Journal of the American Medical Informatics Association, 2024, **31(3)**, 746–761
https://doi.org/10.1093/jamia/ocad222
Advance access publication 9 December 2023
Review

Review

Evaluation framework for conversational agents with artificial intelligence in health interventions: a systematic scoping review

Hang Ding, PhD[1,2,*], Joshua Simmich, PhD[1,2], Atiyeh Vaezipour, PhD[1,2], Nicole Andrews, PhD[1,2,3,4], Trevor Russell, PhD[1,2]

[1]RECOVER Injury Research Centre, Faculty of Health and Behavioural Sciences, The University of Queensland, Brisbane, QLD, Australia, [2]STARS Education and Research Alliance, Surgical Treatment and Rehabilitation Service (STARS), The University of Queensland and Metro North Health, Brisbane, QLD, Australia, [3]The Tess Cramond Pain and Research Centre, Metro North Hospital and Health Service, Brisbane, QLD, Australia, [4]The Occupational Therapy Department, The Royal Brisbane and Women's Hospital, Metro North Hospital and Health Service, Brisbane, QLD, Australia

*Corresponding author: Hang Ding, PhD, RECOVER Injury Research Centre, Faculty of Health and Behavioural Sciences, The University of Queensland, Level 7, Surgical Treatment and Rehabilitation Service (STARS), 296 Herston Road, HERSTON, 4006 Brisbane, QLD, Australia (h.ding@uq.edu.au)

Abstract

Objectives: Conversational agents (CAs) with emerging artificial intelligence present new opportunities to assist in health interventions but are difficult to evaluate, deterring their applications in the real world. We aimed to synthesize existing evidence and knowledge and outline an evaluation framework for CA interventions.

Materials and Methods: We conducted a systematic scoping review to investigate designs and outcome measures used in the studies that evaluated CAs for health interventions. We then nested the results into an overarching digital health framework proposed by the World Health Organization (WHO).

Results: The review included 81 studies evaluating CAs in experimental ($n = 59$), observational ($n = 15$) trials, and other research designs ($n = 7$). Most studies ($n = 72$, 89%) were published in the past 5 years. The proposed CA-evaluation framework includes 4 evaluation stages: (1) feasibility/usability, (2) efficacy, (3) effectiveness, and (4) implementation, aligning with WHO's stepwise evaluation strategy. Across these stages, this article presents the essential evidence of different study designs ($n = 8$), sample sizes, and main evaluation categories ($n = 7$) with subcategories ($n = 40$). The main evaluation categories included (1) functionality, (2) safety and information quality, (3) user experience, (4) clinical and health outcomes, (5) costs and cost benefits, (6) usage, adherence, and uptake, and (7) user characteristics for implementation research. Furthermore, the framework highlighted the essential evaluation areas (potential primary outcomes) and gaps across the evaluation stages.

Discussion and Conclusion: This review presents a new framework with practical design details to support the evaluation of CA interventions in healthcare research.

Protocol registration: The Open Science Framework (https://osf.io/9hq2v) on March 22, 2021.

Key words: chatbot; conversational agent; virtual assistant; healthcare; evaluation; systematic review.

Introduction

Conversational agents (CAs), also known as chatbots or virtual assistants, are software programs that are designed to imitate human conversations.[1,2] Over the past decade, CA technologies and applications have advanced rapidly with emerging artificial intelligence (AI)[3] including natural language processing[4] and machine learning.[5] Several CA applications have already become popular tools in our daily lives, such as ChatGPT,[6] Google Bard,[7] Siri, Google Assistant, and Alexa.[8]

With recent advances, CAs present new opportunities to assist in delivering health interventions.[9,10] For example, researchers have proposed CA-enabled programs in hospitals to provide surgery information,[11] patient triage,[12] inpatient care,[13] and post-discharge follow-ups.[14] Many CA programs have also been studied in community care to improve health education,[15–18] mental health,[19–24] and the self-management of chronic diseases.[25–27] To combat the COVID-19 pandemic, several large national and international health organizations including the World Health Organization (WHO)[28] have implemented CA applications[29] to assist in delivering timely health information[28] or screening the symptoms for early interventions.[28,30,31]

To use CAs in healthcare, rigorous evaluations are essential.[32,33] Conversations in CAs are usually controlled by AI. The use of AI is often associated with poor transparency (known as "the black box effect") because AI-based control mechanisms are normally complex and cannot be well explained.[33,34] A lack of transparency has been the leading concern for using AI-based applications in healthcare.[35,36] In addition, AI studies are often associated with various limitations (or unforeseeable errors) such as ineffective model

Received: July 23, 2023. **Revised:** November 4, 2023. **Editorial Decision:** November 11, 2023. **Accepted:** November 24, 2023

designs, insufficient prior knowledge base, inadequate training data, or inappropriate training processes.[33,37] Because of these limitations, AI-based systems sometimes fail to function as expected.[3,33] The failures often result in poor user experience, low adherence and uptake, ineffective health outcomes, inappropriate care advice, or even unintended harm.[3] These issues have recently been highlighted.[8,33,38–40] Therefore, rigorous clinical evaluations are essential for understanding CA performance,[41] preventing potential risks,[3] and, ultimately, achieving safe, effective, and sustainable interventions in healthcare.[35,36,41] However, effective evaluation of CAs, involving various design methods and strategies, is often complex and challenging.[41] Existing reviews on CA evaluations are limited to a narrow scope, such as technical metrics,[42–44] or simple method descriptions without systematic investigations.[45,46] To support CA evaluations, a comprehensive evaluation framework is needed[47–50] but remains absent.

The objective of the review is to synthesize existing CA evaluation methods and outline an evaluation framework for supporting future CA evaluation studies. We conducted this scoping review to extract the study designs and outcome measures of health-related CA studies. We then categorized the nested data according to an overarching digital health framework by WHO.[51] We finally discussed the findings and knowledge gaps in CA evaluations.

Methods

We conducted the scoping review in accordance with the PRISMA Extension for Scoping Reviews (PRISMA-ScR).[52] Our protocol was prospectively registered in the Open Science Framework on March 22, 2021.[53] We selected the scoping review approach because it allowed us to explore and synthesize complex and diverse evidence in the literature.[54]

Selection criteria

We designed the selection criteria (Table 1) focusing on peer-reviewed journal articles. The review only included the CAs that allowed users to talk or chat in a natural language without any constraints (unconstrained CA).[46] In contrast, some CAs only allowed users to enter predefined text messages, such as answering "Yes" or "No," or select options via forms, menus, or buttons (or *in situ*[55]). We excluded constrained CAs because their conversations and AI functions are often limited.

Search strategy

We searched five databases: CINAHL, Medline via Ovid, Scopus, Embase, and IEEE Xplore, using a search strategy with variants and combinations of search terms relevant to CAs and health interventions (Appendix S1). We included articles published in English from the inception of the databases to January 13, 2021. In addition to the database searches, we also manually identified articles from existing systematic reviews in the CA research field.[9,42–46,56]

Data extraction

We used EndNote (Ver. 20) to export the articles from each database and Covidence[57] to screen and extract the data. A data extraction form was developed according to the review protocol.[53] Two authors (H.D. and J.S.) independently screened the title and abstract of each article. They then conducted full-text reviews to determine the eligible articles. The discrepancies between the 2 authors were resolved by consensus and discussions with the third author (A.V.). We extracted country names according to the recognized members in the United Nations.[58]

Synthesis of results

We identified the design of each study according to published design definitions/descriptions,[59–63] design guide,[59]and overview[64] (Appendix S2).

We extracted the outcome measures from each study. We accordingly identified seven widely used categories: (1) functionality,[51] (2) safety and information quality,[3,65] (3) user experience,[51,66,67] (4) clinical/health outcomes, (5) costs and cost benefits,[68] (6) usage, adherence, and uptake from objective analysis of conversation records, distinct from similar subjective evaluations in "User experience," and (7) user characteristics for implementation.[51]

To generate an evaluation framework, we employed an overarching evaluation framework for digital health interventions, published by the WHO.[51] The framework provides evaluation descriptions and targets across four evaluation stages: (1) "feasibility and usability," (2) "efficacy," (3) "effectiveness," and (4) "implementation".[51] These stages are fundamentally consistent with the 4 widely known phases of clinical trials.[47–50] We accordingly nested the extracted data across these 4 stages. We selected the WHO's framework because it was the state of the art, the most comprehensive, and well-recognized in the digital health research field.

Table 1. The inclusion and exclusion criteria.

Inclusion criteria	Exclusion criteria
At least one objective of the study was to evaluate CA intervention(s), including CA applications, healthcare modes, or programs.	The CA in the study was only one function of a robot, robotic toy, virtual reality, or game-based application and its conversational function was not evaluated independently.
The CA was unconstrained, allowing users to enter text-based messages or conduct voice-based conversations.	The CA mainly sent messages/notifications, asked users to enter data entries, or conversed through predefined responses (buttons, menus, yes, no, etc.)
The evaluation was relevant to primary, secondary, or tertiary prevention of disease or health issues.	The CA application was designed to provide education, training, or knowledge/skill assessments to healthcare providers and/or students.
The evaluation was based on the analysis of data from humans, including testers, participants recruited, or people using the intervention in the real world.	The study was mainly based on AI training data or an exploratory survey to investigate the preferences/perceptions of a CA application.
The article was published in a peer-reviewed journal in the English language.	The study only reported preliminary analysis outcomes, normally in conferences or communications.

Quality assessment

We assessed the reporting quality using the mobile health (mHealth) evidence reporting and assessment (mERA) checklist.[69] This assessment approach has been used in similar reviews,[70,71] and the WHO's digital health evaluation guide.[51] The mERA checklist includes 13 domains. We classify each domain into "Fully reported," "Partially reported," and "Not reported." Two authors (H.D., J.S.) independently assessed each included article, and the discrepancies in the assessment were resolved through discussion.

Results

We retrieved a total of 6350 articles from the search (Figure 1), including 6293 articles from the databases and 57 articles from existing reviews. We then removed 3647 duplicates, 1404 articles through the title-abstract screening, and a further 106 articles through the full-text review. We finally included 81 articles for the data extraction.

Study characteristics

The 81 articles included in this review (Appendix S3) were published between 2009 and 2022 (Figure 2A), with 89% of them ($n = 72$) in the past 5 years. The studies originated from 21 countries (Figure 2B), predominantly from the United States ($n = 31$, 38%). We identified 12 main intervention areas (Figure 2C) with the leading of "Mental, psychological, or cognitive health" ($n = 30$, 37%).

Eight health-related CAs were available to the public, focusing on chronic disease or conditions (DoctorBot[15] and Gia[72]), adolescent health education (on contraception, Layla[73]), mental health (Bunji,[74] ELIZA,[75] Wysa,[76–78]

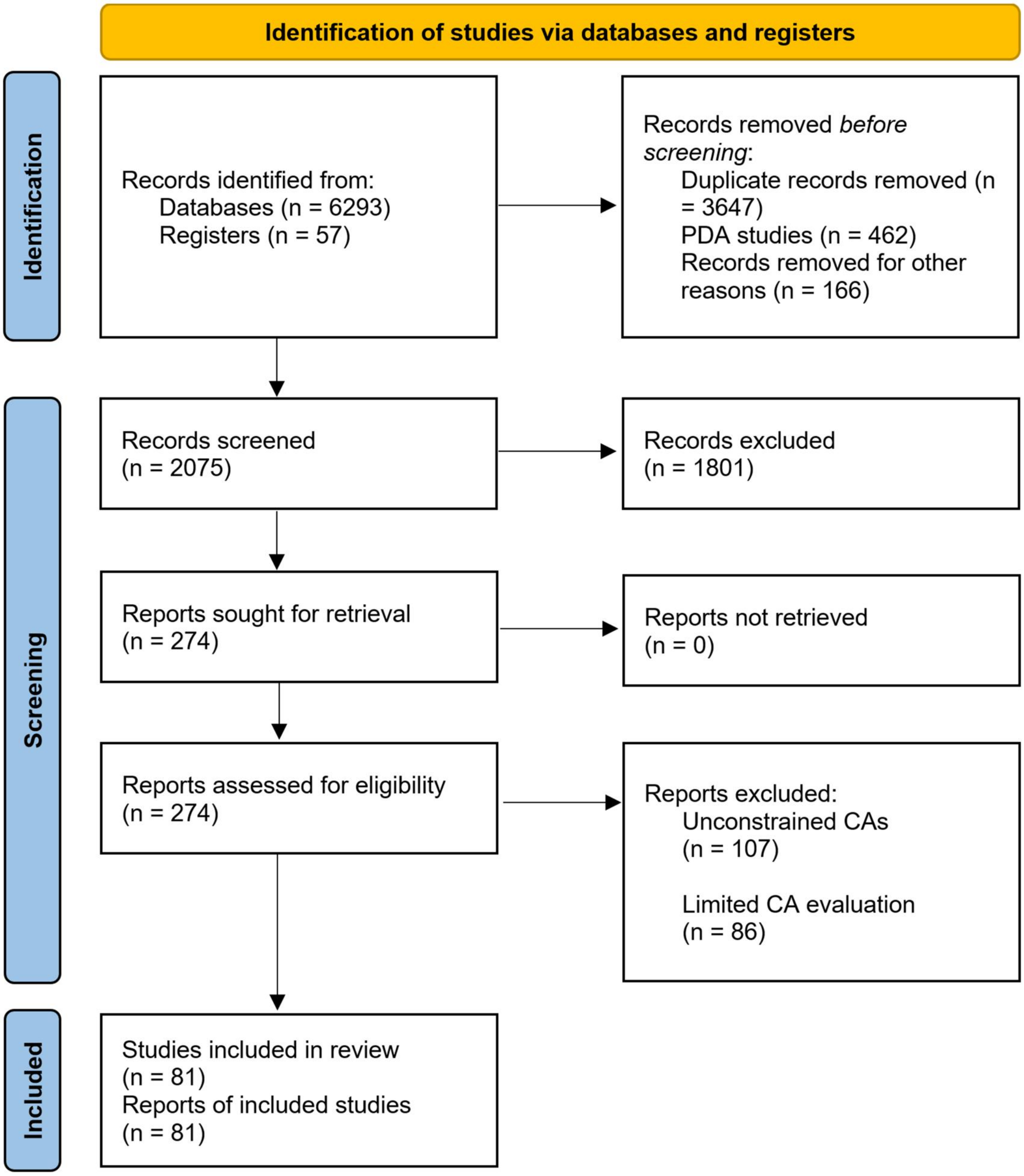

Figure 1. Flow diagram of the bibliographic search results and the included articles through the title-abstract screening and full-text review. CAs, conversational agents; PDA, personal digital assistant—a type of handheld computer irrelevant to CA.

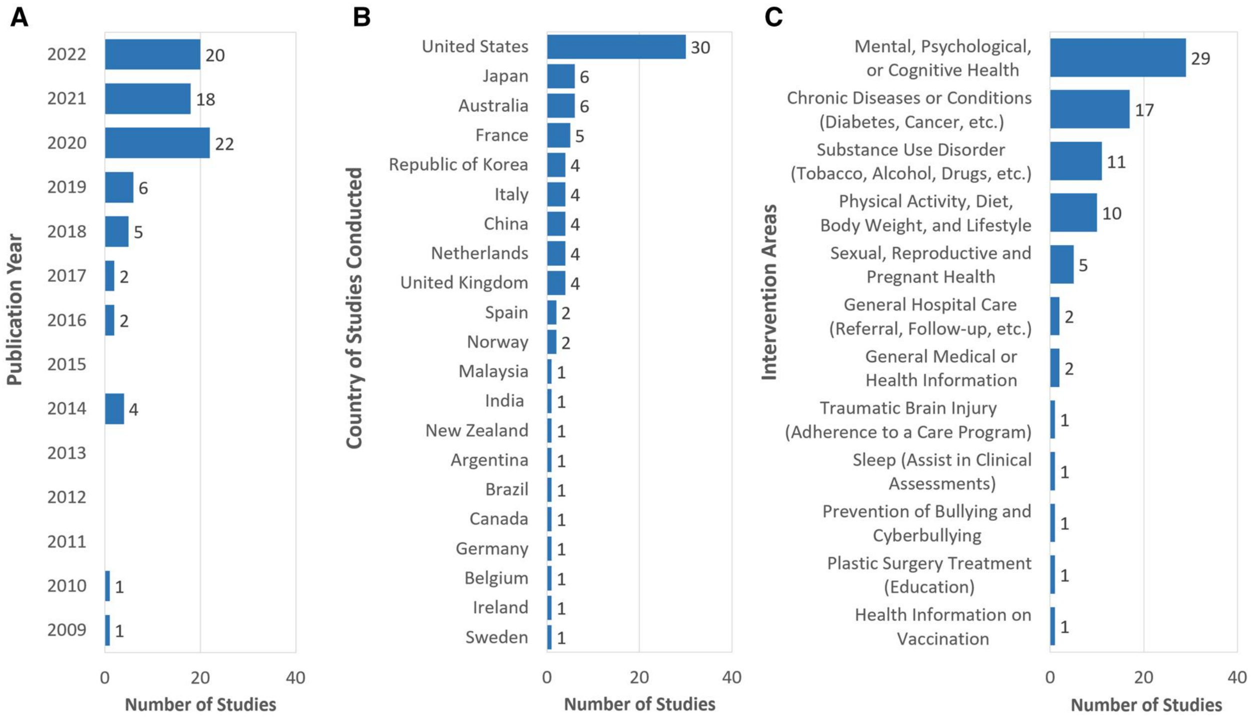

Figure 2. Characteristics of studies in the review with the final bibliographic search in January 2021. (A) Publication year. (B) Country of the studies conducted. (C) Intervention area.

Woebot,[79] and PAT[80]), and substance use disorders (Woebot[81,82]). The general CAs of Siri, Google Assistant, and Alexa were also evaluated for providing health information.[8,17,19,83–87] The remaining CAs were mainly under research.

Study designs

We identified 8 categories of designs (Appendix S3) and summarized them in a hierarchical diagram (Figure 3). There were 15 observational and 59 experimental studies. The observational studies included the cross-sectional ($n = 11$), cohort ($n = 1$), and case-control ($n = 3$) designs. The experimental studies included 20 randomized controlled trials (RCTs) and 38 quasi-experimental trials (or nonrandomized studies). Among the RCTs, there were 20 parallel RCTs and 1 crossover RCT.

We found 7 studies (9%) in which investigators (usually 2 authors) tested CAs using predefined questions and reviewed CAs' responses to determine the safety and information quality of CA interventions.[17,19,83–87] We categorized these studies in a separate design category ("laboratory setting") because the investigators neither assigned the intervention to participants in an experiment/trial (experimental studies[59]) nor observed intervention effects on people in usual clinical practice (observational studies[59]).

We also found 10 other studies (12%) that were not closely related to clinical interventions. One single-arm study evaluated safety and information quality (Siri, Google Assistant, and Alexa).[8] Three studies investigated the differences between 2 CAs (MYLO vs ELIZA, a two-arm parallel RCT)[75] or a CA and traditional search engines (a cross-sectional study).[88] One single-arm study explored the level of self-disclosure (the willingness to answer sensitive or private questions, a single-arm study).[89] Three studies investigated the potential to substitute a CA for conventional clinical assessments (a CA vs a questionnaire, a crossover RCT[40]; a CA vs an interview,[90] single-arm trials; a CA vs a pain questionnaire, a single-arm experimental study[91]). The remaining two studies mainly investigated users' expectations and preferences,92 and the barriers and facilitators of CAs.[80]

Interventions were limited in 23 studies. Many studies ($n = 22$, 27%) had participants performing predefined tasks[23,25,93,94] or conversing with the chatbot for only a single session.[8,11,18,20,22,24,40,75,90,91,95–102] One RCT recruited participants to only review conversation responses (rather than to converse with the CA) from either a CA (Intervention) or a medical committee (Control).[39]

Outcome measures

We identified 285 outcome measures and categorized them into 7 main categories (Appendix S4). Table 2 summarizes the main categories with subcategories and selected outcome measures. The main categories included "Functionality" (Number of outcome measures, $N_{\mathrm{om}} = 44$), "Safety and information quality" ($N_{\mathrm{om}} = 17$), "User experience" ($N_{\mathrm{om}} = 80$), "Clinical/health outcomes" ($N_{\mathrm{om}} = 68$), "Costs and cost benefits" ($N_{\mathrm{om}} = 2$), "Usage, adherence, and uptake" ($N_{\mathrm{om}} = 62$), and "User characteristics for implementation science" ($N_{\mathrm{om}} = 12$).

"Functionality" examines how well CAs functioned as designed. Researchers evaluated how sentence classifications functioned to interpret users' intentions or intents ("Sentence classification performance" with accuracy[25], precision[25,93], etc.) and overall CAs' conversation functions in terms of "Understanding and responses." Some CAs were designed to conduct various small tasks (screening alcohol use[94], collecting symptoms[25], etc.) or engage with users (initiating new topics[26], providing social support[92], etc.) for long-term personalized interventions. These design functions were also evaluated ("User engagement" and "Task achievements and

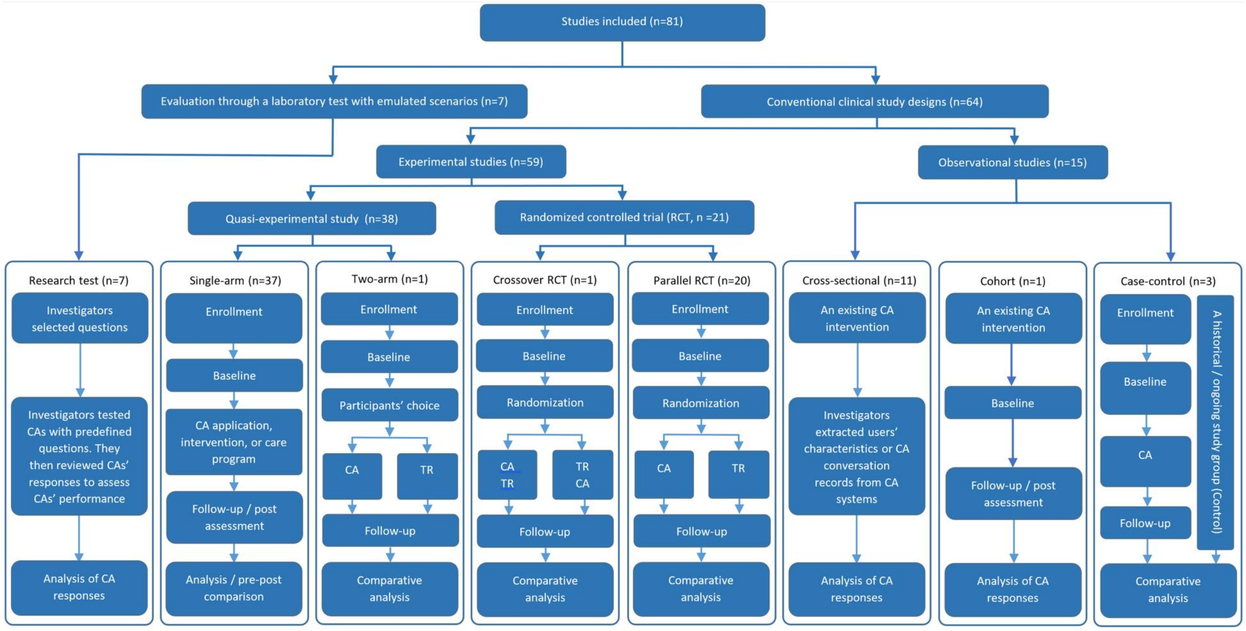
Studies included (n=81)
Evaluation through a laboratory test with emulated scenarios (n=7)
Conventional clinical study designs (n=64)
Experimental studies (n=59)
Observational studies (n=15)
Quasi-experimental study (n=38)
Randomized controlled trial (RCT, n =21)
Research test (n=7)
Investigators selected questions
Investigators tested CAs with predefined questions. They then reviewed CAs' responses to assess CAs' performance
Analysis of CA responses
Single-arm (n=37)
Enrollment
Baseline
CA application, intervention, or care program
Follow-up / post assessment
Analysis / pre-post comparison
Two-arm (n=1)
Enrollment
Baseline
Participants' choice
CA
TR
Follow-up
Comparative analysis
Crossover RCT (n=1)
Enrollment
Baseline
Randomization
CA TR
TR CA
Follow-up
Comparative analysis
Parallel RCT (n=20)
Enrollment
Baseline
Randomization
CA
TR
Follow-up
Comparative analysis
Cross-sectional (n=11)
An existing CA intervention
Investigators extracted users' characteristics or CA conversation records from CA systems
Analysis of CA responses
Cohort (n=1)
An existing CA intervention
Baseline
Follow-up / post assessment
Analysis of CA responses
Case-control (n=3)
Enrollment
Baseline
CA
Follow-up
A historical / ongoing study group (Control)
Comparative analysis

Figure 3. The flowchart of different study designs in the studies. CA, conversational agent (intervention, in comparative studies); TR, a traditional intervention/care program (control, in comparative studies); RCT: randomized controlled trial.

efforts"). For speech-based CAs, two studies evaluated "Voice and device control", such as voice volume, speed, and sound quality.[91] Finally, some CAs were designed for clinical assessments, such as assessing depressive disorders,[40] sleep conditions,[90] and tobacco use disorders.[100] Researchers hence evaluated "Clinical assessment performance" to validate these CAs as clinical assessment tools.

"Safety and information quality" were evaluated, especially for CAs with large language models or contents from the Internet. Researchers examined whether CAs' responses were appropriate and safe ("CA response appropriateness"). They moreover evaluated the "Risk of misinformation." Misinformation refers to a health-related claim of fact that is currently false due to a lack of scientific evidence.[130] Misinformation-related factors were also evaluated, such as resource reliability,[86] evidence-based resources,[85] information accuracy and completeness,[85] and information quality.[84] As CAs could affect users' care decisions, the "Risk of unintended harms or adverse events" was evaluated. "Unintended harms" refers to harmful consequences if the action is not taken timely or appropriately, such as actions for allergic conditions and emergency tasks.[8] "Adverse events" include deaths,[8] suicide attempt, and alcohol/drug overdose.[82] Finally, researchers evaluated "Privacy and trust" using surveys,[55,105] and/or qualitative studies (interviews).[55]

The category of "User experience" included 80 outcome measures in 13 subcategories. None of the outcome measures was specifically designed for evaluating CAs. "Clinical/health outcomes" also included a large number of outcome measures ($N_{\mathrm{om}} = 68$) but were mainly evaluated with validated assessment questionnaires. "Costs and cost benefits" were only reported by 2 studies. Evaluations of "Usage, adherence, and uptake" were diverse with 62 outcome measures. "User characteristics for implementation science" were mainly limited to "Age and gender."

Proposed evaluation framework

The CA evaluation framework comprises 4 essential evaluation stages: (1) Feasibility and usability, (2) efficacy, (3) effectiveness, and (4) implementation (Table 3). For each stage, the framework presents the evaluation descriptions, targets, and illustrative sample sizes, according to WHO's recommendations for digital health interventions[51]. It presents existing study (or trial) designs, outcome measures, and sample sizes. The framework also highlights (in light blue) essential outcome measures. For example, at Stage I, existing studies mainly evaluated CAs using single-arm experimental trials ($n = 30$, 75%) and laboratory tests ($n = 7$, 17.5%). The most widely used sample sizes were in the 10-20 range for single-arm studies and 1 or 2 experts for laboratory tests. The essential outcome measures can be potentially used as primary outcomes. In clinical studies, primary outcomes are aligned with the primary aim to answer important research questions (or hypothesis)[134] and, sometimes, determining study sample sizes,[134] for example, the power analysis in RCTs[135].

Cost-related evaluations are very limited in the framework, with only 2 studies at Stage I and II, respectively. However, studies at Stages I and II are conducted under a research setting, unable to produce generalizable results. We recommend "Costs and health economic analyses" at Stages III and IV. The evaluations potentially include cost-utility analysis, cost-effective analysis, cost-minimization analysis, and cost-benefit analysis.[136]

Quality assessment

We assessed the reporting quality of the included studies using the mERA checklist[69] (Appendix S5). As the studies in a laboratory setting ($n = 7$) did not evaluate the interventions, many mERA criteria were not applicable to these studies. We, therefore, reported the mERA summary for the remaining 74 studies (Figure 4). Most studies provided "fully

Table 2. The summary of the outcome measures nested in the seven main categories and subcategories in the review.

Category and subcategory (number of outcome measures, N_{om})	Selected typical outcome measures (unit, questionnaire or method)
1. Functionality	($N_{\text{om}} = 44$)
• Sentence classification performance ($N_{\text{om}} = 5$)	Precision (%),[25,93] sensitivity (%),[25,93] accuracy (%),[25] and specificity (%),[25] and *F*1 (value)[25,93] of the classifier.
• Understanding and responses ($N_{\text{om}} = 17$)	Response accuracy (%),[17,19,86,103], inquiries unable to answer (%),[86] response completion (%),[86] understanding (scale, survey),[91] etc.
• Engagement functions ($N_{\text{om}} = 7$)	Topics initiated by CA versus participants,[26] attempts to restart conversation (*n*),[8] sentiment (score, coding responses manually),[17] etc.
• Task achievements and efforts ($N_{\text{om}} = 4$)	Conversation tasks completed (%),[94] task failure rate (*n* and %),[8] and task completion (coefficient),[25] and time per task (seconds).[8]
• Voice and device control ($N_{\text{om}} = 2$)	Adequate volume, speed, and sound quality (a survey),[91] and negative technical aspects (qualitative analysis of user's responses).[104]
• Clinical assessment performance ($N_{\text{om}} = 9$)	Accuracy,[40,100] sensitivity,[40,90] specificity of CA-based clinical assessment outcomes (CA vs standard clinical assessments),[40,90] etc.
2. Safety and information quality	($N_{\text{om}} = 17$)
• CA response appropriateness ($N_{\text{om}} = 6$)	Response appropriateness (scale),[19,86,87] appropriate responses (descriptive),[83] etc.
• Risk of misinformation ($N_{\text{om}} = 4$)	Misinformation (%),[17] reliable (%)[86] and evidence-base (%)[85] resources, information accuracy and completeness (%),[85] and quality (descriptive).[84]
• Risk of unintended harms or adverse events ($N_{\text{om}} = 4$)	Responses with risk of unintended harms (*n* and %; eg, medication and emergency tasks),[8] serious adverse events (*n*),[82] and deaths (*n* and %).[8]
• Privacy and trust ($N_{\text{om}} = 3$)	Privacy and trust (a survey),[55] privacy and trust (a qualitative study, interview),[55] and privacy infringement (a survey).[105]
3. User experience	($N_{\text{om}} = 80$)
• Ease of use ($N_{\text{om}} = 2$)	Ease of use (scale, a self-designed questionnaire)[11,88] and learning experience (score, a self-designed questionnaire).[106]
• Engagement ($N_{\text{om}} = 3$)	User engagement (scale, a survey),[95] DBCI engagement (scale),[92] and perceived engagement (scale, a survey).[101]
• Conversation capability ($N_{\text{om}} = 6$)	Response appropriateness (scale, a survey),[11] dialogue performance (score, SASSI),[94] emotional awareness (score, a questionnaire),[106] etc.
• Usefulness/helpfulness ($N_{\text{om}} = 6$)	Usefulness (scale, a survey[21,107] or interview[88]), perceived helpfulness (a survey, open-ended question, or interview),[21,108] etc.
• Perceived quality and trust ($N_{\text{om}} = 5$)	Perceived trust (score, a questionnaire),[89] perceived quality of the answers (score, EORTC QLQ-INFO25),[39] etc.
• Satisfaction ($N_{\text{om}} = 5$)	Satisfaction (scale, a self-defined survey,[8,15,72,99,105,107,109–111] and CSQ-8[81,82,112]), content satisfaction (scale, a survey),[106] etc.
• Feasibility ($N_{\text{om}} = 3$)	Feasibility (score, a self-designed questionnaire).[18,20,26,81]
• Usability ($N_{\text{om}} = 5$)	Usability (scale, SUS),[75,96,107,113,114] usability (open comments, a focus group session),[27] perceived usability (scale),[92,105] etc.
• Acceptance/preference ($N_{\text{om}} = 11$)	Acceptance (scale, a survey),[88] preference of CA (scale, a survey),[88] potential to replace humans (scale, a survey),[11] etc.
• Overall user experience with mixed themes ($N_{\text{om}} = 26$)	Overall user experience (UEQ,[23,25] USE,[96] NPS,[115] URP-I[81,82]), users with positive or negative experience (*n*, the CA prompted the survey),[15] etc.
• Working alliance ($N_{\text{om}} = 1$)	Working alliance (questionnaire, WAI-SR[78,79,81,82,101,112]).
• Suggestions for improvement ($N_{\text{om}} = 4$)	Suggestions for improvements (open-ended question in a survey),[20,93,108,110] good and bad experiences with the CA (a survey),[109] etc.
• Other open comments ($N_{\text{om}} = 3$)	Perceived stress (survey and interview),[24] benefit (focus group study),[27] and feelings of answering sensitive questions (CA vs humans).[89]
4. Clinical/health outcomes	($N_{\text{om}} = 68$)
• Psychological/mental health ($N_{\text{om}} = 34$)	PHQ-9,[21,95,106,109,112,116] QIDS-SR,[105], GAD-7,[21,81,82,95,106,109,112,116–118] SAS,[105] PANAS,[106,,109,112,119] PSYCHLOPS,[21] DASS21,[75,117] PSS-10,[95,105,120] etc.

(continued)

Table 2. (continued)

Category and subcategory (number of outcome measures, N_{om})	Selected typical outcome measures (unit, questionnaire or method)
• Disease conditions ($N_{om} = 3$)	Pain (%, NRS),[82] Parkinson's disease rating scale (MDS-UPDRS),[121] and Parkinson's disease questionnaire.[121]
• Modification of behaviors and risk factors ($N_{om} = 23$)	Behavior modification (score, SQUASH),[122,123] smoking cessation (%, a survey),[124] physical activity (score, AAS),[113,123] etc.
• Knowledge and skills ($N_{om} = 4$)	Knowledge gained (scale, a survey),[16,18] problem solvability (score, a survey),[75] problem resolution (score, a survey),[75] etc.
• Health wellbeing and issues ($N_{om} = 4$)	SWLS,[120] WHO-5-J,[95,125] EQ-5D-5L,[126] and falls (falls per 1000 patient-days).[13]
5. Costs and health economic analyses	($N_{om} = 2$)
• Cost effectiveness ($N_{om} = 1$)	Time spent per 100 patients (hours per 100 patients, an analysis of the conversation logs).[14]
• Costs ($N_{om} = 1$)	Monthly budget (dollars per month, an analysis of running costs of the CA system).[73]
6. Usage, adherence and uptake	($N_{om} = 62$)
• Usage ($N_{om} = 38$)	Conversation duration (second, minute, or hour),[14,15,26,55,72,73,75,81,82,88,115,126–128] exchanges (n),[108–110], CA responses (n),[92,127,129] etc.
• Adherence ($N_{om} = 15$)	Adherence (n),[108,120] dropouts (n, %; conversation dropouts,[15] and dropouts of interventions[15,116,122]), follow-up rate (%),[14] etc.
• Uptake ($N_{om} = 9$)	Completed questionnaires (n),[122] total followers (n),[73] total impressions (n),[73] average daily reach times (times of reach per day),[73] etc.
7. User characteristics for implementation science	($N_{om} = 12$)
• Age and gender ($N_{om} = 2$)	Age (age groups, n, %)[15,72,88,110] and gender (n, %).[15,72,88,110]
• Nationality, ethnicity and religion ($N_{om} = 4$)	Nationality (n),[88] race and ethnicity (%, White, Hispanic, Black),[72] religion (n),[88] and language (%, users in Spanish).[72]
• Education and socioeconomic status ($N_{om} = 2$)	Occupation and education (n, %, self-designed questionnaire),[88] and urbanization levels (n, self-designed questionnaire).[88]
• Health conditions ($N_{om} = 3$)	Users with a personal history of cancer (%),[72] a family history of cancer (%),[72] and risks of different cancers (NCCN criteria, Tyrer–Cuzick criteria).[72]
• Devices used ($N_{om} = 1$)	Mobile users (%).[73]

AAS, The Active Australia Survey; CSQ-8, Client Satisfaction Questionnaire with 8 questions; DASS21, depression, anxiety, and stress scales 21; DBCI, a questionnaire on the Digital Behavior Change Intervention; EORTC QLQ-INFO25, The European Organisation for Research and Treatment of Cancer Quality of Life Group information questionnaire; EQ-5D-5L, health-related quality of life with 5 dimensions: mobility, self-care, usual activities, pain/discomfort, and anxiety/depression; F1, the harmonic mean (average) of the precision and recall; GAD-7, the Generalized Anxiety Disorder scale—7; NCCN criteria, National Comprehensive Cancer Network criteria; NPS, net promoter score; PANAS, the positive and negative affect schedule; PHQ-9, the Patient Health Questionnaire—a 9-item self-report questionnaire that assesses the frequency and severity of depressive symptomatology within the previous 2 weeks; PSS-10, the Perceived Stress Scale; PSYCHLOPS, the psychological outcome profiles; ROC, receiver operating characteristic—a graphical plot to evaluate a binary classifier/decision system across different discrimination thresholds; QIDS-SR, the Quick Inventory of Depressive Symptomatology-Self-report; SAS, the Self-rating Anxiety Scale; SASSI, the Subjective Assessment of System Speech Interfaces—a 7-point Likert scale on accuracy, likeability, cognitive demand, annoyance, habitability, and speed; SQUASH, the Dutch Short Questionnaire to assess health enhancing physical activity; SUS, the system usability scale; SWLS, the Satisfaction with Life Scale; UEQ, the User Experience Questionnaire; URP-I, usage rating profile-intervention with feasibility (6 items) and acceptability (6 items) scales; USE, the Usefulness, Satisfaction, and Ease of Use (USE) Questionnaire Short-Form; WAI-SR, the Working Alliance Inventory-Short Revised (agreement on the tasks of therapy, agreement on the goals of therapy and development of an affective bond); WHO-5-J, HEALTH well-being—5 Well-Being Index (Japanese version).

reported" data for 4 mERA assessment domains, namely "Intervention delivery" (>80%), "Intervention content" (>70%), "User feedback" (>60%), and "Intervention fidelity" (>50%). The report rates for the remaining domains were low (<50%).

Discussion

This review includes 81 evaluation studies on CAs for health interventions. The studies were heterogeneous with regard to evaluation methods, intervention strategies and focus. Most CA studies were reported within the past 5 years ($n = 72$, 89%). In the review, we extracted study designs ($n = 8$), sample sizes, and outcome measures ($n = 285$). Then, we categorized and nested the extracted data according to the overarching framework for digital health evaluations by WHO.[51] We finally outlined a new framework for CA evaluations.

New CA evaluation framework

The new CA evaluation framework synthesizes the existing evidence across 4 evaluation stages. The evidence is rich at Stage I (40 studies, 49%) and gradually becomes limited at Stage IV (8 studies, 10%). Despite the limitations, the framework streamlines evaluation stages, targets, designs, sample sizes, and essential outcome measures (main categories) along the stages. Moreover, the framework includes 2 new essential evaluation aspects: "Functionality" and "Safety and information quality" at Stage I. It also presents the evaluation gaps of "Costs and health economic analyses" at Stages III and IV.

Table 3. The proposed framework for evaluating CAs in healthcare.

	Stage →	1. Feasibility and usability →	2. Efficacy →	3. Effectiveness →	4. Implementation
WHO digital health	Brief description	Feasibility: The ability to work as intended. Usability: The degree of a system being used to achieve specified goals in a specified context of use.	Efficacy: The ability to achieve the intended results in a research setting or trial.	Effectiveness: The ability to achieve the intended results in a real application (nonresearch setting).	Implementation science: To assess the uptake, integration and sustainability of evidence-based digital health interventions for a given context, including policies and practices.
	Evaluation targets	• Stability (system uptime/failure rates) • Performance consistency • Standards adherence (terminology, interoperability, security)	• User satisfaction • Workflow "fit" • Learning curve (design) • Cognitive performance/errors • Reliability	• Changes in care processes (time) • Changes in outcomes (system performance/health)	• Changes in process, outcome, coverage, and costs • Total cost of implementation, and health impact • Error rates • Learning curve of users • Changes in policy, practices attributable to system • Adaptability and extendibility to new use-cases
	Illustrative num of users	10-100	100-1000	10 000 +	1000 000+
Studies reviewed and outcome measures at 4 major evaluation stages aligned with the WHO guide	Studies (n)	40	21	12	8
	Study design and sample size (n)	Single-arm studies (n = 3-10,[107,114,131] 11-20,[21,27,91,95,97,103,111] 21-30,[11,20,23,24,26,96] 31-40,[25,80,90] 41-50,[8,55,98,105], 73,[92] 89,[94] 101,[81] 116,[113] 121,[93] 318[74]; 4390 messages[73]), 2-arm quasi-experimental study(n = 454[125]), laboratory tests (investigators, n = 1,[19] 2[17,83–87]) and RCT (n = 142,[39] and 289[132])	Single-arm studies (n = 28,[108] 31,[123] 44,[18] 47,[89] 61,[77] 128,[99] 139,[100] and 154[22]) Case-control study (n = 153,[76] and 270[14]) Randomized controlled trials (n = 20,[121] 23,[119] 70,[106] 74,[109] 83,[112] 112,[75] 153,[101] 197,[115] 181,[116] 210,[102] and 958[122])	Cross-over study (n = 179[40]) Case-control study (n = 95[13]) Cross-sectional studies (n = 354,[127] 929,[88] 4737[110]) Randomized controlled trials (n = 28,[120] 180,[82] 513,[126] 700,[118] 927,[16] 57 214[124]) Cohort study (n = 3629[117])	Cross-sectional studies (n = 1206,[78] 7099,[104] 16 519,[15] 14 698,[129] 36 070,[79] 61 070,[72] 135 263[128]; 610 conversations[133])
	User characteristics for implementation science	• Devices used.[73]	–	• Age and gender[88,110] • Nationality, ethnicity and religions[88] • Education and socioeconomic status[88]	• Age and gender[15,72] • Nationality, ethnicity and religions[72] • Health conditions[72]
	Usage, adherence and uptake	• Usage, [21,26,55,73,81,92,98,103,113,114,131] • Uptake[27,73]	• Usage,[14,75,77,108,109,115,116,121,123] • Adherence[14,108,112,116,119,120]	• Usage,[82,88,110,117,124,126,127] • Uptake[122] • Adherence[122]	• Usage,[15,72,128,129,133] • Uptake[72,129] • Adherence[15,72,133]
	Costs and health economic analyses	• Costs[73]	• Cost effectiveness[14]	• Health economic analyses, such as cost-utility analysis, cost-effective analysis, cost-minimization analysis, and cost-benefit analysis. (our recommendation)	• Health economic analyses, such as cost-utility analysis, cost-effective analysis, cost-minimization analysis, and cost-benefit analysis. (our recommendation)

(continued)

Table 3. (continued)

Stage →	1. Feasibility and usability →	2. Efficacy →	3. Effectiveness →	4. Implementation
Clinical/health outcomes	• Knowledge and skills[21] • Health wellbeing and issues[95,125] • Psychological/mental health,[21,55,74,81,95,105,125] • Behavioral modification and risk factors[26,81,113]	• Disease conditions[121] • Knowledge and skills[18,75] • Health wellbeing and issues[120] • Psychological/mental health.[75–77,102,106,109,112,115,116,119–121] • Clinical assessment performance,[22,89,100] • Behavioral modification and risk factors[108,123,100]	• Disease condition[82] • Knowledge and skills[16] • Health wellbeing and issues[13,126] • Psychological/mental health[13,16,82,117,118] • Clinical assessment performance[40] • Behavioral modification and risk factors[16,82,110,118,122,124,126]	• Psychological/mental health[78,79] (through short inbuilt questionnaires in CA apps)
User experience	• Usability[27,92,96,103,105,107,113,114] • Feasibility[20,26,81] • Engagement[92,95] • Ease of use[11] • Satisfaction[8,81,91,105,107,111] • Open comments[24,27] • Working alliance[81] • Overall experience[8,21,23–25,55,74,80,81,92,94–98,105,113,114,131,132] • Usefulness/helpfulness[21,55,107] • Acceptance/preference[11,81,98,105,132] • Conversational capability[11,23,90,94,107] • Perceived quality and trust[39,97,132] • Suggestions for improvement[20,24,27,93]	• Usability[75,115] • Feasibility[18] • Ease of use[106] • Satisfaction[99,106,109,112] • Engagement[101] • Working alliance[101,112] • Overall experience[89,99,106,108,112,115,120] • Other open comments[89] • Acceptance/preference[18,75,89,100,122] • Usefulness/helpfulness[108,116] • Conversational capability[106] • Perceived quality and trust[89] • Suggestions for improvement[108,109]	• Ease of use[88] • Satisfaction[82,110,122] • Working alliance[82] • Overall experience[16,82,110,120] • Usefulness/helpfulness[88,118] • Acceptance/preference[82,88] • Conversational capability[88] • Suggestions for improvement[110]	• Satisfaction[15,72] • Working alliance[78,79] • Overall experience[15,78] • Acceptance and preference[104]
Safety and information quality	• Privacy and trust[55,105] • Risk of causing death[8] • CA response capability[17,19,85,86] • Risk of misinformation[17] • Risk of unintended harms[8,85] • CA response appropriateness[19,83,86,87,105] • Resources and contents quality[84–86]		• Risk of unintended harms[82]	
Functionality	• Response speed[8] • Task achievements[8,25,94] • Engagement functions[8,17,26,85,94] • Voice and device control[91] • Classification performance[25,93] • Clinical assessment performance[90] • Understanding and accurate responses[11,19,86,91,98,103,107,111,132]	• Understanding and accurate responses[108]	–	• Voice and device control[104] • Understand and accurate response[133]

The framework demonstrates the included CA evaluation studies ($n = 43$), study designs, outcome measures and sample sizes at four essential evaluation stages. The 4 stages and corresponding evaluation targets were proposed by the WHO. Essential categories, which we identified for each stage, are marked by a light blue.

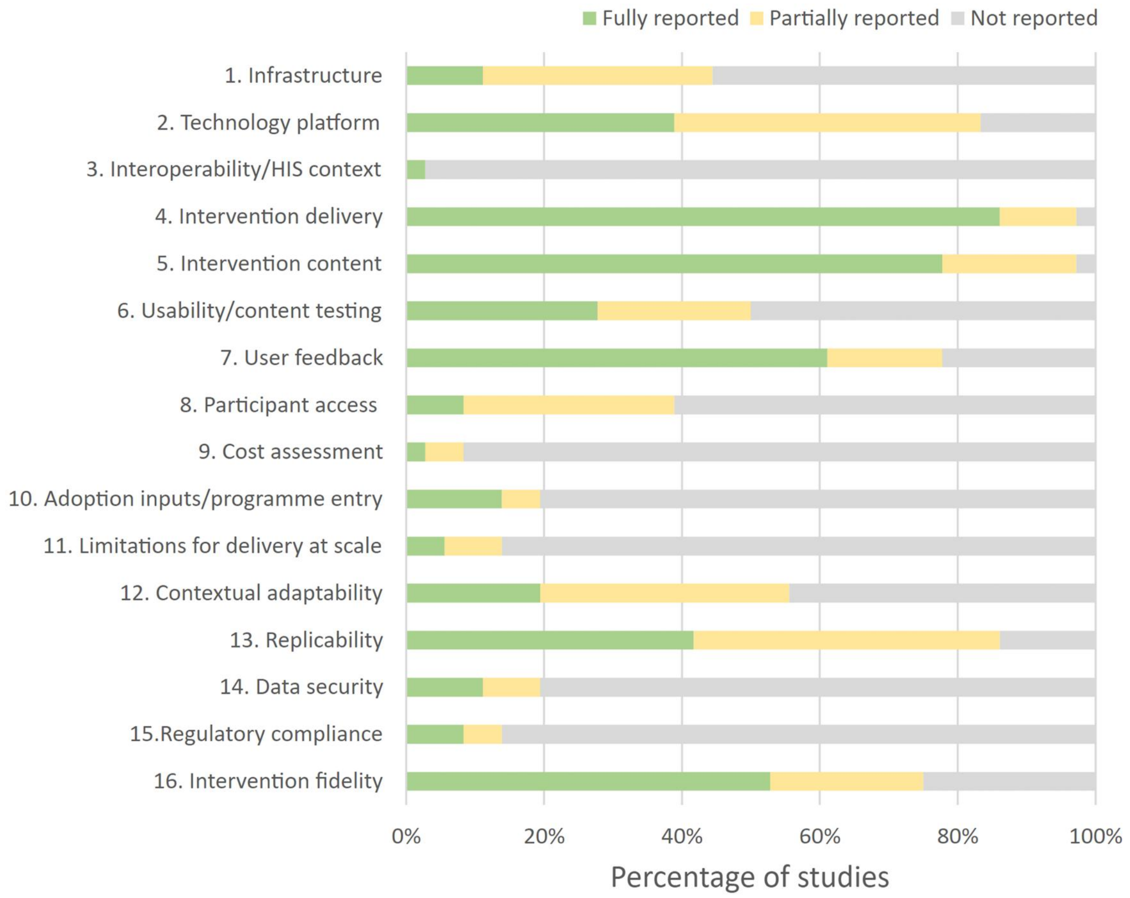

Figure 4. Percentage of articles met the criteria on the mERA checklist. Each mERA criterion is categorized as "Reported," "Partially reported," and "Not reported." HIS, health information systems.

The 4evaluation stages in the CA framework are based on WHO's recommendations. They are fundamentally consistent with the traditional clinical and medical evaluation phases.[47–51] These stages are built upon each other to ensure safety, health benefits, and socioeconomic impacts in the research translation of new interventions into the real world.[47–51]

The framework will, accordingly, help researchers review existing studies, identify knowledge gaps, and develop new methods or recommendations for improving CA evaluations. It practically supports the stepwise evaluation strategy in future CA evaluation studies and encourages new systematic reviews with extensive scopes at different evaluation stages to improve the research translation of CA interventions in the real world. The framework also helps engineers and policymakers understand the complex journey of the research translation. Such understanding is essential for achieving effective and sustainable CA research in healthcare.

Relevant findings and recommendations

We found extensive evaluations of "Functionality", such as classifying sentences, providing responses, engaging with users, achieving various simple tasks, and assessing patients as an independent clinical assessment tool. The functionality evaluations are essential because those functions collectively determine the intervention delivery. In addition, such evaluations help understand the technical potentials and limitations of CA applications for designing effective interventions in the research. We therefore propose functionality evaluations as a main category at Stage I in the CA evaluation framework.

We found 8 studies (10%) evaluated topics with risks, such as medication, suicide attempt, alcohol/drug overdose, and private information, especially for complex CAs with contents from the Internet, such as Siri and Google Assistant. Accordingly, we recommended safety and information evaluations at Stage I in the framework. The findings and recommendation support recent concerns of potential safety and privacy risks when using AI in healthcare,[3,36,137]. They underscore the need for understanding the risks in CA evaluation studies,[138] especially for advanced CAs which enable broad conversations with large knowledge networks or language models[139,140] such as recent ChatGPT[6,141].

By analyzing study designs, we found that the safety and information evaluations were limited to studies using predefined questions in a laboratory setting or a session-based single-arm trial. How to evaluate safety and information quality effectively and reliably in other study designs remains unclear. This finding encourages the development of new strategies, such as large question-answer (QA) databases and experts' reviews of large conversation records, to improve safety and information quality evaluations in CA studies.

We found that the outcome measures for evaluating user experience were diverse ($n = 285$). In addition, none of the measures had been specifically validated for CA interventions. Many evaluation components such as usability, feasibility, satisfaction, and acceptance were inconsistently defined and selected. Because of difficulties in qualitative evaluations, many studies used qualitative evaluation methods such as interviews or open-ended questions to capture in-depth evidence. These findings indicate overwhelming difficulties in evaluating user experience and imply a strong need for developing and validating suitable questionnaires to improve user experience evaluations in future CA studies.

We found that 4 studies explored whether a CA could provide a clinical assessment (not an intervention) equivalent to or better than in-person clinicians[22,40,90,100] ("Clinical assessment performance" in "Functionality"). The evaluations fundamentally differed from the traditional evaluation focusing on comparing an intervention program against usual care in terms of improvements in clinical/health outcomes. The finding indicates a need for extending conventional evaluation frameworks and scopes to accommodate and encourage new evaluation perspectives such as conversation comparisons between CAs and humans in future CA studies.

Only 2 studies evaluated "Costs and health economic analyses"[14,73] at Stages 1 and 2. Cost-related evaluations often determine healthcare policies for improving health interventions. Reporting on cost is also recommended by the mERA checklist. We recommend cost-related evaluations at Stages 3 and 4 in the framework because CAs at these stages are normally implemented in the real world, essential to obtain generalizable results. In addition, clinical trials at these stages are normally larger than those at Stages 1 and 2, essential for achieving reliable evaluation outcomes.

There were only 8 studies found at Stage 4 of the proposed framework. The evidence was insufficient for us to synthesize essential methods for WHO's recommendations, such as "health impact," "Error rates," and "Changes in policy." Obtaining users' data such as age, gender, and conversation records could also be difficult because of privacy and security-related policies.[142] More studies with integrated data retrieval approaches, such as electronic health records and national healthcare systems (eg, Medicare Benefits Schedule in Australia), are needed to effectively address those recommended tasks at Stage 4.

Many CA evaluation studies ($n = 22$, 27%) were limited to predefined tasks, single conversation sessions, or participants' review of a conversation record. We also found that several studies used a crowdsourcing method (the practice of obtaining information or input from paid services of a large number of people via the Internet),[102,132] rather than studies using more traditional methods of recruitment. Crowdsourcing methods help recruit participants quickly,[143] but the results may be inaccurate because of incentive mechanisms and risks of spammers.[144] Therefore, understanding the design details and limitations is essential for the accurate interpretation of evaluation outcomes.

We found that the study sample sizes in this review were generally smaller than the illustrative numbers of users outlined by WHO's framework, especially at Stages 3 and 4. In the research, sample sizes are often estimated carefully[135] according to the intervention, trial design, evaluation stage, and primary outcomes. Therefore, more studies and further investigations are needed for understanding and proposing illustrative sample size ranges for CA interventions in the research.

For individual studies, identification of essential requirements in detail for evaluating safety, information quality, and functionality is often complex because there are various influential factors, such as CA designs (Rule-based dialogues, state-based systems, generative language models, etc.), intervention areas (Mental health, chronic disease management, substance use disorders, etc.), intervention components (Health information, education, clinical assessments, medication, care decision support, etc.), and users' characteristics (Normal healthy adults, women with pregnancy, people with severe mental conditions, seniors, etc.). For example, a complex CA application for mediation intervention to vulnerable people would present a higher level of safety concern than a simple CA for general health promotion in normal adults. However, how to evaluate these 2 CAs differently and effectively to ensure their safety remains unclear. Therefore, expert reviews, from multidisciplinary research fields, would be needed to address the knowledge gap in future studies.

Regarding the reporting quality of CA studies in this review, the mERA results demonstrated low reporting rates across many mERA criteria including "Cost assessment," consistent with recent digital health intervention reviews.[71,145] The results imply that some essential evaluation details might not be reported fully by the authors and, hence, captured in our review. They again indicate a strong need for improving the adherence and update of digital health evaluation guidelines and frameworks in future CA studies.

Strengths

We categorized diverse CA study designs and outcome measures and employed a globally recognized framework to synthesize the evidence. We finally provided a comprehensive evaluation framework for CA interventions and discussed issues and gaps for future studies.

Limitations

The data synthesis was limited to a single overarching digital health evaluation framework. The digital health framework focuses on general web applications or smartphone apps. Its recommendations on AI technologies are limited. Integrating multiple evaluation guidelines or frameworks, such as recommendations for complex interventions[146,147] or AI-related applications,[35,36] would be useful to improve and enrich the CA evaluation framework in future studies.

Conclusion

Evaluation frameworks are essential to achieving safe and effective health and clinical outcomes in CA-based intervention studies, but none yet exists. We synthesized the evidence from 81 CA evaluation studies and outlined an evaluation framework for CA interventions. Our findings provide several important implications for evaluating CA interventions and encourage further investigations to continue to improve CA evaluation frameworks in future research.

Acknowledgments

We thank Ms Jackie Devenish from the library at the University of Queensland for her guidance and support in developing the literature search strategy.

Author Contributions

H.D. and T.R. designed the systematic review. H.D. and J.S. screened titles and abstracts of articles obtained in the databases. H.D. and J.S. reviewed the full-text articles and extracted the data. A.V. resolved data discrepancies. All authors contributed to the interpretation of extracted data and discussion. H.D. contributed to the drafting of the manuscript. All authors contributed to critical revisions and approved the final version of the review.

Supplementary material

Supplementary material is available at *Journal of the American Medical Informatics Association* online.

Funding

Research in this publication was supported by the University of Queensland. The content is solely the responsibility of the authors and does not necessarily represent the official views of the University of Queensland.

Conflict of interest

None declared.

Data availability

The data underlying this article are available in the article and in its online supplementary material.

References

1. Turing AM. I.—Computing machinery and intelligence. *Mind*. 1950;LIX(236):433-460. https://doi.org/10.1093/mind/LIX.236.433
2. Adamopoulou E, Moussiades L. *An Overview of Chatbot Technology*. Springer International Publishing; 2020:373-383.
3. Matheny MT, Ahmed M, Whicher D. *AI in Health Care: The Hope, the Hype, the Promise, the Peril*. National Academy of Medicine; 2019.
4. Zhou M, Duan N, Liu S, Shum HY. Progress in neural NLP: modeling, learning, and reasoning. *Engineering*. 2020;6(3):275-290. https://doi.org/10.1016/j.eng.2019.12.014
5. Waring J, Lindvall C, Umeton R. Automated machine learning: review of the state-of-the-art and opportunities for healthcare. *Artif Intell Med*. 2020;104:101822. https://doi.org/10.1016/j.artmed.2020.101822
6. Editorials. Will ChatGPT transform healthcare? *Nat Med*. 2023;29(3):505-506. doi:10.1038/s41591-023-02289-5.
7. López Espejel J, Yahaya Alassan MS, Chouham EM, Dahhane W, Ettifouri EH. A comprehensive review of state-of-the-art methods for Java code generation from natural language text. *Nat Lang Process J*. 2023;3:100013. https://doi.org/10.1016/j.nlp.2023.100013
8. Bickmore TW, Trinh H, Olafsson S, 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;20(9):e11510. https://doi.org/10.2196/11510
9. Schachner T, Keller R, V Wangenheim F. Artificial intelligence-based conversational agents for chronic conditions: systematic literature review. *J Med Internet Res*. 2020;22(9):e20701. https://doi.org/10.2196/20701
10. Hauser-Ulrich S, Künzli H, Meier-Peterhans D, Kowatsch T. A smartphone-based health care chatbot to promote self-management of chronic pain (SELMA): pilot randomized controlled trial. *JMIR Mhealth Uhealth*. 2020;8(4):e15806-e15806. https://doi.org/10.2196/15806
11. Boczar D, Sisti A, Oliver JD, et al. Artificial intelligent virtual assistant for plastic surgery patient's frequently asked questions: a pilot study. *Ann Plast Surg*. 2020;84(4):e16-e21. https://doi.org/10.1097/sap.0000000000002252
12. The BMJ New. Babylon plans US and Asia expansion as new funders boost company value to $2bn. Accessed June 19, 2023. https://www.bmj.com/content/366/bmj.l5009
13. Bott N, Wexler S, Drury L, et al. A protocol-driven, bedside digital conversational agent to support nurse teams and mitigate risks of hospitalization in older adults: case control pre-post study. *J Med Internet Res*. 2019;21(10):e13440. https://doi.org/10.2196/13440
14. Bian Y, Xiang Y, Tong B, Feng B, Weng X. Artificial intelligence-assisted system in postoperative follow-up of orthopedic patients: exploratory quantitative and qualitative study. *J Med Internet Res*. 2020;22(5):e16896. https://doi.org/10.2196/16896
15. 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;23(1):e19928. https://doi.org/10.2196/19928
16. Maeda E, Miyata A, Boivin J, et al. Promoting fertility awareness and preconception health using a chatbot: a randomized controlled trial. *Reprod Biomed Online*. 2020;41(6):1133-1143. https://doi.org/10.1016/j.rbmo.2020.09.006
17. Ferrand J, Hockensmith R, Houghton RF, Walsh-Buhi ER. Evaluating smart assistant responses for accuracy and misinformation regarding human papillomavirus vaccination: content analysis study. *J Med Internet Res*. 2020;22(8):e19018. https://doi.org/10.2196/19018
18. Harless WG, Zier MA, Harless MG, et al. Evaluation of a virtual dialogue method for breast cancer patient education. *Patient Educ Couns*. 2009;76(2):189-195. https://doi.org/10.1016/j.pec.2009.02.006
19. 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;9(1):e24045. https://doi.org/10.2196/24045
20. Gabrielli S, Rizzi S, Carbone S, Donisi V. A chatbot-based coaching intervention for adolescents to promote life skills: pilot study. *JMIR Hum Factors*. 2020;7(1):e16762. https://doi.org/10.2196/16762
21. Gaffney H, Mansell W, Tai S. Agents of change: understanding the therapeutic processes associated with the helpfulness of therapy for mental health problems with relational agent MYLO. *Digit Health*. 2020;6:2055207620911580. https://doi.org/10.1177/2055207620911580
22. Caballer A, Belmonte O, Castillo A, Gasco A, Sansano E, Montoliu R. Equivalence of chatbot and paper-and-pencil versions of the De Jong Gierveld loneliness scale. *Curr Psychol*. 2020;41(9):6225-6232. 2020/10/13 https://doi.org/10.1007/s12144-020-01117-0
23. Denecke K, Vaaheesan S, Arulnathan A. A mental health chatbot for regulating emotions (SERMO) - Concept and usability test. *IEEE Trans Emerg Topics Comput*. 2021;9(3):1170-1182. https://doi.org/10.1109/TETC.2020.2974478
24. Park S, Choi J, Lee S, et al. Designing a chatbot for a brief motivational interview on stress management: qualitative case study. *J Med Internet Res*. 2019;21(4):e12231. https://doi.org/10.2196/12231
25. Rehman UU, Chang DJ, Jung Y, Akhtar U, Razzaq MA, Lee S. Medical instructed real-time assistant for patient with glaucoma and diabetic conditions. *Appl Scie*. 2020;10(7):2216.
26. Stephens TN, Joerin A, Rauws M, Werk LN. Feasibility of pediatric obesity and prediabetes treatment support through Tess, the AI behavioral coaching chatbot. *Transl Behav Med*. 2019;9(3):440-447. https://doi.org/10.1093/tbm/ibz043
27. Rhee H, Allen J, Mammen J, Swift M. Mobile phone-based asthma self-management aid for adolescents (mASMAA): a

feasibility study. *Patient Prefer Adherence*. 2014;8:63-72. https://doi.org/10.2147/ppa.S53504

28. World Health Organization. WHO launches a chatbot on Facebook Messenger to combat COVID-19 misinformation. Accessed January 19, 2023. https://www.who.int/news-room/feature-stories/detail/who-launches-a-chatbot-powered-facebook-messenger-to-combat-covid-19-misinformation
29. McKillop M, South BR, Preininger A, Mason M, Jackson GP. Leveraging conversational technology to answer common COVID-19 questions. *J Am Med Inform Assoc*. 2021;28(4):850-855. https://doi.org/10.1093/jamia/ocaa316
30. The Centers for Disease Control and Prevention. COVID-19 testing overview. Accessed January 19, 2023. https://www.cdc.gov/coronavirus/2019-ncov/symptoms-testing/testing.html
31. World Health Organization. WHO health alert brings COVID-19 facts to billions via WhatsApp. Accessed January 19, 2023. https://www.who.int/news-room/feature-stories/detail/who-health-alert-brings-covid-19-facts-to-billions-via-whatsapp
32. Yin J, Ngiam KY, Teo HH. Role of artificial intelligence applications in real-life clinical practice: systematic review. *J Med Internet Res*. 2021;23(4):e25759. https://doi.org/10.2196/25759
33. Kelly CJ, Karthikesalingam A, Suleyman M, Corrado G, King D. Key challenges for delivering clinical impact with artificial intelligence. *BMC Med*. 2019;17(1):195. https://doi.org/10.1186/s12916-019-1426-2
34. Cadario R, Longoni C, Morewedge CK. Understanding, explaining, and utilizing medical artificial intelligence. *Nat Hum Behav*. 2021;5(12):1636-1642. https://doi.org/10.1038/s41562-021-01146-0
35. World Health Organization. *Ethics and Governance of Artificial Intelligence for Health. WHO Guidance*. World Health Organization; 2021.
36. Wiens J, Saria S, Sendak M, et al. Do no harm: a roadmap for responsible machine learning for health care. *Nat Med*. 2019;25(9):1337-1340. https://doi.org/10.1038/s41591-019-0548-6
37. Liu Y, Chen P-HC, Krause J, Peng L. How to read articles that use machine learning: users' guides to the medical literature. *JAMA*. 2019;322(18):1806-1816. https://doi.org/10.1001/jama.2019.16489
38. 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;21(4):e12887. https://doi.org/10.2196/12887
39. Bibault JE, Chaix B, Guillemassé A, et al. A chatbot versus physicians to provide information for patients with breast cancer: blind, randomized controlled noninferiority trial. *J Med Internet Res*. 2019;21(11):e15787. https://doi.org/10.2196/15787
40. Philip P, Micoulaud-Franchi JA, Sagaspe P, et al. Virtual human as a new diagnostic tool, a proof of concept study in the field of major depressive disorders. *Sci Rep*. 2017;7:42656. https://doi.org/10.1038/srep42656
41. Tudor Car L, Dhinagaran DA, Kyaw BM, et al. Conversational agents in health care: scoping review and conceptual analysis. *J Med Internet Res*. 2020;22(8):e17158. https://doi.org/10.2196/17158
42. Venkatesh A, Khatri C, Ram A, et al. 2018. On evaluating and comparing open domain dialog systems, arXiv, arXiv:1801.03625, preprint: not peer reviewed. Last Revised December 26, 2018. https://ui.adsabs.harvard.edu/abs/2018arXiv180103625V
43. Radziwill NM, Benton MC. 2017. Evaluating quality of chatbots and intelligent conversational agents, arXiv, arXiv:1704.04579, preprint: not peer reviewed. Submitted April 15, 2017. https://ui.adsabs.harvard.edu/abs/2017arXiv170404579R
44. Abd-Alrazaq A, Safi Z, Alajlani M, Warren J, Househ M, Denecke K. Technical metrics used to evaluate health care chatbots: scoping review. *J Med Internet Res*. 2020;22(6):e18301. https://doi.org/10.2196/18301
45. 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;22(7):e16021. https://doi.org/10.2196/16021
46. Laranjo L, Dunn AG, Tong HL, et al. Conversational agents in healthcare: a systematic review. *J Am Med Inform Assoc*. 2018;25(9):1248-1258. https://doi.org/10.1093/jamia/ocy072
47. Umscheid CA, Margolis DJ, Grossman CE. Key concepts of clinical trials: a narrative review. *Postgrad Med*. 2011;123(5):194-204. https://doi.org/10.3810/pgm.2011.09.2475
48. National Health and Medical Research Council. Phases of clinical trials. Accessed 18 September, 2021. https://www.australianclinicaltrials.gov.au/what-clinical-trial/phases-clinical-trials
49. Evans SR. Fundamentals of clinical trial design. *J Exp Stroke Transl Med*. 2010;3(1):19-27. https://doi.org/10.6030/1939-067x-3.1.19
50. Wright B. Chapter 2 - clinical trial phases. In: Shamley D, Wright B, eds. *A Comprehensive and Practical Guide to Clinical Trials*. Academic Press; 2017:11-15.
51. World Health Organization. *Monitoring and Evaluating Digital Health Interventions: A Practical Guide to Conducting Research and Assessment*. World Health Organization; 2016.
52. Tricco AC, Lillie E, Zarin W, et al. PRISMA extension for scoping reviews (PRISMA-ScR): checklist and explanation. *Ann Intern Med*. 2018;169(7):467-473. https://doi.org/10.7326/m18-0850
53. Ding H, Simmich J, Vaezipour A, Andrews N, Russell T. Evaluation framework for using unconstrained conversational agents in health interventions: a systematic review protocol. The Open Science Framework. Accessed 18 September, 2021. https://osf.io/9hq2v
54. Pham MT, Rajić A, Greig JD, Sargeant JM, Papadopoulos A, McEwen SA. A scoping review of scoping reviews: advancing the approach and enhancing the consistency. *Res Synth Methods*. 2014;5(4):371-385. https://doi.org/10.1002/jrsm.1123
55. Mauriello ML, Tantivasadakarn N, Mora-Mendoza MA, et al. A suite of mobile conversational agents for daily stress management (Popbots): mixed methods exploratory study. *JMIR Form Res*. 2021;5(9):e25294. https://doi.org/10.2196/25294
56. Chattopadhyay D, Ma T, Sharifi H, Martyn-Nemeth P. Computer-controlled virtual humans in patient-facing systems: systematic review and meta-analysis. *J Med Internet Res*. 2020;22(7):e18839. https://doi.org/10.2196/18839
57. Veritas Health Innovation. Covidence systematic review software. Veritas Health Innovation. Accessed September 18, 2021. www.covidence.org
58. United Nations. Member states in the United Nations. Accessed December 11, 2022. https://www.un.org/en/about-us/member-states
59. Grimes DA, Schulz KF. An overview of clinical research: the lay of the land. *Lancet*. 2002;359(9300):57-61. https://doi.org/10.1016/s0140-6736(02)07283-5
60. Mills EJ, Chan A-W, Wu P, Vail A, Guyatt GH, Altman DG. Design, analysis, and presentation of crossover trials. *Trials*. 2009;10(1):27. https://doi.org/10.1186/1745-6215-10-27
61. Sibbald B, Roberts C. Understanding controlled trials. Crossover trials. *BMJ*. 1998;316(7146):1719. https://doi.org/10.1136/bmj.316.7146.1719
62. Evans SR. Clinical trial structures. *J Exp Stroke Transl Med*. 2010;3(1):8-18. https://doi.org/10.6030/1939-067x-3.1.8
63. Dwan K, Li T, Altman DG, Elbourne D. CONSORT 2010 statement: extension to randomised crossover trials. *BMJ*. 2019;366:l4378. https://doi.org/10.1136/bmj.l4378
64. Glasziou P, Heneghan C. A spotter's guide to study designs. *Evid Based Med*. 2009;14(2):37-38. https://doi.org/10.1136/ebm.14.2.37-a
65. Liu X, Cruz Rivera S, Moher D, et al.; SPIRIT-AI and CONSORT-AI Working Group. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the

CONSORT-AI extension. *Nat Med*. 2020;26(9):1364-1374. https://doi.org/10.1038/s41591-020-1034-x

66. Jandoo T. WHO guidance for digital health: what it means for researchers. *Digit Health*. 2020;6:2055207619898984. https://doi.org/10.1177/2055207619898984
67. Agboola SO, Bates DW, Kvedar JC. Digital health and patient safety. *JAMA*. 2016;315(16):1697-1698. https://doi.org/10.1001/jama.2016.2402
68. Higgins AM, Harris AH. Health economic methods: cost-minimization, cost-effectiveness, cost-utility, and cost-benefit evaluations. *Crit Care Clin*. 2012;28(1):11-24, v. https://doi.org/10.1016/j.ccc.2011.10.002
69. Agarwal S, LeFevre AE, Lee J, et al.; WHO mHealth Technical Evidence Review Group. Guidelines for reporting of health interventions using mobile phones: mobile health (mHealth) evidence reporting and assessment (mERA) checklist. *BMJ*. 2016;352:i1174. https://doi.org/10.1136/bmj.i1174
70. Karim H, Choobineh H, Kheradbin N, Ravandi MH, Naserpor A, Safdari R. Mobile health applications for improving the sexual health outcomes among adults with chronic diseases: a systematic review. *Digit Health*. 2020;6:2055207620906956. https://doi.org/10.1177/2055207620906956
71. L'Engle KL, Mangone ER, Parcesepe AM, Agarwal S, Ippoliti NB. Mobile phone interventions for adolescent sexual and reproductive health: a systematic review. *Pediatrics*. 2016;138(3):e20160884. https://doi.org/10.1542/peds.2016-0884
72. Nazareth S, Hayward L, Simmons E, et al. Hereditary cancer risk using a genetic chatbot before routine care visits. *Obstet Gynecol*. 2021;138(6):860-870. https://doi.org/10.1097/AOG.0000000000004596
73. Bonnevie E, Lloyd TD, Rosenberg SD, Williams K, Goldbarg J, Smyser J. Layla's got you: developing a tailored contraception chatbot for Black and Hispanic young women. *Health Educ J*. 2021;80(4):413-424. https://doi.org/10.1177/0017896920981122
74. Rathnayaka P, Mills N, Burnett D, De Silva D, Alahakoon D, Gray R. A mental health chatbot with cognitive skills for personalised behavioural activation and remote health monitoring. *Sensors*. 2022;22(10):3653. https://doi.org/10.3390/s22103653
75. Bennion MR, Hardy GE, Moore RK, Kellett S, Millings A. Usability, acceptability, and effectiveness of web-based conversational agents to facilitate problem solving in older adults: controlled study. *J Med Internet Res*. 2020;22(5):e16794. https://doi.org/10.2196/16794
76. Leo AJ, Schuelke MJ, Hunt DM, Miller JP, Areán PA, Cheng AL. Digital mental health intervention plus usual care compared with usual care only and usual care plus in-person psychological counseling for orthopedic patients with symptoms of depression or anxiety: cohort study. *JMIR Form Res*. 2022;6(5):e36203. https://doi.org/10.2196/36203
77. Leo AJ, Schuelke MJ, Hunt DM, et al. A digital mental health intervention in an orthopedic setting for patients with symptoms of depression and/or anxiety: feasibility prospective cohort study. *JMIR Form Res*. 2022;6(2):e34889. https://doi.org/10.2196/34889
78. Beatty C, Malik T, Meheli S, Sinha C. Evaluating the therapeutic alliance with a free-text CBT conversational agent (Wysa): a mixed-methods study. *Front Digit Health*. 2022;4:847991. https://doi.org/10.3389/fdgth.2022.847991
79. Darcy A, Daniels J, Salinger D, Wicks P, Robinson A. Evidence of human-level bonds established with a digital conversational agent: cross-sectional, retrospective observational study. *JMIR Form Res*. 2021;5(5):e27868. https://doi.org/10.2196/27868
80. Nadarzynski T, Puentes V, Pawlak I, et al. Barriers and facilitators to engagement with artificial intelligence (AI)-based chatbots for sexual and reproductive health advice: a qualitative analysis. *Sex Health*. 2021;18(5):385-393. https://doi.org/10.1071/SH21123
81. Prochaska JJ, Vogel EA, Chieng A, et al. A therapeutic relational agent for reducing problematic substance use (Woebot): development and usability study. *J Med Internet Res*. 2021;23(3):e24850. https://doi.org/10.2196/24850
82. Prochaska JJ, Vogel EA, Chieng A, et al. A randomized controlled trial of a therapeutic relational agent for reducing substance misuse during the COVID-19 pandemic. *Drug & Alcohol Dependence*. 2021;227:108986. https://doi.org/10.1016/j.drugalcdep.2021.108986
83. Nobles AL, Leas EC, Caputi TL, Zhu S-H, 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(1):11. https://doi.org/10.1038/s41746-019-0215-9
84. Boyd M, Wilson N. Just ask Siri? A pilot study comparing smartphone digital assistants and laptop Google searches for smoking cessation advice. *PLoS One*. 2018;13(3):e0194811. https://doi.org/10.1371/journal.pone.0194811
85. 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;176(5):619-625. https://doi.org/10.1001/jamainternmed.2016.0400
86. Schindler-Ruwisch J, Palancia Esposito C. "Alexa, am I pregnant?": A content analysis of a virtual assistant's responses to prenatal health questions during the COVID-19 pandemic. *Patient Educ Couns*. 2021;104(3):460-463. https://doi.org/10.1016/j.pec.2020.12.026
87. Kocaballi AB, Quiroz JC, Rezazadegan D, et al. Responses of conversational agents to health and lifestyle prompts: investigation of appropriateness and presentation structures. *J Med Internet Res*. 2020;22(2):e15823. https://doi.org/10.2196/15823
88. Crutzen R, Peters GJ, Portugal SD, Fisser EM, Grolleman JJ. An artificially intelligent chat agent that answers adolescents' questions related to sex, drugs, and alcohol: an exploratory study. *J Adolesc Health*. 2011;48(5):514-519. doi:10.1016/j.jadohealth.2010.09.002
89. Lee Y-C, 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;4(CSCW1):1-27. https://doi.org/10.1145/3392836
90. Philip P, Bioulac S, Sauteraud A, Chaufton C, Olive J. Could a virtual human be used to explore excessive daytime sleepiness in patients? *Presence*. 2014;23(4):369-376. https://doi.org/10.1162/PRES_a_00197
91. Nam KH, Kim DY, Kim DH, et al. Conversational artificial intelligence for spinal pain questionnaire: validation and user satisfaction. *Neurospine*. 2022;19(2):348-356. https://doi.org/10.14245/ns.2143080.540
92. Maenhout L, Peuters C, Cardon G, Compernolle S, Crombez G, DeSmet A. Participatory development and pilot testing of an adolescent health promotion chatbot. *Front Public Health*. 2021;9:724779. https://doi.org/10.3389/fpubh.2021.724779
93. Almusharraf F, Rose J, Selby P. Engaging unmotivated smokers to move toward quitting: design of motivational interviewing-based chatbot through iterative interactions. *J Med Internet Res*. 2020;22(11):e20251. https://doi.org/10.2196/20251
94. Yasavur U, Lisetti C, Rishe N. Let's talk! Speaking virtual counselor offers you a brief intervention. *J Multimodal User Interfaces*. 2014;8(4):381-398. https://doi.org/10.1007/s12193-014-0169-9
95. Bassi G, Giuliano C, Perinelli A, Forti S, Gabrielli S, Salcuni S. A virtual coach (Motibot) for supporting healthy coping strategies among adults with diabetes: proof-of-concept study. *JMIR Hum Factors*. 2022;9(1):e32211. https://doi.org/10.2196/32211
96. Shah J, DePietro B, D'Adamo L, et al. Development and usability testing of a chatbot to promote mental health services use among individuals with eating disorders following screening. *Int J Eat Disord*. 2022;55(9):1229-1244. https://doi.org/10.1002/eat.23798
97. Figueroa CA, Luo TC, Jacobo A, et al. Conversational physical activity coaches for Spanish and English speaking women: a user

design study. *Front Digit Health*. 2021;3(2010):747153. https://doi.org/10.3389/fdgth.2021.747153

98. Mokmin NAM, Ibrahim NA. The evaluation of chatbot as a tool for health literacy education among undergraduate students. *Educ Inf Technol (Dordr)*. 2021;26(5):6033-6049. https://doi.org/10.1007/s10639-021-10542-y

99. Polignano M, Narducci F, Iovine A, Musto C, Gemmis MD, Semeraro G. HealthAssistantBot: a personal health assistant for the Italian language. *IEEE Access*. 2020;8:107479-107497. https://doi.org/10.1109/ACCESS.2020.3000815

100. Auriacombe M, Moriceau S, Serre F, et al. Development and validation of a virtual agent to screen tobacco and alcohol use disorders. *Drug Alcohol Depend*. 2018;193:1-6. https://doi.org/10.1016/j.drugalcdep.2018.08.025

101. Linwei H, Basar E, Wiers RW, Antheunis ML, Krahmer E, He L. Can chatbots help to motivate smoking cessation? A study on the effectiveness of motivational interviewing on engagement and therapeutic alliance. *BMC Public Health*. 2022;22(1):726-714. https://doi.org/10.1186/s12889-022-13115-x

102. Medeiros L, Bosse T, Gerritsen C. Can a chatbot comfort humans? Studying the impact of a supportive chatbot on users' self-perceived stress. *IEEE Trans Human-Mach Syst*. 2022;52(3):343-353. https://doi.org/10.1109/THMS.2021.3113643

103. Siglen E, Vetti HH, Lunde ABF, et al. Ask Rosa—the making of a digital genetic conversation tool, a chatbot, about hereditary breast and ovarian cancer. *Patient Educ Couns*. 2022;105(6):1488-1494. https://doi.org/10.1016/j.pec.2021.09.027

104. Dosovitsky G, Bunge EL. Bonding with bot: user feedback on a chatbot for social isolation. *Front Digit Health*. 2021;3:735053. https://doi.org/10.3389/fdgth.2021.735053

105. Jang S, Kim J-J, Kim S-J, Hong J, Kim S, Kim E. Mobile app-based chatbot to deliver cognitive behavioral therapy and psychoeducation for adults with attention deficit: a development and feasibility/usability study. *Int J Med Inform*. 2021;150:104440. https://doi.org/10.1016/j.ijmedinf.2021.104440

106. 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;4(2):e19. https://doi.org/10.2196/mental.7785

107. Rahmanti AR, Yang H-C, Bintoro BS, et al. SlimMe, a chatbot with artificial empathy for personal weight management: system design and finding. *Front Nutr*. 2022;9:870775. https://doi.org/10.3389/fnut.2022.870775

108. Davis CR, Murphy KJ, Curtis RG, Maher CA. A process evaluation examining the performance, adherence, and acceptability of a physical activity and diet artificial intelligence virtual health assistant. *Int J Environ Res Public Health*. 2020;17(23):9137. https://doi.org/10.3390/ijerph17239137

109. Fulmer R, Joerin A, Gentile B, Lakerink L, Rauws M. Using psychological artificial intelligence (Tess) to relieve symptoms of depression and anxiety: randomized controlled trial. *JMIR Ment Health*. 2018;5(4):e64. https://doi.org/10.2196/mental.9782

110. Chaix B, Bibault JE, Pienkowski A, et al. When chatbots meet patients: one-year prospective study of conversations between patients with breast cancer and a chatbot. *JMIR Cancer*. 2019;5(1):e12856. https://doi.org/10.2196/12856

111. Kataoka Y, Takemura T, Sasajima M, Katoh N. Development and early feasibility of chatbots for educating patients with lung cancer and their caregivers in Japan: mixed methods study. *JMIR Cancer*. 2021;7(1):e26911. https://doi.org/10.2196/26911

112. Liu H, Peng H, Song X, Xu C, Zhang M. Using AI chatbots to provide self-help depression interventions for university students: a randomized trial of effectiveness. *Internet Interv*. 2022;27:100495. https://doi.org/10.1016/j.invent.2022.100495

113. To QG, Green C, Vandelanotte C. Feasibility, usability, and effectiveness of a machine learning-based physical activity chatbot: quasi-experimental study. *JMIR Mhealth Uhealth*. 2021;9(11):e28577. https://doi.org/10.2196/28577

114. 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;37(3):144-151. https://doi.org/10.1097/HTR.0000000000000770

115. So R, Furukawa TA, Matsushita S, et al. Unguided chatbot-delivered cognitive behavioural intervention for problem gamblers through messaging App: a randomised controlled trial. *J Gambl Stud*. 2020;36(4):1391-1407. https://doi.org/10.1007/s10899-020-09935-4

116. Klos MC, Escoredo M, Joerin A, Lemos VN, Rauws M, Bunge EL. Artificial intelligence-based chatbot for anxiety and depression in university students: pilot randomized controlled trial. *JMIR Form Res*. 2021;5(8):e20678. https://doi.org/10.2196/20678

117. Daley K, Hungerbuehler I, Cavanagh K, Claro HG, Swinton PA, Kapps M. Preliminary evaluation of the engagement and effectiveness of a mental health chatbot. *Front Digit Health*. 2020;2:576361. https://doi.org/10.3389/fdgth.2020.576361

118. Fitzsimmons-Craft EE, Chan WW, Smith AC, et al. Effectiveness of a chatbot for eating disorders prevention: a randomized clinical trial. *Int J Eat Disord*. 2022;55(3):343-353. https://doi.org/10.1002/eat.23662

119. Lavelle J, Dunne N, Mulcahy HE, McHugh L. Chatbot-delivered cognitive defusion versus cognitive restructuring for negative self-referential thoughts: a pilot study. *Psychol Rec*. 2022;72(2):247-261. https://doi.org/10.1007/s40732-021-00478-7

120. Ly KH, Ly A-M, Andersson G. A fully automated conversational agent for promoting mental well-being: a pilot RCT using mixed methods. *Internet Interv*. 2017;10:39-46. https://doi.org/10.1016/j.invent.2017.10.002

121. Ogawa M, Oyama G, Morito K, et al. Can AI make people happy? The effect of AI-based chatbot on smile and speech in Parkinson's disease. *Parkinsonism Relat Disord*. 2022;99:43-46. https://doi.org/10.1016/j.parkreldis.2022.04.018

122. Friederichs S, Bolman C, Oenema A, Guyaux J, Lechner L. Motivational interviewing in a web-based physical activity intervention with an avatar: randomized controlled trial. *J Med Internet Res*. 2014;16(2):e48. https://doi.org/10.2196/jmir.2974

123. Maher CA, Davis CR, Curtis RG, Short CE, Murphy KJ. A physical activity and diet program delivered by artificially intelligent virtual health coach: proof-of-concept study. *JMIR Mhealth Uhealth*. 2020;8(7):e17558. https://doi.org/10.2196/17558

124. Perski O, Crane D, Beard E, Brown J. Does the addition of a supportive chatbot promote user engagement with a smoking cessation app? An experimental study. *Digit Health*. 2019;5:2055207619880676. https://doi.org/10.1177/2055207619880676

125. Suganuma S, Sakamoto D, Shimoyama H. An embodied conversational agent for unguided internet-based cognitive behavior therapy in preventative mental health: feasibility and acceptability pilot trial. *JMIR Ment Health*. 2018;5(3):e10454. https://doi.org/10.2196/10454

126. Olano-Espinosa E, Avila-Tomas JF, Minue-Lorenzo C, et al.; Dejal@ Group. Effectiveness of a conversational chatbot (Dejal@bot) for the adult population to quit smoking: pragmatic, multicenter, controlled, randomized clinical trial in primary care. *JMIR Mhealth Uhealth*. 2022;10(6):e34273. https://doi.org/10.2196/34273

127. Dosovitsky G, Pineda BS, Jacobson NC, Chang C, Escoredo M, Bunge EL. Artificial intelligence chatbot for depression: descriptive study of usage. *JMIR Form Res*. 2020;4(11):e17065. https://doi.org/10.2196/17065

128. Wang H, Gupta S, Singhal A, et al. An artificial intelligence chatbot for young people's sexual and reproductive health in India (SnehAI): instrumental case study. *J Med Internet Res*. 2022;24(1):e29969. https://doi.org/10.2196/29969

129. Verduci E, Vizzuso S, Frassinetti A, et al. Nutripedia: the fight against the fake news in nutrition during pregnancy and early

life. *Nutrients*. 2021;13(9):2998-2998. https://doi.org/10.3390/nu13092998

130. Chou W-YS, Oh A, Klein WMP. Addressing health-related misinformation on social media. *JAMA*. 2018;320(23):2417-2418. https://doi.org/10.1001/jama.2018.16865

131. Miura C, Chen S, Saiki S, Nakamura M, Yasuda K. Assisting personalized healthcare of elderly people: developing a rule-based virtual caregiver system using mobile chatbot. *Sensors*. 2022;22(10):3829. https://doi.org/10.3390/s22103829

132. Pecune F, Callebert L, Marsella S. Designing persuasive food conversational recommender systems with nudging and socially-aware conversational strategies. *Front Robot AI*. 2021;8:733835. https://doi.org/10.3389/frobt.2021.733835

133. Sagstad MH, Morken N-H, Lund A, Dingsør LJ, Nilsen ABV, Sorbye LM. Quantitative user data from a chatbot developed for women with gestational diabetes mellitus: observational study. *JMIR Form Res*. 2022;6(4):e28091. https://doi.org/10.2196/28091

134. Chan A-W, Tetzlaff JM, Gøtzsche PC, et al. SPIRIT 2013 explanation and elaboration: guidance for protocols of clinical trials. *BMJ*. 2013;346:e7586. https://doi.org/10.1136/bmj.e7586

135. Wang X, Ji X. Sample size estimation in clinical research: from randomized controlled trials to observational studies. *Chest*. 2020;158(1):S12-S20. https://doi.org/10.1016/j.chest.2020.03.010

136. Kwee A, Teo ZL, Ting DSW. Digital health in medicine: important considerations in evaluating health economic analysis. *Lancet Reg Health West Pac*. 2022;23:100476. https://doi.org/10.1016/j.lanwpc.2022.100476

137. Quinn TP, Senadeera M, Jacobs S, Coghlan S, Le V. Trust and medical AI: the challenges we face and the expertise needed to overcome them. *J Am Med Inform Assoc*. 2021;28(4):890-894. https://doi.org/10.1093/jamia/ocaa268

138. Kocaballi AB, Sezgin E, Clark L, et al. Design and evaluation challenges of conversational agents in health care and well-being: selective review study. *J Med Internet Res*. 2022;24(11):e38525. https://doi.org/10.2196/38525

139. Korngiebel DM, Mooney SD. Considering the possibilities and pitfalls of generative pre-trained transformer 3 (GPT-3) in healthcare delivery. *NPJ Digit Med*. 2021;4(1):93. https://doi.org/10.1038/s41746-021-00464-x

140. Das A, Verma RM. Can machines tell stories? A comparative study of deep neural language models and metrics. *IEEE Access*. 2020;8:181258-181292. https://doi.org/10.1109/ACCESS.2020.3023421

141. World Health Organization. WHO calls for safe and ethical AI for health. Accessed July 3, 2023. https://www.who.int/news/item/16-05-2023-who-calls-for-safe-and-ethical-ai-for-health

142. Jain P, Gyanchandani M, Khare N. Big data privacy: a technological perspective and review. *J Big Data*. 2016;3(1):25. doi:10.1186/s40537-016-0059-y

143. Zhang J. Knowledge learning with crowdsourcing: a brief review and systematic perspective. *IEEE/CAA J Autom Sin*. 2022;9(5):749-762. doi:10.1109/jas.2022.105434

144. Li Y, Chang L, Li L, Bao X, Gu T. Key research issues and related technologies in crowdsourcing data collection. *Wireless Commun Mobile Comput*. 2021;2021:1. doi:10.1155/2021/8745897

145. Oikonomidi T, Vivot A, Tran VT, Riveros C, Robin E, Ravaud P. A methodologic systematic review of mobile health behavior change randomized trials. *Am J Prev Med*. 2019;57(6):836-843. doi:10.1016/j.amepre.2019.07.008

146. Skivington K, Matthews L, Simpson SA, et al. Framework for the development and evaluation of complex interventions: gap analysis, workshop and consultation-informed update. *Health Technol Assess*. 2021;25(57):1-132. https://doi.org/10.3310/hta25570

147. Skivington K, Matthews L, Simpson SA, et al. A new framework for developing and evaluating complex interventions: update of Medical Research Council guidance. *BMJ*. 2021;374:n2061. https://doi.org/10.1136/bmj.n2061