](https://pubmed.ncbi.nlm.nih.gov/32377576/)]
26. Sezgin E, Huang Y, Ramtekkar U, Lin S. Readiness for voice assistants to support healthcare delivery during a health crisis and pandemic. *NPJ Digit Med* 2020;3:122 [FREE Full text] [doi: [10.1038/s41746-020-00332-0](https://doi.org/10.1038/s41746-020-00332-0)] [Medline: [33015374](https://pubmed.ncbi.nlm.nih.gov/33015374/)]
27. Kocaballi AB, Ijaz K, Laranjo L, Quiroz JC, Rezazadegan D, Tong HL, et al. Envisioning an artificial intelligence documentation assistant for future primary care consultations: A co-design study with general practitioners. *J Am Med Inform Assoc* 2020 Nov 01;27(11):1695-1704 [FREE Full text] [doi: [10.1093/jamia/ocaa131](https://doi.org/10.1093/jamia/ocaa131)] [Medline: [32845984](https://pubmed.ncbi.nlm.nih.gov/32845984/)]
28. 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](https://doi.org/10.2471/BLT.19.237636)] [Medline: [32284654](https://pubmed.ncbi.nlm.nih.gov/32284654/)]
29. Clark L, Cowan B, Edwards J, Munteanu C, Murad C, Aylett M, et al. Mapping Theoretical and Methodological Perspectives for Understanding Speech Interface Interactions. In: *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*. New York, NY: Association for Computing Machinery; 2019 Presented at: 2019 CHI Conference on Human Factors in Computing Systems; May 4-9, 2019; Glasgow, Scotland, UK p. 1-8. [doi: [10.1145/3290607.3299009](https://doi.org/10.1145/3290607.3299009)]
30. Moore RJ, Arar R, Ren GJ, Szymanski MH. Conversational UX Design. In: *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems*. New York, NY: Association for Computing Machinery; 2017 Presented at: 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems; May 6-11, 2017; Denver, CO p. A. [doi: [10.1145/3027063.3027077](https://doi.org/10.1145/3027063.3027077)]
31. Reeves S, Porcheron M, Fischer J, Candello H, McMillan D, McGregor M, et al. Voice-based Conversational UX Studies and Design. In: *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems*. New York, NY: Association for Computing Machinery; 2018 Presented at: 2018 CHI Conference on Human Factors in Computing Systems; April 21-26, 2018; Montréal, QC, Canada p. 1-8. [doi: [10.1145/3170427.3170619](https://doi.org/10.1145/3170427.3170619)]
32. Kocaballi A, Quiroz J, Laranjo L, Rezazadegan D, Kocielnik R, Clark L, et al. Conversational Agents for Health and Wellbeing. In: *CHI '20: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. New York, NY: Association for Computing Machinery; 2020 Presented at: CHI '20: CHI Conference on Human Factors in Computing Systems; April 25-30, 2020; Honolulu, HI p. 1-8. [doi: [10.1145/3334480.3375154](https://doi.org/10.1145/3334480.3375154)]
33. Braun V, Clarke V. Using thematic analysis in psychology. *Qualitative Research in Psychology* 2006 Jan;3(2):77-101. [doi: [10.1191/1478088706qp063oa](https://doi.org/10.1191/1478088706qp063oa)]
34. Clark L, Cowan B, Roper A, Lindsay S, Sheers O. Speech diversity and speech interfaces: considering an inclusive future through stammering. In: *Proceedings of the 2nd Conference on Conversational User Interfaces*. New York, NY: Association for Computing Machinery; 2020 Jul Presented at: CUI '20: 2nd Conference on Conversational User Interfaces; July 22-24, 2020; Bilbao, Spain p. 1-3. [doi: [10.1145/3405755.3406139](https://doi.org/10.1145/3405755.3406139)]
35. Spillane B, Gilmartin E, Saam C, Su K, Cowan B, Lawless S, et al. Introducing ADELE: a personalized intelligent companion. In: *ISIAA 2017: Proceedings of the 1st ACM SIGCHI International Workshop on Investigating Social Interactions with Artificial Agents*. New York, NY: Association for Computing Machinery; 2017 Presented at: ICMI '17: International Conference on Multimodal Interaction; November 13, 2017; Glasgow, UK p. 43-44. [doi: [10.1145/3139491.3139492](https://doi.org/10.1145/3139491.3139492)]
36. Gilmartin E, Saam C, Spillane B, O'Reilly M, Su K, Calvo A, et al. The ADELE Corpus of Dyadic Social Text Conversations: Dialog Act Annotation with ISO 24617-2. In: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. Paris, France: European Language Resources Association; 2018 Presented at: Eleventh International Conference on Language Resources and Evaluation; May 7-12, 2018; Miyazaki, Japan p. e1 URL: <https://aclanthology.org/L18-1633>
37. Spillane B, Gilmartin E, Saam C, Cowan B, Wade V. ADELE: Care and Companionship for Independent Aging. *CEUR Workshop Proceedings*. 2018. URL: <http://ceur-ws.org/Vol-2338/paper2.pdf> [accessed 2022-10-19]



38. Lee S, Yu R, Xie J, Billah S, Carroll J. Opportunities for Human-AI Collaboration in Remote Sighted Assistance. New York, NY: Association for Computing Machinery; 2022 Presented at: IUI '22: 27th International Conference on Intelligent User Interfaces; March 22-25, 2022; Helsinki, Finland p. 63-78. [doi: [10.1145/3490099.3511113](https://doi.org/10.1145/3490099.3511113)]
39. Lee S, Reddie M, Tsai C, Beck J, Rosson M, Carroll J. The Emerging Professional Practice of Remote Sighted Assistance for People with Visual Impairments. In: CHI '20: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems. New York, NY: Association for Computing Machinery; 2020 Presented at: CHI '20: CHI Conference on Human Factors in Computing Systems; April 25-30, 2020; Honolulu, HI p. 1-12. [doi: [10.1145/3313831.3376591](https://doi.org/10.1145/3313831.3376591)]
40. Xie J, Reddie M, Lee S, Billah SM, Zhou Z, Tsai C, et al. Iterative Design and Prototyping of Computer Vision Mediated Remote Sighted Assistance. *ACM Trans. Comput.-Hum. Interact* 2022 Aug 31;29(4):1-40. [doi: [10.1145/3501298](https://doi.org/10.1145/3501298)]
41. Xie J, Yu R, Lee S, Lyu Y, Billah S, Carroll J. Helping Helpers: Supporting Volunteers in Remote Sighted Assistance with Augmented Reality Maps. In: DIS '22: Designing Interactive Systems Conference. New York, NY: Association for Computing Machinery; 2022 Jun Presented at: DIS '22: Designing Interactive Systems Conference; June 13-17, 2022; Virtual Event Australia p. 881-897. [doi: [10.1145/3532106.3533560](https://doi.org/10.1145/3532106.3533560)]
42. Lee Y, Yamashita N, Huang Y. Exploring the Effects of Incorporating Human Experts to Deliver Journaling Guidance through a Chatbot. *Proc. ACM Hum.-Comput. Interact* 2021 Apr 13;5(CSCW1):1-27 [FREE Full text] [doi: [10.1145/3449196](https://doi.org/10.1145/3449196)]
43. Lee Y, Yamashita N, Huang Y. Designing a Chatbot as a Mediator for Promoting Deep Self-Disclosure to a Real Mental Health Professional. *Proc. ACM Hum.-Comput. Interact* 2020 May 28;4(CSCW1):1-27. [doi: [10.1145/3392836](https://doi.org/10.1145/3392836)]
44. Lee Y, Yamashita N, Huang Y, Fu W. "I Hear You, I Feel You": Encouraging Deep Self-disclosure through a Chatbot. In: CHI '20: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems. New York, NY: Association for Computing Machinery; 2020 Apr Presented at: CHI '20: 2020 CHI Conference on Human Factors in Computing Systems; April 25-30, 2020; Honolulu, HI p. 1-12. [doi: [10.1145/3313831.3376175](https://doi.org/10.1145/3313831.3376175)]
45. Smriti D, Shin J, Mujib M, Colosimo M, Kao T, Williams J, et al. TAMICA: Tailorable Autonomous Motivational Interviewing Conversational Agent. In: Proceedings of the 14th EAI International Conference on Pervasive Computing Technologies for Healthcare. New York, NY: Association for Computing Machinery; 2020 Presented at: 14th EAI International Conference on Pervasive Computing Technologies for Healthcare; May 18-20, 2020; Virtual. [doi: [10.1145/3421937.3421963](https://doi.org/10.1145/3421937.3421963)]
46. Smriti D, Kao TA, Rathod R, Shin JY, Peng W, Williams J, et al. Motivational Interviewing Conversational Agent for Parents as Proxies for Their Children in Healthy Eating: Development and User Testing. *JMIR Hum Factors* 2022 Oct 07;9(4):e38908. [doi: [10.2196/38908](https://doi.org/10.2196/38908)] [Medline: [36206036](https://pubmed.ncbi.nlm.nih.gov/36206036/)]
47. Kim J, Park S. Conversational Agents for Health and Wellbeing: Review and Future Agendas. In: CHI EA '20: Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems. New York, NY: Association for Computing Machinery; 2019 Presented at: CHI '20: CHI Conference on Human Factors in Computing Systems; April 25-30, 2020; Honolulu, HI p. 1-8 URL: <https://deepblue.lib.umich.edu/handle/2027.42/151800>
48. Kim J, Muhic J, Robert LP, Park SY. Designing Chatbots with Black Americans with Chronic Conditions: Overcoming Challenges against COVID-19. In: CHI '22: Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems. New York, NY: Association for Computing Machinery; 2022 Apr Presented at: CHI '22: 2022 CHI Conference on Human Factors in Computing Systems; April 29, 2022 to May 5, 2022; New Orleans LA p. 1-17. [doi: [10.1145/3491102.3502116](https://doi.org/10.1145/3491102.3502116)]
49. Silva LM, Epstein DA. Investigating Preferred Food Description Practices in Digital Food Journaling. New York, NY: Association for Computing Machinery; 2021 Presented at: DIS '21: Designing Interactive Systems Conference 2021; June 28, 2021 to July 2, 2021; Virtual Event p. 589-605. [doi: [10.1145/3461778.3462145](https://doi.org/10.1145/3461778.3462145)]
50. Silva LM, Cibrian FL, Epstein DA, Bhattacharya A, Ankras EA, Monteiro E, et al. Adapting Multidevice Deployments During a Pandemic: Lessons Learned From Two Studies. *IEEE Pervasive Comput* 2022 Jan 1;21(1):48-56 [FREE Full text] [doi: [10.1109/MPRV.2021.3104262](https://doi.org/10.1109/MPRV.2021.3104262)]
51. Mitchell EG, Maimone R, Cassells A, Tobin JN, Davidson P, Smaldone AM, et al. Automated vs. Human Health Coaching. *Proc. ACM Hum.-Comput. Interact* 2021 Apr 13;5(CSCW1):1-37. [doi: [10.1145/3449173](https://doi.org/10.1145/3449173)]
52. Mitchell E, Mamykina L. From the Curtain to Kansas: Conducting Wizard-of-Oz Studies in the Wild. In: Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems. New York, NY: Association for Computing Machinery; 2021 Presented at: 2021 CHI Conference on Human Factors in Computing Systems; May 8-13, 2021; Yokohama, Japan p. 1-6. [doi: [10.1145/3411763.3443446](https://doi.org/10.1145/3411763.3443446)]
53. Fan X, Chao D, Zhang Z, Wang D, Li X, Tian F. Utilization of Self-Diagnosis Health Chatbots in Real-World Settings: Case Study. *J Med Internet Res* 2021 Jan 06;23(1):e19928 [FREE Full text] [doi: [10.2196/19928](https://doi.org/10.2196/19928)] [Medline: [33404508](https://pubmed.ncbi.nlm.nih.gov/33404508/)]
54. Richards D, Caldwell P. Improving Health Outcomes Sooner Rather Than Later via an Interactive Website and Virtual Specialist. *IEEE J Biomed Health Inform* 2018 Sep;22(5):1699-1706. [doi: [10.1109/JBHI.2017.2782210](https://doi.org/10.1109/JBHI.2017.2782210)] [Medline: [29990229](https://pubmed.ncbi.nlm.nih.gov/29990229/)]
55. Sezgin E, Noritz G, Lin S, Huang Y. Feasibility of a Voice-Enabled Medical Diary App (SpeakHealth) for Caregivers of Children With Special Health Care Needs and Health Care Providers: Mixed Methods Study. *JMIR Form Res* 2021 May 11;5(5):e25503 [FREE Full text] [doi: [10.2196/25503](https://doi.org/10.2196/25503)] [Medline: [33865233](https://pubmed.ncbi.nlm.nih.gov/33865233/)]

56. Sezgin E, Oiler B, Abbott B, Noritz G, Huang Y. "Hey Siri, Help Me Take Care of My Child": A Feasibility Study With Caregivers of Children With Special Healthcare Needs Using Voice Interaction and Automatic Speech Recognition in Remote Care Management. *Front Public Health* 2022;10:849322 [FREE Full text] [doi: [10.3389/fpubh.2022.849322](https://doi.org/10.3389/fpubh.2022.849322)] [Medline: [35309210](https://pubmed.ncbi.nlm.nih.gov/35309210/)]
57. Kocielnik R, Langevin R, George J, Akenaga S, Wang A, Jones D, et al. Can I Talk to You about Your Social Needs? Understanding Preference for Conversational User Interface in Health. In: CUI '21: Proceedings of the 3rd Conference on Conversational User Interfaces. New York, NY: Association for Computing Machinery; 2021 Presented at: CUI '21: CUI 2021 - 3rd Conference on Conversational User Interfaces; July 27-29, 2021; Bilbao (online), Spain p. 1-10. [doi: [10.1145/3469595.3469599](https://doi.org/10.1145/3469595.3469599)]
58. Kocielnik R, Agapie E, Argyle A, Hsieh DT, Yadav K, Taira B, et al. HarborBot: A Chatbot for Social Needs Screening. *AMIA Annu Symp Proc* 2019;2019:552-561 [FREE Full text] [Medline: [32308849](https://pubmed.ncbi.nlm.nih.gov/32308849/)]
59. Spillane B, Gilmartin E, Saam C, Wade V. Issues relating to trust in care agents for the elderly. New York, NY: Association for Computing Machinery; 2019 Presented at: CUI 2019: 1st International Conference on Conversational User Interfaces; August 22-23, 2019; Dublin, Ireland p. 1-3. [doi: [10.1145/3342775.3342808](https://doi.org/10.1145/3342775.3342808)]
60. Yuksel BF, Collisson P, Czerwinski M. Brains or Beauty. *ACM Trans. Internet Technol* 2017 Mar 06;17(1):1-20 [FREE Full text] [doi: [10.1145/2998572](https://doi.org/10.1145/2998572)]
61. Fan X, Oh S, McNeese M, Yen J, Cuevas H, Strater L, et al. The influence of agent reliability on trust in human-agent collaboration. New York, NY: Association for Computing Machinery; 2008 Presented at: ECCE08: European Conference on Cognitive Ergonomics; September 16-19, 2008; Funchal, Portugal p. 1-8. [doi: [10.1145/1473018.1473028](https://doi.org/10.1145/1473018.1473028)]
62. Benbasat I, Wang W. Trust In and Adoption of Online Recommendation Agents. *JAIS* 2005 Mar;6(3):72-101. [doi: [10.17705/1jais.00065](https://doi.org/10.17705/1jais.00065)]
63. Amazon Inc.. Amazon Web Services (AWS) Comprehend Medical. Amazon Inc. 2022. URL: <https://aws.amazon.com/comprehend/medical/> [accessed 2022-10-19]
64. Errattahi R, El Hannani A, Ouahmane H. Automatic Speech Recognition Errors Detection and Correction: A Review. *Procedia Computer Science* 2018;128:32-37 [FREE Full text] [doi: [10.1016/j.procs.2018.03.005](https://doi.org/10.1016/j.procs.2018.03.005)]
65. Corbett E, Weber A. What can I say? addressing user experience challenges of a mobile voice user interface for accessibility. In: MobileHCI '16: Proceedings of the 18th International Conference on Human-Computer Interaction with Mobile Devices and Services. New York, NY: Association for Computing Machinery; 2016 Presented at: MobileHCI '16: 18th International Conference on Human-Computer Interaction with Mobile Devices and Services; September 6-9, 2016; Florence, Italy p. 72-82. [doi: [10.1145/2935334.2935386](https://doi.org/10.1145/2935334.2935386)]
66. Oviatt S, Bernard J, Levow GA. Linguistic adaptations during spoken and multimodal error resolution. *Lang Speech* 1998;41 ( Pt 3-4):419-442. [doi: [10.1177/002383099804100409](https://doi.org/10.1177/002383099804100409)] [Medline: [10746365](https://pubmed.ncbi.nlm.nih.gov/10746365/)]
67. Branigan HP, Pickering MJ, Pearson J, McLean JF, Brown A. The role of beliefs in lexical alignment: evidence from dialogs with humans and computers. *Cognition* 2011 Oct;121(1):41-57. [doi: [10.1016/j.cognition.2011.05.011](https://doi.org/10.1016/j.cognition.2011.05.011)] [Medline: [21723549](https://pubmed.ncbi.nlm.nih.gov/21723549/)]
68. Luger E, Sellen A. "Like Having a Really Bad PA": The Gulf between User Expectation and Experience of Conversational Agents. In: CHI '16: Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems. New York, NY: Association for Computing Machinery; 2016 May Presented at: CHI'16: CHI Conference on Human Factors in Computing Systems; May 7-12, 2016; San Jose, CA p. 5286-5297. [doi: [10.1145/2858036.2858288](https://doi.org/10.1145/2858036.2858288)]
69. Fischer J, Reeves S, Porcheron M, Sikveland R. Progressivity for voice interface design. In: CUI '19: Proceedings of the 1st International Conference on Conversational User Interfaces. New York, NY: Association for Computing Machinery; 2019 Presented at: CUI 2019: 1st International Conference on Conversational User Interfaces; August 22-23, 2019; Dublin, Ireland p. 1-8. [doi: [10.1145/3342775.3342788](https://doi.org/10.1145/3342775.3342788)]
70. Torres C, Franklin W, Martins L. Accessibility in Chatbots: The State of the Art in Favor of Users with Visual Impairment. In: *Advances in Usability, User Experience and Assistive Technology*. Berlin, Germany: Springer; 2019:623-635.
71. Brewer R, Findlater L, Kaye, Lasecki W, Munteanu C, Weber A. Accessible Voice Interfaces. In: CSCW '18: Companion of the 2018 ACM Conference on Computer Supported Cooperative Work and Social Computing. New York, NY: Association for Computing Machinery; 2018 Presented at: CSCW '18: Computer Supported Cooperative Work and Social Computing; November 3-7, 2018; Jersey City, NJ p. 44-446. [doi: [10.1145/3272973.3273006](https://doi.org/10.1145/3272973.3273006)]
72. Lister K, Coughlan T, Iniesto F, Freear N, Devine P. Accessible conversational user interfaces: considerations for design. In: W4A '20: Proceedings of the 17th International Web for All Conference. New York, NY: Association for Computing Machinery; 2020 Presented at: W4A '20: 17th Web for All Conference; April 20-21, 2020; Taipei, Taiwan p. 1-11. [doi: [10.1145/3371300.3383343](https://doi.org/10.1145/3371300.3383343)]
73. WCAG 2 Overview. W3C Web Accessibility Initiative (WAI). URL: <https://www.w3.org/WAI/standards-guidelines/wcag/> [accessed 2022-10-19]
74. Stara V, Vera B, Bolliger D, Rossi L, Felici E, Di Rosa M, et al. Usability and Acceptance of the Embodied Conversational Agent Anne by People With Dementia and Their Caregivers: Exploratory Study in Home Environment Settings. *JMIR Mhealth Uhealth* 2021 Jun 25;9(6):e25891 [FREE Full text] [doi: [10.2196/25891](https://doi.org/10.2196/25891)] [Medline: [34170256](https://pubmed.ncbi.nlm.nih.gov/34170256/)]

75. ter Stal S, Kramer LL, Tabak M, op den Akker H, Hermens H. Design Features of Embodied Conversational Agents in eHealth: a Literature Review. *International Journal of Human-Computer Studies* 2020 Jun;138:102409. [doi: [10.1016/j.ijhcs.2020.102409](https://doi.org/10.1016/j.ijhcs.2020.102409)]
76. Kramer LL, Ter Stal S, Mulder BC, de Vet E, van Velsen L. Developing Embodied Conversational Agents for Coaching People in a Healthy Lifestyle: Scoping Review. *J Med Internet Res* 2020 Feb 06;22(2):e14058 [FREE Full text] [doi: [10.2196/14058](https://doi.org/10.2196/14058)] [Medline: [32022693](https://pubmed.ncbi.nlm.nih.gov/32022693/)]
77. Lee M, Sin J, Laban G, Kraus M, Clark L, Porcheron M, et al. Ethics of Conversational User Interfaces. In: CHI EA '22: Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems. New York, NY: Association for Computing Machinery; 2022 Presented at: CHI '22: CHI Conference on Human Factors in Computing Systems; April 29, 2022 to May 5, 2022; New Orleans, LA p. 1-7. [doi: [10.1145/3491101.3503699](https://doi.org/10.1145/3491101.3503699)]
78. Clark L, Doyle P, Garaialde D, Gilmartin E, Schlögl S, Edlund J, et al. The State of Speech in HCI: Trends, Themes and Challenges. *Interact Comput* 2019;31(4):349-371. [doi: [10.1093/iwc/iwz016](https://doi.org/10.1093/iwc/iwz016)]
79. Murad C, Munteanu C, Cowan B, Clark L, Porcheron M, Candello H, et al. Let's talk about CUIs: Putting conversational user interface design into practice. In: Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems. New York, NY: Association for Computing Machinery; 2021 Presented at: 2021 CHI Conference on Human Factors in Computing Systems; May 8-13, 2021; Yokohama, Japan p. 1-6. [doi: [10.1145/3411763.3441336](https://doi.org/10.1145/3411763.3441336)]
80. Ethical guidelines for voice experiences a case for inclusivity and trustworthiness. openvoicenetwork. 2022. URL: [https://openvoicenetwork.org/documents/ovn\\_ethical\\_guidelines\\_voice\\_experiences.pdf](https://openvoicenetwork.org/documents/ovn_ethical_guidelines_voice_experiences.pdf) [accessed 2022-08-01]
81. Seehausen M, Kazzer P, Bajbouj M, Prehn K. Effects of empathic paraphrasing - extrinsic emotion regulation in social conflict. *Front Psychol* 2012;3:482 [FREE Full text] [doi: [10.3389/fpsyg.2012.00482](https://doi.org/10.3389/fpsyg.2012.00482)] [Medline: [23162516](https://pubmed.ncbi.nlm.nih.gov/23162516/)]
82. Jain M, Kumar P, Kota R, Patel S. Evaluating and Informing the Design of Chatbots. In: Proceedings of the 2018 Designing Interactive Systems Conference. New York, NY: Association for Computing Machinery; 2018 Presented at: 2018 Designing Interactive Systems Conference; June 9-13, 2018; Hong Kong, China p. 895-906. [doi: [10.1145/3196709.3196735](https://doi.org/10.1145/3196709.3196735)]
83. Fraser H, Coiera E, Wong D. Safety of patient-facing digital symptom checkers. *Lancet* 2018 Nov 24;392(10161):2263-2264. [doi: [10.1016/S0140-6736\(18\)32819-8](https://doi.org/10.1016/S0140-6736(18)32819-8)] [Medline: [30413281](https://pubmed.ncbi.nlm.nih.gov/30413281/)]

## Abbreviations

**ASR:** automatic speech recognition  
**CA:** conversational agent  
**MCI:** mild cognitive impairment  
**NLP:** natural language processing

*Edited by R Kukafka; submitted 06.04.22; peer-reviewed by R Ahmad, V Stara, M Rampioni, J Kim, C Bérubé; comments to author 03.06.22; revised version received 17.09.22; accepted 26.09.22; published 15.11.22*

### *Please cite as:*

Kocaballi AB, Sezgin E, Clark L, Carroll JM, Huang Y, Huh-Yoo J, Kim J, Kocielnik R, Lee YC, Mamykina L, Mitchell EG, Moore RJ, Murali P, Mynatt ED, Park SY, Pasta A, Richards D, Silva LM, Smriti D, Spillane B, Zhang Z, Zubatiy T  
*Design and Evaluation Challenges of Conversational Agents in Health Care and Well-being: Selective Review Study*  
*J Med Internet Res* 2022;24(11):e38525  
 URL: <https://www.jmir.org/2022/11/e38525>  
 doi: [10.2196/38525](https://doi.org/10.2196/38525)  
 PMID:

©Ahmet Baki Kocaballi, Emre Sezgin, Leigh Clark, John M Carroll, Yungui Huang, Jina Huh-Yoo, Junhan Kim, Rafal Kocielnik, Yi-Chieh Lee, Lena Mamykina, Elliot G Mitchell, Robert J Moore, Prasanth Murali, Elizabeth D Mynatt, Sun Young Park, Alessandro Pasta, Deborah Richards, Lucas M Silva, Diva Smriti, Brendan Spillane, Zhan Zhang, Tamara Zubatiy. Originally published in the *Journal of Medical Internet Research* (<https://www.jmir.org/>), 15.11.2022. This is an open-access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the *Journal of Medical Internet Research*, is properly cited. The complete bibliographic information, a link to the original publication on <https://www.jmir.org/>, as well as this copyright and license information must be included.