# Improving the Usability of Randomized, Controlled Trials of Artificial Intelligence-based Chatbots in Healthcare: Results of a Systematic Literature Review

**János Tibor Czere**

Doctoral School of Innovation Management, Obuda University, Bécsi út 96/B, H-1034 Budapest, Hungary; PSI CRO Hungary LLC, Dózsa György út 84/B, H-1068 Budapest, Hungary, czere.janos@phd.uni-obuda.hu

**Áron Hölgyesi**

Health Economics Research Center, University Research and Innovation Center, Obuda University, Bécsi út 96/B, H-1034 Budapest, Hungary, holgyesi.aron@uni-obuda.hu

**Márta Péntek**

Health Economics Research Center, University Research and Innovation Center, Doctoral School of Innovation Management, Obuda University, Bécsi út 96/B, H-1034 Budapest, Hungary, pentek.marta@uni-obuda.hu

*Abstract: Artificial Intelligence (AI)-powered chatbots and virtual assistants (VAs) have gained importance in healthcare to support communication and decision-making. The value of **R**andomized **C**ontrolled **T**rials (RCTs) providing evidence on their effectiveness and safety depends on the quality of their design, conducting and reporting. The CONSORT-AI guideline has been developed to guide researchers in designing, conducting and reporting RCTs for AI interventions. The aim of this research is to identify and analyze RCTs on AI-based chatbots in healthcare, with a special focus on their compliance with the AI-extension part of the CONSORT-AI. A **S**ystematic **L**iterature **R**eview (SLR) was conducted and identified 28 relevant RCTs. Most AI-chatbot application studies were performed in mental health, public health and cancer care. An increase in the number of RCTs, by time, was observed, however, none of the RCTs fully adhered to the CONSORT-AI guideline. Weak reporting quality on input data, inclusion criteria, AI versioning, performance errors and human-AI interactions were common.*

*Our study highlights that, in parallel with the development of AI-chatbots, from simple statistical models, to advanced neural network architectures like transformers, more emphasis is needed on standardizing their clinical research and reporting to ensure more robust, transparent and ethical evidence for their application in healthcare.*

*Keywords: artificial intelligence; natural language processing; chatbot; randomized controlled trial; health; reporting guideline*

# 1 Introduction

Computer technology has realized rapid growth over the last three decades, with Artificial Intelligence (AI) emerging as a key area of innovation [1]. AI-based communication tools like chatbots and virtual assistants (VAs) are becoming increasingly important, evolving from simple statistical models to transformer-based models. Bidirectional Encoder Representations from Transformers (BERT), Generative Pre-trained Transformer (GPT), and T5 have revolutionized natural language processing (NLP), enabling tasks such as translation, sentiment analysis, and text generation. Despite these advances, statistical models like FastText still contribute to linguistic research and problem-solving [2-4]. Another task of the NLP is the language recognition in the online communication [5].

The use of chatbots has grown exponentially in healthcare, particularly during the COVID-19 pandemic. Organizations like WHO and CDC utilized chatbots to provide timely information, while tools like IBM Watson supported government and healthcare institutions [6] [7]. OpenAI's ChatGPT, introduced in 2022, showcases advanced transformer architecture, facilitating human-like conversations, detecting fake news, and toxicity in communication [8-11]. It has been proved to be valuable in enhancing patient communication and supporting clinical decision-making in the medical field [12] [13].

The chatbots' ability to understand and generate human-like responses relies on NLP [14]. However, the "black box" nature of AI models raises concerns about transparency, as the internal workings of these algorithms remain challenging to interpret [15].

Transparency is critical in the use of AI-based chatbots in medical decision-making and healthcare, where decisions must be grounded in scientific evidence. Randomized controlled trials (RCTs), regarded as the gold standard for clinical evidence, compare outcomes between randomized test and control groups. The test group receives the treatment under study, while the control group receives a standard or placebo treatment, allowing evaluation of treatment efficacy and safety through comparative analysis [16].

To minimize bias, RCT design and reporting should adhere to specific guidelines. Standardized reporting tools, such as narratives or checklists, enhance experiment reproducibility [17] [18]. By directing attention to critical factors and ensuring consistency, checklists have demonstrated effectiveness in various fields, notably reducing perioperative morbidity and mortality in surgery [19] [20]. The CONSORT statement is a reporting guideline for assessing RCTs, shown to improve trial outcomes when properly applied [21]. With the growth of AI applications in healthcare, the need for standardized reporting has emerged. While several checklists for AI-based medical solutions exist, their adoption remains limited [17] [22]. These include standalone AI guidelines (e.g., MI-CLAIM) and AI-specific extensions to existing clinical trial guidelines [19, 21, 22]. The CONSORT-AI extension, designed for AI interventions in clinical trials, promotes transparency, reduces bias and aids in evaluating AI's impact on patient outcomes [23].

Literature evidence regarding the effectiveness of chatbots in healthcare have been collected and analyzed by Milne-Ives et al. and Laranjo et al [24] [25]. Authors highlighted methodological weaknesses of the studies and thus their limited usability for health care. However, the reviews did not analyze specifically the AI aspects of the trials. Moreover, the field of AI-based chatbots for health purposes is rapidly evolving. Therefore, the aim of our study is to update and systematically review the literature for RCTs on AI-based chatbots in healthcare and to analyze the quality of the studies in terms AI reporting and compliance with the CONSORT-AI guideline [23].

# 2 Methods

## 2.1 Literature Search: Data Sources and Search Terms

We rely on previous SLRs by Milne-Ives et al. [24] and Laranjo et al. [25] that searched PubMed, Medline (Ovid), EMBASE (Excerpta Medica dataBASE), CINAHL (Cumulative Index to Nursing and Allied Health Literature), Web of Science, PsycInfo, and the Association for Computing Machinery Digital Library databases. Based on our prior experience and research, we hypothesized that the IEEE Xplore database would provide valuable, additional information in the topic. To this end, our search strategy was designed to query six key databases: PubMed, EMBASE, ACM Digital Library, Web of Science, Scopus, and IEEE Xplore. The SLRs by Laranjo et al. [25] and Milnes-Ives et al. [24] covered the literature up to November 29, 2019. In these two SLRs altogether eight RCTs on healthcare chatbots were identified all of which fulfilled our eligibility criteria (detailed below), hence these were included in our current review. We conducted a

complementary literature search for the period between Mar 01, 2018, and August 11, 2022.

Another branch of the literature search involved identifying SLRs and examining their citations. If an SLR met our eligibility criteria (NLP-based chatbot, English full text, within the specified time frame), the cited articles were further assessed according to our criteria.

Using the Cochrane Library's suggested search terms [26], which are widely recognized as a gold standard for evidence-based healthcare information, the search failed to identify an extensive list of synonyms for the term "chatbot". Therefore, an alternative strategy was built to ensure a thorough and comprehensive examination of relevant literature for our research question [27] (https://osf.io/4x57r). These terms were: "Avatar"; "Chat Bot"; "Chatbot"; "Conversational Agent"; "Conversational Interface"; "Dialog System"; "Digital Assistant"; "Digital Characters"; "Digital interlocutors"; "Embodied Agent"; "Embodied Conversational Agent"; "Intelligent Agent"; "Interactive Agent"; "Natural Language System"; "Relational Agent"; "Virtual Agent"; "Virtual Assistant"; "Virtual Coach" and "Virtual Human".

To identify RCTs, we applied the highly sensitive Cochrane search strategy for RCTs (2008 revision) in the PubMed format [26]. The exact search terms are provided in electronic supplementary material (https://osf.io/4x57r).

## 2.2 Selection of Studies: Selection Process, Inclusion and Exclusion Criteria

Study selection was performed in two phases. First, two reviewers (JTC, ÁH) independently screened the records for inclusion by title and abstract. Second, the selected publications were reviewed by the same two reviewers independently for inclusion by full text. In both phases disagreements between the two reviewers were discussed, remaining disagreements were solved by involving a third reviewer (MP). An Excel database was developed to record the decision process and final decisions.

All RCTs investigating the effectiveness of AI-based chatbots in healthcare were included regardless of the condition. Studies involving any human participants were included (e.g., patients, citizens attending healthcare, informal caregivers etc.) without any age limit. Studies involving AI-based chatbot that applies natural language processing as a communication method to be used for any health-related intervention (e.g., screening, treatment, patient management etc.) were included. Chatbots that were based on predefined response options (rule-based chatbots) were excluded. Any comparator or control was accepted, including the "to do nothing" option. No limits were set regarding the outcomes of the studies. Only RCTs were included in the analysis, observational studies and other forms of

experimental studies, such as non-randomized and quasi-experimental studies, were excluded. Only original studies published in English and available in full text were considered. Conference abstracts, proceedings, editorials, letters to the editor, and opinion papers were excluded. Systematic reviews were examined to identify relevant original studies. Any study context was accepted (e.g., hospital, outpatient).

## 2.3 Data Extraction and the CONSORT-AI Checklist

The CONSORT-AI is an extension of the widely used CONSORT statement, which provides guidelines for reporting RCTs [23]. It allows researchers to ensure that AI-based interventions in clinical trials are properly reported, enabling other researchers to replicate or build upon the results. The guideline emphasizes the need for clear and detailed descriptions of the AI intervention, its integration into the trial setting, and its performance during the study. The extension part of the CONSORT-AI checklist includes 14 items specifically designed to address the unique aspects of RCTs involving AI interventions. The 14 items are related to the Title abstract (2 items), Introduction (1 item), Methods (9 items), Results (1 item) and Funding (1 item) part of the study report [23]. (The items are detailed in the Results section below.)

First, an Excel spreadsheet database was developed, where the columns contained general article information, study data from the publication in terms of healthcare area, patient information, interventions, chatbot names, additional details, and the CONSORT-AI statements along with their explanatory fields. Each row represented one included publication. One researcher (JTC) performed the data extraction, one third of the studies was randomly checked by an independent reviewer (MP). Disagreements between individual judgements were resolved by discussions and, if necessary, a third researcher (ÁH) was involved.

## 2.4 Data Analysis

We analyzed the studies based on publication year, journal, clinical area of disease, characteristics of the patient samples, interventions and controls and the use of chatbot.

The evaluation focused on the quality and compliance of the articles with the AI part of the CONSORT-AI guideline. The 14 AI-extension items of the CONSORT-AI checklist [17] [23] were applied and compliance with each item was assessed in a binary (‘yes’ or ‘no’) format Additionally, an explanatory field was included to facilitate understanding of how each statement was applied within the articles.

# 3 Results

## 3.1 Search Results

The searches in ACM DL, Embase, PubMed, Scopus, Web of Science, and IEEE databases yielded a total of 1166 records after automated (Endnote function) deduplication (N=107). Number of excluded publications and the reasons for exclusion are summarized in a PRISMA flowchart (Figure 1) [28].

During the citation screening process, we identified 2205 records from the SLR citations. We excluded duplicate records, those outside the search period, studies that were not RCTs, and those not involving chatbots. As a result, a total of 2204 records were excluded. The last record was excluded under the full text review, because there was no evidence for RCT.

## 3.2 Main Characteristics of the Randomized Controlled Trials

Main characteristics of the publications and RCTs are summarized in Table 1. The earliest study dates back to 2010 while a continuous growing in the number of studies was observed by time, achieving seven published RCTs both in 2021 and 2022. The most RCTs (N=14) were published in Journal of Medical Internet Research (JMIR) journals (JMIR, JMIR mHealth and uHealth, JMIR Mental Health, JMIR Formative Research), the other papers were published in the Internet Interventions journal (N=2) and single papers in further different journals. Regarding the clinical area in which the chatbot was used, mental health was dominant (N=7), followed by public health (N=4) and cancer (N=3), but other patient groups were also involved in other studies. The characteristics of the patient (sample) groups were highly heterogeneous [29-56].

In most cases, adult patients constituted the target population; however, two studies focused on children, one involving child aged 3-6 years [56] and the other comprising children diagnosed with social anxiety disorder [31]. Three studies targeted university or college students who self-reported anxiety and depression [35, 37, 43]. The target populations of the remaining studies varied, with each study focusing on a distinct group. For example, one study included individuals engaged in hazardous drinking behaviors [34], while another concentrated-on patients with breast cancer [33].

The studies investigated the use of chatbots for various interventions across different health contexts (Table 1). The research also explored the role of chatbots in enhancing data collection, facilitating behavior change (e.g., smoking cessation or physical activity), and improving mental health outcomes, with a focus on assessing their effectiveness relative to established practices. Chatbots were often

compared to traditional methods such as usual care, human counseling, providing information only, putting patients on waitlists, predefined expert responses, or no intervention at all. In most cases, the chatbot used during the intervention was applied specifically for that particular study. Only two chatbots, Tess [37] [42] and Woebot [35] [51] (along with one of its derivatives), were utilized in two studies each. However, in nine articles, the virtual assistant was not explicitly named.

## 3.3 Compliance with AI-specific Items of the CONSORT-AI Checklist

We provide an item-by-item analysis of the 28 RCTs using the AI extension part of the CONSORT-AI guideline in Supplementary file (https://osf.io/y2t8r) and summarize the main findings in text. We present the analysis of the 10 best performing RCTs in Table 2.

None of the RCTs complied fully with the AI extension part of the CONSORT-AI checklist, highlighting that adherence to the reporting standards proposed by the checklist remains sub-optimal in the current literature. Among the 28 RCTs, the number of "Yes" responses, indicating compliance, ranged from 4 to 11, and compliance percentages ranged from 28.6% to 78.6%. Studies such as those utilizing Pegasys-VR [31] and Dejal@bot [49] showed higher adherence to reporting standards, with compliance rates of 71.4% and 78.6%, respectively. Conversely, several studies demonstrated limited adherence, achieving compliance rates below 30% [29, 34-36, 42, 47]. Key areas of weakness include inadequate descriptions of inclusion and exclusion criteria at the data level, insufficient reporting of AI versioning, and limited analyses of performance errors.

In the Title and Abstract section, nearly all articles (96.4%) successfully indicated that the intervention involved AI or machine learning and specified the type of model in the title or abstract. Only one article failed to meet this requirement [29]. Additionally, 92.9% of the articles stated the intended use of the AI intervention within the trial in a clear manner, with two studies lacking this essential information [29] [35]. Likewise, the Introduction section demonstrated strong compliance, with 96.4% of articles effectively explaining the intended use of the AI intervention, in the context of the clinical pathway, detailing the purpose and specifying the target audience. Only one article didn’t satisfy the requirement [29].

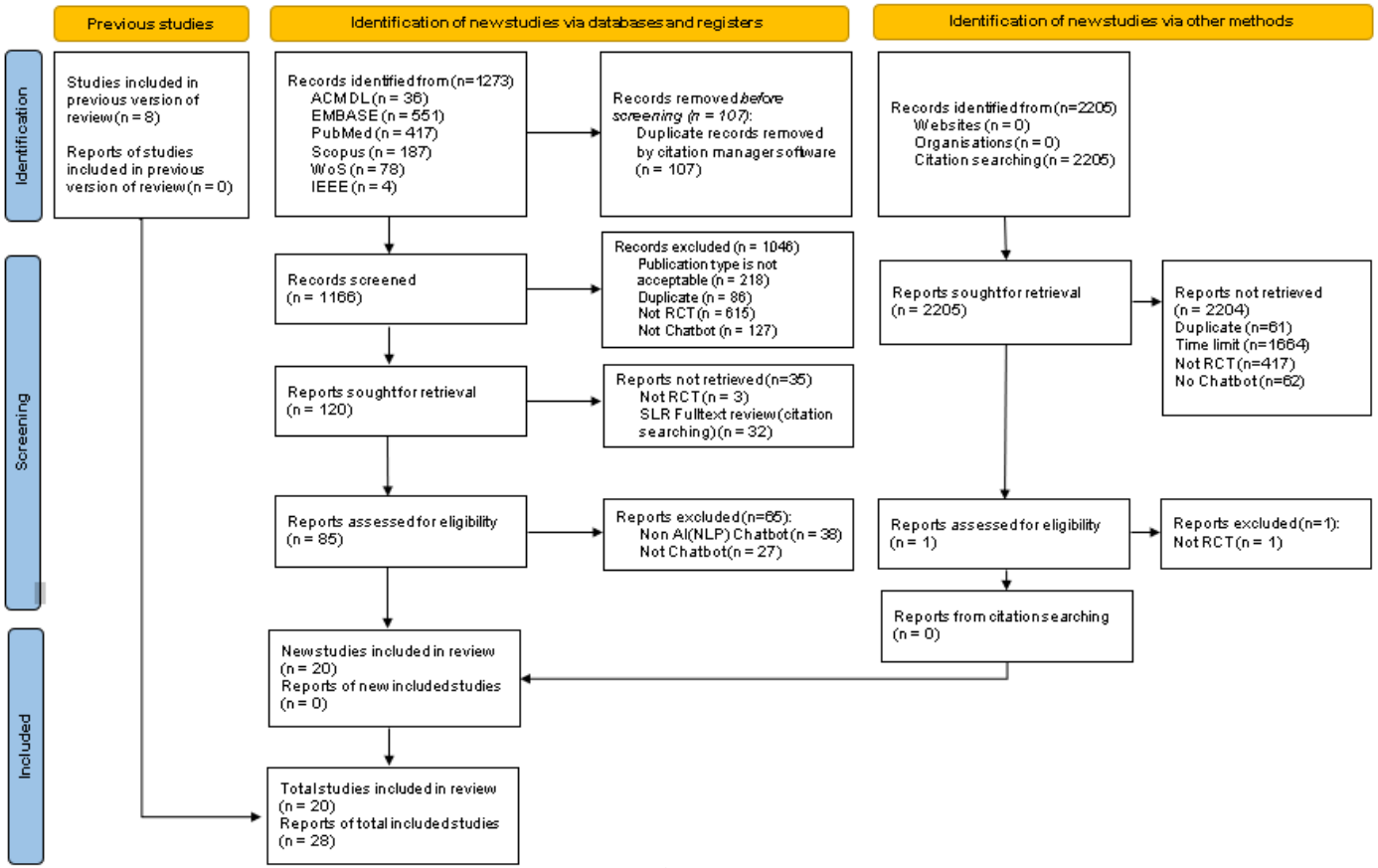


Figure 1
PRISMA Flowchart

In the Methods section, there was a notable discrepancy in reporting standards. The inclusion and exclusion criteria at the participant level were well articulated in 96.4% of the articles, with only one study failing to provide this information [53]. However, reporting on the inclusion and exclusion criteria for input data was significantly lacking, with only 3.6% of the articles addressing this aspect [40]. This substantial gap suggests a critical area for improvement, as input data criteria are essential for understanding the applicability and reliability of AI interventions. Furthermore, 71.4% of the studies described how the AI intervention was integrated into the trial setting, but eight articles did not provide sufficient detail on integration, highlighting a need for more comprehensive explanations.

The Interventions section revealed several areas requiring improvement. Only 21.4% of the studies specified the version of the AI algorithm used [43, 46, 49, 50, 53, 54], and fewer than one-third (28.6%) described how input data were acquired and selected [31, 38, 40, 46, 48, 49, 54, 56]. The handling of poor quality or unavailable data was inadequately reported, with just 17.7% of articles addressing this concern [31, 44, 48, 49, 55]. Moreover, only 10.7% of studies explained human-AI interactions and the level of expertise required from users, indicating a significant gap in detailing the operational aspects of the intervention [31, 49, 55].

Despite these shortcomings, 60.7% of the studies did specify the output of the AI intervention, yet only 28.6% articulated how these outputs contributed to decision-making in clinical practice [29, 31, 38, 46, 48, 51, 54, 55]. This highlights a need for clearer reporting on the practical implications and usability of AI outputs.

The Results section demonstrated a particularly concerning deficiency. Only 7.1% of articles discussed performance errors or how errors were identified, indicating a widespread lack of error analysis [53] [54]. This gap is crucial as understanding performance errors is essential for evaluating the reliability and safety of AI interventions. Lastly, the Other Information category, specifically concerning funding and accessibility, revealed that only 10.7% of articles mentioned whether and how the AI intervention or its code could be accessed [33, 39, 49]. This lack of transparency in reporting access and reuse restrictions, suggests a need for more consistent and detailed information to promote reproducibility and ethical sharing of AI resources.

Table 1

Summary of the randomized controlled trials on AI-based chatbots involved

| **First author (ref.), publication year** | **Journal** | **Country of the study** | **Clinical area / disease** | **Patients** | **Intervention (I), Comparator (C)** | **Chatbot** |
|---|---|---|---|---|---|---|
| Simon et al [52], 2010 | Arch Intern Med | USA | Colorectal Cancer | Members of Harvard Pilgrim Health Care | To participate in colorectal screening:<br>I - Single telephone outreach with speech recognition;<br>C - Usual care | Not defined |
| Adams et al [29], 2014 | Pediatrics | USA | Pediatric primary care | Parents of children aged 4 months - 11 years | To prepare next visit:<br>I - Phone call, tailored counseling IVR system;<br>C - Phone call, 18-question Framingham Safety Survey Phone call, tailored counseling IVR system | Personal Health Partner (PHP) |
| Friederichs et al [36], 2014 | J Med Internet Res | The Netherlands | Public health (physical activity) | Dutch adults | To motivate for physical activity:<br>I - Motivational interview with an avatar;<br>C1 - Intervention without an avatar;<br>C2 - No intervention | AVATAR |
| Heyworth et al [40], 2014 | Osteoporos Int | USA | Osteoporosis (OP) | Women at OP risk | To encourage participation in OP screening:<br>I - Usual care plus IVR;<br>C1 - Usual care;<br>C2 - Usual care plus mailed educational materials | Not defined |
| Fitzpatrick et al [35],2017 | JMIR Ment Health | USA | Mental health (Depression) | College students with self-reported anxiety and depression | To deliver CBT:<br>I - CBT with a conversation agent;<br>C - information only CBT with a conversation agent | WoeBot |
| Fulmer et al [37], | JMIR Ment Health | USA | Mental health (Depression) | College students with self-reported anxiety and | To reduce symptoms:<br>I - Chatbot 2 / 4 weeks; | Tess |

| | | | | | | |
|---|---|---|---|---|---|---|
| 2018 | | | | depression | C - Information only | |
| Bibault et al [33], 2019 | J Med Internet Res | France | Breast Cancer | Breast cancer patients | To inform patients about breast cancer:<br>I - Chatbot.<br>C - Predefined experts' responses Chatbot | Vik |
| Greer et al [39], 2019 | JMIR Mhealth Uhealth | USA | Cancer | Young adult patients after cancer treatment | To promote positive psychology and well-being:<br>I - Chatbot;<br>C - emotional ratings and chatbot only after | Vivibot |
| Ly et al [45], 2019 | Internet Interv | Sweden | Mental health (mental well-being) | Adults | To promote mental well-being:<br>I - Delivering positive psychology and CBT via an automated chatbot.<br>C - Put on waitlist. Delivering positive psychology and CBT via an automated chatbot | Shim |
| Tanana et al [54], 2019 | J Med Internet Res | USA | Therapists' training (psychotherapists) | Non-therapists | To train basic counseling skills:<br>I - Real-life feedback via a chatbot.<br>C - No feedback | Client-Bot |
| Bennion et al [32], 2020 | J Med Internet Res | United Kingdom | Elderly health | Members (aged 50+) of the University of the Third Age | To facilitate problem solving:<br>I - Chatbot type 1.<br>C - Chatbot type 2 | MYLO and ELIZA |
| Gong et al [38], 2020 | J Med Internet Res | Australia | Diabetes (Type 2) | Adults with Type 2 diabetes | To support self-management:<br>I - Chatbot;<br>C - Usual care | My Diabetes Coach (MDC) |
| Maeda et al [46], 2020 | Reprod Biomed Online | Japan | Fertility | Women 20-34 years old | To promote fertility awareness and preconception:<br>I - Fertility education chatbot.<br>C2 - A book about fertility and preconception health.<br>C3 - A document about an irrelevant topic | Not defined |
| Piao et al [50], 2020 | JMIR Mhealth Uhealth | South Korea | Public health (physical activity) | Office workers | Stair-climbing habit formation:<br>I - Chatbot;<br>C - Intervention started only on the fifth week | Not defined |
| Beidel et al [31], | Behavior Therapy | USA | Childhood social anxiety | Children with social anxiety disorder | Social effectiveness therapy:<br>I - Web-based Artificial Intelligence VR treatment. | Pegasys-VR |

| | | | | | | |
|---|---|---|---|---|---|---|
| 2021 | | | disorder | | C - Usual social effectiveness therapy | |
| Jang et al [41], 2021 | Int J Med Inform | South Korea | Attention deficit | Adults with attention deficit | To alleviate attention deficit symptoms:<br>I - Chatbot;<br>C - Information only (book) | Todaki |
| Klos et al [42], 2021 | JMIR Form Res | Argentina | Mental health (Depression, anxiety) | University students | To support mental health:<br>I - Chatbot;<br>C - Information only (book) | Tess |
| Loveys et al [44], 2021 | JMIR Ment Health | New Zealand | Mental health (menatl well-being) | Adults at greater risk for COVID-19 | Remote loneliness and stress intervention:<br>I - Cognitive behavioral and positive psychology exercises with chatbot;<br>C – Waitlist | Bella |
| Prochaska et al [51], 2021 | Drug Alcohol Depend | USA | Substance use | Adults with substance use disorder | To reduce substance misuse:<br>I - Chatbot;<br>C – Waitlist | Woebot-SUDs |
| Söderström et al [53], 2021 | Behav Res Methods | Australia | Online surveying | Adults | To improve quality of data online data collection:<br>I - Chatbot-guided survey;<br>C - Self-guided survey | Not defined |
| Tsai et al [55], 2021 | Psychol Mark | USA | Vaccination | College students with self-reported anxiety and depression | Health marketing communication about HPV:<br>I - Interact with a chatbot;<br>C - Interact with a human representative | Not defined |
| Beaman et al [30], 2022 | Journal of Medical Systems | USA | Mental health | Patients newly admitted to behavioral medicine clinic | Completing the Patient Health Module-9 (PHM-9) questionnaire:<br>I - Paper-based then Interactive Voice respond system-based format.<br>C - Interactive Voice respond system-based then Paper-based format | Amazon Alexa |
| Dulin et al [34], 2022 | JMIR Form Res | USA | Public health (drinking) | Hazardous drinking persons | To reduce alcohol consumption:<br>I - Step Away chatbot.<br>C1 - Step Away app;<br>C2 - Assessment-only | Step Away |

| | | | | | | |
|---|---|---|---|---|---|---|
| Liu et al [43], 2022 | Internet Interv | China | Mental health (Depression) | University students | Self-help intervention:<br>I - Chatbot;<br>C – Bibliotherapy | XiaoNan |
| Nißen et al [47], 2022 | J Med Internet Res | Germany | Chatbot development in healthcare | Participants of an online panel | To test the social role of a chatbot:<br>I - No choice on chatbot;<br>C - Choice on chatbot persona | Not defined |
| Ogawa et al [48], 2022 | Parkinsonism Relat Disord | Japan | Parkinson's disease | Patients with Parkinson's disease | To improve smile and symptoms of PD:<br>I - Chatbot and video conferencing sessions.<br>C - Video conferencing sessions | Not defined |
| Olano-Espinosa et al [49], 2022 | JMIR Mhealth Uhealth | Spain | Public health (smoking) | Patients visiting primary care | To quit smoking:<br>I - Chatbot;<br>C - Usual clinical practice | Dejal@bot |
| Xu et al [56], 2022 | Child Dev | USA | Development of children's skills | Children aged 3-6 years | To improve children's story comprehension and engagement:<br>I - Chatbot;<br>C - Adult reading partner | Not defined |

CBT=cognitive Behavior therapy, C= Comparator, I=Intervention group, IVR= Interactive Voice Response, OP= Osteoporosis

# 4 Discussion

We analyzed the available clinical evidence on AI-based chatbots in healthcare, focusing on RCTs and their compliance with the AI extension part of the CONSORT-AI guideline. Overall, reporting quality and adherence to the AI part of the CONSORT-AI was suboptimal. Common weaknesses included limited description of inclusion criteria, inadequate reporting of AI versioning, and sparse analyses of performance errors. Furthermore, only a few studies detailed human-AI interaction and algorithm outputs in decision-making. The Results section of the RCTs was particularly deficient, with minimal attention to error reporting, while information about funding and code accessibility was largely absent.

According to our best knowledge, to date, three SLRs have dealt with RCTs on AI-based chatbots. However, two of these (by Milne-Ives et al. [24] and Laranjo et al. [25]) did not analyze the quality of the AI part of the publications. In contrast, Martindale et al. analyzed the completeness of publications on RCTs for AI interventions (not restricted to any device or clinical area) that have been published since the publication of the CONSORT-AI guideline [57].

Table 2

Compliance of the 10 best performing AI-based chatbot RCTs with the AI-specific items of the CONSORT-AI checklist

| **Section** | **CONSORT-AI extension item** | **Olano-Espinosa et al., 2022 [49]** | **Tanana et al., 2019 [54]** | **Beidel et al., 2021 [31]** | **Ogawa et al., 2022 [48]** | **Maeda et al., 2020 [46]** | **Gong et al., 2020 [38]** | **Heyworth et al., 2014 [40]** | **Xu et al., 2022 [56]** | **Tsai et al., 2021 [55]** | **Prochaska et al., 2021 [51]** |
|---|---|---|---|---|---|---|---|---|---|---|---|
| TITLE AND ABSTRACT | | | | | | | | | | | |
| Identification as an RCT | 1a | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Structured summary (design, methods, results, conclusions) | 1b | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| INTRODUCTION | | | | | | | | | | | |
| Backgrounds and objectives | 2a | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| METHODS | | | | | | | | | | | |
| Participants: eligibility | 4a (i) | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| criteria | | | | | | | | | | | |
| | 4a (ii) | x | x | x | x | x | x | ✓ | x | x | x |
| Participants: settings and locations of data collection | 4b | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | x | ✓ |
| Interventions | 5 (i) | ✓ | ✓ | x | x | ✓ | x | x | x | x | x |
| | 5 (ii) | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | x | x |
| | 5 (iii) | ✓ | x | ✓ | ✓ | x | x | x | x | ✓ | x |
| | 5 (iv) | ✓ | x | ✓ | x | x | x | x | x | ✓ | x |
| | 5 (v) | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | x | ✓ |
| | 5 (vi) | x | ✓ | ✓ | ✓ | ✓ | ✓ | x | x | ✓ | ✓ |
| RESULTS | | | | | | | | | | | |
| Harms | 19 | x | ✓ | x | x | x | x | x | x | x | x |
| OTHER INFORMATION | | | | | | | | | | | |
| Funding | 25 | ✓ | x | x | x | x | x | x | x | x | x |
| ***Compliance*** | | ***0.8*** | ***0.7*** | ***0.7*** | ***0.6*** | ***0.6*** | ***0.6*** | ***0.6*** | ***0.5*** | ***0.5*** | ***0.5*** |

Notes: 1a – "(i) Indicate that the intervention involves artificial intelligence/machine learning in the title and/or abstract and specify the type of model."; 1b – "(ii) State the intended use of the AI intervention within the trial in the title and/or abstract."; 2a (i) – "Explain the intended use of the AI intervention in the context of the clinical pathway, including its purpose and its intended users (e.g. healthcare professionals, patients, public).";4a (i) – "State the inclusion and exclusion criteria at the level of participants."; 4a (ii) – "State the inclusion and exclusion criteria at the level of the input data."; 4b – "Describe how the AI intervention was integrated into the trial setting, including any onsite or offsite requirements."; 5 (i) – "State which version of the AI algorithm was used.";5 (ii) – "Describe how the input data were acquired and selected for the AI intervention.";5 (iii) – "Describe how poor quality or unavailable input data were assessed and handled.";5 (iv) – "Specify whether there was human-AI interaction in the handling of the input data, and what level of expertise was required of users."; 5 (v) – "Specify the output of the AI intervention."; 5 (vi) – "Explain how the AI intervention's outputs contributed to decision-making or other elements of clinical practice."; 19 – "Describe results of any analysis of performance errors and how errors were identified, where applicable. If no such analysis was planned or done, justify why not."; 25 – "State whether and how the AI intervention and/or its code can be accessed, including any restrictions to access or re-use." [23].

Their research interval spanned from September 2020 to September 2022 and among others, they identified only seven RCTs on AI-based chatbots. Our findings are in line with their results, showing that AI-specific items (e.g., algorithm version, accessibility of the AI intervention, and input data handling) were often poorly reported. Additionally, performance error analysis and data input criteria were identified as weak points in the reporting. Nonetheless, we found that eight out of the 10 best quality publications have been published since 2020 (Table 2).

To assess the trends of compliance with CONSORT-AI, we have grouped the articles by publication year (from 2010 to 2019 and 2020 to 2022) a slight improvement was observed, which aligns with our overall findings.

Among the strengths of our study, we would like to highlight that we have conducted a comprehensive search to identify all RCTs on AI-based chatbots in healthcare. Validated search terms are available to detects RCTs, however no such search term set is available for chatbots. While previous SLRs used a limited number of terms for chatbots, we developed and applied a broad set of terms that have contributed to the completeness of our SLR. (For instance, using this search term set provided 417 records in PubMed, while inserting the single search term 'chatbot' instead resulted only 130 hits.) However, we assume that SLRs on healthcare chatbots will gain increasing importance in the future. At the time of our study chatbot use was not widespread (e.g., ChatGPT was launched in November 2022) and we have witnessed its incredible evolution and impact on the chatbot market and functionality in the past three years. Therefore, it would be important to develop a validated (sensitivity, precision) search term set of chatbots for use in diverse literature databases considering the new chatbot synonyms and brand names as well. Another strength of our study is that we have provided a detailed analysis of the AI part of the RCTs based on the CONSORT-AI, so that researchers can make use of these experiences in their work.

Some limitations of our study need to be noted. Chatbot development is a rapidly evolving area, therefore, a significant number of articles may have been published in this field since the closure of our search (2022), potentially adhering more closely to the CONSORT-AI guidelines. The final version of the CONSORT-AI guideline was published in September 2020. Several studies included in our analysis were conducted before this date and may have been based on a preliminary version of the reporting guideline or without using any AI reporting guideline. Nevertheless. among the 10 most compliant publications with the AI part of the CONSORT-AI one was published in 2014 (Heyworth et al., [40]) far before the establishment of CONSORT-AI.

The shortcomings that we have identified in the RCTs underscore the need for further research to advance the field of AI-based chatbot interventions. One potential research direction could focus on chatbots that leverage advanced AI models, such as generative transformers, multi-modal systems, or AI-augmented virtual reality tools. Another direction could emphasize a root cause analysis of the possible gaps, focusing on the timeframe from 2022 to the present. Above all, we would consider it important, that journals require the use of relevant reporting guidelines (checklists) prior to submission, but at the latest during the review process. Better quality RCT publications could make better, evidence-based use of AI-based chatbots to improve patient care. We aimed to support this goal with our work.

## Conclusions

A notable increase in the number of RCTs on AI chatbots across various healthcare domains, can be observed. However, the usability of the studies for medical-decision making is hampered by the suboptimal reporting quality of the AI part. Key gaps include inconsistent reporting of input data inclusion and exclusion criteria, limited exploration of human-AI interaction in data handling, insufficient descriptions of performance errors, and a lack of transparency regarding code availability.

We recommend researchers use the CONSORT-AI guideline, during both the design and reporting phases of AI chatbot RCTs. Adopting this approach, both by researchers and journal editors, would enhance transparency, improve standardization, and facilitate the generation of more robust, evidence based RCTs on AI chatbots. This would increase the usability of AI chatbots in supporting medical decision-making and patient care.

Also, we encourage future research on the development and the validation of literature search strategy, for chatbots that include the terms, Large Language Model and ChatGPT. The usage of these terms has grown exponentially in recent years and they are increasingly regarded as synonymous with chatbots and virtual assistants, in contemporary literature.

## Acknowledgements

The authors wish to express their sincere gratitude to Dr. László Berek, Director of Library of Óbuda University, for his support in data analysis.

This work was supported by the National Research, Development, and Innovation Fund of Hungary, financed under the TKP2021-NKTA-36 funding scheme.

## References

[1] E. Adamopoulou and L. Moussiades, “An Overview of Chatbot Technology”, Artificial Intelligence Applications and Innovations. AIAI 2020. IFIP Advances in Information and Communication Technology, Vol. 584, pp. 373-383, 2020, doi: 10.1007/978-3-030-49186-4_31

[2] Z. Yang and T. Váradi, "Training Experimental Language Models with Low Resources, for the Hungarian Language", Acta Polytechnica Hungarica, Vol. 20, No. 5, pp. 169-188, 2023, doi: 10.12700/APH.20.5.2023.5.11

[3] L. J. Laki and Z. G. Yang, "Sentiment Analysis with Neural Models for Hungarian", Acta Polytechnica Hungarica, Article Vol. 20, No. 5, pp. 109-128, 2023, doi: 10.12700/APH.20.5.2023.5.8

[4] V. V. Bochkarev, S. V. Khristoforov, A. V. Shevlyakova, and V. D. Solovyev, "Comparison of the Three Algorithms for Concreteness Rating Estimation of English Words", Acta Polytechnica Hungarica, Article Vol. 19, No. 10, pp. 99-121, 2022, doi: 10.12700/APH.19.10.2022.10.7

[5] S. H. Lakshmaiah, F. Balouchzahi, M. D. Anusha, and G. Sidorov, "CoLI-Machine Learning Approaches for Code-mixed Language Identification at the Word Level in Kannada-English Texts", Acta Polytechnica Hungarica, Article Vol. 19, No. 10, pp. 123-141, 2022, doi: 10.12700/APH.19.10.2022.10.8

[6] A. Miner, L. Laranjo, and A. B. Kocaballi, "Chatbots in the fight against the COVID-19 pandemic", npj Digital Medicine, Vol. 3, p. 65, 2020, doi: 10.1038/s41746-020-0280-0

[7] H. Karen, "The pandemic is emptying call centers. AI chatbots are swooping in", ed. https://www.technologyreview.com/2020/05/14/1001716/ai-chatbots-take-call-center-jobs-during-coro-navirus-pandemic/: MIT Technology Review online, 2020

[8] https://chatgpt.com

[9] M. Sarnovský, V. Maslej-Krešňáková, and K. Ivancová, "Fake news detection related to the COVID-19 in Slovak language using deep learning methods", Acta Polytechnica Hungarica, Article Vol. 19, No. 2, pp. 43-57, 2022

[10] M. Amjad, S. Butt, A. Zhila, G. Sidorov, L. Chanona-Hernandez, and A. Gelbukh, "Survey of Fake News Datasets and Detection Methods in European and Asian Languages", Acta Polytechnica Hungarica, Article Vol. 19, No. 10, pp. 185-204, 2022, doi: 10.12700/APH.19.10.2022.10.11

[11] K. Machová, M. Mach, and M. Vasilko, "Recognition of Toxicity of Reviews in Online Discussions", Acta Polytechnica Hungarica, Article Vol. 19, No. 4, pp. 7-26, 2022, doi: 10.12700/APH.19.4.2022.4.1

[12] C. Chakraborty, S. Pal, M. Bhattacharya, S. Dash, and S. S. Lee, "Overview of Chatbots with special emphasis on artificial intelligence-enabled ChatGPT in medical science", Front Artif Intell, Vol. 6, p. 1237704, 2023, doi: 10.3389/frai.2023.1237704

[13] M. Osváth, Z. G. Yang, and K. Kósa, "Analyzing Narratives of Patient Experiences: A BERT Topic Modeling Approach", Acta Polytechnica Hungarica, Article Vol. 20, No. 7, pp. 153-171, 2023, doi: 10.12700/APH.20.7.2023.7.9

[14] D. Khurana, A. Koli, K. Khatter, and S. Singh, "Natural language processing: state of the art, current trends and challenges", Multimed Tools Appl, Vol. 82, No. 3, pp. 3713-3744, 2023, doi: 10.1007/s11042-022-13428-4

[15] B. Kim, J. Park, and J. Suh, "Transparency and accountability in AI decision support: Explaining and visualizing convolutional neural networks for text information", Decision Support Systems, Vol. 134, p. 113302, 04/01 2020, doi: 10.1016/j.dss.2020.113302

[16] J. Grossman and F. J. Mackenzie, "The randomized controlled trial: gold standard, or merely standard?", Perspectives in biology and medicine, Vol. 48, No. 4, pp. 516-534, doi: 10.1353/pbm.2005.0092

[17] https://www.equator-network.org

[18] C. Zednik, "Solving the Black Box Problem: A Normative Framework for Explainable Artificial Intelligence", Philosophy & Technology, Vol. 34, 06/01 2021, doi: 10.1007/s13347-019-00382-7

[19] S. A.-O. Han et al., "A checklist is associated with increased quality of reporting preclinical biomedical research: A systematic review", Plos one, Vol. 12, No. 9, e0183591, doi: 10.1371/journal.pone.0183591

[20] J. Bergs et al., "Systematic review and meta-analysis of the effect of the World Health Organization surgical safety checklist on postoperative complications," (in eng), Br J Surg, Vol. 101, No. 3, pp. 150-8, Feb 2014, doi: 10.1002/bjs.9381

[21] K. Schulz, D. Altman, and D. Moher, "CONSORT 2010 statement: updated Guidelines for Reporting Parallel Group Randomized Trials," BMC medicine, Vol. 8, p. 18, 03/01 2010, doi: 10.1186/1741-7015-8-18

[22] B. Norgeot et al., "Minimum information about clinical artificial intelligence modeling: the MI-CLAIM checklist," (in eng), Nat Med, Vol. 26, No. 9, pp. 1320-1324, Sep 2020, doi: 10.1038/s41591-020-1041-y

[23] X. Liu et al., "Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension," Nature Medicine, Vol. 26, No. 9, pp. 1364-1374, 2020/09/01 2020, doi: 10.1038/s41591-020-1034-x

[24] M. Milne-Ives et al., "The Effectiveness of Artificial Intelligence Conversational Agents in Health Care: Systematic Review," (in eng), J Med Internet Res, Vol. 22, No. 10, p. e20346, Oct 22 2020, doi: 10.2196/20346

[25] L. Laranjo et al., "Conversational agents in healthcare: a systematic review," (in eng), J Am Med Inform Assoc, Vol. 25, No. 9, pp. 1248-1258, Sep 1 2018, doi: 10.1093/jamia/ocy072

[26] J. Shuster, "Review: Cochrane handbook for systematic reviews for interventions, Version 5.1.0, published 3/2011. Julian P.T. Higgins and Sally Green, Editors," Research Synthesis Methods, Vol. 2, 06/01 2011, doi: 10.1002/jrsm.38

[27] https://www.chatbots.org/synonyms/#all

[28] https://www.prisma-statement.org/prisma-2020-flow-diagram

[29] W. G. Adams, B. D. Phillips, J. D. Bacic, K. E. Walsh, C. W. Shanahan, and M. K. Paasche-Orlow, "Automated conversation system before pediatric primary care visits: a randomized trial," (in eng), Pediatrics, Vol. 134, No. 3, pp. e691-9, Sep 2014, doi: 10.1542/peds.2013-3759

[30] J. Beaman, L. Lawson, A. Keener, and M. L. Mathews, "Within Clinic Reliability and Usability of a Voice-Based Amazon Alexa Administration of the Patient Health Questionnaire 9 (PHQ 9)," (in English), Journal of Medical Systems, Article Vol. 46, No. 6, 2022, doi: 10.1007/s10916-022-01816-0

[31] D. C. Beidel, P. W. Tuerk, J. Spitalnick, C. A. Bowers, and K. Morrison, "Treating Childhood Social Anxiety Disorder With Virtual Environments and Serious Games: A Randomized Trial," Behavior Therapy, Vol. 52, No. 6, pp. 1351-1363, NOV 2021, doi: 10.1016/j.beth.2021.03.003

[32] M. R. Bennion, G. E. Hardy, R. K. Moore, S. Kellett, and A. Millings, "Usability, Acceptability, and Effectiveness of Web-Based Conversational Agents to Facilitate Problem Solving in Older Adults: Controlled Study," (in eng), J Med Internet Res, Vol. 22, No. 5, p. e16794, May 27, 2020, doi: 10.2196/16794

[33] J. E. Bibault et al., "A Chatbot Versus Physicians to Provide Information for Patients With Breast Cancer: Blind, Randomized Controlled Noninferiority Trial," (in eng), J Med Internet Res, Vol. 21, No. 11, p. e15787, Nov 27 2019, doi: 10.2196/15787

[34] P. Dulin, R. Mertz, A. Edwards, and D. King, "Contrasting a Mobile App With a Conversational Chatbot for Reducing Alcohol Consumption: Randomized Controlled Pilot Trial," (in eng), JMIR Form Res, Vol. 6, No. 5, p. e33037, May 16 2022, doi: 10.2196/33037

[35] K. K. Fitzpatrick, A. Darcy, and M. Vierhile, "Delivering Cognitive Behavior Therapy to Young Adults With Symptoms of Depression and Anxiety Using a Fully Automated Conversational Agent (Woebot): A Randomized Controlled Trial," (in eng), JMIR Ment Health, Vol. 4, No. 2, p. e19, Jun 6 2017, doi: 10.2196/mental.7785

[36] S. A. Friederichs, A. Oenema, C. Bolman, and L. Lechner, "Long term effects of self-determination theory and motivational interviewing in a web-based physical activity intervention: randomized controlled trial," (in eng), Int J Behav Nutr Phys Act, Vol. 12, p. 101, Aug 18 2015, doi: 10.1186/s12966-015-0262-9

[37] R. Fulmer, A. Joerin, B. Gentile, L. Lakerink, and M. Rauws, "Using Psychological Artificial Intelligence (Tess) to Relieve Symptoms of Depression and Anxiety: Randomized Controlled Trial," (in eng), JMIR Ment Health, Vol. 5, No. 4, p. e64, Dec 13 2018, doi: 10.2196/mental.9782

[38] E. Y. Gong et al., "My Diabetes Coach, a Mobile App-Based Interactive Conversational Agent to Support Type 2 Diabetes Self-Management: Randomized Effectiveness-Implementation Trial," Journal of Medical Internet Research, Vol. 22, No. 11, NOV 5 2020, Art no. e20322, doi: 10.2196/20322

[39] S. Greer, D. Ramo, Y. J. Chang, M. Fu, J. Moskowitz, and J. Haritatos, "Use of the Chatbot "Vivibot" to Deliver Positive Psychology Skills and Promote Well-Being Among Young People After Cancer Treatment: Randomized Controlled Feasibility Trial," (in eng), JMIR Mhealth Uhealth, Vol. 7, No. 10, p. e15018, Oct 31 2019, doi: 10.2196/15018

[40] L. Heyworth et al., "Comparison of interactive voice response, patient mailing, and mailed registry to encourage screening for osteoporosis: a randomized controlled trial," (in eng), Osteoporos Int, Vol. 25, No. 5, pp. 1519-26, May 2014, doi: 10.1007/s00198-014-2629-1

[41] S. Jang, J. J. Kim, S. J. Kim, J. Hong, S. Kim, and E. Kim, "Mobile app-based chatbot to deliver cognitive behavioral therapy and psychoeducation for adults with attention deficit: A development and feasibility/usability study," (in eng), Int J Med Inform, Vol. 150, p. 104440, Jun 2021, doi: 10.1016/j.ijmedinf.2021.104440

[42] M. C. Klos, M. Escoredo, A. Joerin, V. N. Lemos, M. Rauws, and E. L. Bunge, "Artificial Intelligence-Based Chatbot for Anxiety and Depression in University Students: Pilot Randomized Controlled Trial," (in eng), JMIR Form Res, Vol. 5, No. 8, p. e20678, Aug 12 2021, doi: 10.2196/20678

[43] H. Liu, H. Peng, X. Song, C. Xu, and M. Zhang, "Using AI chatbots to provide self-help depression interventions for university students: A randomized trial of effectiveness," (in eng), Internet Interv, Vol. 27, p. 100495, Mar 2022, doi: 10.1016/j.invent.2022.100495

[44] K. Loveys, M. Sagar, and E. Broadbent, "The Effect of Multimodal Emotional Expression on Responses to a Digital Human during a Self-Disclosure Conversation: a Computational Analysis of User Language," (in eng), J Med Syst, Vol. 44, No. 9, p. 143, Jul 22 2020, doi: 10.1007/s10916-020-01624-4

[45] K. H. Ly, A. M. Ly, and G. Andersson, "A fully automated conversational agent for promoting mental well-being: A pilot RCT using mixed methods," (in eng), Internet Interv, Vol. 10, pp. 39-46, Dec 2017, doi: 10.1016/j.invent.2017.10.002

[46] E. Maeda et al., "Promoting fertility awareness and preconception health using a chatbot: a randomized controlled trial," (in eng), Reprod Biomed Online, Vol. 41, No. 6, pp. 1133-1143, Dec 2020, doi: 10.1016/j.rbmo.2020.09.006

[47] M. Nißen et al., "The Effects of Health Care Chatbot Personas With Different Social Roles on the Client-Chatbot Bond and Usage Intentions: Development of a Design Codebook and Web-Based Study," (in eng), J Med Internet Res, Vol. 24, No. 4, p. e32630, Apr 27 2022, doi: 10.2196/32630

[48] M. Ogawa et al., "Can AI make people happy? The effect of AI-based chatbot on smile and speech in Parkinson's disease," (in eng), Parkinsonism

Relat Disord, Vol. 99, pp. 43-46, Jun 2022, doi: 10.1016/j.parkreldis.2022.04.018

[49] E. Olano-Espinosa et al., "Effectiveness of a Conversational Chatbot (Dejal@bot) for the Adult Population to Quit Smoking: Pragmatic, Multicenter, Controlled, Randomized Clinical Trial in Primary Care," (in eng), JMIR Mhealth Uhealth, Vol. 10, No. 6, p. e34273, Jun 27 2022, doi: 10.2196/34273

[50] M. Piao, H. Ryu, H. Lee, and J. Kim, "Use of the Healthy Lifestyle Coaching Chatbot App to Promote Stair-Climbing Habits Among Office Workers: Exploratory Randomized Controlled Trial," (in eng), JMIR Mhealth Uhealth, Vol. 8, No. 5, p. e15085, May 19 2020, doi: 10.2196/15085

[51] J. J. Prochaska et al., "A randomized controlled trial of a therapeutic relational agent for reducing substance misuse during the COVID-19 pandemic," (in eng), Drug Alcohol Depend, Vol. 227, p. 108986, Oct 1 2021, doi: 10.1016/j.drugalcdep.2021.108986

[52] S. R. Simon et al., "Failure of automated telephone outreach with speech recognition to improve colorectal cancer screening: a randomized controlled trial," (in eng), Arch Intern Med, Vol. 170, No. 3, pp. 264-70, Feb 8 2010, doi: 10.1001/archinternmed.2009.522

[53] A. Söderström, A. Shatte, and M. Fuller-Tyszkiewicz, "Can intelligent agents improve data quality in online questiosnnaires? A pilot study," (in eng), Behav Res Methods, Vol. 53, No. 5, pp. 2238-2251, Oct 2021, doi: 10.3758/s13428-021-01574-w

[54] M. J. Tanana, C. S. Soma, V. Srikumar, D. C. Atkins, and Z. E. Imel, "Development and Evaluation of ClientBot: Patient-Like Conversational Agent to Train Basic Counseling Skills," (in eng), J Med Internet Res, Vol. 21, No. 7, p. e12529, Jul 15 2019, doi: 10.2196/12529

[55] W. S. Tsai, D. Lun, N. Carcioppolo, and C. H. Chuan, "Human versus chatbot: Understanding the role of emotion in health marketing communication for vaccines," (in eng), Psychol Mark, Vol. 38, No. 12, pp. 2377-2392, Dec 2021, doi: 10.1002/mar.21556

[56] Y. Xu, J. Aubele, V. Vigil, A. S. Bustamante, Y. S. Kim, and M. Warschauer, "Dialogue with a conversational agent promotes children's story comprehension via enhancing engagement," (in eng), Child Dev, Vol. 93, No. 2, pp. e149-e167, Mar 2022, doi: 10.1111/cdev.13708

[57] A. P. L. Martindale et al., "Concordance of randomized controlled trials for artificial intelligence interventions with the CONSORT-AI reporting guidelines," Nature Communications, Vol. 15, No. 1, p. 1619, 2024/02/22 2024, doi: 10.1038/s41467-024-45355-3