2020-10

The effectiveness of artificial intelligence conversational agents in healthcare: a systematic review

Milne-Ives, Madison

10.2196/20346
Journal of Medical Internet Research
Journal of Medical Internet Research

Systematic Review

Madison Milne-Ives, BAS, MSc,[1] Caroline de Cock, BSc, MSc,[1] Ernest Lim, BSc, MBBS, [2,3] Melissa Harper Shehadeh, MSc, PhD,[4] Nick de Pennington, MA, BM BCh FRCS[3,5] Guy Mole, MBBS, MSc,[3,5] Edward Meinert, MA, MSc, MBA, MPA, PhD CEng FBCS EUR ING[1,6,7]

[1]Digitally Enabled PrevenTative Health (DEPTH) Research Group, Department of Paediatrics, University of Oxford, Oxford, United Kingdom, OX3 9DU

[2]Imperial College Healthcare NHS Trust, Western Eye Hospital, London, United Kingdom, NW1 5QH

[3]Ufonia Limited, c/o Oxford University Innovation, Buxton Court, 3 West Way, Oxford, United Kingdom, OX2 0JB

[4]Be He@lthy Be Mobile Initiative, World Health Organization, Avenue Appia 20, 1202 Genève, Switzerland

[5]Oxford University Hospitals NHS Foundation Trust, John Radcliffe Hospital, Oxford, United Kingdom, OX3 9DU

[6]Centre for Health Technology, Faculty of Health, University of Plymouth, Plymouth, United Kingdom PL4 8AA

[7]Department of Primary Care and Public Health, School of Public Health, Imperial College London, London, United Kingdom, W6 8RP

Corresponding author:
Edward Meinert, MA, MSc, MBA, MPA, PhD
e.meinert14@imperial.ac.uk ; edward.meinert@plymouth.ac.uk

The effectiveness of artificial intelligence conversational agents in healthcare: a systematic review

Abstract

Background: High demand on healthcare services and the growing capability of artificial intelligence has led to the development of conversational agents designed to support a variety of health-related activities - including behaviour change, treatment support, health monitoring, training, triage, and screening support. Automation of these tasks could free clinicians to focus on more complex work and increase accessibility to healthcare services for the general public. An overarching assessment of the acceptability, usability, and effectiveness of these agents in healthcare is needed to collate the evidence so that future development can target areas for improvement and potential for sustainable adoption.

Objective: This systematic review aimed to assess the effectiveness and usability of conversational agents in healthcare and identify the elements that users like and dislike, to inform future research and development of these agents.

Methods: PubMed, Medline (Ovid), EMBASE, CINAHL, Web of Science, and ACM Digital Library were systematically searched for articles published since 2008 that evaluated unconstrained natural language processing conversational agents used in healthcare. Endnote (version X9; Clarivate Analytics) reference management software was used for initial screening, then full-text screening was conducted by one reviewer. Data was extracted and risk of bias was assessed by one reviewer and validated by another.

Results: A total of 31 studies were selected and included a variety of conversational agents - 14 chatbots (two of which were voice chatbots), 6 embodied conversational agents, 3 each of interactive voice response calls, virtual patients, and speech recognition screening systems, as well as one contextual question answering agent and one voice recognition triage system. Overall, the evidence reported was mostly positive or mixed. Usability and satisfaction performed well (27/30 and 26/31) and positive or mixed effectiveness was found in three quarters of the studies (23/30), but there were several limitations of the agents highlighted in specific qualitative feedback.

Conclusions: The studies generally reported positive or mixed evidence for the effectiveness, usability, and satisfactoriness of the conversational agents investigated, but qualitative user perceptions were more mixed. The quality of many of the studies was limited, and improved study design and reporting is necessary to more accurately evaluate the usefulness of the agents in healthcare and identify key areas for improvement. Further research should also analyse the cost-effectiveness, privacy, and security of the agents.

Keywords

Speech Recognition Software (MeSH); Natural Language Processing (MeSH); Artificial Intelligence (MeSH); Telemedicine (MeSH); Medical Informatics (MeSH); Health Services (MeSH); Health Communication (MeSH); Delivery of Healthcare (MeSH); Patient Acceptance of Health Care (MeSH); Mental Health (MeSH); Cell Phone (MeSH); Internet (MeSH); Conversational Agent; Chatbot

Introduction

Background

Conversational agents are among the many digital technologies being introduced into the health sector to address current healthcare challenges, such as shortages of healthcare providers reducing the availability and accessibility of healthcare services [1–3]. Conversational agents use artificial intelligence - including machine learning (a statistical means of training models with data so that they can make predictions based on a variety of features) and natural language processing (NLP; the ability to recognize and analyse verbal and written language) - to interact with humans, via speech, text, or other input and output on mobile, web-based, or audio-based platforms [1,4]. Many of these agents are designed to use NLP so that users can speak or write to the agent as they would to a human. The agent can then analyze the input and respond appropriately in a conversational manner [5].

Conversational agents first emerged as a tool in healthcare in 1966, with the development of a virtual psychotherapist (ELIZA) that could provide pre-determined answers to text-based user input [6]. In the decades since, the capabilities of natural language processing have significantly progressed and aided the development of more advanced artificial intelligence agents. Many different types of conversational agents that use NLP have been developed - including chatbots, embodied conversational agents, and virtual patients - and are accessible by telephone, mobile phone, computer, and many other digital platforms [7–10]. The types of input that conversational agents can receive and interpret has also expanded, with some conversational agents capable of analysing movement - including gestures, facial expressions, and eye movements [11,12].

Conversational agents have been developed for many different aspects of the health sector to support healthcare professionals and the general public. Specific uses include screening for health conditions, triage, counselling, at-home health management support, and training for healthcare professionals [8,13–15]. With phone, mobile, and online platforms widely accessible, conversational agents can support populations with limited access to health care or poor health literacy [16,17]. They also have the potential to be affordably scaled-up to reach large proportions of a population [3] Because of this accessibility, conversational agents are also a

promising tool for the advancement of patient-centred care, and can support users' involvement in their own healthcare [17,18]. Personalizable features have the potential to further improve usability and satisfaction, though more research is needed to evaluate their effectiveness - in achieving their stated health outcomes and reducing costs - and ensure that there are no negative consequences for decision-making or privacy [10].

Despite the large body of research concerning the application of conversational agents to healthcare, most reviews have limited their focus to a particular health area, agent type, or function [10,19–22]. Though there are a few recent systematic reviews that have examined a more comprehensive scope, these have presented an overall synthesis of the body of knowledge. One review developed a taxonomy that described the architecture and functions of conversational agents in healthcare and the state of the field, but did not evaluate the effectiveness, usability, or implications for users [5]. Another systematic review did investigate the outcome measures of the studies of conversational agents but limited the inclusion criteria to agents that used natural language input and had been tested with human participants [2]. Additionally, their initial database searches only retrieved 1531 articles, which raises the concern that some relevant articles may have been overlooked [2]. Their search was updated in February 2018, but given the rapid pace of technological development, there is a need to provide an update and expansion to these previous systematic reviews.

For conversational agents to be successful in healthcare, it is crucial to understand the effectiveness of current agents in achieving their intended outcomes. However, it is just as important to understand how users feel about and relate to these agents, because the adoption of new health technologies depends on users' perceptions of them (for instance, whether they trust the technology, find it easy to use, and feel privacy and data security are being respected) [23]. User-identified problems will need to be addressed if conversational agents are to have a significant impact in healthcare, because their impact depends on people being willing to use them, and preferring to use them over alternatives. The information gathered in this review identifies the current issues with conversational agents that need to be overcome and can be used to help determine which elements of the agents are most likely to be successful and useful in various aspects of healthcare. As conversational agents are often touted as having the potential to reduce burdens on healthcare resources, evaluations of the implications of the agents for improved healthcare provision and reduced resource demand also need to be assessed.

Objectives

The primary objectives of the review are to describe the scope of conversational agents currently being used for healthcare activities (by patients, healthcare providers, or the general public), examine users' perceptions of these agents, and evaluate their effectiveness. Three main research questions were developed to address these objectives. First, are the conversational agents investigated effective at achieving their intended health-related outcomes, and does the effectiveness vary depending on the type of agent? Second, how do users rate the usability and satisfactoriness of the conversational agents and what specific elements of the agents do they like

and dislike? Finally, what are the current limitations and gaps in the utility of conversational agents in healthcare? These objectives build on previous systematic reviews while widening the scope of included studies to update the body of knowledge on conversational agents in healthcare to inform future research and development.

Methods

Database Search

The full methods for this review have been published in detail in a systematic review protocol [24]. The Participant, Intervention, Comparison, Outcome (PICO) framework [25] was used to develop the search strategy, which was performed following the Preferred Reporting Items for Systematic Reviews and Meta-Analyses Protocols (PRISMA-P) [26]. No study design filter was used; any type of study was eligible for inclusion. The search strategy was finalised and tailored to the different databases in consultation with a medical librarian. PubMed, Medline (Ovid), Embase (Ovid), Cumulative Index of Nursing and Allied Health Literature (CINAHL), Web of Science, and ACM Digital Library databases were searched. The search terms were grouped into three themes - conversational agents, health application, and outcome assessment - to capture all studies that fit the key inclusion criteria: evaluating conversational agents used in healthcare. These themes were subsequently searched with the structure: conversational agent (MeSH OR Keywords) AND health application (MeSH OR Keywords) AND outcome assessment (MeSH OR Keywords). The full search strategy can be found in Multimedia Appendix A. The search was completed on November 29th, 2019.

Inclusion and Exclusion Criteria

This systematic review aimed to assess conversational agents designed for healthcare purposes. Studies that evaluated at least one conversational agent were included. Studies targeting any population group, geographical location, and mental or physical health-related function (eg. screening, education, training, self-management) were included. These broad inclusion criteria were established to enable an assessment of the wide range of applications of conversational agents. There was no restriction on study type, as long as the conversational agent was evaluated, so intervention and observational studies, such as cross-sectional surveys, cohort studies, and qualitative studies, were included. Intervention studies were not required to have a specific, or any, comparator.

During the screening process, studies of conversational agents that were not capable of interacting with human users via unconstrained natural language processing (NLP) were excluded. These included conversational agents that only allowed users to select from predefined options or agents with pre-recorded responses which did not adapt to subsequent user responses. The basis of this exclusion is without capability of using NLP computational methods, technologies were rudimentary and not advancing the aims of artificial intelligence for

autonomous computational agents. As many studies did not explicitly state whether the investigated agent was capable of NLP, a description in the paper of the conversational agent allowing free text or free speech input was used as an indicator for NLP and these studies were included. Studies that did not report the architecture of the agent were excluded.

Due to the number of conversational agents in development and/or those that do not progress to evaluation stages of development, studies which were solely descriptive were excluded. Furthermore, due to the pace at which conversational agents have developed over recent decades, studies were limited to those published during or after 2008. 2008 was the year the first iPhone was released and marks an increase in the prevalence and capabilities of digital technology. Only studies published in English were included to ensure accurate interpretation by the authors. Conference publications were also excluded to focus the review on peer-reviewed literature.

Outcomes

The primary objective of this review was to provide an overview of the use of NLP conversational agents in healthcare. Therefore, the primary outcomes evaluated were the effectiveness of conversational agents in achieving their intended health-related outcomes and user perceptions of the agents (including but not limited to acceptability, usability, satisfaction, and specific qualitative feedback). Secondary outcomes included improvement in healthcare provision and resource implications for the healthcare system.

Screening and Study Selection

All studies retrieved from the databases were stored in the reference management software Endnote (version X9; Clarivate Analytics), which automatically eliminated duplicates. Due to time constraints, the Endnote search function was used to extract relevant studies prior to screening of the citations against the inclusion and exclusion criteria by two independent reviewers. Where duplicates or publications from the same study were identified, the more recent publication or the one with the most detail was selected for inclusion in the review. Any disagreements were discussed and if a consensus was not reached a third reviewer was consulted. Full Endnote (version X9; Clarivate Analytics) search details can be seen in Multimedia Appendix B.

The full-texts of the articles thought to meet the inclusion criteria were screened by one of the reviewers. Fifty-eight of the screened articles deemed eligible for inclusion were conference or meeting abstracts, did not have full-texts available, and were excluded. This highlights the early developmental stages of many of these agents.

Data Extraction

Data was extracted by one reviewer and key data points from the studies that were specified in the protocol and identified on further study of the publications were recorded in a

spreadsheet and validated by a second reviewer. The data extraction form was based on the minimum requirements as recommended by the Cochrane Handbook for Systematic Reviews [27]. The types of data extracted from the studies can be seen in Table 1.

Table 1. Data that were extracted from the studies

Article Information	**Data Extracted**
General study information	
	Title of publication
	Year of publication
	Authors
Study characteristics	
	Study design
	Country of study
	Study population
	Analysed sample size
	Comparator(s)
	Study duration
Characteristics of the conversational agent(s)	
	Name of conversational agent(s)
	Architecture
	Device/platform on which agent is accessed
	Intended user
	Primary purpose
Intended outcome(s) of the conversational agent(s)	
	Health objective (general)
	Health outcomes (specific)
Evaluation	
	Effectiveness in achieving intended purpose
	Health literacy
	Improvement in healthcare provision
	Healthcare resource implications
	Usability
	Acceptability / Satisfaction
	User perceptions qualitative feedback
	Conclusions
	Implications for future study

Risk of Bias / Quality Assessment

All quality assessments were conducted by two independent reviewers, with disagreements being resolved by consensus. If this was not possible, the opinion of a third reviewer was sought. As there was a wide variety of study designs, the study types were classified by one reviewer and validated by a second reviewer, with disagreements being resolved by discussion with a third reviewer. Because of the broad inclusion criteria that was intended to capture all relevant studies, a few of the included studies used implementation models for AI research that were beyond the scope of classic public health design methods. This resulted in some study designs being categorized as 'other'.

The Cochrane Collaboration Risk of Bias tool was used to evaluate the risk of bias in randomized controlled trials (RCTs) [28]. The Critical Appraisal Skills Programme tools for cohort and qualitative studies were used for the respective studies [29] and the AXIS tool was used to assess the quality of cross-sectional survey studies [30]. Studies that were coded as 'other' design types were also assessed using the AXIS tool, which was deemed to be the most rigorous and appropriate tool because it systematically evaluates elements of the introduction, methods, results, and discussion sections and is not limited to the RCT specific questions used in the Risk of Bias tool.

The results of the Cochrane Collaboration Risk of Bias Tool were summarized using RevMan 5.3. Critical Appraisal Skills Programme and AXIS scores were calculated using yes = 1, no = 0, can't tell / don't know = 0 for each question. The scores for each question were summed to provide a score for each study, which were averaged according to study type and presented in the results.

Data Analysis and Synthesis

Due to the variability in populations, interventions, outcomes and study designs, a meta-analysis of the studies was not possible. Therefore, we report a structured analysis of the findings to draw conclusions about the effectiveness and user perceptions of conversational agents in healthcare. For the purpose of this review, the agent was considered effective if there was a statistically significant ($P<.05$) improvement in a given outcome as compared to a comparator or control, or over time. If no significance was reported or the difference was non-significant or significantly worse between groups or over time the agent was considered to have no significant evidence supporting it. Limitations and future directions for research were also summarised.

The Synthesis Framework for the Assessment of Health Information Technology (SF/HIT) was used to structure the evaluation of the studies because it included a whole system set of outcome variables [31]. These included effectiveness, satisfaction and perceived ease of use / usefulness, among others. According to the framework, evidence for each of the outcome

variables was coded as ‘positive or mixed’ or ‘neutral or negative.’ If the study did not address the outcome in question, it was coded as ‘neutral or negative’.

Finally, when qualitative user feedback was reported, it was examined to extract common themes by the sections of original text that discussed the qualitative perceptions, reducing them to key themes, and then comparing those key themes across the different studies.

Results

Included Studies

Overall, 9441 studies were retrieved from the six databases of which 2782 were duplicates. The reference management software Endnote (version X9; Clarivate Analytics) was used for initial screening, with keywords based on the original search categories and irrelevant studies identified from preliminary viewing used to exclude studies that did not meet the criteria. After six passes, 957 citations remained for abstract screening. The primary reasons for exclusion at screening stage were that the study did not include an interactive, responsive conversational agent (n=470), was a review article (n=65), was not health-related (n=48) or did not report any evaluation of the conversational agent (n=46). Of these 957 citations, 293 were selected for full-text review. Thirty-one papers were included in the final review. The reasons for the exclusion at full-text review are detailed in Figure 1, with the most common being that the conversational agent did not use NLP (n=81), no full text was available (n=71), or there was no conversational agent in the paper (n=51).

Figure 1. PRISMA flow diagram

Study Characteristics

The characteristics of the 31 included studies are summarized in Multimedia Appendix C. Of these studies, 45% (14/31) evaluated conversational agents that had some type of audio or speech element. 45% (14/31) of the agents were chatbots (including two voice chatbots and one chatbot that also used a wizard), 19% (6/31) were embodied conversational agents (ECA, including one virtual doctor), 10% each (3/31) were interactive voice response (IVR) phone calls, virtual patients, and speech recognition screening systems, and the final two were a contextual question answering agent and a voice recognition triage system. Of the 26 studies that reported the device their conversational agent was used on, 35% (9/26) were computers, 27% (7/26) were web-based, 23% (6/26) were mobile phone apps, and 15% (4/26) were telephone calls, and one study used a tablet (the percentages do not add up to 100% because one agent could be used on a computer or telephone).

Figure 2. The number of studies examining agents that were designed for use on certain platforms (n = 26, one agent could be used on a computer or telephone)

There were a wide variety of areas of healthcare targeted by the conversational agents of the included studies. The greatest number (39%, 12/31) addressed mental health issues [13,32–42], with 19% (6/31) each providing some form of clinical decision or triage support [8,12,40,42–44] and treatment support (including encouraging users to get screened) [9,45–49], 10% (3/31) each were used to support training of healthcare students [15,41,50] and the screening or diagnosis of users [14,38,51], 7% (2/31) each targeted physical health [52,53] and layperson medical education [54,55], and one agent was designed to help monitor users' speech [56]. The percentages do not add up to a hundred because some of the studies that addressed mental health also fit in one of the other categories.

Figure 3. The number of studies examining conversational agents designed for certain healthcare areas or functions (n = 31, some studies counted in two categories)

The study designs also varied widely, with 29% (9/31) using cross-sectional designs, 26% (8/31) using randomized controlled trials, 23% (7/21) using qualitative methods, 19% (6/31) using cohort studies, and one using a cluster crossover design. The full data extraction table is available in Multimedia Appendix D.

Overall Evaluation of Conversational Agents

Overall, about three quarters of studies (22/30) reported positive or mixed results for the majority of outcomes. Eight of the studies were coded as reporting positive or mixed evidence for 10 or more of the 11 outcomes specified in the SF/HIT; analysis for this review was limited to interpretation of impact as reported by study authors to reflect evaluation outcomes. Excluding one study, which was an acceptability study only and did not assess the other outcomes, the average number of outcomes that were coded as 'positive or mixed' was 7 (67%, SD = 2.5). However, the number of outcomes met per study ranged from 1-11 (9-100%). Perceived ease of use / usefulness (27/30, 90%), process of service delivery / performance (26/30, 87%), appropriateness (24/30, 80%), and satisfaction (26/31, 84%) were the outcomes that had the most support from the studies. Just over three quarters (23/30) of the studies also reported positive or mixed evidence of effectiveness.

However, very few studies discussed cost effectiveness (5/30, 17% coded as 'positive or mixed') or the safety, privacy, and security (14/30, 47% coded as 'positive or mixed') outcomes for the agents being evaluated. Just over a quarter of studies (8/30) had neither positive nor mixed reported evidence for more than half of the SF/HIT outcomes. The evaluation of the SF/HIT outcomes is summarized in Table 2.

Table 2. Summary of the studies based on the evaluation outcomes from the SF/HIT [31]

First Author	Preventive care	Adherence/ Attendance	Efficiency	Perceived ease of use/ Usefulness	Effectiveness	Performance	Safety/ Privacy/ Security	Acceptability	Cost effectiveness	Appropriateness	Satisfaction	N (%)
Adams [9]	1	1	1	1	1	1	1	1	0	1	1	10 (91)
Bibault [46]	1	1	1	1	1	1	1	1	0	1	1	10 (91)
Borja-Harta [50]	0	1	1	1	1	1	1	0	0	1	0	7 (64)
Cameron [32]	0	0	1	1	0	1	0	1	0	0	1	5 (45)
Chaix [45]	1	0	1	1	1	1	1	0	0	1	1	8 (73)
Chang [8]	0	1	0	1	1	0	1	1	0	1	1	7 (64)
Crutzen [54]	0	1	1	1	1	1	1	1	0	1	1	9 (82)
Dimeff [42]	1	0	1	1	1	1	1	1	1	1	1	10 (91)
Elmasri [33]	0	0	0	1	0	1	1	0	0	1	1	5 (45)
Fitzpatrick [13]	1	1	1	1	1	1	1	1	0	1	1	10 (91)
Friederichs [53]	0	0	0	1	0	1	0	1	0	0	1	4 (36)
Fulmer [34]	1	1	0	0	1	1	1	0	0	0	1	6 (55)
Galescu [52]	0	0	1	1	0	1	0	0	0	0	0	3 (27)
Ghosh [44]	1	1	1	1	1	1	0	1	0	1	1	9 (82)
Havik [14]	1	1	1	1	1	1	0	1	1	1	1	10 (91)
Heyworth [47]	0	1	1	1	1	1	1	1	0	1	0	8 (73)
Hudlicka [35]	1	1	1	1	1	1	1	1	1	1	1	11 (100)
Inkster [36]	1	1	1	1	1	1	0	1	0	1	1	9 (82)
Ireland [56]											1	1 (100)

Isaza-Restrepo [15]	1	1	1	1	1	1	0	1	1	1	1	10 (91)
Ly [37]	0	1	0	1	0	1	0	0	0	1	1	5 (45)
Nakagawa [12]	1	0	1	1	1	1	0	0	0	1	1	7 (64)
Philip (2014) [51]	1	1	1	1	1	1	1	1	0	1	1	10 (91)
Philip (2017) [38]	1	1	1	1	1	1	0	1	0	1	1	9 (82)
Rhee [48]	1	1	1	1	1	1	0	1	0	1	1	9 (82)
Simon [49]	0	1	0	1	0	1	1	1	0	1	1	7 (64)
Spänig [43]	0	0	1	0	1	1	0	1	0	1	1	6 (55)
Washburn [41]	1	0	0	1	1	1	0	0	1	0	0	5 (45)
Wong [55]	0	0	0	1	0	0	0	0	0	0	0	1 (9)
Xu [40]	1	0	1	0	1	0	0	0	0	1	1	5 (45)
Yasavur [39]	0	1	1	1	1	0	0	1	0	1	1	7 (64)
N (%)	17 (57)	19 (63)	22 (73)	27 (90)	23 (77)	26 (87)	14 (47)	20 (67)	5 (17)	24 (80)	26 (84)	

*The impact ‘positive or mixed’ has been coded as 1 and the outcome ‘neutral or negative’ as 0

When grouped by the agent’s healthcare scope, studies of certain types of agents appear to do better than others (see Table 3). Studies examining screening or diagnosis agents and treatment support agents had the highest average number of positive or mixed outcomes (mean = 10, SD = 0.6 and mean = 9, SD = 1.2, respectively). Treatment support agents had primary functions that included empowering patients to engage more fully in clinical appointments, encouraging attending screenings for healthcare conditions, and supporting patient self-management. In contrast, mental health agents focused on addressing challenges related to depression, anxiety, and alcohol abuse, among others. However, given the small number of studies for each category of agent, this should be interpreted with caution.

Table 3. Summary of evaluation outcomes by the area of healthcare addressed by the conversational agent

Agent focus	Number of studies	Average number of outcomes coded 'positive or mixed' n (%)	Range of scores	Standard deviation
Mental health [13,32–42]	12	7 (66)	5-11	2.4
Clinical decision / triage support [8,12,40,42–44]	6	7 (67)	5-10	1.9
Treatment support [9,45–49]	6	9 (79)	7-10	1.2
Healthcare training (students) [15,41,50]	3	7 (67)	5-10	2.5
Screening / diagnosis [14,38,51]	3	10 (88)	9-10	0.6
Healthcare education (laypeople) [54,55]	2	5 (45)	1-9	5.7
Physical health [52,53]	2	4 (32)	3-4	0.7

*The number of studies does not add up to 31 because some studies fit into two categories, and the study on monitoring speech was not included because it only addressed one of the eleven outcomes. The percentages associated with the average number of outcomes varies slightly due to rounding.

Qualitative User Perceptions

18 of the 31 studies included more specific user feedback. The most frequently raised issue with the conversational agents (in 9 studies) was poor understanding due to limited vocabulary, voice recognition accuracy, or error management of word inputs [13,32–37,41,52]. Related to this issue, as the conversational agents often had to ask questions more than once to be able to process the response, users in three studies noted disliking that conversations with the agents were repetitive [13,36,37]. These are both key areas of improvement for future research and development of conversational agents because they represent limitations in the usability of the agents in a real-world context.

Feedback from users in five studies expressed a preference for interactivity, with users in one study noting that they liked the interactivity of the chatbot [35,37], and users in the other four studies expressing a desire for greater interactivity or relational skills in the conversational agent [14,32,34,53]. Similarly, users in four studies reported liking that the agent had a personality and/or showed empathy [13,32,34,42] while users in other studies reported disliking the lack of personal connection or difficulty empathizing with the agent [35,37,50] or its limited conversation and responses [35,56].

Due to the wide variety of conversational agents, and their aims and healthcare contexts, much of the qualitative user perception data concerned distinct aspects of the agents. However,

several studies reported feedback concerned with customisation or availability of feature options - with two studies commenting on it positively (e.g. having both voice and touch modes to allow hands-free work and rapid data input on a triage system for nurses) [8,35], and three studies desiring more features and more control [33,37,48]. Additionally, users in two studies suggested that better integration of the agent with EHR systems (for a virtual doctor [42]) or healthcare providers (for an asthma self-management chatbot [48]) would be useful.

Other features of the agents that users reported liking were reminders and the assistance in forming routines [37,48], that they provided accountability [13,34,48], that they facilitated learning [13,34,37], and that they were easy to learn and use [8,15]. Three of the conversational agents in the included studies were virtual patients, and users in all three studies reported liking that it provided a platform for risk-free learning, because they were not practicing on real patients [15,41,50].

Several of the studies reported user feedback that was specific to that conversational agent. This included a preference for telephone IVR over web-based pediatric care guidance [9] and for a simple avatar with a computer-generated voice over a more life-like agent with a recorded voice [42]. Users in one study reported liking that the agent initiated conversations [37] but there was opposite feedback in two studies about the format of response, with users preferring pre-formatted options for one chatbot [36] while some users preferred the free-text responses for a diagnostic chatbot because it allowed them to provide contextual information but others found that it more difficult to know how to respond so the agent would understand [14].

Other agent-specific negative feedback included that the virtual doctor did not have capability to go deep enough or provide access to other materials [42], that too much information was provided [13,33] or the interaction was too long [13], the use of non-verbal expressions on the avatar [35], and a lack of clarity regarding the aim of the chatbot [37]. Some students who used the virtual patients also reported that it was difficult to empathize [50] and that the agent did not sufficiently encompass real situational complexity [15]. The variety of specific feedback reported demonstrates the importance of examining usability for individual conversational agents and tailoring the design to the intended population. While there were some preferences and complaints that were frequently reported, much of the feedback was agent-dependent. A summary of the thematic analysis is included in Multimedia Appendix E.

Implications for healthcare provision and resources

Unfortunately, only a few of the studies discussed any improvement in healthcare provision or implications for resources. Two of the studies that suggested improvement in healthcare provision were evaluating virtual patients [41,50], and students reported (in one study, significantly) increased confidence in their clinical skills and ability to interview patients. Over 80% of users also reported that the agents helped them follow their treatment more effectively [45] and be more prepared for pediatric visits [9]. In a study of an embodied conversational agent (ECA) for sleep disorder screening, 65% of users reported thinking that the agent could provide significant assistance to physicians [51]. As for resource implications, the study of a preparatory

IVR phone call before pediatric visits found that visit time was significantly reduced in the IVR group compared to the control [9]. The use of an ECA to screen for depression [38] and a virtual doctor for suicidal patients in emergency departments (ED) [42] were suggested by the authors to have the potential to save physicians time and reduce the costs associated with ED visits for suicidal ideation, but these outcomes were not evaluated. Likewise, another study suggested that mindfulness meditation could be of more use with more cost-effective training made available via a virtual coach [35].

Suggestions such as this - that the conversational agents have the potential to improve healthcare provision, save healthcare providers' time, and reduce costs - were frequent among the studies. However, as demonstrated above, very few studies quantified these claims, and even fewer measured these outcomes with objective measures. This is a limitation of the studies as a whole - even though many were in early stages of testing, claims about potential value to the healthcare system in terms of time or money should be substantiated. However, as evidenced by the number of 'neutral or negative' codings on the evaluation, many of the studies were not considering whole system implementation outcomes. It will be important for future development of conversational agents to consider outcomes like these from the beginning, so that agents can be built that are not only acceptable and usable, but also provide value to the healthcare system.

Risk of Bias and Quality Assessments

There were a variety of study types included in this review, so several different quality assessment tools were used to assess the risk of bias in and quality of the 31 included studies. Six of the studies could not be classified as RCTs, cohort, qualitative, or cross-sectional studies, and their study design was coded as 'other' [12,39,40,44,52,55]. Most of these were papers describing the development and initial evaluation of conversational agents, and half of them did not use participants [40,44,55]. Initially, studies that did not have an explicit design were classified as qualitative / interpretative studies. However, upon further analysis, many of the studies did not fit the criteria of qualitative studies as evaluating subjective, thematic, non-numerical data because they evaluated performance metrics such as word error rates [52], accuracy [12,39,40,52,55], precision [44], and user experience quantified on Likert scales [39]. Therefore, these studies were coded as 'other' and assessed using the AXIS tool for cross-sectional studies, which was deemed to provide the most systematic evaluation of the various elements of the studies [30]. The quality of these studies was assessed as best as possible, however, the judgments should be considered in the context of these limitations.

Overall, the quality was poor to moderate. On average, the randomized controlled trials (RCTs) [9,13,34,37,46,47,49,53] and qualitative studies [41,48,56] evaluated were generally determined to have the highest quality and lowest risk of bias, with none of the other three study types meeting more than half of the criteria for the quality assessment. The evaluation of risk of bias for the eight RCTs was conducted using the Cochrane Collaboration Risk of Bias tool [28], and the results were summarized using the RevMan 5.3 software [57]. Overall, the RCTs performed fairly well in the risk of bias assessment. About half of the studies were assessed as

having a low risk of selection bias because of proper random sequence generation (5/8) and allocation concealment (4/8), and a low risk of reporting bias (4/8), as outcomes reported could be compared to *a priori* protocols or trial registrations. Most studies reported blinding of outcome assessors (7/8) and a low risk of attrition bias because of low or equal dropout across groups or the use of intention-to-treat analyses (6/8). The majority of studies (5/8) had a high risk of performance bias, but this was predominantly because blinding was not possible given the intervention.

Figure 4. Risk of bias summary: review authors' judgements about each risk of bias item for each included study

Figure 5. Risk of bias graph: review authors' judgements about each risk of bias item presented as percentages across all included studies

The cohort (n = 9) and qualitative (n = 3) studies assessed using the Critical Appraisal Skills Programme checklists met on average 5/12 (range: 1-10) and 7/10 (range: 4-9) criteria, respectively [29]. Of the cohort studies, the questions with the best performance were “Did the study address a clearly focused issue?” (8/9 yes), “Was the follow up long enough?” (8/9 yes), and “Do the results of this study fit with other available evidence?” (6/9 yes). Studies performed the worst - either through failing to meet the criteria or failing to report it - on questions about cohort recruitment (1/9 yes), identifying and accounting for confounding factors (1/9 yes), accurate exposure and outcome measurement (2/9 and 3/9 yes, respectively), and the applicability of results to the local population (3/9 yes). The qualitative studies, on the other hand, performed best on the questions about whether a qualitative methodology was appropriate, the consideration of ethical issues, clear statements of findings, and if the results will help locally (3/3 yes for each). None of the three studies reported any consideration of the relationship between researcher and participant, and also performed poorly on questions about sample recruitment, data collection, and data analysis (1/3 yes for each).

The cross-sectional (n = 5) and ‘other’ (n = 6) studies assessed using the AXIS tool met on average 50% (range: 26-80%) and 42% (range: 29-70%) of the criteria, respectively [30]. Percentages are reported instead of the exact number of criteria because several of the questions were not applicable to the studies, so the total number of criteria assessed per study was not the same (averages: 19 and 16, ranges: 18-20 and 10-19, respectively). Overall, the cross-sectional studies performed best on questions about the clarity of aims (5/5 yes), appropriate outcome variables for the aims (5/5 yes), internal consistency (5/5 yes), and adequate description of basic data (4/5 yes). They performed worst on questions about sample selection - if it was taken from an appropriate base to represent the population (1/5 yes) and whether the process was likely to select a representative sample (0/5 yes) - the use of appropriate outcome measures (previously assessed; 0/5 yes), whether the methods were adequately described for replication (1/5 yes), and conflicts of interest (1/5 no, most did not report).

The ‘other’ studies performed best on the questions about whether the study design was appropriate for the aims and if the conclusions were justified by the results (6/6 yes for both), and also did well overall on appropriate choice of outcome variables and internal consistency (5/6 yes for both). However, all the ‘other’ studies, for whom the questions were applicable, performed poorly on questions about the justification of sample size (0/5 yes), whether the selection process was likely to get a representative sample (0/5 yes), addressing non-responders (0/2 yes), adequate description of basic data (0/4 yes), concerns about non-response bias (0/3 yes), the presentation of results for all the analyses described in the methods (0/6 yes, although this was mostly because analyses were not adequately described in the methods), and conflicts of interest (0/6 yes, again because nothing was reported). Furthermore, only one study adequately addressed the questions about the use of previously assessed outcome measures (1/5 yes), sufficient description of the methods for replication (1/6 yes), and discussion of study limitations (1/6 yes). It should be noted that the AXIS tool used to assess the ‘other’ studies was designed for cross-sectional studies, and does not fit exactly with the designs of these studies. Therefore, it is possible these studies would perform better when assessed by a tool specific to their study type. Tables depicting the judgments for each question of the CASP cohort and qualitative checklists and the AXIS tool for the cross-sectional and ‘other’ studies are included in Multimedia Appendices F-I.

Discussion

Principal Findings

In this systematic review, we examined 31 studies that evaluated the effectiveness and usability of conversational agents in healthcare. Overall, studies reported a moderate amount of evidence supporting the effectiveness, usability, and positive user perceptions of the agents. On average, two thirds of the studies (67%) reported positive or mixed evidence for each evaluation outcome. However, this ranged significantly, with usability, agent performance, and satisfaction having the most support across the studies, and cost-effectiveness receiving hardly any. It should also be noted that the definitions of ‘effectiveness’ were highly varied and, as evidenced by the methodological limitations identified in the quality assessment, rarely evaluated with the scrutiny expected for medical devices. While this is promising for the use of conversational agents in healthcare, there are a number of limitations in both the studies analysed and the structure of this review that question the validity of this finding.

With regard to qualitative user perceptions of the agents, specific feedback was very mixed. Users highlighted many positive factors of the agents, particularly its personality and ability to provide empathy and emotional support, that it supports learning, that it’s easy to use and access, and that it helps them be accountable, all of which support the generally positive evaluations of usability and satisfaction outcomes. However, there were a number of limitations of the agents that were consistently raised across the studies that reported qualitative feedback. These included that the agents had difficulty understanding them, that they were repetitive and

not sufficiently interactive, and that the users had difficulty forming personal connections with the agents. This suggests that despite the generally positive usability reported by the studies, there are a number of barriers to successful use of conversational agents in healthcare that will need to be addressed before they can achieve the greatest impact. It should be noted that this review only included studies of conversational agents that NLP, and that free-text inputs are likely to present greater difficulties for comprehension.

The results of this systematic review are largely consistent with the literature, particularly the previous systematic review evaluating conversational agents in healthcare [2]. They also found a limited quality of design and evidence in the included studies, with inconsistent reporting of study methods (including methods of selection, attrition, and a lack of validated outcome measures) and conflicts of interest [2]. They identified that high-quality evidence of effectiveness and patient safety was limited, which was also observed in this review. Likewise, they noted that high overall satisfaction was generally reported by the studies but that the most common issues with the conversational agents related to language understanding or poor dialogue management, which is consistent with our findings [2]. Some of this similarity in results is likely due to the overlap in included studies - 7 of their 17 studies were also included in our review [2].

Quality of the Evidence

As noted in the previous systematic review [2], there were significant issues with the quality of many of the included studies. One of the consistent issues among many of them was the risk of selection bias. A large proportion of the studies relied on volunteers for the study, many of whom were recruited via self-selection means such as flyers and emails or downloading the app being studied. The risk with self-selection recruitment is that participants who elect to take part in the study are already more positively predisposed to new technologies than those who don't participate, which would tend to weight the evaluation of the technology more positively. To make matters worse, several of the studies also did not sufficiently report their recruitment strategies, so their potential selection bias cannot be accurately evaluated. In research such as this, where user perceptions are a main outcome, this is a serious concern. Future studies should take care to implement recruitment strategies that minimize the risk of this selection bias or balance the potential bias in evaluations by actively recruiting participants who are less inclined towards new technology.

Another limitation of many of the studies is small sample sizes. Almost two thirds of the studies (19/31) used samples of less than 100 participants or items of analysis (voice clips, clinical scenarios) with a median sample size of 48 across all the studies. Many also did not sufficiently report demographic data or whether their sample was representative of their target population. While many of these studies were early feasibility and usability trials, this is an important issue to address in future research testing these agents, to determine whether an agent will be used - and used effectively - by its target population.

Limitations

The validity of the evidence extracted from the included studies was also affected by limitations in the structure of the review. The Synthesis Framework for the Assessment of Health Information Technology (SF/HIT) was used to provide a structured set of whole system implementation outcomes on which to evaluate the conversational agents [31]. However, an issue with the use of this framework that was discovered during analysis was that many of the included studies were describing system innovation. Therefore, they did not address or provide evidence for many of the outcomes described by the SF/HIT. Additionally, as the included data indicated self-reported impact in the studies of effectiveness, the study effectiveness is biased favorably to authors reporting of impact.

This limitation in the use of the framework for this review also highlights a limitation in many of these studies; namely, that they are not thinking about whole system implementation from the early stages of agent design, development, and testing. It is possible that lack of evaluation about the implications of the agents for healthcare provision and resources was due to an emphasis on technology development and evaluation, rather than system integration. This is a pervasive issue in technological innovation; so much so that it drove the development of the non-adoption, abandonment, scale-up, spread, and sustainability (NASSS) framework as a means of predicting and assessing the success of new health technologies [58] the development and evaluation of new conversational agents, to ensure that these later-stage implications of healthcare provision, cost-effectiveness, and privacy and security are being sufficiently considered from the early stages of innovation. They must also be properly evaluated with a large sample of users, rather than simply presented as unsubstantiated claims that the agent will reduce costs and save healthcare providers time.

Additionally, in accordance with their framework, the outcomes' impact on each outcome was coded as 'positive or mixed' or 'neutral or negative'. However, this combination of positive and mixed outcomes reduces the granularity of the results. During the coding process, several outcomes were distinctly coded as 'positive' or as 'mixed', and collating the two outcome impacts into one reduces the precision of the information presented to the readers. Additionally, studies that did not assess the outcome in question were coded as 'neutral or negative' because they did provide explicit support for the outcome. In the analysis, outcomes were initially coded separately as positive, mixed, positive or mixed (for studies that reported a positive outcome but did not provide sufficient statistical evidence), and neutral or negative. This table is available in Multimedia Appendix J. Positive and mixed outcomes were combined for the final presentation of the data in line with the framework. However, it might be more useful to distinguish between studies that attempted to find significant evidence for an outcome but did not, and those that did not attempt it. This would provide a clearer picture of which outcomes are not being supported by the evidence and should be targeted for improvement, and which outcomes still need to be examined. In future, it would be worth evaluating whether the coding system should be adjusted to provide a more detailed and informative summary of the evidence.

Further limitations of this review are that we limited the focus to include only unconstrained natural language processing and interaction. This was chosen as a focus because of the advantages NLP poses for simulating human-to-human interaction, however, it will have necessarily excluded studies of relevant conversational agents that could be satisfactory, useful, and effective in addressing current healthcare challenges. Additionally, no spidering searches were used to identify potentially relevant studies in the references of the included studies that were missed in the initial search. The exclusion of conference abstracts might also have missed relevant papers that were classed as abstracts; however, a previous systematic review that included conference abstracts in their search only had one included in their final selection [2]. The inclusion of only studies published in English also likely excluded relevant research on conversational agents conducted in other countries. These limitations should be corrected in future studies in order to ensure that the full body of relevant literature is examined.

Future Directions

Future reviews of conversational agents in healthcare could be extended to include constrained NLP and non-NLP conversational agents. A synthesis of the evidence identified here with other types of conversational agent in healthcare - perhaps structured according to the taxonomy suggested by Montenegro et al. [5] - could be used to examine overall trends, and provide a better picture of what is being used, what works, and what doesn't, to further guide the development of the conversational agents that are most likely to be successful.

Future research should also include more qualitative evaluations of the features that users like and dislike. Only just over half (18/31) of the studies included in this review reported specific user feedback, despite the fact that 7 of the remaining 13 studies conducted some measure of usability or user perceptions. It will be important to identify all of the structural, physical, and psychological barriers to use if conversational agents are to achieve their potential for improving healthcare provision and reducing the strain on healthcare resources. To this aim, it would be useful for future studies to structure their evaluation of conversational agents around a behavioural change framework (for example, the Behaviour Change Wheel framework [59]). This is important not just when evaluating the effectiveness of behaviour change focused conversational agents, but when determining whether and how the adoption of new conversational agent technology will be successful.

It will be important for future studies of conversational agents to take care to properly structure and report their studies to improve the quality of the evidence. Without high-quality evidence, it is difficult to assess the current state of conversational agents in healthcare, what is working, and what needs to be improved to make them a more useful tool. Likewise, there is a gap in the evidence regarding the health economics of these agents. Very few studies in this review even discussed the cost analysis of the agent in questions, let alone providing substantive evidence about its cost-effectiveness. The evaluation of costs and outcomes of new technologies, as well as their privacy, security, and interoperability, will be necessary to advance value-based healthcare [60]. However, there was very little evidence to suggest that the conversational agents

examined in this review considered or addressed these concerns. User feedback on two of the studies even noted that better interoperability between the agent and EHRs or healthcare providers would improve its usefulness.

Conclusions

The objective of this systematic review was to provide a synthesis of the evidence of conversational agents' usability, effectiveness, and satisfactoriness in healthcare. Although the studies generally reported positive outcomes relating to the agents' usability and effectiveness, the quality of the evidence was not sufficient to provide strong evidence to support these claims. This study extended the literature by expanding the summary of the literature to examine a whole system set of evaluation outcomes - including cost-effectiveness and privacy and security, which have not been systematically examined in previous reviews. Additionally, it provides a distinct contribution by conducting a thematic analysis of qualitative user perceptions of the agents. Further research is needed to examine the cost-effectiveness and value of these agents in healthcare - both in their current and potential states. Higher quality studies - with more consistent reporting of design methods and better sample selection - are also needed to more accurately assess the usefulness of, and identify the key areas of improvement for, current conversational agents. A more holistic approach to the design, development, and evaluation of conversational agents will help drive innovation and improve their value in healthcare.

Acknowledgements

We would like to thank the outreach librarians Liz Callow (University of Oxford) and Kirsten Elliot (Imperial College London) for their assistance in developing search terms and reviewing search strategy.

Specific funding for this work has not been acquired. EM's work on digital health solutions is currently supported by the Sir David Cooksey Fellowship in Healthcare Translation at the University of Oxford. The conclusions drawn in the paper are made by the authors and not necessarily supported by the University of Oxford. The funding body had no role in the design, execution, or analysis of this systematic review.

Author Contributions

CdC and EM conceived the study topic and designed the review protocol. CdC and MMI screened the studies. CdC conducted the data extraction, which was validated by MMI, and MMI conducted the risk of bias and quality assessments, which were validated by EM. MMI and EM analysed the data extracted. The methods section was drafted by CdC and the rest of the review was written by MMI with revisions from EM. MHS, EL, NdP, and GM provided feedback on the final drafted text. EM supervised the study execution.

We confirm that we have followed all appropriate research reporting guidelines. The PRISMA checklist for systematic reviews has been uploaded as Multimedia Appendix K along with other relevant materials.

Conflicts of Interest

EL, NdP, and GM are all employees of Ufonia Limited, a voice artificial intelligence company. However, the paper was funded by the Sir David Cooksey Fellowship in Healthcare Translation at the University of Oxford and Ufonia had no editorial influence on the final drafting and their contribution was limited to feedback given their applied voice AI expertise, therefore no conflict of interest is identified.

References

1. Bibault J-E, Chaix B, Nectoux P, Pienkowsky A, Guillemasse A, Brouard B. Healthcare ex Machina: Are conversational agents ready for prime time in oncology? Clin Transl Radiat Oncol 2019 May;16:55–59. PMID:31008379

2. Laranjo L, Dunn AG, Tong HL, Kocaballi AB, Chen J, Bashir R, Surian D, Gallego B, Magrabi F, Lau AYS, Coiera E. Conversational agents in healthcare: a systematic review. J Am Med Inform Assoc 2018 Sep 1;25(9):1248–1258. PMID:30010941

3. Luxton DD. Ethical implications of conversational agents in global public health. Bull World Health Organ 2020 Apr 1;98(4):285–287. PMID:32284654

4. Davenport T, Kalakota R. The potential for artificial intelligence in healthcare. Future Healthcare Journal 2019;6(2):94–98.

5. Montenegro JLZ, da Costa CA, da Rosa Righi R. Survey of conversational agents in health [Internet]. Expert Systems with Applications. 2019. p. 56–67. [doi: 10.1016/j.eswa.2019.03.054]

6. Weizenbaum J. ELIZA --- a computer program for the study of natural language communication between man and machine [Internet]. Communications of the ACM. 1966. p. 23–28. [doi: 10.1145/357980.357991]

7. Campillos-Llanos L, Thomas C, Bilinski É, Zweigenbaum P, Rosset S. Designing a virtual patient dialogue system based on terminology-rich resources: Challenges and evaluation [Internet]. Natural Language Engineering. 2020. p. 183–220. [doi: 10.1017/s1351324919000329]

8. Chang P, Sheng Y-H, Sang Y-Y, Wang D-W. Developing a wireless speech- and touch-based intelligent comprehensive triage support system. Comput Inform Nurs 2008 Jan;26(1):31–38. PMID:18091619

9. Adams WG, Phillips BD, Bacic JD, Walsh KE, Shanahan CW, Paasche-Orlow MK. Automated conversation system before pediatric primary care visits: a randomized trial. Pediatrics 2014 Sep;134(3):e691–9. PMID:25092938

10. Kocaballi AB, Berkovsky S, Quiroz JC, Laranjo L, Tong HL, Rezazadegan D, Briatore A, Coiera E. The Personalization of Conversational Agents in Health Care: Systematic Review. J Med Internet

Res 2019 Nov 7;21(11):e15360. PMID:31697237

11. Sun R, Aldunate R, Ratnam R, Jain S, Morrow D, Sosnoff J. VALIDITY AND USABILITY OF AN AUTOMATED FALL RISK ASSESSMENT TOOL FOR OLDER ADULTS [Internet]. Innovation in Aging. 2018. p. 362–362. [doi: 10.1093/geroni/igy023.1338]

12. Nakagawa S, Enomoto D, Yonekura S, Kanazawa H, Kuniyoshi Y. A Telecare System that Estimates Quality of Life through Communication [Internet]. 2018 5th IEEE International Conference on Cloud Computing and Intelligence Systems (CCIS). 2018. [doi: 10.1109/ccis.2018.8691360]

13. Fitzpatrick KK, Darcy A, Vierhile M. Delivering Cognitive Behavior Therapy to Young Adults With Symptoms of Depression and Anxiety Using a Fully Automated Conversational Agent (Woebot): A Randomized Controlled Trial. JMIR Ment Health 2017 Jun 6;4(2):e19. PMID:28588005

14. Håvik R, Wake JD, Flobak E, Lundervold A, Guribye F. A Conversational Interface for Self-screening for ADHD in Adults [Internet]. Internet Science. 2019. p. 133–144. [doi: 10.1007/978-3-030-17705-8_12]

15. Isaza-Restrepo A, Gómez MT, Cifuentes G, Argüello A. The virtual patient as a learning tool: a mixed quantitative qualitative study [Internet]. BMC Medical Education. 2018. [doi: 10.1186/s12909-018-1395-8]

16. van Heerden A, Ntinga X, Vilakazi K. The potential of conversational agents to provide a rapid HIV counseling and testing services [Internet]. 2017 International Conference on the Frontiers and Advances in Data Science (FADS). 2017. [doi: 10.1109/fads.2017.8253198]

17. Bickmore TW, Pfeifer LM, Byron D, Forsythe S, Henault LE, Jack BW, Silliman R, Paasche-Orlow MK. Usability of conversational agents by patients with inadequate health literacy: evidence from two clinical trials. J Health Commun 2010;15 Suppl 2:197–210. PMID:20845204

18. Zhang Z, Bickmore T. Medical Shared Decision Making with a Virtual Agent [Internet]. Proceedings of the 18th International Conference on Intelligent Virtual Agents. 2018. [doi: 10.1145/3267851.3267883]

19. Vaidyam AN, Wisniewski H, Halamka JD, Kashavan MS, Torous JB. Chatbots and Conversational Agents in Mental Health: A Review of the Psychiatric Landscape. Can J Psychiatry 2019 Jul;64(7):456–464. PMID:30897957

20. Russo A, D'Onofrio G, Gangemi A, Giuliani F, Mongiovi M, Ricciardi F, Greco F, Cavallo F, Dario P, Sancarlo D, Presutti V, Greco A. Dialogue Systems and Conversational Agents for Patients with Dementia: The Human-Robot Interaction. Rejuvenation Res 2019 Apr;22(2):109–120. PMID:30033861

21. Xing Z, Yu F, Qanir YAM, Guan T, Walker J, Song L. Intelligent Conversational Agents in Patient Self-Management: A Systematic Survey Using Multi Data Sources. Stud Health Technol Inform 2019 Aug 21;264:1813–1814. PMID:31438357

22. Provoost S, Lau HM, Ruwaard J, Riper H. Embodied Conversational Agents in Clinical Psychology: A Scoping Review. J Med Internet Res 2017 May 9;19(5):e151. PMID:28487267

23. Safi S, Thiessen T, Schmailzl KJ. Acceptance and Resistance of New Digital Technologies in

Medicine: Qualitative Study. JMIR Res Protoc 2018 Dec 4;7(12):e11072. PMID:30514693

24. de Cock C, Milne-Ives M, van Velthoven MH, Alturkistani A, Lam C, Meinert E. Effectiveness of Conversational Agents (Virtual Assistants) in Health Care: Protocol for a Systematic Review. JMIR Res Protoc 2020 Mar 9;9(3):e16934. PMID:32149717

25. Schardt C, Adams MB, Owens T, Keitz S, Fontelo P. Utilization of the PICO framework to improve searching PubMed for clinical questions. BMC Med Inform Decis Mak 2007 Jun 15;7:16. PMID:17573961

26. Shamseer L, Moher D, Clarke M, Ghersi D, Liberati A, Petticrew M, Shekelle P, Stewart LA, PRISMA-P Group. Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015: elaboration and explanation. BMJ 2015 Jan 2;350:g7647. PMID:25555855

27. Higgins JPT. Cochrane Handbook for Systematic Reviews of Interventions. 2019. ISBN:9781119536628

28. Higgins JPT, Altman DG, Gotzsche PC, Juni P, Moher D, Oxman AD, Savovic J, Schulz KF, Weeks L, Sterne JAC, Cochrane Bias Methods Group, Cochrane Statistical Methods Group. The Cochrane Collaboration's tool for assessing risk of bias in randomised trials [Internet]. BMJ. 2011. p. d5928–d5928. [doi: 10.1136/bmj.d5928]

29. CASP Checklists [Internet]. CASP (Critical Appraisal Skills Programme). [cited 2020 May 8]. Available from: https://casp-uk.net/casp-tools-checklists/

30. Downes MJ, Brennan ML, Williams HC, Dean RS. Development of a critical appraisal tool to assess the quality of cross-sectional studies (AXIS). BMJ Open 2016 Dec 8;6(12):e011458. PMID:27932337

31. Christopoulou SC, Kotsilieris T, Anagnostopoulos I. Assessment of Health Information Technology Interventions in Evidence-Based Medicine: A Systematic Review by Adopting a Methodological Evaluation Framework. Healthcare (Basel) [Internet] 2018 Aug 31;6(3). PMID:30200307

32. Cameron G, Cameron D, Megaw G, Bond R, Mulvenna M, O'Neill S, Armour C, McTear M. Assessing the Usability of a Chatbot for Mental Health Care. In: Bodrunova S. et al. Internet Science., editor. Lecture Notes in Computer Science, vol 11551 Springer, Cham; 2019.

33. Elmasri D, Maeder A. A Conversational Agent for an Online Mental Health Intervention [Internet]. Brain Informatics and Health. 2016. p. 243–251. [doi: 10.1007/978-3-319-47103-7_24]

34. 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 Dec 13;5(4):e64. PMID:30545815

35. Hudlicka E. Virtual training and coaching of health behavior: example from mindfulness meditation training. Patient Educ Couns 2013 Aug;92(2):160–166. PMID:23809167

36. Inkster B, Sarda S, Subramanian V. An Empathy-Driven, Conversational Artificial Intelligence Agent (Wysa) for Digital Mental Well-Being: Real-World Data Evaluation Mixed-Methods Study. JMIR Mhealth Uhealth 2018 Nov 23;6(11):e12106. PMID:30470676

37. 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 Dec;10:39–46. PMID:30135751

38. Philip P, Micoulaud-Franchi J-A, Sagaspe P, De Sevin E, Olive J, Bioulac S, Sauteraud A. Virtual human as a new diagnostic tool, a proof of concept study in the field of major depressive disorders [Internet]. Scientific Reports. 2017. [doi: 10.1038/srep42656]

39. Yasavur U, Lisetti C, Rishe N. Let's talk! speaking virtual counselor offers you a brief intervention. Journal on Multimodal User Interfaces 2014;8:381–398.

40. Xu R, Mei G, Zhang G, Gao P, Judkins T, Cannizzaro M, Li J. A voice-based automated system for PTSD screening and monitoring. Stud Health Technol Inform 2012;173:552–558. PMID:22357057

41. Washburn M, Bordnick P, Rizzo AS. A pilot feasibility study of virtual patient simulation to enhance social work students' brief mental health assessment skills. Soc Work Health Care 2016 Oct;55(9):675–693. PMID:27552646

42. Dimeff LA, Jobes DA, Chalker SA, Piehl BM, Duvivier LL, Lok BC, Zalake MS, Chung J, Koerner K. A novel engagement of suicidality in the emergency department: Virtual Collaborative Assessment and Management of Suicidality. Gen Hosp Psychiatry 2018;63:119–126. PMID:29934033

43. Spänig S, Emberger-Klein A, Sowa J-P, Canbay A, Menrad K, Heider D. The virtual doctor: An interactive clinical-decision-support system based on deep learning for non-invasive prediction of diabetes. Artif Intell Med 2019 Sep;100:101706. PMID:31607340

44. Ghosh S, Bhatia S, Bhatia A. Quro: Facilitating User Symptom Check Using a Personalised Chatbot-Oriented Dialogue System. Stud Health Technol Inform 2018;252:51–56. PMID:30040682

45. Chaix B, Bibault J-E, Pienkowski A, Delamon G, Guillemassé A, Nectoux P, Brouard B. When Chatbots Meet Patients: One-Year Prospective Study of Conversations Between Patients With Breast Cancer and a Chatbot. JMIR Cancer 2019;5(1):e12856.

46. Bibault J-E, Chaix B, Guillemassé A, Cousin S, Escande A, Perrin M, Pienkowski A, Delamon G, Nectoux P, Brouard B. A Chatbot Versus Physicians to Provide Information for Patients With Breast Cancer: Blind, Randomized Controlled Noninferiority Trial. J Med Internet Res 2019 Nov 27;21(11):e15787. PMID:31774408

47. Heyworth L, Kleinman K, Oddleifson S, Bernstein L, Frampton J, Lehrer M, Salvato K, Weiss TW, Simon SR, Connelly M. Comparison of interactive voice response, patient mailing, and mailed registry to encourage screening for osteoporosis: a randomized controlled trial. Osteoporos Int 2014 May;25(5):1519–1526. PMID:24566584

48. 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 Jan 7;8:63–72. PMID:24470755

49. Simon SR, Zhang F, Soumerai SB, Ensroth A, Bernstein L, Fletcher RH, Ross-Degnan D. Failure of automated telephone outreach with speech recognition to improve colorectal cancer screening: a randomized controlled trial. Arch Intern Med 2010 Feb 8;170(3):264–270. PMID:20142572

50. Borja-Hart NL, Spivey CA, George CM. Use of virtual patient software to assess student confidence and ability in communication skills and virtual patient impression: A mixed-methods approach. Curr Pharm Teach Learn 2019 Jul;11(7):710–718. PMID:31227094

51. Philip P, Bioulac S, Sauteraud A, Chaufton C, Olive J. Could a Virtual Human Be Used to Explore Excessive Daytime Sleepiness in Patients? [Internet]. Presence: Teleoperators and Virtual Environments. 2014. p. 369–376. [doi: 10.1162/pres_a_00197]

52. Galescu L, Allen J, Ferguson G, Quinn J, Swift M. Speech recognition in a dialog system for patient health monitoring [Internet]. 2009 IEEE International Conference on Bioinformatics and Biomedicine Workshop. 2009. [doi: 10.1109/bibmw.2009.5332111]

53. 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 [Internet]. Journal of Medical Internet Research. 2014. p. e48. [doi: 10.2196/jmir.2974]

54. Crutzen R, Peters G-JY, 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 May;48(5):514–519. PMID:21501812

55. Wong W, Thangarajah J, Padgham L. Contextual question answering for the health domain [Internet]. Journal of the American Society for Information Science and Technology. 2012. p. 2313–2327. [doi: 10.1002/asi.22733]

56. Ireland D, Atay C, Liddle J, Bradford D, Lee H, Rushin O, Mullins T, Angus D, Wiles J, McBride S, Vogel A. Hello Harlie: Enabling Speech Monitoring Through Chat-Bot Conversations. Stud Health Technol Inform 2016;227:55–60. PMID:27440289

57. The Cochrane Collaboration. Copenhagen: The Nordic Cochrane Centre. Review Manager (RevMan) [Internet]. 2014. Available from: https://community.cochrane.org/help/tools-and-software/revman-5

58. Greenhalgh T, Wherton J, Papoutsi C, Lynch J, Hughes G, A'Court C, Hinder S, Fahy N, Procter R, Shaw S. Beyond Adoption: A New Framework for Theorizing and Evaluating Nonadoption, Abandonment, and Challenges to the Scale-Up, Spread, and Sustainability of Health and Care Technologies. J Med Internet Res 2017 Nov 1;19(11):e367. PMID:29092808

59. Michie S, van Stralen MM, West R. The behaviour change wheel: a new method for characterising and designing behaviour change interventions. Implement Sci 2011 Apr 23;6:42. PMID:21513547

60. Meinert E, Alturkistani A, Brindley D, Knight P, Wells G, Pennington N de. The technological imperative for value-based health care. Br J Hosp Med 2018 Jun 2;79(6):328–332. PMID:29894248

Appendices

Multimedia Appendix A. Search queries and number of results for each database

Database	Search terms	Retrieved
PubMed	((((Speech recognition software[mh] or "Conversational agent*"[tiab] or "embodied conversational agent*"[tiab] or chatbot*[tiab] or avatar*[tiab] or "dialog* system"[tiab] or "speech recognition software"[tiab] or "voice recognition software"[tiab] or "virtual assistan*"[tiab] or "virtual nurs*"[tiab] or "virtual patient"[tiab] or "virtual coach*"[tiab] or "virtual agent"[tiab] or "relation* agent"[tiab] or "assistance technol*"[tiab] or "intelligent assistan*"[tiab] or "digital assistan*"[tiab] or "natural language interface"[tiab] or "interactive computer agent"[tiab] or "computer-assisted instruction"[tiab] or "natural language communication"[tiab] or "natural language understanding"[tiab] or "unconstrained natural language processing"[tiab])) AND (Health facilities[mh] or Health communication[mh] or Health services[mh] or exp health services accessibility[mh] or Delivery of healthcare[mh] or exp Health behavior[mh] or Exercise[mh] or Simulation training[mh] or Health education[mh] or health literacy[mh] or "patient acceptance of healthcare"[mh] or health knowledge, attitudes, practice[mh] or "treatment adherence or compliance"[mh] or asthma[mh] or sex education[mh] or exp aged[mh] or exp counseling[mh] or smoking cessation[mh] or exp diet[mh] or exp education, medical[mh] or exp substance-related disorders[mh] or social skills[mh] or autism spectrum disorder[mh] or patient education as topic[mh] or diabetes mellitus[mh] or cardiovascular disease[mh] or pulmonary disease, chronic obstructive[mh] or "healthcare delivery"[tiab] or "healthcare access"[tiab] or health[tiab] or healthcare[tiab] or behavio?r[tiab] or exerci*[tiab] or diet[tiab] or "simulation training"[tiab] or education[tiab] or "elderly care"[tiab] or "sex* education"[tiab] or "health literacy"[tiab] or "counsel?ing"[tiab] or "well-being"[tiab] or "smoking cessation"[tiab] or "cognitive dysfunction"[tiab] or "mental health"[tiab] or "social skills"[tiab] or "autism spectrum disorder"[tiab] or diabetes[tiab] or "heart health"[tiab] or "chronic obstructive pulmonary disease"[tiab] or "COPD"[tiab] or "sun protection"[tiab] or "physical activity"[tiab]))) AND ("Outcome Assessment (Health Care)"[mh] or program evaluation[mh] or feasibility studies[mh] or pilot projects[mh] or "diffusion of innovation"[mh] or cost-benefit analysis[mh] or "Reproducibility of results"[mh] or Feasib*[tiab] or usab*[tiab] or evaluat*[tiab] or outcome*[tiab] or acceptability[tiab] or acceptance[tiab] or "treatment adherence"[tiab] or effectiv*[tiab] or adoption[tiab] or assess*[tiab] or "user experience*"[tiab] or efficacy[tiab] or utility[tiab] or utili?ation[tiab] or "patient* acceptance"[tiab] or "patient* acceptability"[tiab] or "user* acceptance"[tiab] or "user* acceptability"[tiab] or "user* perce*"[tiab] or "patient* perce*"[tiab] or "user* perspective*"[tiab] or "patient* perspective*"[tiab] or "user* view*"[tiab] or "patient* view*"[tiab] or cost*[tiab])	1065

Medline (OVID)	(Speech recognition software/ or ((Conversational adj1 agent*) or (embodied adj2 agent*) or chatbot* or avatar* or (dialog* adj1 system) or speech recognition software or voice recognition software or (virtual adj1 (assistan* or nurs* or patient or coach* or agent)) or (relation* adj1 agent) or assistance technol* or (intelligent adj2 assistan*) or (digital adj2 assistan*) or natural language interface or interactive computer agent or computer-assisted instruction or natural language communication or natural language understanding or unconstrained natural language processing).ti,ab.) AND (Health facilities/ or Health communication/ or Health services/ or exp health services accessibility/ or Delivery of healthcare/ or exp Health behavior/ or Exercise/ or Simulation training/ or Health education/ or health literacy/ or "patient acceptance of healthcare"/ or health knowledge, attitudes, practice/ or "treatment adherence or compliance"/ or asthma/ or sex education/ or exp aged/ or exp counseling/ or smoking cessation/ or exp diet/ or exp education, medical/ or exp substance-related disorders/ or social skills/ or autism spectrum disorder/ or patient education as topic/ or diabetes mellitus/ or cardiovascular disease/ or pulmonary disease, chronic obstructive/ or (healthcare delivery or healthcare access or health or healthcare or behavio?r or exerci* or diet or simulation training or education or elderly care or sex* education or health literacy or counsel?ing or well-being or smoking cessation or cognitive dysfunction or mental health or social skills or autism spectrum disorder or diabetes or heart health or chronic obstructive pulmonary disease or COPD or sun protection or physical activity).ti,ab.) AND ("Outcome Assessment (Health Care)"/ or program evaluation/ or feasibility studies/ or pilot projects/ or "diffusion of innovation"/ or cost-benefit analysis/ or "Reproducibility of results"/ or (Feasib* or usab* or evaluat* or outcome* or acceptability or acceptance or treatment adherence or effectiv* or adoption or assess* or user experience* or efficacy or utility or utili?ation or patient* acceptance or patient* acceptability or user* acceptance or user* acceptability or user* perce* or patient* perce* or user* perspective* or patient* perspective* or user* view* or patient* view* or cost*).ti,ab.)	1599

Embase (OVID)	(Automatic speech recognition/ or ((Conversational adj1 agent*) or (embodied adj2 agent*) or chatbot* or avatar* or (dialog* adj1 system) or (dialog* adj1 agent) or speech recognition software or voice recognition software or (virtual adj1 (assistan* or nurs* or patient or coach* or agent)) or (relation* adj1 agent) or assistance technol* or (intelligent adj2 assistan*) or (digital adj2 assistan*) or natural language interaction or interactive computer agent or computer-assisted instruction or natural language communication or natural language understanding or unconstrained natural language processing).ti,ab.) AND (Health care facility/ or medical information/ or Health service/ or exp healthcare access/ or healthcare delivery/ or exp Health behavior/ or Exercise/ or Simulation training/ or Health education/ or health literacy/ or patient attitude/ or attitude to health/ or patient compliance/ or asthma/ or sexual education/ or exp aged/ or exp counseling/ or smoking cessation/ or exp diet/ or exp medical education/ or exp drug dependence/ or social competence/ or autism/ or patient education/ or diabetes mellitus/ or cardiovascular disease/ or chronic obstructive lung disease/ or (healthcare delivery or healthcare access or health or healthcare or behavio?r or exerci* or diet or simulation training or education or elderly care or sex* education or health literacy or counsel?ing or well-being or smoking cessation or cognitive dysfunction or mental health or social skills or autism spectrum disorder or diabetes or heart health or chronic obstructive pulmonary disease or COPD or sun protection or physical activity).ti,ab.) AND (Outcome assessment/ or program evaluation/ or feasibility study/ or pilot study/ or mass communication/ or cost benefit analysis/ or reproducibility/ or (feasib* or usab* or evaluat* or outcome* or acceptability or acceptance or treatment adherence or effectiv* or adoption or assess* or user experience* or efficacy or utility or utili?ation or patient* acceptance or patient* acceptability or user* acceptance or user* acceptability or user* perce* or patient* perce* or user* perspective* or patient* perspective* or user* view* or patient* view* or cost*).ti,ab.)	2145

CINAHL	((MH Voice recognition systems) OR TI ((Conversational n1 agent*) or (embodied n2 agent*) or chatbot* or avatar* or (dialog* n1 system) or speech recognition software or voice recognition software or (virtual n1 (assistan* or nurs* or patient or coach* or agent)) or (relation* n1 agent) or assistance technol* or (intelligent n2 assistan*) or (digital n2 assistan*) or natural language interface or interactive computer agent or computer-assisted instruction or natural language communication or natural language understanding or unconstrained natural language processing) OR AB ((Conversational n1 agent*) or (embodied n2 agent*) or chatbot* or avatar* or (dialog* n1 system) or speech recognition software or voice recognition software or (virtual n1 (assistan* or nurs* or patient or coach* or agent)) or (relation* n1 agent) or assistance technol* or (intelligent n2 assistan*) or (digital n2 assistan*) or natural language interface or interactive computer agent or computer-assisted instruction or natural language communication or natural language understanding or unconstrained natural language processing)) AND ((MH “Health facilities”) or (MH “Communication”) or (MH “health services accessibility+”) or (MH “Health behavior”) or (MH “exercise”) or (MH “Computerized clinical simulation testing) or (MH “health education”) or (MH “health literacy”) or (MH “Patient attitudes”) or (MH “Attitude to health”) or (MH “patient compliance”) or (MH “asthma”) or (MH “sex education”) or (MH “Aged+”) or (MH “Counseling+”) or (MH “smoking cessation”) or (MH “diet+”) or (MH “Education, medical+”) or (MH “substance dependence+) or (MH “social skills training”) or (MH “Autistic disorder”) or (MH “patient education”) or (MH "diabetes mellitus") or (MH "cardiovascular diseases") or (MH "Pulmonary Disease, Chronic Obstructive") OR TI (healthcare delivery or healthcare access or health or healthcare or behavio?r or exerci* or diet or simulation training or education or elderly care or sex* education or health literacy or counsel?ing or well-being or smoking cessation or cognitive dysfunction or mental health or social skills or autism spectrum disorder or diabetes or heart health or chronic obstructive pulmonary disease or COPD or sun protection or physical activity) OR AB (healthcare delivery or healthcare access or health or healthcare or behavio?r or exerci* or diet or simulation training or education or elderly care or sex* education or health literacy or counsel?ing or well-being or smoking cessation or cognitive dysfunction or mental health or social skills or autism spectrum disorder or diabetes or heart health or chronic obstructive pulmonary disease or COPD or sun protection or physical activity)) AND ((MH “Outcome assessment”) or (MH “Program evaluation”) or (MH “pilot studies”) or (MH “Diffusion of innovation”) or (MH “Cost benefit anaylsis”) or (MH “reproducibility of results”) OR TI (feasib* or usab* or evaluat* or outcome* or acceptability or acceptance or treatment adherence or effectiv* or adoption or assess* or user experience* or efficacy or utility or utili?ation or patient* acceptance or patient* acceptability or user* acceptance or user* acceptability or user* perce* or patient* perce* or user* perspective* or patient* perspective* or user* view* or patient* view* or cost*) OR AB (feasib* or usab* or evaluat* or outcome* or acceptability or acceptance or treatment adherence or effectiv* or adoption or assess* or user experience* or efficacy or utility or utili?ation or patient* acceptance or patient* acceptability or user* acceptance or user* acceptability or user* perce* or patient* perce* or user* perspective* or patient* perspective* or user* view* or patient* view* or cost*))	935

Web of Science	((MH Voice recognition systems) OR TI ((Conversational n1 agent*) or (embodied n2 agent*) or chatbot* or avatar* or (dialog* n1 system) or speech recognition software or voice recognition software or (virtual n1 (assistan* or nurs* or patient or coach* or agent)) or (relation* n1 agent) or assistance technol* or (intelligent n2 assistan*) or (digital n2 assistan*) or natural language interface or interactive computer agent or computer-assisted instruction or natural language communication or natural language understanding or unconstrained natural language processing) OR AB ((Conversational n1 agent*) or (embodied n2 agent*) or chatbot* or avatar* or (dialog* n1 system) or speech recognition software or voice recognition software or (virtual n1 (assistan* or nurs* or patient or coach* or agent)) or (relation* n1 agent) or assistance technol* or (intelligent n2 assistan*) or (digital n2 assistan*) or natural language interface or interactive computer agent or computer-assisted instruction or natural language communication or natural language understanding or unconstrained natural language processing)) AND ((MH “Health facilities”) or (MH “Communication”) or (MH “health services accessibility+”) or (MH “Health behavior”) or (MH “exercise”) or (MH “Computerized clinical simulation testing) or (MH “health education”) or (MH “health literacy”) or (MH “Patient attitudes”) or (MH “Attitude to health”) or (MH “patient compliance”) or (MH “asthma”) or (MH “sex education”) or (MH “Aged+”) or (MH “Counseling+”) or (MH “smoking cessation”) or (MH “diet+”) or (MH “Education, medical+”) or (MH “substance dependence+) or (MH “social skills training”) or (MH “Autistic disorder”) or (MH “patient education”) or (MH "diabetes mellitus") or (MH "cardiovascular diseases") or (MH "Pulmonary Disease, Chronic Obstructive") OR TI (healthcare delivery or healthcare access or health or healthcare or behavio?r or exerci* or diet or simulation training or education or elderly care or sex* education or health literacy or counsel?ing or well-being or smoking cessation or cognitive dysfunction or mental health or social skills or autism spectrum disorder or diabetes or heart health or chronic obstructive pulmonary disease or COPD or sun protection or physical activity) OR AB (healthcare delivery or healthcare access or health or healthcare or behavio?r or exerci* or diet or simulation training or education or elderly care or sex* education or health literacy or counsel?ing or well-being or smoking cessation or cognitive dysfunction or mental health or social skills or autism spectrum disorder or diabetes or heart health or chronic obstructive pulmonary disease or COPD or sun protection or physical activity)) AND ((MH “Outcome assessment”) or (MH “Program evaluation”) or (MH “pilot studies”) or (MH “Diffusion of innovation”) or (MH “Cost benefit anaylsis”) or (MH “reproducibility of results”) OR TI (feasib* or usab* or evaluat* or outcome* or acceptability or acceptance or treatment adherence or effectiv* or adoption or assess* or user experience* or efficacy or utility or utili?ation or patient* acceptance or patient* acceptability or user* acceptance or user* acceptability or user* perce* or patient* perce* or user* perspective* or patient* perspective* or user* view* or patient* view* or cost*) OR AB (feasib* or usab* or evaluat* or outcome* or acceptability or acceptance or treatment adherence or effectiv* or adoption or assess* or user experience* or efficacy or utility or utili?ation or patient* acceptance or patient* acceptability or user* acceptance or user* acceptability or user* perce* or patient* perce* or user* perspective* or patient* perspective* or user* view* or patient* view* or cost*))	2954

ACM digital library	(recordAbstract:(+Speech +recognition +software) OR recordAbstract:(+conversational +agent) OR recordAbstract:(+embodied +agent) OR recordAbstract:(chatbot*) OR recordAbstract:(avatar*) OR recordAbstract:(+dialog* +system) OR recordAbstract:(+voice +recognition +software) OR recordAbstract:(+virtual +assistan*) OR recordAbstract:(+virtual +nurs*) OR recordAbstract:(+virtual +patient) OR recordAbstract:(+virtual +coach*) OR recordAbstract:(+virtual +agent) OR recordAbstract:(+relation* +agent) OR recordAbstract:(+assistance +technol*) OR recordAbstract:(+intelligent +assistan*) OR recordAbstract:(+digital +assistan*) OR recordAbstract:(+natural +language +interface) OR recordAbstract:(+interactive +computer +agent) OR recordAbstract:(+computer +assisted +instruction) OR recordAbstract:(+natural +language +communication) OR recordAbstract:(+natural +language +understanding) OR recordAbstract:(+unconstrained +natural +language +processing)) AND ((recordAbstract:(+health +facilities) OR recordAbstract:(+health +communication) OR recordAbstract:(+health +services) OR recordAbstract:(+health +access) OR recordAbstract:(+healthcare +delivery) OR recordAbstract:(+health +behavior) OR recordAbstract:(+health +behaviour) OR recordAbstract:(+exerci*) OR recordAbstract:(+simulation +training) OR recordAbstract:(+health +education) OR recordAbstract:(+health +literacy) OR recordAbstract:(+health +knowledge) OR recordAbstract:(+health +practice) OR recordAbstract:(+health +attitudes) OR recordAbstract:(+treatment +compliance) OR recordAbstract:(+treatment +adherence) OR recordAbstract:(+asthma) OR recordAbstract:(+sex* +education) OR recordAbstract:(+elderly +care) OR recordAbstract:(+counseling) OR recordAbstract:(+counselling) OR recordAbstract:(+smoking +cessation) OR recordAbstract:(+diet) OR recordAbstract:(+medical +education) OR recordAbstract:(+substance-related +disorders) OR recordAbstract:(+social +skills) OR recordAbstract:(+autism) OR recordAbstract:(+patient +education) OR recordAbstract:(+health) OR recordAbstract:(+healthcare) OR recordAbstract:(+education) OR recordAbstract:(+health +literacy) OR recordAbstract:(+wellbeing) OR recordAbstract:(+cognitive +dysfunction) OR recordAbstract:(+mental +health) OR recordAbstract: (diabetes) OR recordAbstract: (+cardiovascular +disease) OR recordAbstract: (+sun +protection) OR recordAbstract: (+chronic +obstructive +pulmonary +disease) OR recordAbstract: (COPD) OR recordAbstract: (+physical +activity)) AND ((recordAbstract:(+outcome*) OR recordAbstract:(+program +evaluation) OR recordAbstract:(+feasibility +stud*) OR recordAbstract:(+pilot +stud*) OR recordAbstract:(+cost*) OR recordAbstract:(+reproducibility) OR recordAbstract:(+feasib*) OR recordAbstract:(+usab*) OR recordAbstract:(+evaluat*) OR recordAbstract:(+effectiv*) OR recordAbstract:(+adoption) OR recordAbstract:(+assess*) OR recordAbstract:(+user +experience*) OR recordAbstract:(+efficacy) OR recordAbstract:(+utility) OR recordAbstract:(+utilisation) OR recordAbstract:(+utilization) OR recordAbstract:(+patient* +accept*) OR recordAbstract:(+user* +accept*) OR recordAbstract:(+user* +perce*) OR recordAbstract:(+patient* +perce*) OR recordAbstract:(+user* +perspective) OR recordAbstract:(+patient* +perspective*) OR recordAbstract:(+user* +view*) OR recordAbstract:(+patient* +view*))	743

Multimedia Appendix B

The retrieval was conducted in a series of searches due to limitation in the number of search criteria that could be specified. These were conducted in line with the original search categories. The exclusion searches were derived from characteristics of irrelevant studies identified. Pass refers to a search following which the subsequent search was conducted on the subset of studies retrieved in the previous pass.

First pass: Any field = Conversation* OR chat* OR virtual OR interactive OR relational OR speech OR voice OR natural language
Second pass: Any field = health*
Third pass: Any field = Outcome OR evaluat* OR effect* OR efficacy OR feasib* OR usab* OR accepta* OR perce*
Fourth pass: Title = NOT (review OR protocol OR guidelines)
Fifth pass: Any field = NOT (surgery OR surgical OR ecol* OR animal OR industr* OR transcription OR imaging OR librar* OR social media)
Sixth pass: Year = greater than or equal to 2008

Multimedia Appendix C. Summary of study characteristics

Authors (year)	Study design	Country of study	Study population	N (study arms)	Conversational agent
Adams et al (2014) [9]	RCT	USA	Children aged 4 months to 11 years who had an RHCM or well-child visit	475 (Personal Health Partner: n=293; single automated call: n=182)	Personal Health Partner
Bibault et al (2019) [46]	RCT	France	Patients with breast cancer and their relatives	142 (Vik: n=71; physician: n=71)	Vik
Borja-Harta et al (2019) [50]	Pre-post	USA	Pharmacy students	203	Shadow Health
Cameron et al (2019) [32]	Cross-sectional	UK	Employees from a mental health enterprise	7	iHelpr
Chaix B et al (2019) [45]	Cross-sectional	France	Patients with breast cancer and their relatives	958	Vik
Chang et al (2008) [8]	Cross-sectional	Taiwan	Emergency department patients	30	Speech and touch based intelligent comprehensive triage support system
Crutzen et al (2011) [54]	Cross-sectional	The Netherlands	Adolescents	929	Bzz
Dimeff et al (2018) [42]	Cross-sectional	USA	Emergency department patients admitted due to acute suicidal crisis	24	Dr Dave
Elmasri & Maeder (2016) [33]	Cross-sectional	Australia	Young adults (18-25) at low to medium risk of alcoholism	17	Chatbot to address substance abuse
Fitzpatrick et al (2017) [13]	RCT	USA	University students who self-identified with depression and anxiety	70 (Woebot: n = 34, e-book control: n = 36)	Woebot
Friederichs et al (2014) [53]	RCT	The Netherlands	Adults (18-70)	958; 500 at follow up 2 (avatar: n=162; text: n=146; control: n=192)	AVATAR
Fulmer et al (2018) [34]	RCT	USA	Students	75 (intervention: n=50 [group 1: n=24; group 2: n=26]; control: n=25)	Tess

Galescu et al (2009) [52]	Cross-sectional	USA	Patients with chronic heart failure	14	Computer assistant for robust dialogue interaction and care (CARDIAC)
Ghosh et al (2018) [44]	Qualitative	Australia	Clinical scenarios	30	Quro
Havik et al (2019) [14]	Cohort	Norway	General population	11	ROB
Heyworth et al (2014) [47]	RCT	United States	Women 50-64 with risk factors for osteoporosis	4685 (IVR call: n = 1565, usual care: n = 1558, usual care + mailed info: n = 1562)	Interactive Voice Response (IVR) phone call
Hudlicka (2013) [35]	Cohort	Not reported	Students	32 (coach vs. written and audio materials, n not stated)	“Chris” (virtual mindfulness coach)
Inkster et al (2018) [36]	Cohort	Global	General population	129	Wysa app
Ireland et al (2016) [56]	Qualitative	Australia	General population	33	Harlie
Isaza-Restrepo et al (2018) [15]	Pre-post	Columbia	Undergraduate medical students	20	Virtual Patient
Ly et al (2017) [37]	RCT	Sweden	Non-clinical population	28 (chatbot: n = 14, wait-list control: n = 14)	Shim
Nakagawa et al (2018) [12]	Qualitative	Japan	General population	14	Telecare system that estimates QoL
Philip et al (2014) [51]	Cohort	France	Sleep clinic patients and health controls	62 (patients: n = 32, controls: n = 30)	Virtual physician (Embodied Conversational Agent)
Philip et al (2017) [38]	Cluster crossover	France	Sleep clinic patients aged 18-65	179	ECA for diagnosing MDD
Rhee et al (2014) [48]	Qualitative	United States	Adolescent-parent dyads	15 dyads	Mobile phone-based asthma self-management aid for adolescents (mASMAA)
Simon et al (2010) [49]	RCT	United States	Men and women aged 50 to 64	20,938 (ATO-SR: n = 10,432, usual care: n = 10,506)	Automated telephone outreach with speech recognition (ATO-SR)
Spänig et al (2019) [43]	Prospective, cross-sectional	Germany	University students	320	Virtual doctor
Washburn et al (2016) [41]	Qualitative	United States	Medical students	5	Virtual patient software

Wong et al (2012) [55]	Qualitative	Australia	N/A	N/A	enquireMe
Xu et al (2012) [40]	Qualitative	None	Voice clips of US soldiers	10	Tele-PTSD Monitor
Yasavur et al (2014) [39]	Cross-sectional	United States	University students	89 (52 training system, 37 testing system)	Virtual alcohol counsellor (ECA)