**Digital Health Trends Journal.** 2026;2(1):23-33.
doi:10.34172/dhtj.10
https://dhtjournal.ae

Digital Health Trends Journal

Systematic Review

# Chatbots and Virtual Health Assistants in Healthcare: From Initial Triage to Patient Follow-up

**Mandana Akbari Marandi[1*], Reza Shoja Ghiass[2]**

[1]Deggendorf Institute of Technology, Bavaria, Germany
[2]Synapsis Health AI, Ontario, Canada

***Corresponding author:** Mandana Akbari Marandi, Email: Makbarimarandi@gmail.com

**Abstract**

**Background:** Chatbots and virtual health assistants (VHAs) are emerging tools in healthcare, supporting triage, symptom assessment, diagnosis, follow-up, treatment adherence, and remote monitoring. Despite the rapid adoption, comprehensive evidence on their effectiveness, limitations, and clinical utility is limited. Thus, this review evaluates current evidence regarding their applications, performance, and impact across the patient care continuum.

**Methods:** A systematic search of PubMed, Scopus, and Web of Science (2010–2025) identified studies on chatbots and VHAs in healthcare. Eligible studies addressed triage, symptom checking, diagnosis support, follow-up, treatment adherence, and remote monitoring. The extracted data covered design, sample size, artificial intelligence (AI) architecture, platform, usability, engagement, and outcomes. Study quality was assessed using Mixed Methods Appraisal Tool (MMAT). Given heterogeneous designs and outcomes, a narrative synthesis approach was employed to summarize the findings.

**Results:** Forty-six studies from 17 countries were identified, including randomized controlled trials, cross-sectional studies, pilot projects, and retrospective analyses. The reported diagnostic accuracy ranged from 33% to 93%. Usability was high (SUS 68–85), with strong potential to enhance engagement, adherence, and access to care, especially in chronic disease and mental health management. Large language model-based and retrieval-augmented systems outperformed traditional rule-based chatbots in complex tasks. Limitations included short follow-ups, simulated cases, inconsistent evaluation metrics, and scarce evidence of long-term impact.

**Conclusion:** Chatbots and VHAs represent effective complementary healthcare tools, improving patient support, workflow efficiency, and accessibility, although they should not be independently used in diagnostic/therapeutic decision-making. Future research should emphasize large-scale real-world evaluations, standardized metrics, long-term outcomes, and secure, patient-centered implementations.

**Keywords:** Chatbots, Virtual health assistants, Healthcare, Triage, Remote monitoring, Treatment adherence

**Received:** September 9, 2025, **Revised:** November 1, 2025, **Accepted:** November 12, 2025, **ePublished:** November 22, 2025

## Background

Artificial intelligence (AI) has rapidly emerged as a transformative domain in healthcare, offering tools that mimic human cognition through machine learning, deep learning, and natural language processing (NLP).[1,2] In addition, recent advancements, particularly in large language models (LLMs), have demonstrated remarkable potential in enhancing diagnostic accuracy, personalizing treatment, and streamlining healthcare operations.[2] These technological shifts address critical challenges (e.g., clinician shortages, long waiting times, and the growing burden of chronic diseases), positioning AI-based conversational agents as one of the most promising solutions.[3]

Conversational agents, also referred to as chatbots or virtual health assistants (VHAs), are intelligent systems capable of engaging in dialogue with patients through text or voice interfaces.[4,5] While the earliest chatbot, ELIZA (1966), offered only rudimentary interactions,[6] subsequent advances in NLP and deep learning have enabled modern chatbots to perform complex functions, including symptom assessment, triage, counseling, and remote monitoring.[5,7-11] They are now deployed across diverse platforms (e.g., smartphones, web applications, and telehealth systems), supporting both patients and healthcare professionals.[12,13]

In healthcare, chatbots have shown potential to act as first-contact automation tools, assisting with screening, gathering medical histories, providing psychological support, and delivering health recommendations.[14,15] Some studies highlight that these systems can supplement human caregivers, improve adherence to treatment, and empower patients through real-time, personalized support.[14,16] Additionally, they offer scalability, affordability, and accessibility, helping extend healthcare services to underserved and hard-to-reach populations.[4,17]

Despite these advantages, several challenges persist. Concerns about patient safety, accuracy of medical advice,

and the protection of sensitive health data are consistently observed in the literature.[18-20] Moreover, sustaining long-term patient engagement with chatbots remains difficult, and ethical questions regarding trust, transparency, and the potential replacement of human interaction require careful consideration.[2,21,22]

Nevertheless, the potential of chatbots and VHAs to transform healthcare delivery is substantial. They have been applied in a wide range of use cases, including initial triage, symptom checking, diagnosis support, patient follow-up, treatment adherence, and lifestyle coaching.[5,14,23] Furthermore, their integration into healthcare systems can help alleviate resource constraints, provide patient-centered care, and enhance health outcomes at scale.[24-26]

Given the fast-paced evolution of AI technologies and the growing body of research in this domain, there is a pressing need for systematic reviews that comprehensively evaluate the applications, effectiveness, and limitations of chatbots and VHAs across the continuum of patient care. Accordingly, the present review seeks to fill this gap by synthesizing current evidence on the role of these tools in healthcare, from initial triage to long-term follow-up.

## Methods

### *Study Design*

This study was conducted as a systematic review following the guidelines of the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA). The aim was to comprehensively examine the use of chatbots and VHAs in healthcare, ranging from initial triage to patient follow-up.

### *Search Strategy*

A systematic literature search was performed across three major electronic databases: PubMed, Scopus, and Web of Science. In addition, the search included studies published from January 2010 to June 2025, restricted to the English language.

Further, the search strategy combined terms related to conversational agents with healthcare applications, as follows:

((Chatbot* OR (Conversational agent*) OR (Virtual health assistant*) OR (Intelligent virtual assistant*) AND (Triage OR (Symptom checker) OR (Diagnosis support) OR (Patient follow-up) OR (Treatment adherence) OR (Remote monitoring)))

It is noteworthy that Boolean operators, truncation, and quotation marks were applied according to the specifications of each database to ensure high sensitivity.

Additionally, the reference lists of the included studies and relevant reviews were manually screened to identify further eligible publications.

### *Eligibility Criteria*

#### *Inclusion Criteria*

English-language publications and studies reporting chatbots or VHAs used in healthcare settings were included in this study, in addition to applications addressing initial triage, symptom checking, diagnosis support, patient follow-up, treatment adherence, or remote monitoring.

#### *Exclusion Criteria*

Non-English publications, conference abstracts, editorials, commentaries, and opinion papers were excluded from the investigation, along with studies focusing on non-healthcare domains (e.g., education, marketing, and customer service) and studies lacking empirical data (e.g., conceptual papers without implementation or evaluation).

### *Study Selection*

All retrieved articles were imported into EndNote X9 for reference management and duplicate removal.

Two independent reviewers (reviewer A and reviewer B) screened titles and abstracts against the eligibility criteria. Then, full texts of potentially relevant articles were assessed independently by both reviewers. In addition, discrepancies were resolved through discussion or consultation with a third reviewer (reviewer C).

### *Data Extraction*

A standardized data extraction form was developed in Microsoft Excel to collect relevant information, including:

- Bibliographic details: Author(s), year, country, and journal
- Study characteristics: Study design, sample size, and healthcare setting
- Technology details: Type of chatbot or VHA, platform, and applied AI techniques (e.g., rule-based, machine learning, and NLP)
- Application domain: Triage, symptom checker, diagnosis support, patient follow-up, treatment adherence, and remote monitoring
- Outcomes: Clinical effectiveness, patient satisfaction, usability, accuracy, and engagement metrics.

### *Limitations Reported by the Authors*

The required data were extracted independently by the two reviewers. Any disagreements were discussed and resolved by consensus.

### *Quality Assessment*

The quality of the included studies was assessed using the Mixed Methods Appraisal Tool (MMAT) for primary studies, including quantitative, qualitative, and mixed-methods research. The findings were interpreted, considering the risk of bias.

### *Data Synthesis*

A narrative synthesis approach was adopted due to the heterogeneity of study designs, outcomes, and chatbot applications. The findings were organized according to

the healthcare application, including initial triage (1), symptom checking and diagnosis support (2), patient follow-up and treatment adherence (3), and remote monitoring (4).

Quantitative data were summarized using descriptive statistics (frequencies and percentages), while qualitative findings were synthesized thematically. In addition, performance metrics (e.g., diagnostic accuracy and patient engagement rates) were reported when feasible.

### *Flow Diagram of the Preferred Reporting Items for Systematic Reviews and Meta-Analyses*

A PRISMA flow diagram was used to illustrate the study selection process (the number of records identified, screened, excluded, and included).

## Results

### *Study Selection*

A total of 1,064 records were identified through database searches, with 355 duplicates removed automatically. Following title and abstract screening of 396 studies, 277 were excluded due to irrelevance. Then, 119 full-text articles were assessed for eligibility, of which 49 were excluded because of not meeting the inclusion criteria. Ultimately, 46 studies were included in this review (Figure 1).

### *Characteristics of Included Studies*

The 46 included studies were published between 2018 and 2025, encompassing 17 countries across North America, Europe, Asia, and Australia.

The extracted articles varied in design, including cross-sectional evaluations (n = 20), randomized controlled trials (RCTs, n = 6), pilot/feasibility studies (n = 8), observational retrospective analyses (n = 7), and comparative accuracy studies (n = 5).

Moreover, sample sizes ranged from 3 patients in a pediatric chatbot feasibility trial to 129,400 patients in a large-scale National Health Service (NHS) mental health deployment. In addition, platforms included web-based conversational agents, mobile health applications, and integrated telemedicine systems. Further, chatbot architectures varied across rule-based, NLP-driven, and LLM-based generative AI systems. Detailed characteristics of the included studies are presented in Table S1 (Supplementary file 1).

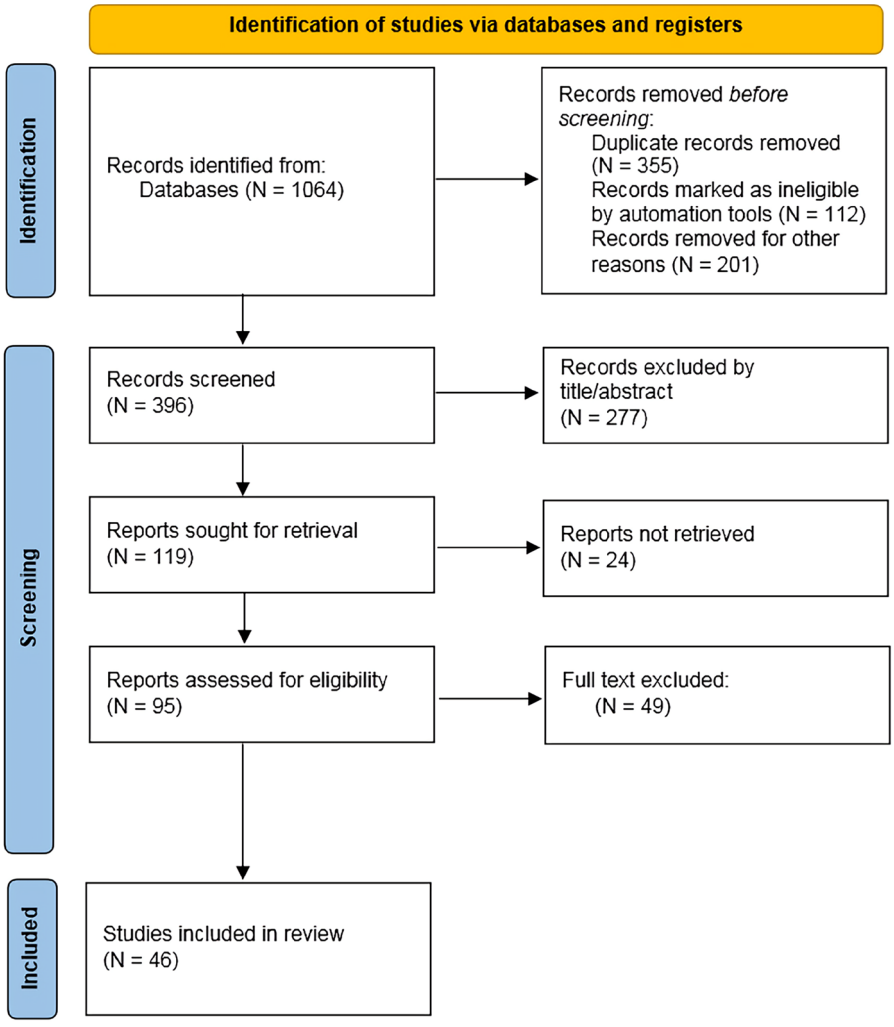


**Figure 1.** PRISMA Flow Diagram of the Study Selection Process. *Note*. PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses

### *Performance Across Application Domains*

#### *Initial Triage and Symptom Checking*

Seventeen studies evaluated chatbots for initial triage and self-assessment.

- Diagnostic accuracy ranged from 33% (WebMD) to 93% (ChatGPT, ophthalmology).
- Ophthalmology-focused chatbots revealed high agreement with clinical graders (Cohen's kappa: 0.74–0.79).
- Emergency triage performance varied:
  - High-acuity sensitivity reached 76.2% with ChatGPT in simulated ESI-1/ESI-2 cases.
  - However, general emergency scenarios had specificities as low as 34.5%, reflecting over-triage or under-triage risks.
- Commercially available symptom checkers (e.g., Kahun) outperformed others in clinical data gathering (recall of 0.51 vs. average of 0.32), but overall diagnostic comprehensiveness remained suboptimal.

In summary, while AI-powered triage chatbots demonstrate promising sensitivity in critical conditions, variability across platforms and contexts limits their standalone deployment in emergency medicine.

#### *Diagnostic and Treatment Decision Support*

Twenty-one studies investigated diagnostic assistance across ophthalmology, oncology, dental medicine, dermatology, and internal medicine. Key findings are as follows:

- GPT-4-assisted oral lesion diagnosis achieved high sensitivity for malignancies (no malignant lesions missed), although specificity remained below specialist-level performance.
- In gastrointestinal radiology, a retrieval-augmented GPT-4 model achieved 78% correct primary differentials, outperforming generic GPT-4 (54%).
- Cancer treatment recommendation chatbots yielded mixed performance:
  - ChatGPT achieved 95% accuracy in first-line therapy recommendations for head and neck malignancies but failed in 55% of staging assessments.
  - Some chatbots generated unsafe treatment suggestions, emphasizing the need for clinician oversight.
- Comparative studies confirmed that ChatGPT-4.0 outperformed Google Bard and Gemini in ocular diagnostic scenarios (accuracy of 89.2% vs. 59.5% and 40.5%).

In brief, diagnostic support chatbots approach human-level performance in certain structured contexts but remain error-prone, especially in cancer staging and rare condition management.

#### *Patient Education and Health Literacy*

Fourteen studies assessed the role of chatbots in disseminating patient information and improving understanding of treatment plans:

- DISCERN quality scores varied widely, from average (~40) in urogenital cancer chatbots to high (>70) for orthopedic education platforms (e.g., Microsoft Copilot).
- Readability metrics (i.e., Coleman-Liau index and SMOG) frequently exceeded college-level complexity, thereby limiting accessibility for patients with low health literacy.
- Personalized NHS self-referral chatbots significantly increased accessibility to mental health care; referral rates improved by 15% versus 6% in control groups and disproportionately benefited minority populations.

Briefly, while chatbots expand patient education access, optimizing content readability and ensuring source transparency remain essential for patient safety.

#### *Remote Monitoring and Patient Follow-up*

Eleven studies focused on chatbots integrated into remote patient monitoring (RPM) and chronic disease management systems:

- In postoperative cardiac patients, a text-based chatbot paired with portable coagulometers improved therapeutic time-in-range and yielded high patient satisfaction.
- During the coronavirus disease 2019 pandemic, chatbots enabled large-scale follow-up, with over 6,000 patient-generated comments processed via semi-automated topic modeling, demonstrating feasibility for high-volume care coordination.
- In oncology, integration with smartphone-based platforms improved medication adherence and real-time complication detection, with compliance rates reaching 86%.

In short, although RPM-integrated chatbots effectively support adherence and early complication detection, long-term outcome data are still limited.

#### *Treatment Adherence and Behavioral Interventions*

Eight studies evaluated chatbots for promoting lifestyle changes and treatment compliance:

- In eating disorder management, an RCT of the Eating Disorder Electronic Single-Session Intervention demonstrated significant symptom reductions ($P = 0.003$) and improved treatment initiation (93% vs. 70%, $P = 0.042$).
- Mental health support chatbots providing behavioral activation therapy could improve patient motivation and sustained engagement.
- In pediatric cancer patients, GPT-4-based chatbots reduced anxiety in 80% of participants and facilitated the disclosure of sensitive information previously unreported to clinicians.

Overall, chatbots reveal strong potential as scalable behavioral health tools, but further controlled trials are needed to confirm the durability of effects.

### *Cross-Study Synthesis*

- *Accuracy:* High variability, 33–98%, depending on task complexity.
- *Engagement:* Sustained usage observed in most longitudinal studies, but drop-off rates reached > 50% in poorly optimized chatbots.
- *Usability:* System usability scores (SUS) ranged from 68 to 85; rule-based systems generally underperformed LLM-driven platforms.
- *Equity impact:* Chatbots facilitated broader access to care, particularly in mental health and underserved populations.

### *Summary of Limitations*

Across studies, key limitations included

- Predominant reliance on scenario-based or simulated cases rather than real-world clinical data,
- Insufficient long-term follow-ups on clinical outcomes,
- Inconsistent evaluation metrics across platforms, limiting meta-analytic synthesis,
- High variability in chatbot architectures, ranging from rule-based to GPT-4-driven systems, complicating direct comparison.

### *Quality Assessment*

The quality of the included studies was assessed using the MMAT. Overall, the methodological rigor of the 46 studies varied considerably:

*High-quality studies:* Approximately one-third (33%) of the included studies (mostly RCTs and large-scale observational designs) demonstrated clear research questions, appropriate sampling, transparent reporting of outcomes, and low risk of bias. Generally, these studies provided robust evidence regarding usability, patient engagement, and short-term effectiveness of chatbots.[27-38]

*Moderate-quality studies:* Half of the studies fell into this category. They frequently lacked standardized outcome measures, relied on simulated case vignettes instead of real-world patient data, or had limited sample sizes. While informative, their findings require cautious interpretation.[24,39-64]

*Low-quality studies:* A minority (17%) of studies (mainly preliminary pilots and descriptive analyses) suffered from serious methodological limitations, including extremely small sample sizes, the absence of control groups, unclear AI model specifications, or insufficient reporting of statistical results.[65-72]

Table 1 provides the detailed quality classification of all included studies.

### *Key quality concerns across studies were as follows*

- Heavy reliance on scenario-based or simulated patient data, thus reducing generalizability to clinical practice.
- Short follow-up durations, thereby limiting insights into long-term outcomes (e.g., adherence, morbidity, or mortality).
- Inconsistent evaluation metrics (e.g., accuracy, usability, engagement, DISCERN, and SUS), thus preventing robust cross-study comparisons.
- Limited transparency regarding chatbot architectures, training datasets, and prompt engineering, thereby making reproducibility difficult.

In a nutshell, while a subset of well-designed randomized and observational studies provides encouraging evidence on the utility of chatbots and VHAs in healthcare, the overall quality of evidence remains heterogeneous. Hence, stronger study designs with standardized evaluation frameworks and longer-term follow-ups are needed to confirm clinical effectiveness and safety.

## Discussion

This systematic review synthesized evidence on the clinical applications, performance, and limitations of conversational agents, including chatbots and VHAs, across diverse healthcare contexts. These systems demonstrated potential in symptom assessment, triage, diagnostic decision support, patient education, remote monitoring, and behavioral interventions. However, the findings revealed significant heterogeneity in study design, performance metrics, and outcome reporting, limiting the generalizability of results. The reported accuracy rates varied widely, ranging from 33% to 98%, with usability, user satisfaction, and engagement similarly inconsistent across studies.

### *Interpretation by Application Domain*

#### *Initial Triage and Symptom Assessment*

In structured domains (e.g., ophthalmology), chatbots displayed performance approaching that of human evaluators, with reported inter-rater reliability ranging from $\kappa \approx 0.74$ to 0.79 (27, 31). However, performance substantially varied across general emergency settings; while sensitivity for high-acuity cases (ESI-1/ESI-2) reached 76.2%, specificity dropped as low as 34.5%,

**Table 1.** Quality Assessment of Included Studies Using the Mixed Methods Appraisal Tool

| Quality category | Number of studies | References | Type of study and key features |
|---|---|---|---|
| High quality | 12 | 27-38 | RCTs or large-scale observational studies<br>• Robust design, clear reporting, and low risk of bias |
| Moderate quality | 26 | 39-64 | Cross-sectional, scenario-based, or comparative studies<br>• Informative but limited generalizability |
| Low quality | 8 | 65-72 | Pilot or feasibility studies<br>• Very small sample sizes, descriptive analyses, and lack of control groups |

*Note*. RCT: Randomized controlled trial.

thereby raising concerns regarding both over-triage and under-triage risks.[66] However, most studies in this area were of moderate or low quality, relying on simulated scenarios or small samples.[65-69] Only a few high-quality studies (e.g., large-scale analyses of consumer symptom checker apps) provided stronger evidence for usability and functionality.[27,30]. Collectively, these findings highlight that while chatbots may offer value in recognizing critical scenarios, their reliability diminishes when facing complex, ambiguous, or unstructured cases.[39,41,66]

### *Diagnostic and Therapeutic Decision Support*

LLM-based models, particularly those leveraging GPT-4, yielded promising results when used within well-structured diagnostic frameworks. For example, in oral mucosal lesion detection, GPT-4 achieved high sensitivity in identifying lesions suspicious for malignancy without missing any malignant cases, but specificity remained lower compared to expert clinicians.[64] Similarly, Rau et al concluded that a retrieval-augmented GPT-4 model in gastrointestinal radiology achieved a 78% diagnostic accuracy, outperforming the baseline GPT-4 performance of 54%.[38]

Conversely, moderate-quality studies in oncology underscored substantial risks; while ChatGPT correctly recommended first-line therapies in 95% of head and neck cancer cases, it produced unsafe errors in staging and surgical planning.[50] Likewise, a high-quality multicenter study reported poor reliability in rare laryngology conditions.[37] Consequently, chatbots may serve as educational aids or case-preparation tools, but independent clinical decision-making without expert supervision remains unsafe.[38,50,64] Taken together, high-quality evidence confirms potential in structured tasks, but moderate-quality and low-quality evidence underlines significant safety concerns, particularly in complex treatment planning.

### *Patient Education and Health Literacy*

The quality, readability, and reliability of patient-facing information varied considerably. For instance, Microsoft Copilot produced the highest DISCERN scores among orthopedic platforms; nonetheless, in other contexts (e.g., urological cancer information), chatbots generally generated low DISCERN scores and content readability levels exceeding patient comprehension thresholds, as indicated by high Coleman-Liau and SMOG indices.[40,43] Consequently, without adjustments to improve clarity, accuracy, and evidence attribution, many chatbot outputs remain inaccessible or potentially unsafe for general patient populations.[43]

Chatbot-generated health information demonstrated high variability in both quality and readability. While Microsoft Copilot achieved superior DISCERN scores in orthopedic information delivery, other contexts, particularly urological oncology, produced generally poor readability scores and inconsistent accuracy, often exceeding patient comprehension thresholds based on Coleman-Liau and SMOG indices.[40,43] These findings indicate that without evidence attribution, language simplification, and source verification mechanisms, patient-directed chatbot outputs may exacerbate misinterpretation and health literacy inequities. Importantly, a high-quality large-scale observational study from NHS mental health services[53] revealed that self-referral chatbots improved accessibility and equity, particularly for minority populations. This indicates that while moderate-quality studies highlight concerns about readability and accuracy, stronger evidence supports the role of chatbots in expanding equitable access to care.

### *Remote Patient Monitoring and Treatment Adherence*

The integration of chatbots into RPM systems displayed encouraging preliminary results. In post-cardiac surgery management, for instance, the combination of chatbots and CoaguChek could improve time-in-therapeutic range.[52] Similarly, Piau et al reported 86% adherence rates in outpatient oncology cohorts when chatbots were paired with mobile monitoring platforms.[70] Despite these findings, the majority of evidence comes from small-scale, short-term studies, and there remains a significant gap in long-term outcome data regarding hospital readmissions, morbidity, and mortality.[52,70]

Our findings align with those of the study by Geoghegan et al, evaluating automated conversational agents for post-intervention follow-ups.[73] In their review, chatbots were primarily deployed for postoperative monitoring and follow-ups after a number of interventions, such as cancer treatment, hypertension management, asthma care, orthopedic procedures, ureteroscopy, and varicose vein surgery. Engagement rates ranged from 31% to 97% response rates, indicating strong patient receptivity in certain contexts but inconsistent adoption overall. In line with our study, the mentioned researchers reported no studies assessing patient safety outcomes, highlighting a critical evidence gap that persists across both reviews.[73] In contrast, high-quality studies demonstrated more robust evidence; for example, the analysis of over 6,000 coronavirus disease 2019 patient-generated comments confirmed the scalability and feasibility of chatbot-assisted monitoring,[29] while integration with wound-care telemedicine achieved high diagnostic accuracy.[33] Thus, while preliminary low-quality evidence is promising, reliable conclusions about long-term outcomes primarily rely on the fewer but stronger high-quality studies.

### *Behavioral Interventions and Mental Health*

Chatbots designed for behavioral interventions revealed modest but promising effects in RCTs. High-quality RCTs provided some of the strongest evidence for chatbot effectiveness in behavioral health. For instance, Vivibot reduced anxiety symptoms in young cancer survivors,[31] and the Eating Disorder Electronic Single-Session Intervention significantly improved eating disorder pathology and treatment uptake.[57] Moderate-quality or

low-quality studies[46,48] further demonstrated feasibility, user satisfaction, and reductions in anxiety in pediatric oncology and weight management settings, but their limited scale and lack of controls restrict generalizability. Overall, high-quality RCTs confirm the role of chatbots as scalable, complementary tools in mental health and behavioral interventions, while lower-quality evidence highlights feasibility and acceptability rather than clinical effectiveness.

Consistent with these findings, symptom reductions were observed in managing eating disorders and promoting treatment initiation, but short follow-up durations and limited sample sizes further limit generalizability.[57,70] Conversational agents indicated growing utility in mental health interventions. Gaffney et al reviewed 13 studies, including four RCTs, and reported reductions in psychological distress following chatbot interventions, with five controlled studies showing significant improvements compared to inactive controls.[74] However, studies comparing chatbots to active treatment modalities failed to display superiority, suggesting potential complementary rather than replacement roles.

These results corroborate the findings of a review study of 29 RCTs conducted by Yang et al, focusing on physical and psychological symptom management. Their findings highlighted significant improvements in 22 studies, particularly for depression, anxiety, and pain-related symptoms, with median recruitment and completion rates of 72% and 79%, respectively. Nevertheless, 17 studies exhibited a high risk of bias, limiting generalizability.[75] Collectively, these reviews and our synthesis confirmed that conversational agents are effective in enhancing short-term mental health and symptom outcomes, but heterogeneous designs, small sample sizes, and limited follow-up durations prevent definitive conclusions regarding long-term clinical benefits.

### *Medical History-Taking and Workflow Integration*

The potential of chatbots to streamline medical history-taking is increasingly recognized. The findings of a systematic review of 18 studies (including three RCTs) performed by Hindelang et al revealed that conversational agents can collect structured patient histories through targeted queries and automated data capture, enhancing both efficiency and patient engagement.[76] Notably, their review emphasized the advantages of 24/7 accessibility and electronic health record integration, which may reduce clinician workload and optimize resource utilization. However, usability challenges, privacy concerns, and limited empathic capacity remain barriers to clinical integration. Bias assessments showed that only 33% of observational studies were high quality, underscoring the need for rigorous validation prior to large-scale implementation.

### *Overall Effectiveness and User Acceptance*

To broadly assess chatbot effectiveness, Milne-Ives et al reviewed 31 studies across diverse healthcare applications. Conversational agents demonstrated positive or mixed effectiveness in 75% of studies, with high usability (27/30 studies) and strong patient satisfaction (26/31 studies). These findings conform to our synthesis, suggesting that conversational agents are generally well-received and capable of supporting various health-related tasks, including triage, training, monitoring, and behavior change. However, similar to our conclusion, Milne-Ives et al emphasized significant methodological variability and highlighted the need for robust evaluations of cost-effectiveness, privacy, and safety prior to routine clinical integration.[77]

Overall, our findings are consistent with prior systematic reviews and extend them simultaneously.

Geoghegan et al underlined strong engagement in post-intervention monitoring,[73] which is in line with our findings and is expanded by integrating evidence on remote patient adherence.

Likewise, Gaffney et al and Yang et al support our conclusion that conversational agents are effective in improving psychological outcomes, though they caution against overinterpreting short-term gains without long-term validation.[74,75]

The results of Hindelang et al also align with our findings on the efficiency of history-taking chatbots while underscoring gaps in usability and personalization.[76]

Moreover, Milne-Ives et al provided a comprehensive overview of effectiveness and usability, reinforcing that while conversational agents are promising, safety evaluations, standardized metrics, and cost-effectiveness analyses remain critical gaps,[77] which matches our findings.

### *Overall Synthesis Across Quality Levels*

A clear gradient emerges when results are interpreted in light of methodological quality:

- High-quality studies (n = 12) consistently supported chatbot utility in several domains, such as mental health interventions, large-scale monitoring, radiology decision support, and patient self-referral.[27-38]
- Moderate-quality studies (n = 26) provided informative but less generalizable results, often based on scenario-based evaluations or limited samples.[10,24,39-44,46-60,62-64]
- Low-quality studies (n = 8) largely assessed feasibility with extremely small cohorts or descriptive analyses, offering limited evidence for clinical integration.[65-72]

This distribution underscores that while certain applications, particularly in mental health, education, and remote monitoring, are supported by stronger evidence, many domains (e.g., acute triage and complex oncology decision-making) still heavily depend on lower-quality or simulated studies. Therefore, future work should prioritize large-scale, high-quality randomized and real-world studies with standardized evaluation metrics and

long-term follow-ups to establish safety and clinical effectiveness.

### *Clinical and Policy Implications*

Current evidence supports the supervised integration of chatbots into clinical workflows but does not justify autonomous deployment for diagnostic or therapeutic decision-making.[36,50] Accordingly, safe implementation in healthcare systems requires:

- Source transparency and evidence-linked responses to improve clinician trust,[43]
- Improved readability and health literacy adaptation for patient-directed content,[23,43]
- Secure integration with remote monitoring and patient education platforms, which currently represent the most promising short-term use cases.[52,70]

### *Strengths and Limitations of the Evidence*

#### *Strengths*

- *Comprehensive scope:* Our review synthesized findings from diverse clinical domains, integrating RCTs, observational studies, and pilot implementations to provide a multidimensional overview.
- *Structured quality assessment:* Some tools (e.g., AMSTAR-2 and MMAT) enabled the systematic evaluation of study quality and supported a cautious interpretation of findings.

#### *Limitations*

- A substantial proportion of studies relied on simulated cases or structured vignettes, with limited real-world clinical evaluations.
- Some studies applied inconsistent metrics (i.e., accuracy, SUS, DISCERN, SMOG, and usability scales), showing high methodological heterogeneity and preventing robust meta-analysis.
- Many studies featured small sample sizes, short follow-ups, and unclear reporting standards, leading to potential reporting bias.
- Some studies demonstrated limited transparency regarding model architectures, training datasets, and prompt designs, resulting in reduced reproducibility and a lack of independent validation.

### *Research Priorities and Future Directions*

- Large, multicenter RCTs assessing real-world clinical outcomes (e.g., hospitalizations, mortality, and quality of life)
- Standardization of evaluation metrics, establishing a minimal core set of benchmarks (accuracy, explainability, usability, and safety) for comparative studies
- Implementation science studies to assess real-world feasibility, workflow integration, cost-effectiveness, and regulatory implications
- Data security and privacy evaluations before widespread commercial deployment
- Development of retrieval-augmented architectures, which have already shown superior performance in radiology and other structured domains.[38,64]

## Conclusion

Chatbots and VHAs are emerging as promising tools across multiple healthcare domains, including initial triage, diagnostic support, patient education, remote monitoring, and behavioral interventions. Evidence from 46 studies indicated that LLM-based and retrieval-augmented systems can approach human-level performance in structured tasks, while rule-based platforms generally show lower accuracy and engagement. Chatbots have demonstrated potential to improve patient access, adherence, and engagement, particularly in mental health and chronic disease management, and to support healthcare workflows (e.g., history-taking and follow-up).

However, significant challenges remain, including variability in diagnostic accuracy, limited long-term outcome data, inconsistent evaluation metrics, and usability and health literacy barriers. Current evidence supports the supervised integration of chatbots into clinical workflows rather than autonomous decision-making. Hence, future research should focus on large-scale real-world trials, standardized performance metrics, data privacy, and integration into healthcare systems to ensure safe, effective, and equitable deployment.

In summary, chatbots represent a scalable adjunct to traditional healthcare delivery, capable of enhancing patient care and system efficiency. Nonetheless, their implementation must be guided by evidence, clinician oversight, and patient-centered design.

### Authors' Contribution

**Conceptualization:** Mandana Akbari Marandi, Reza Shoja Ghiass.
**Data curation:** Mandana Akbari Marandi.
**Formal analysis:** Reza Shoja Ghiass.
**Investigation:** Mandana Akbari Marandi, Reza Shoja Ghiass.
**Methodology:** Reza Shoja Ghiass.
**Project administration:** Mandana Akbari Marandi.
**Resources:** Reza Shoja Ghiass.
**Software:** Reza Shoja Ghiass.
**Supervision:** Reza Shoja Ghiass.
**Validation:** Mandana Akbari Marandi, Reza Shoja Ghiass.
**Visualization:** Mandana Akbari Marandi.
**Writing–original draft:** Mandana Akbari Marandi.
**Writing–review & editing:** Reza Shoja Ghiass.

### Competing Interests

None.

### Data Availability Statement

All data generated or analyzed during this review are included in this article.

### Ethical Approval

No ethical approval was required for this systematic review, as it analyzed and synthesized data from previously published studies.

### Funding

This study received no financial support.

## Intelligence Use Disclosure

The authors used Copilot for grammar correction and language editing to improve manuscript readability. All AI-generated language suggestions were reviewed and edited by the authors.

## Supplementary Files

Supplementary file 1 contains Table S1.

## References


1. Udegbe FC, Ebulue OR, Ebulue CC, Ekesiobi CS. The role of artificial intelligence in healthcare: a systematic review of applications and challenges. Int Med Sci Res J. 2024;4(4):500-8. doi: 10.51594/imsrj.v4i4.1052.
2. Tetteh SG, Azupwah L, Agyemana AY, Adjei SK, Twumasi AP, Mohammed-Nurudeen S. Artificial intelligence in healthcare: a systematic review of virtual healthcare assistants. Asian J Probab Stat. 2025;27(7):43-62. doi: 10.9734/ajpas/2025/v27i7776.
3. Damingo LA, Igulu KT, Saturday NR, Asunogie TO, Kizzy EN. Virtual health assistants: AI in patient engagement. In: Singh TP, Kumar CJ, Abraham A, Igulu KT, eds. Revolutionizing Healthcare: Impact of Artificial Intelligence on Diagnosis, Treatment, and Patient Care. Cham: Springer; 2025. p. 243-52. doi: 10.1007/978-3-031-80813-5_16.
4. 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;9(3):e16934. doi: 10.2196/16934.
5. Bates M. Health care chatbots are here to help. IEEE Pulse. 2019;10(3):12-4. doi: 10.1109/mpuls.2019.2911816.
6. Weizenbaum J. ELIZA—a computer program for the study of natural language communication between man and machine. Commun ACM. 1966;9(1):36-45.
7. Owens OL, Felder T, Tavakoli AS, Revels AA, Friedman DB, Hughes-Halbert C, et al. Evaluation of a computer-based decision aid for promoting informed prostate cancer screening decisions among African American men: iDecide. Am J Health Promot. 2019;33(2):267-78. doi: 10.1177/0890117118786866.
8. 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. Nat Lang Eng. 2020;26(2):183-220. doi: 10.1017/S1351324919000329.
9. Tanaka H, Negoro H, Iwasaka H, Nakamura S. Embodied conversational agents for multimodal automated social skills training in people with autism spectrum disorders. PLoS One. 2017;12(8):e0182151. doi: 10.1371/journal.pone.0182151.
10. Zhang H, Zheng J. The application analysis of medical chatbots and virtual assistant. Front Soc Sci Technol. 2021;3(2):11-6. doi: 10.25236/fsst.2021.0302.
11. Madhu D, Jain CN, Sebastain E, Shaji S, Ajayakumar A. A novel approach for medical assistance using trained chatbot. In: 2017 International Conference on Inventive Communication and Computational Technologies (ICICCT). Coimbatore, India: IEEE; 2017. p. 243-6. doi: 10.1109/icicct.2017.7975195.
12. Oliven A, Nave R, Gilad D, Barch A. Implementation of a web-based interactive virtual patient case simulation as a training and assessment tool for medical students. Stud Health Technol Inform. 2011;169:233-7. doi: 10.3233/978-1-60750-806-9-233.
13. Park S, Choi J, Lee S, Oh C, Kim C, La S, et al. Designing a chatbot for a brief motivational interview on stress management: qualitative case study. J Med Internet Res. 2019;21(4):e12231. doi: 10.2196/12231.
14. Burri SR, Ghorpade VV, Dutt V, Lipi K. The rise of virtual health assistants: chatbot-based healthcare support and counseling using recurrent neural networks (RNNs). In: 2023 3rd International Conference on Technological Advancements in Computational Sciences (ICTACS). Tashkent, Uzbekistan: IEEE; 2023. p. 811-6. doi: 10.1109/ictacs59847.2023.10390207.
15. Curtis RG, Bartel B, Ferguson T, Blake HT, Northcott C, Virgara R, et al. Improving user experience of virtual health assistants: scoping review. J Med Internet Res. 2021;23(12):e31737. doi: 10.2196/31737.
16. Maher C, Singh B, Wylde A, Chastin S. Virtual health assistants: a grand challenge in health communications and behavior change. Front Digit Health. 2024;6:1418695. doi: 10.3389/fdgth.2024.1418695.
17. Abd-Alrazaq AA, Alajlani M, Ali N, Denecke K, Bewick BM, Househ M. Perceptions and opinions of patients about mental health chatbots: scoping review. J Med Internet Res. 2021;23(1):e17828. doi: 10.2196/17828.
18. Gahelot P, Sarangi PK, Saxena M, Jha J, Vajpayee A, Sahoo AK. Hog features based handwritten Bengali numerals recognition using SVM classifier: a comparison with Hopfield implementation. In: 2022 IEEE International Conference on Current Development in Engineering and Technology (CCET). Bhopal, India: IEEE; 2022. p. 1-6. doi: 10.1109/ccet56606.2022.10080015.
19. Dolianiti F, Tsoupouroglou I, Antoniou P, Konstantinidis S, Anastasiades S, Bamidis P. Chatbots in healthcare curricula: the case of a conversational virtual patient. In: International Conference on Brain Function Assessment in Learning. Cham: Springer International Publishing; 2020. p. 137-47. doi: 10.1007/978-3-030-60735-7_15.
20. Branley-Bell D, Brown R, Coventry L, Sillence E. Chatbots for embarrassing and stigmatizing conditions: could chatbots encourage users to seek medical advice? Front Commun. 2023;8:1275127. doi: 10.3389/fcomm.2023.1275127.
21. Eton DT, Ridgeway JL, Linzer M, Boehm DH, Rogers EA, Yost KJ, et al. Healthcare provider relational quality is associated with better self-management and less treatment burden in people with multiple chronic conditions. Patient Prefer Adherence. 2017;11:1635-46. doi: 10.2147/ppa.S145942.
22. Harari R, Al-Taweel A, Ahram T, Shokoohi H. Explainable AI and augmented reality in transesophageal echocardiography (TEE) imaging. In: 2024 IEEE International Conference on Artificial Intelligence and eXtended and Virtual Reality (AIxVR). Los Angeles, CA: IEEE; 2024. p. 306-9. doi: 10.1109/AIxVR59861.2024.00051.
23. Siontis KC, Attia ZI, Asirvatham SJ, Friedman PA. ChatGPT hallucinating: can it get any more humanlike? Eur Heart J. 2024;45(5):321-3. doi: 10.1093/eurheartj/ehad766.
24. Singh B, Olds T, Brinsley J, Dumuid D, Virgara R, Matricciani L, et al. Systematic review and meta-analysis of the effectiveness of chatbots on lifestyle behaviours. NPJ Digit Med. 2023;6(1):118. doi: 10.1038/s41746-023-00856-1.
25. Bickmore TW, Pfeifer LM, Byron D, Forsythe S, Henault LE, Jack BW, et al. Usability of conversational agents by patients with inadequate health literacy: evidence from two clinical trials. J Health Commun. 2010;15 Suppl 2:197-210. doi: 10.1080/10810730.2010.499991.
26. Zhang Z, Bickmore T. Medical shared decision making with a virtual agent. In: Proceedings of the 18th International Conference on Intelligent Virtual Agents. New York: Association for Computing Machinery; 2018. p. 113-8. doi: 10.1145/3267851.3267883.
27. You Y, Gui X. Self-diagnosis through AI-enabled chatbot-based symptom checkers: user experiences and design considerations. AMIA Annu Symp Proc. 2020;2020:1354-63.
28. Ye BJ, Kim JY, Suh C, Choi SP, Choi M, Kim DH, et al. Development of a chatbot program for follow-up management of workers' general health examinations in Korea: a pilot study. Int J Environ Res Public Health. 2021;18(4):2170. doi: 10.3390/ijerph18042170.

29. Melnik T, Thompson JA, Vasilakes J, Annis T, Zhou S, Schutte D, et al. Semi-automated clinical content curation of COVID-19 chatbot remote patient monitoring solution. AMIA Annu Symp Proc. 2022;2022:756-65.
30. Ben-Shabat N, Sharvit G, Meimis B, Ben Joya D, Sloma A, Kiderman D, et al. Assessing data gathering of chatbot based symptom checkers - a clinical vignettes study. Int J Med Inform. 2022;168:104897. doi: 10.1016/j.ijmedinf.2022.104897.
31. Greer S, Ramo D, Chang YJ, Fu M, Moskowitz J, Haritatos J. Use of the chatbot "vivibot" to deliver positive psychology skills and promote well-being among young people after cancer treatment: randomized controlled feasibility trial. JMIR Mhealth Uhealth. 2019;7(10):e15018. doi: 10.2196/15018.
32. Luo MX, Lyle A, Bennett P, Albertson D, Sirohi D, Maughan BL, et al. Artificial intelligence chatbot vs pathology faculty and residents: real-world clinical questions from a genitourinary treatment planning conference. Am J Clin Pathol. 2024;162(6):541-3. doi: 10.1093/ajcp/aqae078.
33. Chen CC, Wei CJ, Tseng TY, Chiu MC, Chang CC. Applying object detection and large language model to establish a smart telemedicine diagnosis system with chatbot: a case study of pressure injuries diagnosis system. Telemed J E Health. 2024;30(6):e1705-12. doi: 10.1089/tmj.2023.0715.
34. Gur Ozcan SG, Erkan M. Reliability and quality of information provided by artificial intelligence chatbots on post-contrast acute kidney injury: an evaluation of diagnostic, preventive, and treatment guidance. Rev Assoc Med Bras (1992). 2024;70(11):e20240891. doi: 10.1590/1806-9282.20240891.
35. Sharp G, Dwyer B, Xie J, McNaney R, Shrestha P, Prawira C, et al. Co-design of a single session intervention chatbot for people on waitlists for eating disorder treatment: a qualitative interview and workshop study. J Eat Disord. 2025;13(1):46. doi: 10.1186/s40337-025-01225-x.
36. Chen S, Kann BH, Foote MB, Aerts H, Savova GK, Mak RH, et al. Use of artificial intelligence chatbots for cancer treatment information. JAMA Oncol. 2023;9(10):1459-62. doi: 10.1001/jamaoncol.2023.2954.
37. Dronkers EA, Geneid A, Al Yaghchi C, Lechien JR. Evaluating the potential of AI chatbots in treatment decision-making for acquired bilateral vocal fold paralysis in adults. J Voice. 2025;39(4):871-81. doi: 10.1016/j.jvoice.2024.02.020.
38. Rau S, Rau A, Nattenmüller J, Fink A, Bamberg F, Reisert M, et al. A retrieval-augmented chatbot based on GPT-4 provides appropriate differential diagnosis in gastrointestinal radiology: a proof of concept study. Eur Radiol Exp. 2024;8(1):60. doi: 10.1186/s41747-024-00457-x.
39. Lyons RJ, Arepalli SR, Fromal O, Choi JD, Jain N. Artificial intelligence chatbot performance in triage of ophthalmic conditions. Can J Ophthalmol. 2024;59(4):e301-8. doi: 10.1016/j.jcjo.2023.07.016.
40. Kring T, Prasad S, Dadi S, Sokhn E, Franzmann E. A comparison of quality and readability of artificial intelligence chatbots in triage for head and neck cancer. Am J Otolaryngol. 2025;46(5):104710. doi: 10.1016/j.amjoto.2025.104710.
41. Schumacher I, Ferro Desideri L, Bühler VM, Sagurski N, Subhi Y, Bhardwaj G, et al. Performance analysis of an emergency triage system in ophthalmology using a customized chatbot. Digit Health. 2025;11:20552076251320298. doi: 10.1177/20552076251320298.
42. Rathnayaka P, Mills N, Burnett D, De Silva D, Alahakoon D, Gray R. A mental health chatbot with cognitive skills for personalised behavioural activation and remote health monitoring. Sensors (Basel). 2022;22(10):3653. doi: 10.3390/s22103653.
43. Erkan A, Koc A, Barali D, Satir A, Zengin S, Kilic M, et al. Can patients with urogenital cancer rely on artificial intelligence chatbots for treatment decisions? Clin Genitourin Cancer. 2024;22(6):102206. doi: 10.1016/j.clgc.2024.102206.
44. Ali SR, Dobbs TD, Whitaker IS. Using a chatbot to support clinical decision-making in free flap monitoring. J Plast Reconstr Aesthet Surg. 2022;75(7):2387-440. doi: 10.1016/j.bjps.2022.04.072.
45. Singh A, Schooley B, Patel N. Effects of user-reported risk factors and follow-up care activities on satisfaction with a COVID-19 chatbot: cross-sectional study. JMIR Mhealth Uhealth. 2023;11:e43105. doi: 10.2196/43105.
46. Stephens TN, Joerin A, Rauws M, Werk LN. Feasibility of pediatric obesity and prediabetes treatment support through Tess, the AI behavioral coaching chatbot. Transl Behav Med. 2019;9(3):440-7. doi: 10.1093/tbm/ibz043.
47. Qazi F, Shaheen O, Andrabi WI, Arif M, Begum F, Mansoor M. Evaluating the incidence of co-existing injuries in anterior talofibular ligament injuries a magnetic resonance imaging study: co-existing injuries in anterior talofibular ligament. Pak J Health Sci. 2025;6(3):115-20. doi: 10.54393/pjhs.v6i3.2667.
48. Hasei J, Hanzawa M, Nagano A, Maeda N, Yoshida S, Endo M, et al. Empowering pediatric, adolescent, and young adult patients with cancer utilizing generative AI chatbots to reduce psychological burden and enhance treatment engagement: a pilot study. Front Digit Health. 2025;7:1543543. doi: 10.3389/fdgth.2025.1543543.
49. Fan X, Chao D, Zhang Z, Wang D, Li X, Tian F. Utilization of self-diagnosis health chatbots in real-world settings: case study. J Med Internet Res. 2021;23(1):e19928. doi: 10.2196/19928.
50. Washington CJ, Abouyared M, Karanth S, Braithwaite D, Birkeland A, Silverman DA, et al. The use of chatbots in head and neck mucosal malignancy treatment recommendations. Otolaryngol Head Neck Surg. 2024;171(4):1062-8. doi: 10.1002/ohn.818.
51. Shapiro J, Lyakhovitsky A. Revolutionizing teledermatology: exploring the integration of artificial intelligence, including generative pre-trained transformer chatbots for artificial intelligence-driven anamnesis, diagnosis, and treatment plans. Clin Dermatol. 2024;42(5):492-7. doi: 10.1016/j.clindermatol.2024.06.020.
52. Guede-Fernández F, Silva Pinto T, Semedo H, Vital C, Coelho P, Oliosi ME, et al. Enhancing postoperative anticoagulation therapy with remote patient monitoring: a pilot crossover trial study to evaluate portable coagulometers and chatbots in cardiac surgery follow-up. Digit Health. 2024;10:20552076241269515. doi: 10.1177/20552076241269515.
53. Habicht J, Viswanathan S, Carrington B, Hauser TU, Harper R, Rollwage M. Closing the accessibility gap to mental health treatment with a personalized self-referral chatbot. Nat Med. 2024;30(2):595-602. doi: 10.1038/s41591-023-02766-x.
54. Ejaz H, Ali SR, Berner JE, Dobbs TD, Whitaker IS. The "flapbot": a global perspective on the validity and usability of a flap monitoring chatbot. J Reconstr Microsurg. 2025;41(3):227-36. doi: 10.1055/a-2355-3970.
55. Jungmann SM, Klan T, Kuhn S, Jungmann F. Accuracy of a chatbot (Ada) in the diagnosis of mental disorders: comparative case study with lay and expert users. JMIR Form Res. 2019;3(4):e13863. doi: 10.2196/13863.
56. Coelho MS, Piva GB, Vasconcelos RA, Toia CC, Santos Zambon L, Brenelli S. Chatbot versus lecture in the teaching of endodontic diagnosis for undergraduate students-a pilot study. J Dent Educ. 2025:e13940. doi: 10.1002/jdd.13940.
57. Sharp G, Dwyer B, Randhawa A, McGrath I, Hu H. The effectiveness of a chatbot single-session intervention for people on waitlists for eating disorder treatment: randomized controlled trial. J Med Internet Res. 2025;27:e70874. doi: 10.2196/70874.
58. Zhang S, Song J. A chatbot based question and answer system for the auxiliary diagnosis of chronic diseases based on large language model. Sci Rep. 2024;14(1):17118. doi: 10.1038/

s41598-024-67429-4.
59. Pushpanathan K, Lim ZW, Er Yew SM, Chen DZ, Hui'En Lin HA, Lin Goh JH, et al. Popular large language model chatbots' accuracy, comprehensiveness, and self-awareness in answering ocular symptom queries. iScience. 2023;26(11):108163. doi: 10.1016/j.isci.2023.108163.
60. Guler R, Yalcin E. Performance of AI chatbots in preliminary diagnosis of maxillofacial pathologies. Med Sci Monit. 2025;31:e949076. doi: 10.12659/msm.949076.
61. Rebelo N, Sanders L, Li K, Chow JCL. Learning the treatment process in radiotherapy using an artificial intelligence-assisted chatbot: development study. JMIR Form Res. 2022;6(12):e39443. doi: 10.2196/39443.
62. Giammanco PA, Collins CE, Zimmerman J, Kricfalusi M, Rice RC, Trumbo M, et al. Evaluating the quality and readability of information provided by generative artificial intelligence chatbots on clavicle fracture treatment options. Cureus. 2025;17(1):e77200. doi: 10.7759/cureus.77200.
63. DeFrancisis JS, Richa P, Oar D, Henwood L, Buchman ZJ, Grewal G. Assessing the accuracy of artificial intelligence chatbots in the diagnosis and management of meniscal tears. Cureus. 2025;17(5):e84124. doi: 10.7759/cureus.84124.
64. Grinberg N, Whitefield S, Kleinman S, Ianculovici C, Wasserman G, Peleg O. Assessing the performance of an artificial intelligence based chatbot in the differential diagnosis of oral mucosal lesions: clinical validation study. Clin Oral Investig. 2025;29(4):188. doi: 10.1007/s00784-025-06268-7.
65. Wolmer S, Shauly O. Evaluating plastic surgery chatbot performance: insights into medical triage, classification accuracy and escalation trends. Aesthet Surg J. 2025. doi: 10.1093/asj/sjaf123.
66. Sarbay İ, Berikol GB, Özturan İ U. Performance of emergency triage prediction of an open access natural language processing based chatbot application (ChatGPT): a preliminary, scenario-based cross-sectional study. Turk J Emerg Med. 2023;23(3):156-61. doi: 10.4103/tjem.tjem_79_23.
67. Tsui JC, Wong MB, Kim BJ, Maguire AM, Scoles D, VanderBeek BL, et al. Appropriateness of ophthalmic symptoms triage by a popular online artificial intelligence chatbot. Eye (Lond). 2023;37(17):3692-3. doi: 10.1038/s41433-023-02556-2.
68. Tan TC, Roslan NE, Li JW, Zou X, Chen X, Santosa A. Patient acceptability of symptom screening and patient education using a chatbot for autoimmune inflammatory diseases: survey study. JMIR Form Res. 2023;7:e49239. doi: 10.2196/49239.
69. de Moura JD, Fontana CE, da Silva Lima VH, de Souza Alves I, de Melo Santos PA, de Almeida Rodrigues P. Comparative accuracy of artificial intelligence chatbots in pulpal and periradicular diagnosis: a cross-sectional study. Comput Biol Med. 2024;183:109332. doi: 10.1016/j.compbiomed.2024.109332.
70. Piau A, Crissey R, Brechemier D, Balardy L, Nourhashemi F. A smartphone chatbot application to optimize monitoring of older patients with cancer. Int J Med Inform. 2019;128:18-23. doi: 10.1016/j.ijmedinf.2019.05.013.
71. Ghosh S, Bhatia S, Bhatia A. Quro: facilitating user symptom check using a personalised chatbot-oriented dialogue system. In: Connecting the System to Enhance the Practitioner and Consumer Experience in Healthcare. IOS Press; 2018. p. 51-6. doi: 10.3233/978-1-61499-890-7-51.
72. Hirosawa T, Harada Y, Yokose M, Sakamoto T, Kawamura R, Shimizu T. Diagnostic accuracy of differential-diagnosis lists generated by generative pretrained transformer 3 chatbot for clinical vignettes with common chief complaints: a pilot study. Int J Environ Res Public Health. 2023;20(4):3378. doi: 10.3390/ijerph20043378.
73. Geoghegan L, Scarborough A, Wormald JC, Harrison CJ, Collins D, Gardiner M, et al. Automated conversational agents for post-intervention follow-up: a systematic review. BJS Open. 2021;5(4):zrab070. doi: 10.1093/bjsopen/zrab070.
74. Gaffney H, Mansell W, Tai S. Conversational agents in the treatment of mental health problems: mixed-method systematic review. JMIR Ment Health. 2019;6(10):e14166. doi: 10.2196/14166.
75. Yang Q, Cheung K, Zhang Y, Zhang Y, Qin J, Xie YJ. Conversational agents in physical and psychological symptom management: a systematic review of randomized controlled trials. Int J Nurs Stud. 2025;163:104991. doi: 10.1016/j.ijnurstu.2024.104991.
76. Hindelang M, Sitaru S, Zink A. Transforming health care through chatbots for medical history-taking and future directions: comprehensive systematic review. JMIR Med Inform. 2024;12:e56628. doi: 10.2196/56628.
77. Milne-Ives M, de Cock C, Lim E, Shehadeh MH, de Pennington N, Mole G, et al. The effectiveness of artificial intelligence conversational agents in health care: systematic review. J Med Internet Res. 2020;22(10):e20346. doi: 10.2196/20346.