Transforming Healthcare through Chatbots for Medical History-Taking and Future Directions: A Comprehensive Systematic Review

Michael Hindelang, Sebastian Sitaru, Alexander Zink

Submitted to: JMIR Medical Informatics
on: January 22, 2024

Table of Contents

Transforming Healthcare through Chatbots for Medical History-Taking and Future Directions: A Comprehensive Systematic Review

Michael Hindelang[1, 2, 3]; Sebastian Sitaru[1]; Alexander Zink[1]

[1]Technical University of Munich, TUM School of Medicine and Health, Department of Dermatology and Allergy Munich DE
[2]Pettenkofer School of Public Health Munich DE
[3]Institute for Medical Information Processing, Biometry and Epidemiology (IBE), Faculty of Medicine Ludwig-Maximilian University, LMU Munich DE

Corresponding Author:
Michael Hindelang
Technical University of Munich, TUM School of Medicine and Health, Department of Dermatology and Allergy
Biedersteiner str. 22
Munich
DE

Abstract

Background: The integration of artificial intelligence and chatbot technology in healthcare has attracted significant attention due to its potential to improve patient care and streamline history-taking. As AI-driven conversational agents, chatbots offer the opportunity to revolutionise history-taking, necessitating a comprehensive examination of their impact on medical practice.

Objective: This systematic review comprehensively assesses the role, efficacy, usability, and patient acceptance of chatbots for healthcare history-taking. It also explores potential challenges and future opportunities for integration into clinical practice.

Methods: This systematic review includes 18 studies and focuses on chatbots for healthcare history-taking to support diagnosis and treatment decisions by capturing detailed patient information. All study designs, except conference papers, were eligible to evaluate the feasibility, acceptability, and effectiveness of chatbot-based history-taking. A systematic search included PubMed, Embase, Medline (via Ovid), CENTRAL, Scopus, and Open Science and covered studies through August 2023. The quality of observational studies was classified using the STROBE criteria, while the RoB 2 tool assessed areas of bias in randomised clinical trials (RCTs).

Results: The review included 15 observational studies and 3 randomised clinical trials (RCTs) and synthesised evidence from different medical fields and populations. Chatbots systematically collect information through targeted queries and data retrieval, improving patient engagement and satisfaction. They also demonstrated the potential to improve healthcare efficiency and accessibility through automated data collection 24/7.

Conclusions: This systematic review provides critical insights into the potential benefits and challenges of using chatbots for history-taking. Chatbots can potentially increase patient engagement, streamline data collection, and improve healthcare decision-making. However, the limitations of the studies stem from the different designs and methodological variations, which limits the validity of the results. Clinical Trial: PROSPERO database for systematic reviews, registration number: CRD42023410312

(JMIR Preprints 22/01/2024:56628)
DOI: https://doi.org/10.2196/preprints.56628

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✓ **Please make my preprint PDF available to anyone at any time (recommended).**
Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.
Only make the preprint title and abstract visible.
No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**

Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v
Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http

Original Manuscript

Transforming Healthcare through Chatbots for Medical History-Taking and Future Directions: A Comprehensive Systematic Review

Michael Hindelang[1,2,3], Sebastian Sitaru[1], Alexander Zink[1,4]

[1] Technical University of Munich, School of Medicine and Health, Department of Dermatology and Allergy, Munich, Germany

[2] Pettenkofer School of Public Health, Munich, Germany

[3] Institute for Medical Information Processing, Biometry and Epidemiology (IBE), Faculty of Medicine, Ludwig-Maximilian University, LMU Munich

[4] Division of Dermatology and Venereology, Department of Medicine Solna, Karolinska Institute, Stockholm, Sweden

Corresponding author
Michael Hindelang
Department of Dermatology and Allergy
School of Medicine
Technical University of Munich
Biedersteiner Str 29
Munich, 80802
Tel 08941403061
Germany
Email: Michael.hindelang@tum.de

Abstract count
Word count: 295 (Abstract), 2817 (Text w/o tables); Figure count 6; Table count: 2,

Keywords
Chatbots in Healthcare; Artificial Intelligence; Systematic Review; Medical History-Taking; Patient Engagement; Usability of Healthcare Technology; Patient Data Security; Natural Language Processing (NLP); Healthcare Access; AI Ethical Considerations

Degress
Alexander Zink, MD, MPH, PhD
Sebastian Sitaru, MD
Michael Hindelang, M.Sc. Public Health, Phd Candidate – Medical Research

Declarations

Availability of data and materials

All data generated or analysed during this study are included in this published article.

All aggregate data collected for this review are available from the corresponding

author upon reasonable request.

Registration and protocol

PROSPERO, registration number: CRD42023410312

Conflict of interest

The Authors declare no competing interests.

Authors' contributions

M.H. conceptualised and designed the analysis, collected the data, performed the screening and analysis, and was the primary author of the article. S.S. served as the second reviewer for screening and quality appraisal. A.Z. critically reviewed and provided feedback on the paper.

Abstract

Background

The integration of artificial intelligence and chatbot technology in healthcare has attracted significant attention due to its potential to improve patient care and streamline history-taking. As AI-driven conversational agents, chatbots offer the opportunity to revolutionise history-taking, necessitating a comprehensive examination of their impact on medical practice.

Objectives

This systematic review aims to assess the role, effectiveness, usability, and patient acceptance of chatbots in medical history taking. It also examines potential challenges and future opportunities for integration into clinical practice.

Methods

A systematic search included PubMed, Embase, Medline (via Ovid), CENTRAL, Scopus, and Open Science and covered studies through July 2024. The inclusion and exclusion criteria for the studies reviewed were based on the PICOS framework. The population included individuals using healthcare chatbots for medical history-taking. Interventions focused on chatbots designed to facilitate medical history-taking. Chatbots designed only as 'symptom-checkers' or used in non-healthcare contexts were excluded. However, if a study with a 'symptom-checker' provided more comprehensive functionality than only checking symptoms, such as facilitating detailed medical history-taking, it was included. The outcomes of interest were the feasibility, acceptance, and usability of chatbot-based medical history taking. Studies not reporting on these outcomes were excluded. All study designs except conference papers were eligible for inclusion. Only English-language articles were considered. There were no specific restrictions on study duration. Key search terms included "Chatbot*", "conversational agent*", "virtual assistant", "artificial intelligence chatbot", "medical history", and "history-taking". The quality of observational studies was classified

using the STROBE criteria (e.g. sample size, design, data collection, follow-up). The RoB 2 tool assessed areas and the levels of bias in randomised clinical trials (RCTs).

Results

The review included 15 observational studies and 3 RCTs and synthesised evidence from different medical fields and populations. Chatbots systematically collect information through targeted queries and data retrieval, improving patient engagement and satisfaction. The results show that chatbots have great potential for history taking and that the efficiency and accessibility of the healthcare system can be improved by 24/7 automated data collection. Bias assessments revealed mixed methodological quality. Among the 16 observational studies, 6 (37.5%) had a high quality. 5 studies (31.25%) had moderate and another 5 (31.25%) had a low quality. For the RCTs, 1 exhibited a high risk due to lack of blinding and incomplete outcome data, highlighting significant biases in study design and execution.

Conclusion

This systematic review provides critical insights into the potential benefits and challenges of using chatbots for medical history taking. The included studies showed that chatbots can increase patient engagement, streamline data collection, and improve healthcare decision-making. For effective integration into clinical practice, it is crucial to design user-friendly interfaces, ensure robust data security, and maintain empathetic patient-physician interactions. Future research should focus on refining chatbot algorithms, improving their emotional intelligence, and extending their application to different healthcare settings to realise their full potential in modern medicine.

Funding

The systematic review was funded by the Department of Dermatology and Allergology, Technical University of Munich, Germany. Funding did not influence the review process or results.

Introduction

Medical history taking is an important aspect of patient care and crucial for accurate diagnosis and treatment planning in healthcare. Complete patient information is key for informed clinical decisions, and this information needs to be accurate and comprehensive [1]. This is done through face-to-face interviews or paper-based questionnaires, which can have inherent limitations in terms of efficiency and accuracy and may result in low patient participation [2]. Technological developments have led to a greater exploration of innovative alternatives, such as chatbots, which have the potential to transform the process of taking historical records [3]. Chatbots are computer programs that mimic human-like communication using artificial intelligence (AI) and natural language processing (NLP) [4], [5], [6]. The integration of AI and chatbot technology in healthcare promises significant improvements in patient care [7], [8]. It can increase the accuracy and efficiency of data collection. This can lead to better diagnostic results, more personalised treatment plans and greater patient involvement. Despite the widespread use in other sectors such as entertainment, gaming and customer service [9] as well as in security systems and emergency communications [10], [11], [12], there is a notable lack of comprehensive systematic reviews assessing the effectiveness, usability and patient acceptability of chatbots specifically for medical history taking. Research to date has primarily focussed on niche applications of chatbots without providing a holistic view of their broader potential.

In the past, the proliferation of sophisticated AI systems was limited due to their high cost and complexity, making them accessible only to a few. With the development of new accessible AI models and large language models (LLM) such as ChatGPT [13], [14], [15], [16], these capabilities are now being made available to a wide audience.

LLMs can analyse large amounts of medical literature and patient data in seconds, supporting clinical decisions that are both data-driven and potentially more accurate than those made under

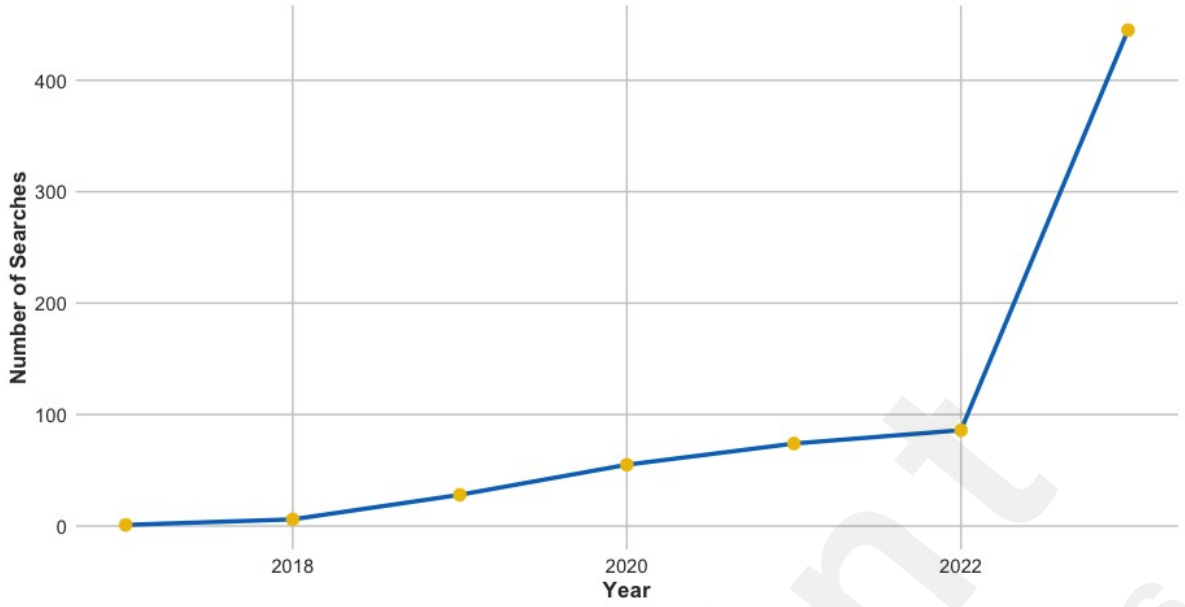
time pressure [17]. They can also enable more personalized medicine by adjusting interactions based on the analysis of individual patient data, which can lead to better treatment outcomes [18]. In addition, their ability to operate continuously can improve healthcare by ensuring that expert-level advice is always available, improving access to high-quality care, particularly in underserved areas [18], [19]. However, these benefits must be balanced by robust measures to ensure that the use of AI in healthcare improves, rather than undermines, patient care and trust [20].

Despite the promise of chatbots, important considerations must be made, particularly in healthcare. Cybersecurity is paramount, as chatbots handle sensitive medical information that must be protected from unauthorised access or data breaches [21], [22]. Furthermore, despite the remarkable capabilities of chatbots in effectively processing and generating responses through predefined algorithms, they often lack the empathetic understanding and emotional intelligence inherent in human interactions [23]. This limitation can affect relationship-building and patient trust, especially during sensitive medical conversations [20].

Recent data highlighted the growing interest in the interplay between chatbots and medicine. An analysis of articles from the first article in 2017 to 2024 with the search query "chatbot*" AND “medicine" shows a significant increase, especially in 2022, with the trend rising from a single article in 2017 to 445 in 2023 (Figure 1).

Figure 1 Number of articles over the last years: chatbot* AND medicine

Legend 1 This chart shows the increasing trend in publications on chatbots in medicine from 2017 to 2023, with a notable rise in 2022, indicating growing research interest and advancements in AI and healthcare.



Chatbots rely on advanced algorithms and AI-supported natural language processing for their technical function. These techniques enable chatbots to analyse user input, provide relevant information in response and adapt their interactions based on context and user behaviour [24], [25], [26]. Machine learning techniques, which include data-driven learning and pattern recognition, can be used to improve the effectiveness and accuracy of chatbots.

Considering the potential benefits and difficulties associated with chatbots, thorough research is needed to assess their impact on the medical history-taking process. Although there is already research that has explored the practicality and acceptability of chatbots in specific healthcare settings, such as mental health or genetic counselling, a systematic literature review is required for a comprehensive understanding of chatbot-based history-taking [27], [28], [29].

The primary objective of this systematic review is to comprehensively assess the role, efficacy, usability, and patient acceptance of chatbots in medical history taking. This systematic review also aims to explore the impact and future directions of integrating chatbots into clinical settings by assessing data accuracy, patient interaction levels, healthcare provider efficiency and patient outcomes. The use of chatbots could transform the medical history-taking process by supporting the accurate collection of patient data. In addition, this has the potential to increase productivity and

improve the quality and delivery of healthcare services.

Methods

The systematic review was conducted according to the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines [30]. The protocol was registered in the PROSPERO database with the registration number CRD42023410312 (National Institute for Health Research (NHS)). [32]

Eligibility Criteria

The selection criteria for the study were based on the PICOS (participants, interventions, comparators, outcomes, study design) [33]. The studies included were investigating healthcare chatbots, with the interventions focusing on history-taking facilitated by a chatbot. The scope was restricted to chatbots aiming to collect medical histories and detailed patient information, aiding medical professionals in accurate diagnosis and treatment decisions. In contrast, chatbots designed exclusively as 'symptom-checkers,' such as standalone apps providing rapid assessments and potential diagnoses, were excluded. This exclusion was made to focus on tools that facilitate comprehensive medical history-taking rather than immediate symptom-based advice. There were no limitations on the modality of chatbot input and output. The comparators were not subjected to any specific restrictions. The outcomes of interest included the feasibility, acceptability, and efficacy of chatbot-based history-taking interventions. There were no restrictions on study design, except for conference papers, which were excluded to ensure the inclusion of studies with rigorous peer review and substantial data reporting. The review was limited to English-language articles because resources were limited.

Information Sources

PubMed, CENTRAL, Embase, Medline (through Ovid), Scopus, and Open Science were searched to identify relevant studies. In addition, reference lists of relevant articles were screened manually.

Search Strategy

For each database, we developed a search strategy that included keywords, subject headings, mesh terms (in PubMed), filters and restrictions to find relevant studies. The search terms focused on chatbots, anamnesis, history-taking, and related concepts: ("Chatbot*" OR "conversational agent*" OR "chatterbot*" OR "virtual assistant" OR "intelligent virtual agent" OR "artificial intelligence chatbot" OR "AI chatbot" OR "conversational AI" OR "dialogue system") AND ("anamnesis" OR "medical history" OR "history-taking" OR "medical interview" OR "patient interview" OR "medical questionnaire" OR "patient questionnaire"). The last search was done in July 2024 (Appendix 1). Additionally, a reference list search was conducted.

Selection Process

The selection process was done by two authors (MH, SS) independently screening the titles and abstracts of the identified studies based on the predetermined eligibility criteria. Potentially relevant studies were retrieved in full text and further assessed for eligibility. The full-text assessment was also performed independently (MH, SS). Any disagreements between the two authors were resolved through discussion, focusing on the eligibility criteria and study relevance. If consensus could not be reached, the involvement of a third author (AZ) was sought when necessary.

Data Collection Process

Data from the selected studies were extracted independently (MH, SS) using a data extraction form based on the PICO criteria (STROBE - Strengthening the Reporting of Observational Studies in Epidemiology) [33], [34]. The extracted data included information such as the first author, number of authors, country, year, title of the scientific journal, topics and type of journal, impact factor, and main results focused on history-taking (anamnesis). Additional data collected encompassed study design, setting, sample size, type of participants, female percentage, mean age (range), and results. Outcomes extracted focused on key aspects such as feasibility, acceptability, and efficacy.

When full-text access was unavailable, the corresponding author was contacted by e-mail. Data were visualised using the R-package for creating alluvial diagrams [35]. Any discrepancies in data extraction were resolved through discussion between the two authors (MH and SS),

Quality assessment

The methodological quality of the included observational studies was assessed using the STROBE criteria [34]. Each study was evaluated based on the fulfilment of the STROBE criteria. The studies were categorised into three categories: Category A, if more than 80% of the STROBE criteria were fulfilled; Category B, if 50-80% were met; and Category C if less than 50% of the criteria were fulfilled [36]. For example, Category A studies provided comprehensive details on study objectives, participant selection, and statistical analysis. Category B had adequate but incomplete information. Category C studies frequently lacked critical details such as clear definitions of eligibility criteria or thorough data collection methods.

In addition, the randomised controlled trials (RCTs) included in this review were evaluated for risk of bias using the Risk of Bias tool and the robvis R-package [37], [38]. The RoB 2 tool assesses various domains of bias, including randomisation, allocation concealment, blinding, incomplete outcome data, selective reporting, and other potential sources of bias. The overall risk of bias score was determined for each study based on the number of criteria for high risk of bias met. Studies are considered to have a low risk of bias if no domains are rated as high risk and most domains are rated as low risk. Studies with some concerns in one or more domains but no high-risk ratings are considered to have some concerns. If any domain is rated as high risk, the study is considered to have a high risk of bias.

Software and tools

Data were managed and analyzed using R version 4.2.1. The ggplot2 package [32]was used for data visualization and the robvis R-package was utilized for risk of bias charts [38]. The alluvial R package [35] was used to create alluvial diagrams.

Results

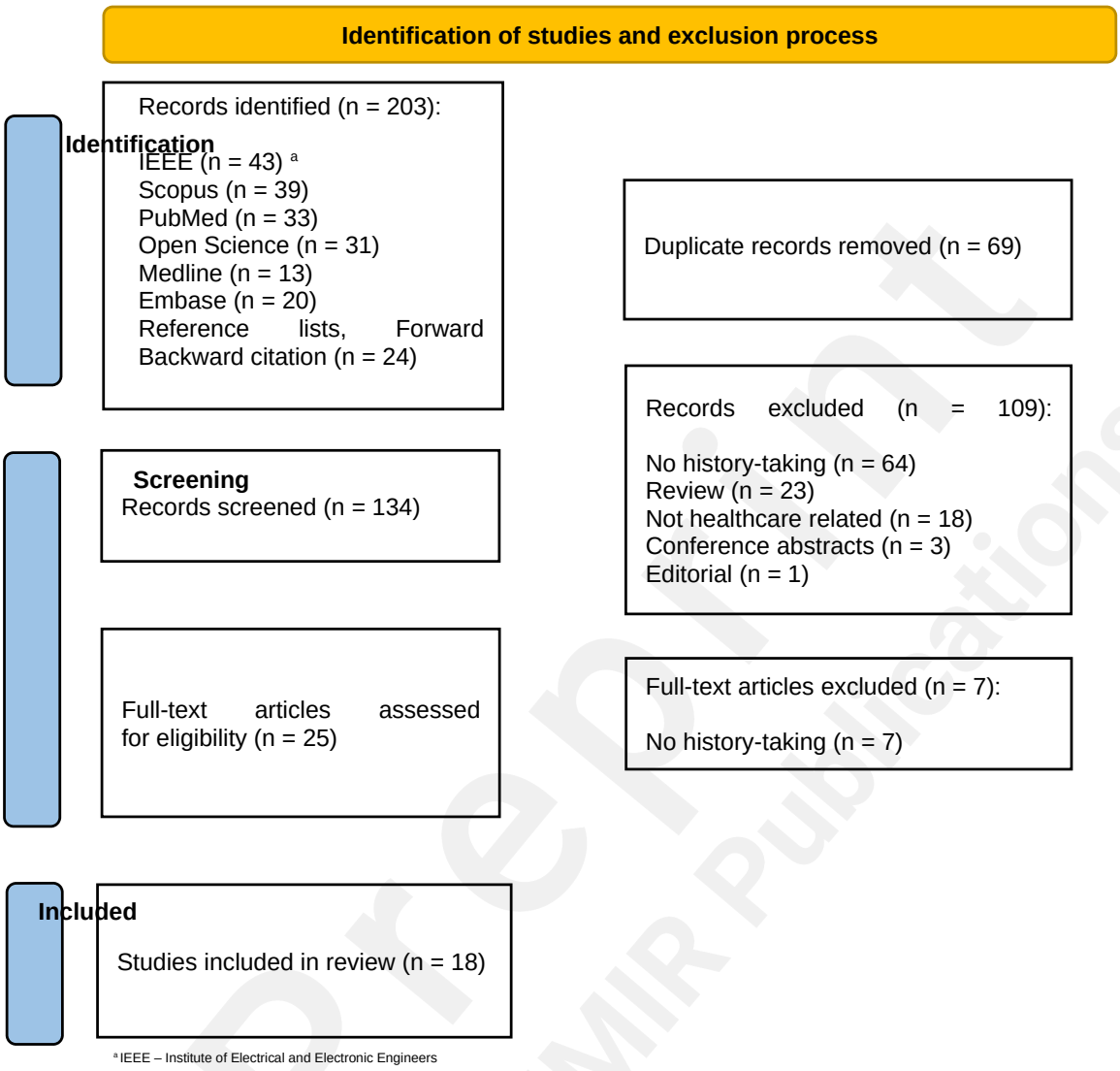
Study selection

The initial literature search yielded 203 records. After removing 69 duplicate articles, a total of 134 unique records were screened based on titles and abstracts. Of these, 109 articles did not meet the eligibility criteria and were excluded. Subsequently, 25 full-text articles were screened, resulting in 18 studies being included in the review (see Figure 2).

Figure 2 Flow chart of the article search and inclusion

Legend 2 This flow chart details the systematic process of selecting studies for the review, starting from 203 records and narrowing down to 18 studies after removing

duplicates and applying eligibility criteria.



Study characteristics

The studies investigated the use of chatbots for history-taking across diverse patient populations and sample sizes (range: n = 5 – 61,070) and were mostly published in scientific health technology journals with varying impact factors (mean: 4·52, range: 0·14 - 14·71), Table 1). The studies employed different research designs, including nine cross-sectional studies, three case-control studies, two observational studies, and three randomised controlled trials (Appendix 2, Tables 1, 2, 3).

Figure 3 Alluvial diagram of the publication date, country, and area of articles

Legend 3 The alluvial diagram illustrates the distribution of studies by year, country, and medical area from 2015 to 2023, highlighting increased publications in 2020 and 2022, with contributions from Germany, the US, and Switzerland across various medical fields.

The alluvial diagram (Figure 3) shows an overview of the literature over time, indicating the year, the country of origin and the medical area of focus for each study. The included studies were published from 2015 to 2023. Most of the studies were published in 2020 and 2022. The included studies (Figure 3, 4) were conducted in Switzerland [39], [40], [41], Germany [42], [43], [44], [45], [46], [47], the United States [27], [29], [48], [49], [50], [51], Australia [28], [52], and New Zealand [53]. The studies cover a diverse range of medical areas: general medicine [39], [42], [43], [52], [53] genetics [28], [29], [49], [50] cancer research [27], [51], family medicine [48], mental health [44], [47], radiology [40], surgery [45], Allergy [46], music therapy [41].

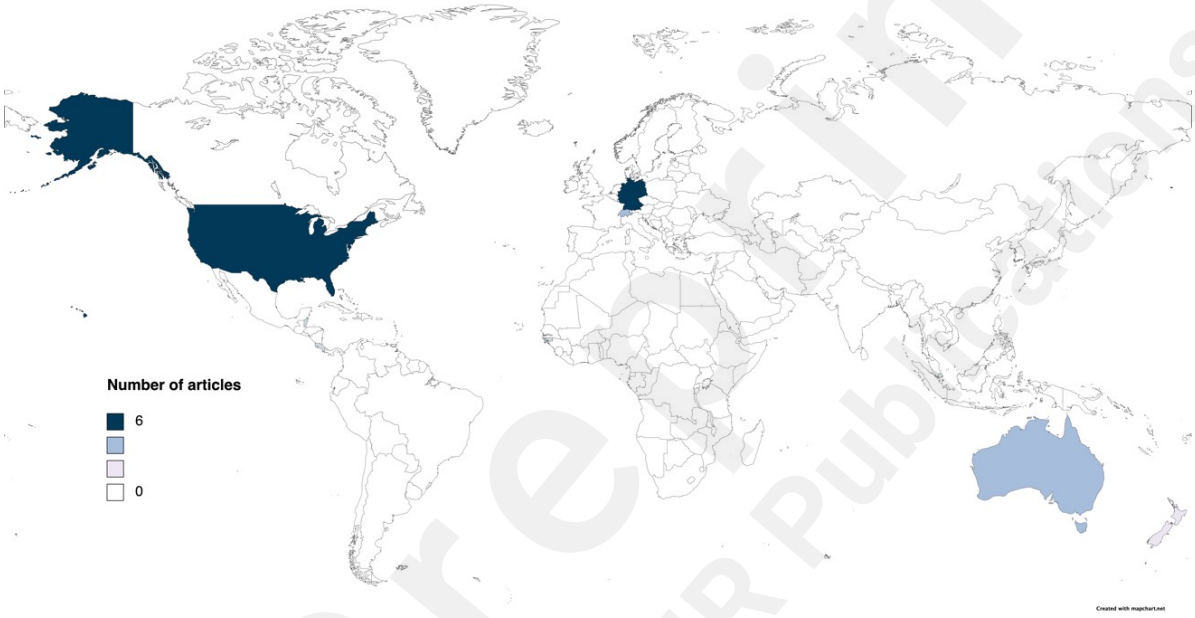


Figure 4 World map showing the number of articles published in each country

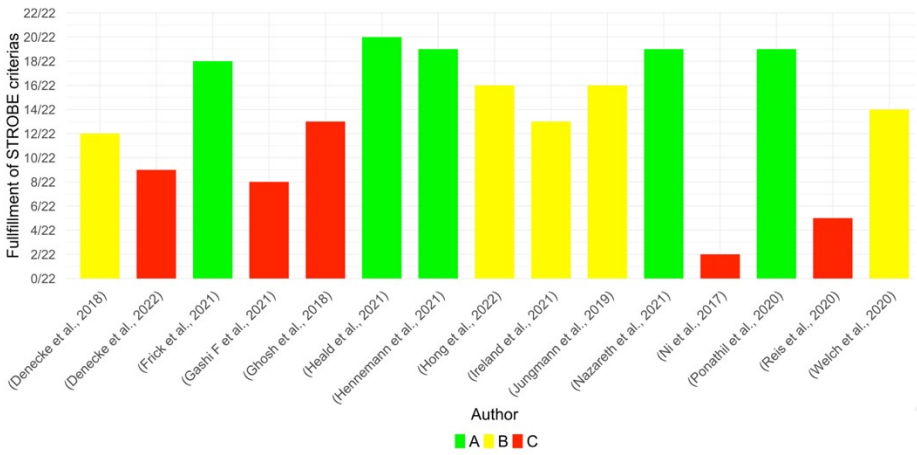
Legend 4 This map shows the geographical distribution of the studies, with most research originating from Germany and the US.

Quality appraisal of the included studies

Among the 16 observational studies, 6 (37·5%) were classified as category A [27], [42], [44], [49], [50], indicating high methodological quality with more than 80% of the STROBE criteria fulfilled (Appendix 1: Quality assessment of the included studies). Five studies (31·25%) were classified as category B [28], [41], [47], [48], [51], meeting 50-80% of the STROBE criteria and five studies (31·25%) were classified as category C [39], [40], [43], [52], [53], meeting less than 50% of the STROBE criteria (Figure 5). The lack of adherence to STROBE criteria in observational studies can have a significant impact on the quality. Missing elements, such as clear definitions of eligibility criteria or participants or detailed methods, lead to biases that reduce validity and reliability. For example, the study of Denecke et al. [39] showed a high risk of selection bias due to a small, non-representative sample and lack of eligibility criteria, limiting the generalizability of their findings. Gashi et al. [39] faced biases from the absence of a control group and unclear eligibility criteria. This could impact the validity of the effectiveness results. Ghosh et al. [52] showed high bias from simulated scenarios without real patient interactions. This could lead to overestimated accuracy and applicability in real-world settings.

Figure 5 Fulfilment of STROBE criteria and categorisation

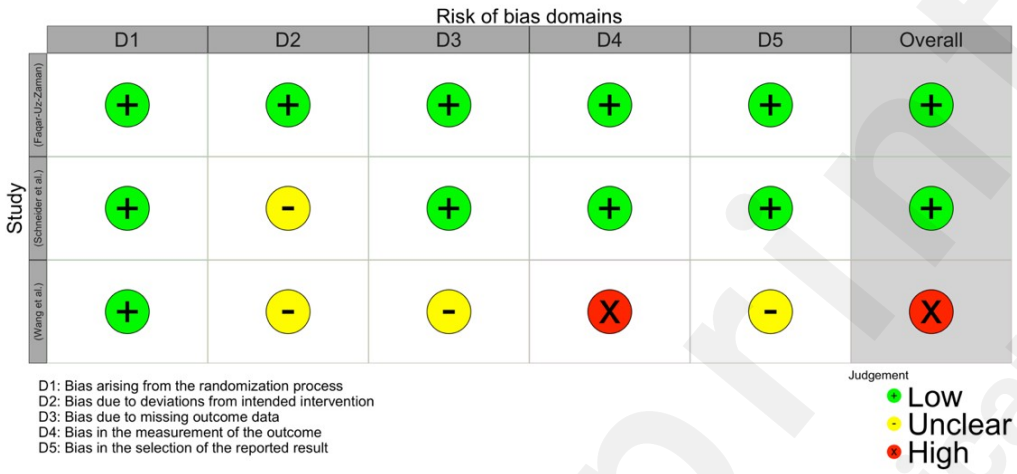
Legend 5 This bar chart categorizes observational studies by their adherence to STROBE criteria, showing 37.5% in high-quality (Category A), and an even split between moderate (Category B) and lower quality (Category C).



The studies of Schneider et al. [46] and Faqar-Uz-Zaman et al. [45] showed a low risk of bias according to the RoB Tool, with detailed methodology and statistical analysis. In contrast, the study of Wang et al. [29] showed a risk of bias due to the absence of intention-to-treat analysis and participants being aware of the intervention (Appendix 3, Figure 6), which could skew results by excluding non-completers and altering participant behavior.

Figure 6 Risk of bias domains (RoB-tool) for RCTs

Legend 6 This chart presents the risk of bias assessment for RCTs using the RoB 2 tool, categorizing studies as "low risk," "some concerns," or "high risk" based on various bias domains.



Summary of Statistical Analyses

The studies included in this systematic review used a variety of statistical methods. Descriptive statistics summarised demographics and usability ratings. Comparative analyses used t-tests and chi-square tests to compare diagnostic accuracy and user engagement. Kappa statistics measured agreement between chatbot and expert diagnoses. Precision and accuracy metrics were assessed using precision, recall and F1 scores. Non-parametric tests such as the Mann-Whitney U-test showed significant reductions in anamnesis duration. Confidence intervals and p-values were reported where relevant to clarify the strength of the evidence.

Usability and User Experience of Chatbots

Five studies focused on the usability and user experience of chatbots in history-taking (Table 2, 3). Denecke et al. [40], [41] found that chatbots were well-received by participants and showed potential for history-taking. Usability scores were high, between 90 and 100 (average 96). Ponathil et al. [50] found that using a voice-controlled assistant interface for taking family health history significantly reduced history-taking duration. Ghosh et al. [52] implemented a medical chatbot that

assists with automated patient pre-assessment through symptom analysis, demonstrating the possibility of avoiding form-based data entry. The chatbot correctly identified at least one of the top three conditions in 83.3% of cases and two out of three conditions in 66.6% of cases. Welch et al. [51] found high engagement and interest in chatbots, suggesting the potential for gathering family health history information at the population level in the US. Of the over 14,000 who participated in the assessment of the study, 54.4% of users went beyond the consent step, and 22.7% completed the full assessment.

Table 1: General characteristics of the included articles

Legend Table 1: This table summarizes the number of authors, countries, and journal topics of the studies, showing most research from Germany and the US, and a focus on Health Informatics and Technology.

Authors	**Numbers of authors**	1-3	4 (22.22%)
		4-6	8 (44.44%)
		>6	6 (33.33%)
	Countries	Germany	6 (33·33%)
		United States	6 (33·33%)
		Switzerland	3 (16·67%)
		Australia	2 (11·11%)
		New Zealand	1 (5·56%)
Scientific journals	**Topics of scientific journals**	Health Informatics and Technology	12 (66·67%)
		Medical Imaging and Radiology	2 (11·11%)

		Genetics and Genetic Counselling	2 (11·11%)
		Surgical Procedures and Techniques	1 (5·56%)
		Mental Health and Psychology	1 (5·56%)

Table 2: Study characteristics

Legend Table 2: This table details study characteristics including author, year, design, sample size, participant type, and key findings, highlighting diverse participant demographics and study outcomes.

Reference		Participants				Methods and result	
First author, year	**Study design**	**n**	**Type of participants**	**Female %**	**Mean age in years (range)**	**Type of measurement**	**Relevant results**
(Deneke et al., 2018)	Cross–sectional study	22	Music therapy patients	41	39 (19–73)	Usability test of the tool and corresponding questionnaire	CUI–based self anamnesis app well-received, potential for collecting anamnesis data.
(Deneke et al., 2022)	Cross–sectional study	5	Radiology patients	40	39·2 (17–73)	System Usability Scale	Digital medical interview assistant with good usability.
(Faqar–Uz–Zaman et al., 2022)	RCT	450	Patients with abdominal pain in ER	52·2	44 (18 – 97)	Accuracy of Diagnosis by ER Doctor and Ada–App According to the Final Diagnosis	Classic patient-physician interaction superior to AI-based tool, but

							AI benefits diagnostic efficacy.
(Frick et al., 2021)	Cross–sectional study	148	German participants	53	33·32 (SD 12·59)	Scales for disclosure and concealment of medical information	Patients prefer disclosing to physicians over chatbots. No significant difference in concealment.
(Gashi et al., 2021)	Cross–sectional study	N/A	N/A	N/A	N/A	N/A	AnCha chatbot improves patient-doctor communication, enhances diagnostic process.
(Ghosh et al., 2018)	Case–control study	30 scenarios	Not specified	N/A	N/A	True positives and false positives, precision	Medical chatbot helps with automated patient preassessment.
(Heald et al., 2021)	Feasibility study	506	Various types of care	58	56·6 (SD 12·5)	Colon Cancer Risk Assessment Tool	Chatbot feasible for increasing genetic screening in at-risk individuals.
(Hennemann et al., 2022)	Observational study	49	Adult patients from an outpatient psychotherapy clinic	61	33·41 (SD 12·79)	Interviews, questionnaires, diagnostic software	Chatbot shows moderate-to-good accuracy for

							condition suggestions.
(Hong et al., 2022)	Cross–sectional study	20	Primary care patients	60	50	Web–based survey	Patients believe chatbot helps clinicians better understand their health.
(Ireland et al., 2021)	Cross–sectional study	83	Adults who had whole exome sequencing for genetic condition diagnosis	53	23·2 – 80·4	Transcript analysis	Chatbot enhances genetic counseling by providing genomic information.
(Jungmann et al., 2019)	Case–control study	6	Psychotherapists, psychology students, and laypersons	50	40 (Therapists) 22 (Students)	Case vignettes, health app comparison	Chatbot shows moderate diagnostic agreement, improvement needed for childhood disorders.
(Nazareth et al., 2021)	Retrospective, observational study	61,070	Women's health	96	N/A	Genetic Testing Results	Chatbot helps identify patients at high risk for hereditary cancer syndromes.
(Ni et al., 2017)	Cross–sectional study/Proof–of–concept	11	Patients with chest pain, respiratory infections, headaches, and dizziness	N/A	N/A	Question accuracy, prediction accuracy	Chatbot generates medical reports with varying

							accuracy based on disease category.
(Ponathil et al., 2020)	Cross–sectional study	50	Adults	50	N/A	NASA Task Load Index workload instrument IBM Usability Questionnaire Technology Acceptance Model questionnaire	Chatbot interface saves time, preferred for collecting family health history.
(Reis et al., 2020)	Case–control study	16	Physicians	35	35·51 years	N/A	Failure of cognitive agent highlights need for managing resistance and transparency.
(Schneider et al., 2023)	RCT	30	Hymenoptera venom allergic patients	N/A	38·93 (SD 12·56)	Standardized questionnaire	Chatbot-supported anamnesis saves time, potential for allergology assessments.
(Wang et al., 2015)	RCT, hospital	70	Majority of patients from underserved populations (low–income families, elders, people with disabilities, and immigrant)	60	Majority in age group 45–54	Interview, questions	Technological support for documenting family history risks is accepted and feasible.

(Welch et al., 2020)	Cross–sectional study	3,204	General population	100	49·4 (SD 7·1)	Standardized questionnaire	Chatbot engages users, potential for gathering family health history at population level.

RCT: Randomized controlled trial
ER: Emergency Room
AI: Artificial intelligence
SD: Standard deviation
CUI: Conversational user interface
N/A: Not applicable

Chatbots and Patient-Doctor Communication

One study highlighted the potential of chatbots to improve patient-doctor communication. Gashi et al. [39] reported that using a chatbot could reduce patient nervousness, allow patients to respond more thoughtfully, and give physicians a more comprehensive picture of the patient's condition.

Diagnostic Accuracy and Efficacy of Chatbots

Nazareth et al. [49] found that a chatbot can help identify high-risk patients for hereditary cancer syndromes. 27.2% of the chatbot users met the criteria for genetic testing, and 5.6% had a pathogenic variant. Ni et al. [53] reported that Mandy, a chatbot, automates history-taking, understands symptoms expressed in natural language, and generates comprehensive reports for further medical investigations, with varying degrees of accuracy depending on the disease category. Hennemann et al. [44] reported that the app-based symptom checker with an AI chatbot showed agreement with therapist diagnoses in 51% of cases for the first condition suggestion and in 69% of cases for the top five condition suggestions. Jungmann et al. [47] tested a health app's diagnostic agreement with case vignettes for mental disorders, pointing to the need for improvement in diagnostic accuracy, especially for mental disorders in childhood and adolescence.

Patient Perceptions and Acceptance of Chatbots

Hong et al. [48] reported that most primary care patients believed that chatbots could help clinicians better understand their health and identify health risks. Ireland et al. [28] found that the development of the Edna tool, an AI-based chatbot that interacts with patients via speech-to-text, signifies progress toward creating digital health processes that are accessible, acceptable, and well-supported, enabling patients to make informed decisions about additional findings. Heald et al. [27] highlighted the feasibility of using chatbots for increasing genetic screening and testing in individuals at risk of hereditary colorectal cancer syndromes.

Challenges and Limitations of Chatbots

Reis et al. [43] noted the importance of managing user resistance and fostering realistic expectations when implementing AI-based history-taking tools. Frick et al. [42] found that patients preferred to disclose medical information to a physician rather than a conversational agent (CA).

Effectiveness on Chatbots

Faqar–Uz–Zaman et al. [45] found that classic patient-physician interaction was superior to an AI-based diagnostic tool applied by patients. However, they also noted that AI tools can benefit clinicians' diagnostic efficacy and improve the quality of care. Schneider et al. [46] found that a chatbot-supported anamnesis could save significant time by 57.3%, in assessing Hymenoptera venom allergies with high completeness (73.3%) and patient satisfaction (75%). Wang et al. [29] demonstrated that technological support for documenting family history risks can be highly accepted, feasible, and effective.

Table 3: Chatbot characteristics

Legend Table 3: This table outlines the chatbots used in the studies, including their name, goa techniques, outcomes, user preference, and challenges, showcasing varied applications and te approaches in healthcare.

First Author, Year	Name	Goal	Modality	Techniques	Main Outcomes	User Preference	Cha
Denecke et al., 2018	Ana	Collect medical history for music therapy	Mobile app: Text input	AIML, rule-based	Comprehensive data collection, usability	Engaging, intuitive	Integ diver inter comp
Denecke et al., 2022	Not specified	Improve radiological diagnostics	Telegram CUI	RiveScript (rule-based)	Enhanced knowledgeability, diagnostic quality	User-friendly	Clini integ secu
Faqar-Uz-Zaman et al., 2022	Ada	Evaluate diagnostic accuracy in ER	iPad app	AI questionnaire, ML	Increased diagnostic accuracy	Not specified	Phys integ diag varia
Frick et al., 2021	Not specified	Elicit truthful medical disclosure	Online survey	Common CA technologies	Disclosure vs. concealment	Prefer physicians	Infor accu
Gashi et al., 2021	AnCha	Collect pre-visit medical history	IBM Watson, web-based	Rule-based tree	Efficient data collection	Reduces pre-visit anxiety	Clini integ secu
Ghosh et al., 2018	Quro	User symptom check, personalized assessments	Web interface	NLP, ML	Precision in condition prediction	High engagement	Data accu pred
Heald et al., 2021	Not specified	Screen for heritable cancer syndromes	Web-based, text-based	AI conversation, NLP	Efficient risk assessment, facilitated testing	High engagement, completion rates	Worl integ gene unde
Henneman n et al., 2022	ADA	Diagnose mental disorders	App-based symptom checker	AI analysis, NLP	Moderate diagnostic accuracy	Mixed preferences	Diag perf user depe
Hong et al., 2022	Genie	Collect detailed medical histories	Web-based, AI speech-to-text	AI, NLP	Improved history collection	Helpful for PCPs	Ease use
Ireland et al., 2021	Edna	Support genomic findings	Mobile, tablet, PC	NLP, Sentiment Analysis	Enhanced patient agency, informed	Ease of access, supports	Emp comp inter

		decision-making			decisions	consent	priva
Jungmann et al., 2019	Ada	Diagnose mental disorders	Mobile app	AI symptom analysis	Moderate diagnostic agreement	Not specified	Accu comp
Nazareth et al., 2021	Gia	Hereditary cancer risk triage	Web-based, mobile	NLP	Automated risk triage, educational interactions	High engagement	Work integ priva need
Ni et al., 2017	Mandy	Automate patient intake	Mobile app	NLP, data-driven analysis	Reduced staff workload, privacy maintenance	Improves physician efficiency	Full integ priva inter
Ponathil et al., 2020	VCA	Collect family health history	Web-based chat	Not specified	Higher satisfaction, lower workload	Preferred by most users	Multi exte inter
Reis et al., 2020	Cognitive Agent	Automate anamnesis-diagnosis-treatment	Voice-based AI chatbot	ML, NLP, Speech Recognition	Reduced documentation time	Reduces non-billable activities	Phys resis conc over
Schneider et al., 2023	Not specified	Standardize allergy history taking	HTML-based, online	HTML, Java scripting	Time-efficient, accurate history taking	High satisfaction	Ques spec
Wang et al., 2015	VICKY	Collect family health histories	Touch-screen tablet	Speech recognition, decision trees	High satisfaction, effective identification	Easy to use, recommended	Data issue ques
Welch et al., 2020	It Runs In My Family	Assess hereditary cancer risk	Web-based chatbot	NLP	High engagement, thorough assessments	Prefer chatbot to web forms	Data inter dem reac

AIML: Artificial intelligence markup language
CUI: Conversational user interface
AI: Artificial intelligence
ML: Machine learning
NLP: Natural language processing
CA: Conversational agent
PCP: Primary care physician
ER: Emergency room
HTML: Hypertext markup language

Table format based on Schachner et al. [54]

Discussion

Principal Results

This systematic review highlights that the use of chatbots can improve medical history taking. Results of the included studies have shown that chatbots can facilitate data collection while increasing patient engagement and satisfaction [41], [53]. Chatbots show value, especially in collecting structured data such as family history [29], [50], [51]. As highlighted, the collection of family history benefits significantly from chatbot automation due to the simple nature of their queries, which typically require binary responses. This area contrasts with the challenges of collecting data on undiagnosed symptoms, where patient responses are inherently more nuanced and variable. The inherent abilities of chatbots to handle yes/no questions efficiently and without misinterpretation make them particularly valuable in this context, minimising human error and optimising the data collection process. Several studies have highlighted that chatbots provide a more engaging patient interaction, often perceived as less intimidating than traditional face-to-face conversations [27], [48]. This interaction is crucial as it motivates patients to disclose more comprehensive health information, which can lead to better health outcomes. However, they may underperform in scenarios requiring nuanced understanding and empathy. Studies have discussed that face-to-face communication is necessary to build trust when talking about more sensitive issues [42]. In addition, chatbots provide accessibility and efficiency as they are available 24/7 and allow patients to share their medical history in the comfort and privacy of their own homes. Increased accessibility can improve healthcare services, especially in situations where immediate access to accurate patient data can greatly impact outcomes [45], [51]. In addition, chatbots support healthcare professionals by providing organised and comprehensive patient information, leading to more accurate and faster clinical decisions [39], [46]. This support is crucial in situations where healthcare professionals need to make quick decisions based on extensive data.

These findings are consistent with previous research [3], [22], which emphasises the advantages of chatbots in ensuring systematic and comprehensive information collection. Chatbots' interactive and conversational nature contributes to higher patient engagement and satisfaction [4], [50], enhancing the quality of patient histories. Additionally, the review supports the findings that chatbots can improve healthcare efficiency and accessibility by automating data collection processes [4]. Integrating chatbots into history-taking can streamline and optimise patient care, aligning with the emphasis on the importance of history-taking in medical diagnosis and management [21], [41].

While chatbots already promise success in supporting diagnostic processes, the required level of accuracy must be achieved for complex medical scenarios that require in-depth understanding and sound clinical judgment. The limitations of current systems are highlighted in the studies by Hennemann et al. [44] and Jungmann et al. [47], highlighting the need to improve the algorithms and decision-making processes to manage complex health conditions.

In addition, successfully integrating chatbots into clinical workflows requires secure data infrastructures and user-friendly interfaces. Denecke et al. [49] and Nazareth et al. [49] reported that this can drive adoption by both healthcare providers and patients. In order to meet the needs of different patient groups and healthcare facilities, customised chatbots are required. Addressing these different needs can increase patient engagement and satisfaction [49], [50].

[55][56], [57][56], [57], [58][56]However, there are some aspects that need to be considered when developing chatbot technologies [55]. Rapid deployment and development of chatbots without adequate refinement and validation can lead to inaccuracies and potentially harmful outcomes [56], [57]. These chatbots may not understand complex medical scenarios and

risk misdiagnosis or incorrect treatment recommendations [56], [57], [58]. The use of chatbots must be done with caution and focus on rigorous testing and validation to minimise these risks [56].

Limitations

There are limitations to consider in this systematic review. The exclusion of non-English-speaking studies leads to a linguistic bias and limits the generalisability of our results. This may lead to important research from non-English-speaking regions being overlooked. Future reviews could include multilingual studies to gain a more comprehensive understanding of chatbot applications around the world. The variability of study designs, patient groups and healthcare contexts makes it difficult to draw definitive conclusions. Different studies, such as those by Denecke et al. [41] and Faqar-Uz-Zaman et al. [45], focused on different settings and patient groups, which influences the results. Cross-sectional studies provide snapshots of usability, while RCTs provide robust evidence. Heterogeneity in demographics and health status also affects generalizability, as seen in the studies by Welch et al. [51] and Wang et al. [29]. Bias assessment frequently showed unmet STROBE criteria. Clear eligibility criteria and detailed methods could influence reliability. For example, Gashi et al. [39] lacked defined selection criteria, and Jungmann et al. [47] had a selection bias. Inconsistent reporting and lack of blinding in some RCTs, such as Wang et al. [29], impaired internal validity."

The methodological quality of the included studies varied. At the same time, most observational studies demonstrated satisfactory quality, and a significant proportion fulfilled only some of the STROBE criteria. Additionally, the risk of bias assessment of the RCTs revealed a high risk of bias in one of the studies [45]. It is important to consider these limitations when interpreting the data and trying to understand how they relate to clinical practice. In addition, only published research has been included in this systematic review, which may lead to publication bias as studies with positive results are more likely to be published.

[45]

Future Directions

Based on the findings and limitations of this systematic review, future research should focus on conducting more standardised and well-designed studies in this field. Emphasising rigorous study designs, such as randomised controlled trials (RCTs), with larger sample sizes and standardised outcome measures will enhance the scientific validity of the research and provide more substantial evidence of the effectiveness of chatbots in history-taking. Standardised outcome measures between studies are crucial for better comparability. Future studies should use measures such as diagnostic accuracy, patient satisfaction, engagement and usability ratings. Instruments such as the System Usability Scale (SUS) or the Technology Acceptance Model (TAM) could be used. Further investigation is needed to explore the specific contexts and patient populations where chatbots for history-taking may be most effective [29], [50], [51]. Different medical areas and health situations may present special considerations and challenges that could influence the implementation and acceptance of chatbot-based systems for taking medical histories, such as in the case of older people due to a more limited technical affinity or long medical histories in people with chronic illnesses.

Moreover, future research should address the challenges and limitations identified in this review. Efforts should be made to minimise bias and improve the methodological quality of studies. Conducting studies with more homogeneous patient populations and utilising consistent outcome measures would enhance the comparability and generalizability of the findings [41].

Finally, it would be valuable to explore the integration of chatbots with other technologies or interventions to optimise the history-taking process. The integration of chatbots with modern technologies such as NLP, machine learning algorithms and decision support systems has the potential to significantly improve history-taking [21], [43], [48]. NLP could improve the

ability to understand and interpret patient responses to the chatbot. The interactions will be more fluid and intuitive. Machine learning algorithms can be used to continuously improve chatbot responses based on patient interactions. This could lead to more accurate and personalised information. The integration of decision support systems can provide healthcare providers with real-time evidence-based recommendations. Research designs to investigate these integrations could include comparative studies for measuring differences in diagnostic accuracy, patient satisfaction and efficiency between two groups. One group could use a simple chatbot, and another group could use an advanced chatbot with integrated NLP and machine learning.

Conclusion

Overall, this systematic review provides important information on the benefits of using chatbots in medical history taking. They improve patient engagement by increasing data completeness and user satisfaction. They capture accurate and comprehensive patient data, which is crucial for diagnosis. In addition, chatbots can increase efficiency by significantly reducing data collection time. For example, chatbots can be used in primary care to record a patient's medical history before a consultation. This could reduce the workload of medical staff and enable more targeted interaction between patient and physician. The efficiency gains from using chatbots can also help in emergency departments, where fast and accurate data capture is crucial.

Future research should focus on different areas to improve the use of chatbots for medical history taking. Larger studies and RCTs are essential for adequate validation. The use of chatbots needs to be investigated in different healthcare settings and with different patient groups. For example, in patients with chronic diseases, mental illness or elderly patients and in people who are not tech-savvy. Another area that needs to be considered is analysing the impact of chatbots on workflows in clinics or practices and the change in the doctor-patient relationship. In addition, data protection and security issues must be clarified to ensure the

protection of patient data. Especially considering the latest developments in AI models. These offer new opportunities for more precise and personalised interactions. Research should optimise these models for history taking and integrate them into decision support systems for real-time evidence-based recommendations. If these areas are addressed, chatbots can significantly transform healthcare by improving efficiency, accuracy, and patient engagement, especially for underserved patient populations, as well as chronic disease management and real-time symptom assessment.

Multimedia Appendix: Search strategies conducted, overview of studies, quality assessment of included studies

[1] F. J. Fowler, C. A. Levin, and K. R. Sepucha, ‘Informing and involving patients to improve the quality of medical decisions’, *Health Aff (Millwood)*, vol. 30, no. 4, pp. 699–706, Apr. 2011, doi: 10.1377/HLTHAFF.2011.0003.

[2] J. R. Hampton, M. J. G. Harrison, J. R. A. Mitchell, J. S. Prichard, and C. Seymour, ‘Relative contributions of history-taking, physical examination, and laboratory investigation to diagnosis and management of medical outpatients’, *Br Med J*, vol. 2, no. 5969, pp. 486–489, May 1975, doi: 10.1136/BMJ.2.5969.486.

[3] L. Laranjo *et al.*, ‘Conversational agents in healthcare: a systematic review’, *J Am Med Inform Assoc*, vol. 25, no. 9, pp. 1248–1258, Sep. 2018, doi: 10.1093/JAMIA/OCY072.

[4] K. Denecke, R. May, and Y. Deng, ‘Towards Emotion-Sensitive Conversational User Interfaces in Healthcare Applications’, *Stud Health Technol Inform*, vol. 264, pp. 1164–1168, Aug. 2019, doi: 10.3233/SHTI190409.

[5] G. I. Hess, G. Fricker, and K. Denecke, ‘Improving and Evaluating eMMA’s Communication Skills: A Chatbot for Managing Medication’, *Stud Health Technol Inform*, vol. 259, pp. 101–104, 2019, doi: 10.3233/978-1-61499-961-4-101.

[6] M. das G. B. Marietto *et al.*, ‘Artificial Intelligence MArkup Language: A Brief Tutorial’, *International Journal of Computer Science & Engineering Survey*, vol. 4, no. 3, pp. 1–20, Jul. 2013, doi: 10.5121/ijcses.2013.4301.

[7] N. Rebelo, L. Sanders, K. Li, and J. C. L. Chow, ‘Learning the Treatment Process in Radiotherapy Using an Artificial Intelligence-Assisted Chatbot: Development Study’, *JMIR Form Res*, vol. 6, no. 12, Dec. 2022, doi: 10.2196/39443.

[8] H. S. J. Chew, ‘The Use of Artificial Intelligence–Based Conversational Agents (Chatbots) for Weight Loss: Scoping Review and Practical Recommendations’, *JMIR Med Inform 2022;10(4):e32578 https://medinform.jmir.org/2022/4/e32578*, vol. 10, no. 4, p. e32578, Apr. 2022, doi: 10.2196/32578.

[9] Y. Xu, J. Zhang, and G. Deng, ‘Enhancing customer satisfaction with chatbots: The influence of communication styles and consumer attachment anxiety’, *Front Psychol*, vol. 13, Jul. 2022, doi: 10.3389/FPSYG.2022.902782.

[10] P. Amiri and E. Karahanna, ‘Chatbot use cases in the Covid-19 public health response’, *Journal of the American Medical Informatics Association*, vol. 29, no. 5, pp. 1000–1010, Apr. 2022, doi: 10.1093/JAMIA/OCAC014.

[11] M. Almalki and F. Azeez, ‘Health Chatbots for Fighting COVID-19: a Scoping Review’, *Acta Inform Med*, vol. 28, no. 4, pp. 241–247, Dec. 2020, doi: 10.5455/AIM.2020.28.241-247.

[12] T. J. Judson *et al.*, ‘Implementation of a digital chatbot to screen health system employees during the COVID-19 pandemic’, *Journal of the American Medical Informatics Association*, vol. 27, no. 9, pp. 1450–1455, Sep. 2020, doi: 10.1093/JAMIA/OCAA130.

[13] H. Else, ‘Abstracts written by ChatGPT fool scientists’, *Nature*, vol. 613, no. 7944, p. 423, Jan. 2023, doi: 10.1038/D41586-023-00056-7.

[14] T. Someya and M. Amagai, ‘Toward a new generation of smart skins’, *Nat Biotechnol*, vol. 37, no. 4, pp. 382–388, Apr. 2019, doi: 10.1038/S41587-019-0079-1.

[15] The Lancet Digital Health, ‘ChatGPT: friend or foe?’, *Lancet Digit Health*, vol. 5, no. 3, p. e102, Mar. 2023, doi: 10.1016/S2589-7500(23)00023-7.

[16] A. Gilson *et al.*, ‘How Does ChatGPT Perform on the United States Medical Licensing Examination? The Implications of Large Language Models for Medical Education and Knowledge Assessment’, *JMIR Med Educ*, vol. 9, 2023, doi: 10.2196/45312.

[17] K. Singhal *et al.*, ‘Large language models encode clinical knowledge’, *Nature 2023 620:7972*, vol. 620, no. 7972, pp. 172–180, Jul. 2023, doi: 10.1038/s41586-023-06291-2.

[18] E. C. Stade *et al.*, ‘Large language models could change the future of behavioral healthcare: a

proposal for responsible development and evaluation', *npj Mental Health Research 2024 3:1*, vol. 3, no. 1, pp. 1–12, Apr. 2024, doi: 10.1038/s44184-024-00056-z.

[19] T. Dave, S. A. Athaluri, and S. Singh, 'ChatGPT in medicine: an overview of its applications, advantages, limitations, future prospects, and ethical considerations', *Front Artif Intell*, vol. 6, 2023, doi: 10.3389/FRAI.2023.1169595.

[20] O. Asan, A. E. Bayrak, and A. Choudhury, 'Artificial Intelligence and Human Trust in Healthcare: Focus on Clinicians', *J Med Internet Res*, vol. 22, no. 6, Jun. 2020, doi: 10.2196/15154.

[21] L. Xu, L. Sanders, K. Li, and J. C. L. Chow, 'Chatbot for Health Care and Oncology Applications Using Artificial Intelligence and Machine Learning: Systematic Review', *JMIR Cancer*, vol. 7, no. 4, Oct. 2021, doi: 10.2196/27850.

[22] M. Milne-Ives *et al.*, 'The Effectiveness of Artificial Intelligence Conversational Agents in Health Care: Systematic Review', *J Med Internet Res*, vol. 22, no. 10, Oct. 2020, doi: 10.2196/20346.

[23] R. Li, A. Kumar, and J. H. Chen, 'How Chatbots and Large Language Model Artificial Intelligence Systems Will Reshape Modern Medicine: Fountain of Creativity or Pandora's Box?', *JAMA Intern Med*, vol. 183, no. 6, pp. 596–597, Jun. 2023, doi: 10.1001/JAMAINTERNMED.2023.1835.

[24] 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', *JMIR Ment Health*, vol. 5, no. 4, p. e9782, Dec. 2018, doi: 10.2196/MENTAL.9782.

[25] Y. J. Oh, J. Zhang, M. L. Fang, and Y. Fukuoka, 'A systematic review of artificial intelligence chatbots for promoting physical activity, healthy diet, and weight loss', *International Journal of Behavioral Nutrition and Physical Activity*, vol. 18, no. 1, pp. 1–25, Dec. 2021, doi: 10.1186/S12966-021-01224-6/TABLES/4.

[26] T. W. Bickmore *et al.*, 'A randomized controlled trial of an automated exercise coach for older adults', *J Am Geriatr Soc*, vol. 61, no. 10, pp. 1676–1683, Oct. 2013, doi: 10.1111/JGS.12449.

[27] B. Heald *et al.*, 'Using chatbots to screen for heritable cancer syndromes in patients undergoing routine colonoscopy', *J Med Genet*, vol. 58, no. 12, pp. 807–814, Dec. 2021, doi: 10.1136/JMEDGENET-2020-107294.

[28] D. Ireland *et al.*, 'Introducing Edna: A trainee chatbot designed to support communication about additional (secondary) genomic findings', *Patient Educ Couns*, vol. 104, no. 4, pp. 739–749, Apr. 2021, doi: 10.1016/J.PEC.2020.11.007.

[29] C. Wang *et al.*, 'Acceptability and feasibility of a virtual counselor (VICKY) to collect family health histories', *Genetics in Medicine 2015 17:10*, vol. 17, no. 10, pp. 822–830, Jan. 2015, doi: 10.1038/gim.2014.198.

[30] D. Moher *et al.*, 'Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement', *PLoS Med*, vol. 6, no. 7, p. e1000097, Jul. 2009, doi: 10.1371/JOURNAL.PMED.1000097.

[31] National Institute for Health Research (NHS), 'PROSPERO – International prospective register of systematic reviews.' Accessed: Apr. 02, 2023. [Online]. Available: https://www.crd.york.ac.uk/prospero/

[32] H. Wickham, 'Ggplot2: Elegant graphics for data analysis (2nd ed.)', *Springer International Publishing*, 2016, doi: 10.1007/978-3-319-24277-4.

[33] S. A. Miller and J. L. Forrest, 'Enhancing your practice through evidence-based decision making: PICO, learning how to ask good questions', *Journal of Evidence Based Dental Practice*, vol. 1, no. 2, pp. 136–141, Oct. 2001, doi: 10.1016/S1532-3382(01)70024-3.

[34] E. von Elm, D. G. Altman, M. Egger, S. J. Pocock, P. C. Gøtzsche, and J. P. Vandenbroucke, 'The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE)

statement: guidelines for reporting observational studies', *J Clin Epidemiol*, vol. 61, no. 4, pp. 344–349, Apr. 2008, doi: 10.1016/J.JCLINEPI.2007.11.008.

[35] M. Bojanowski and R. Edwards, 'alluvial: R package for creating alluvial diagrams'. R Package Version 0.1 2, 2016.

[36] A. Mendy *et al.*, 'Endotoxin Exposure and Childhood Wheeze and Asthma: A Meta-Analysis of Observational Studies', *J Asthma*, vol. 48, no. 7, pp. 685–693, Sep. 2011, doi: 10.3109/02770903.2011.594140.

[37] J. A. C. Sterne *et al.*, 'RoB 2: a revised tool for assessing risk of bias in randomised trials', *BMJ*, vol. 366, 2019, doi: 10.1136/BMJ.L4898.

[38] L. A. McGuinness and J. P. T. Higgins, 'Risk-of-bias VISualization (robvis): An R package and Shiny web app for visualizing risk-of-bias assessments', *Res Synth Methods*, vol. 12, no. 1, pp. 55–61, Jan. 2021, doi: 10.1002/JRSM.1411.

[39] F. Gashi, S. F. Regli, R. May, P. Tschopp, and K. Denecke, 'Developing Intelligent Interviewers to Collect the Medical History: Lessons Learned and Guidelines', *Stud Health Technol Inform*, vol. 279, pp. 18–25, May 2021, doi: 10.3233/SHTI210083.

[40] K. Denecke, P. Lombardi, and K. Nairz, 'Digital Medical Interview Assistant for Radiology: Opportunities and Challenges', *Stud Health Technol Inform*, vol. 293, pp. 39–46, May 2022, doi: 10.3233/SHTI220345.

[41] K. Denecke, S. L. Hochreutener, A. Pöpel, and R. May, 'Self-Anamnesis with a Conversational User Interface: Concept and Usability Study', *Methods Inf Med*, vol. 57, no. 5–06, pp. 243–252, 2018, doi: 10.1055/S-0038-1675822.

[42] N. R. J. Frick, F. Brünker, B. Ross, and S. Stieglitz, 'Comparison of disclosure/concealment of medical information given to conversational agents or to physicians', *Health Informatics J*, vol. 27, no. 1, 2021, doi: 10.1177/1460458221994861.

[43] L. Reis, C. Maier, J. Mattke, M. Creutzenberg, and T. Weitzel, 'Addressing User Resistance Would Have Prevented a Healthcare AI Project Failure', *MIS Quarterly Executive*, vol. 19, no. 4, Dec. 2020, Accessed: Jul. 25, 2023. [Online]. Available: https://aisel.aisnet.org/misqe/vol19/iss4/8

[44] S. Hennemann, S. Kuhn, M. Witthöft, and S. M. Jungmann, 'Diagnostic Performance of an App-Based Symptom Checker in Mental Disorders: Comparative Study in Psychotherapy Outpatients', *JMIR Ment Health*, vol. 9, no. 1, Jan. 2022, doi: 10.2196/32832.

[45] S. F. Faqar-Uz-Zaman *et al.*, 'The Diagnostic Efficacy of an App-based Diagnostic Health Care Application in the Emergency Room: eRadaR-Trial. A prospective, Double-blinded, Observational Study', *Ann Surg*, vol. 276, no. 5, pp. 935–942, Nov. 2022, doi: 10.1097/SLA.0000000000005614.

[46] S. Schneider *et al.*, 'Successful usage of a chatbot to standardize and automate history taking in Hymenoptera venom allergy', *Allergy: European Journal of Allergy and Clinical Immunology*, 2023, doi: 10.1111/ALL.15720.

[47] S. M. Jungmann, T. Klan, S. Kuhn, and F. Jungmann, 'Accuracy of a Chatbot (Ada) in the Diagnosis of Mental Disorders: Comparative Case Study With Lay and Expert Users', *JMIR Form Res*, vol. 3, no. 4, Oct. 2019, doi: 10.2196/13863.

[48] G. Hong, M. Smith, and S. Lin, 'The AI Will See You Now: Feasibility and Acceptability of a Conversational AI Medical Interviewing System', *JMIR Form Res*, vol. 6, no. 6, Jun. 2022, doi: 10.2196/37028.

[49] S. Nazareth *et al.*, 'Hereditary Cancer Risk Using a Genetic Chatbot Before Routine Care Visits', *Obstetrics and gynecology*, vol. 138, no. 6, pp. 860–870, Dec. 2021, doi: 10.1097/AOG.0000000000004596.

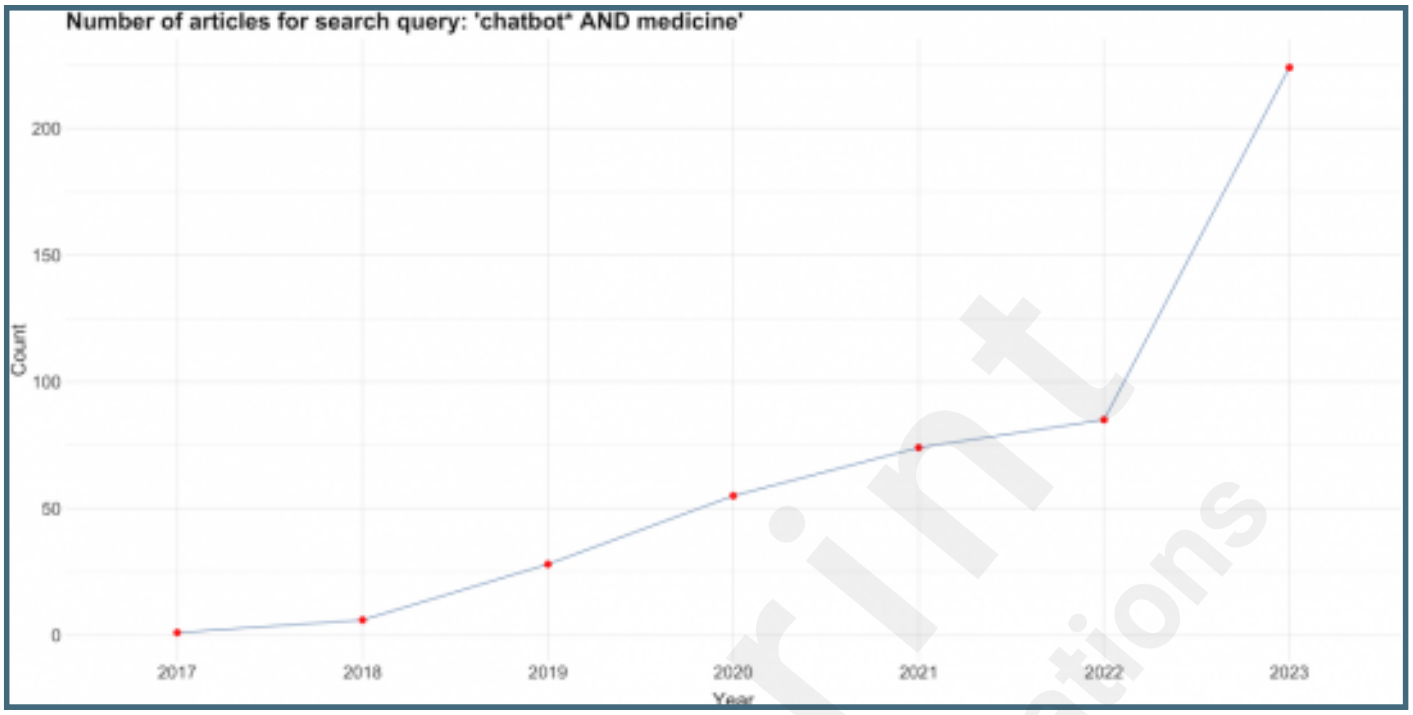
[50] A. Ponathil, F. Ozkan, B. Welch, J. Bertrand, and K. Chalil Madathil, 'Family health history collected by virtual conversational agents: An empirical study to investigate the efficacy of this approach', *J Genet Couns*, vol. 29, no. 6, pp. 1081–1092, Dec. 2020, doi:

10.1002/JGC4.1239.
[51] B. M. Welch, C. G. Allen, J. B. Ritchie, H. Morrison, C. Hughes-Halbert, and J. D. Schiffman, ‘Using a Chatbot to Assess Hereditary Cancer Risk’, *JCO Clin Cancer Inform*, vol. 4, no. 4, pp. 787–793, Nov. 2020, doi: 10.1200/CCI.20.00014.
[52] S. Ghosh, S. Bhatia, and A. Bhatia, ‘Quro: Facilitating User Symptom Check Using a Personalised Chatbot-Oriented Dialogue System’, pp. 51–56, 2018, doi: 10.3233/978-1-61499-890-7-51.
[53] L. Ni, C. Lu, N. Liu, and J. Liu, ‘MANDY: Towards A Smart Primary Care Chatbot Application’, *Department of Computer Science, The University of Auckland, New Zealand*, 2017, Accessed: Jul. 20, 2023. [Online]. Available: https://deepmind.com/applied/deepmind-health/,
[54] T. Schachner, R. Keller, and F. v. Wangenheim, ‘Artificial Intelligence-Based Conversational Agents for Chronic Conditions: Systematic Literature Review’, *J Med Internet Res*, vol. 22, no. 9, Sep. 2020, doi: 10.2196/20701.
[55] Z. Ni *et al.*, ‘Implementation of Chatbot Technology in Health Care: Protocol for a Bibliometric Analysis.’, *JMIR Res Protoc*, vol. 13, no. 1, p. e54349, Feb. 2024, doi: 10.2196/54349.
[56] C. Wang, S. Liu, H. Yang, J. Guo, Y. Wu, and J. Liu, ‘Ethical Considerations of Using ChatGPT in Health Care’, *J Med Internet Res*, vol. 25, 2023, doi: 10.2196/48009.
[57] P. P. Ray, ‘ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope’, *Internet of Things and Cyber-Physical Systems*, vol. 3, pp. 121–154, Jan. 2023, doi: 10.1016/J.IOTCPS.2023.04.003.
[58] L. Wilson and M. Marasoiu, ‘The Development and Use of Chatbots in Public Health: Scoping Review’, *JMIR Hum Factors*, vol. 9, no. 4, Oct. 2022, doi: 10.2196/35882.

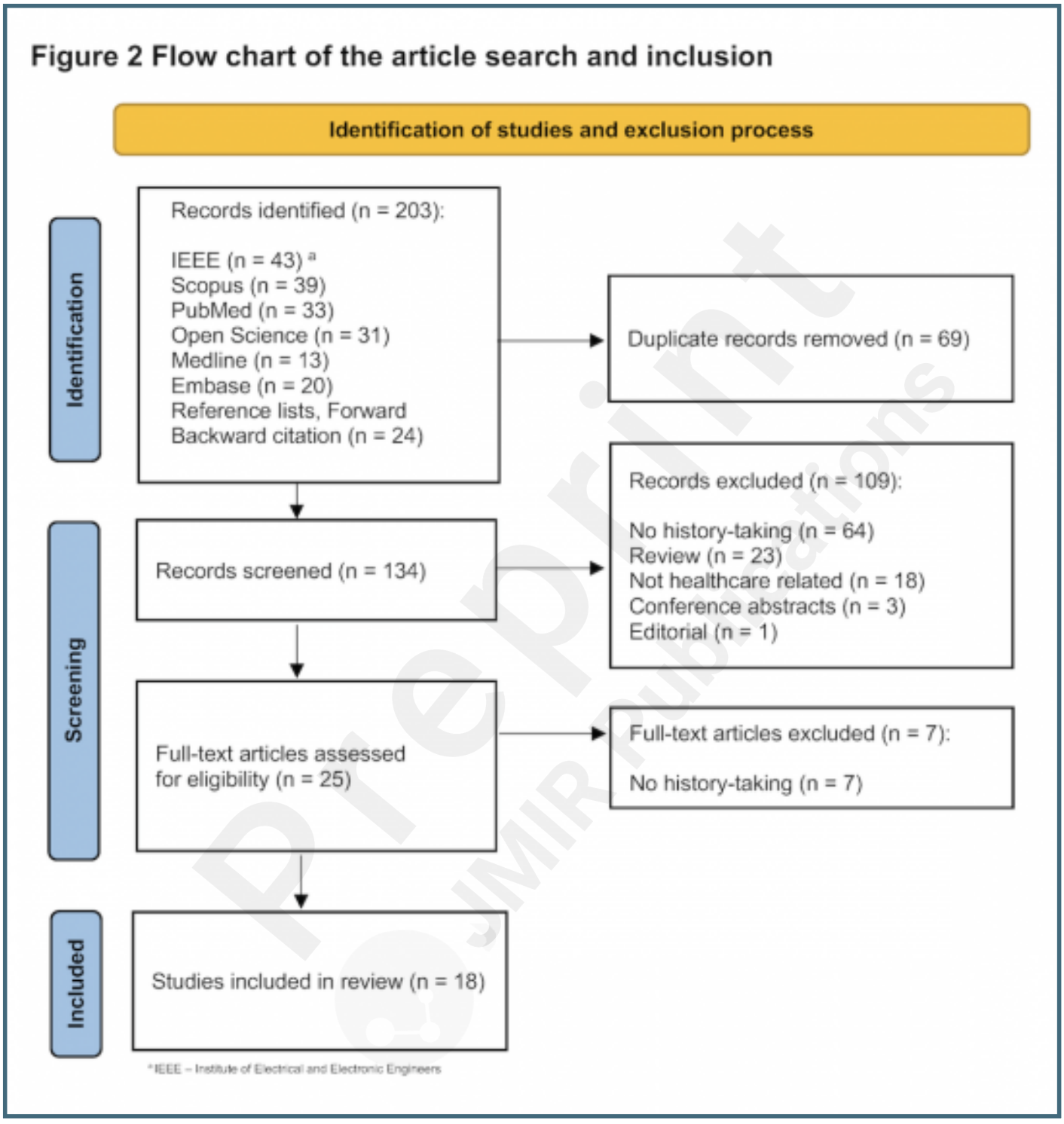
Supplementary Files

Figures

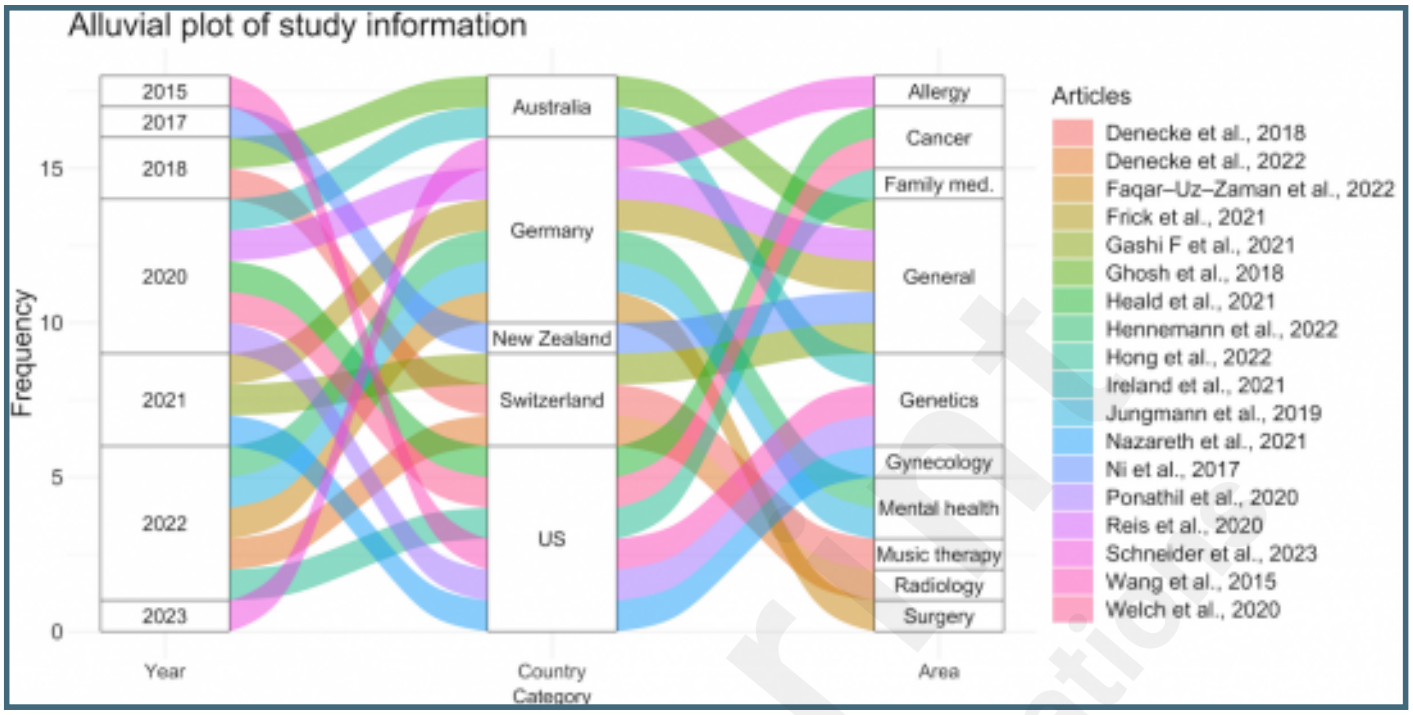
Number of articles over the last years: chatbot* AND medicine.



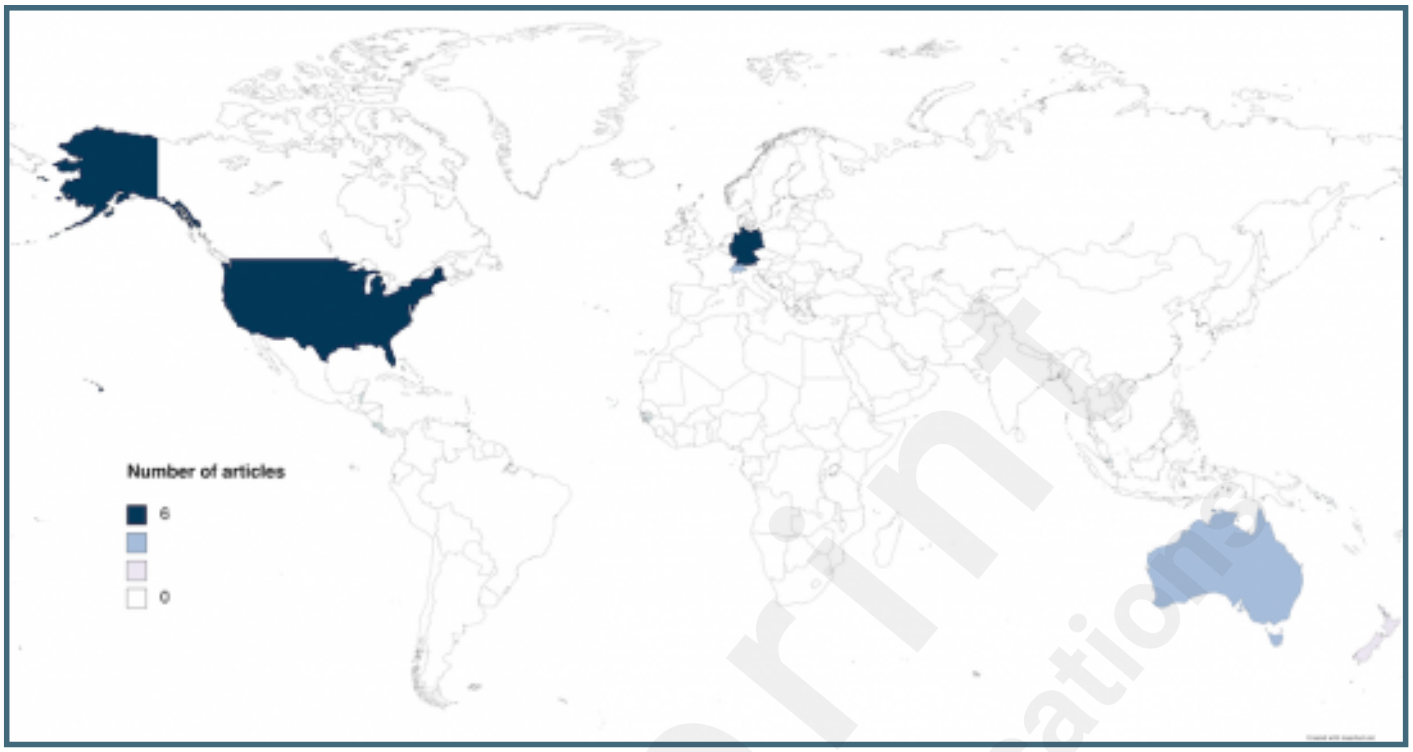
Flow chart of the article search and inclusion.



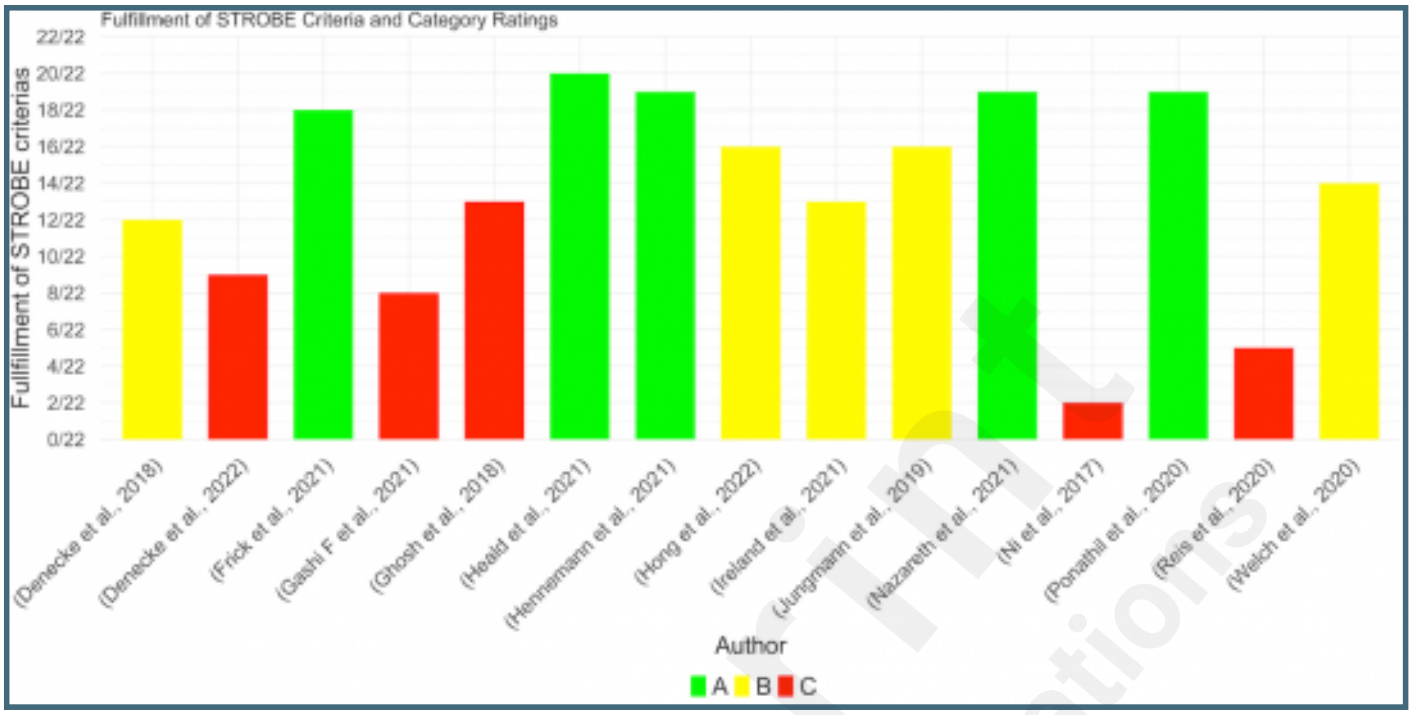
Alluvial diagram of the publication date, country, and area of articles.



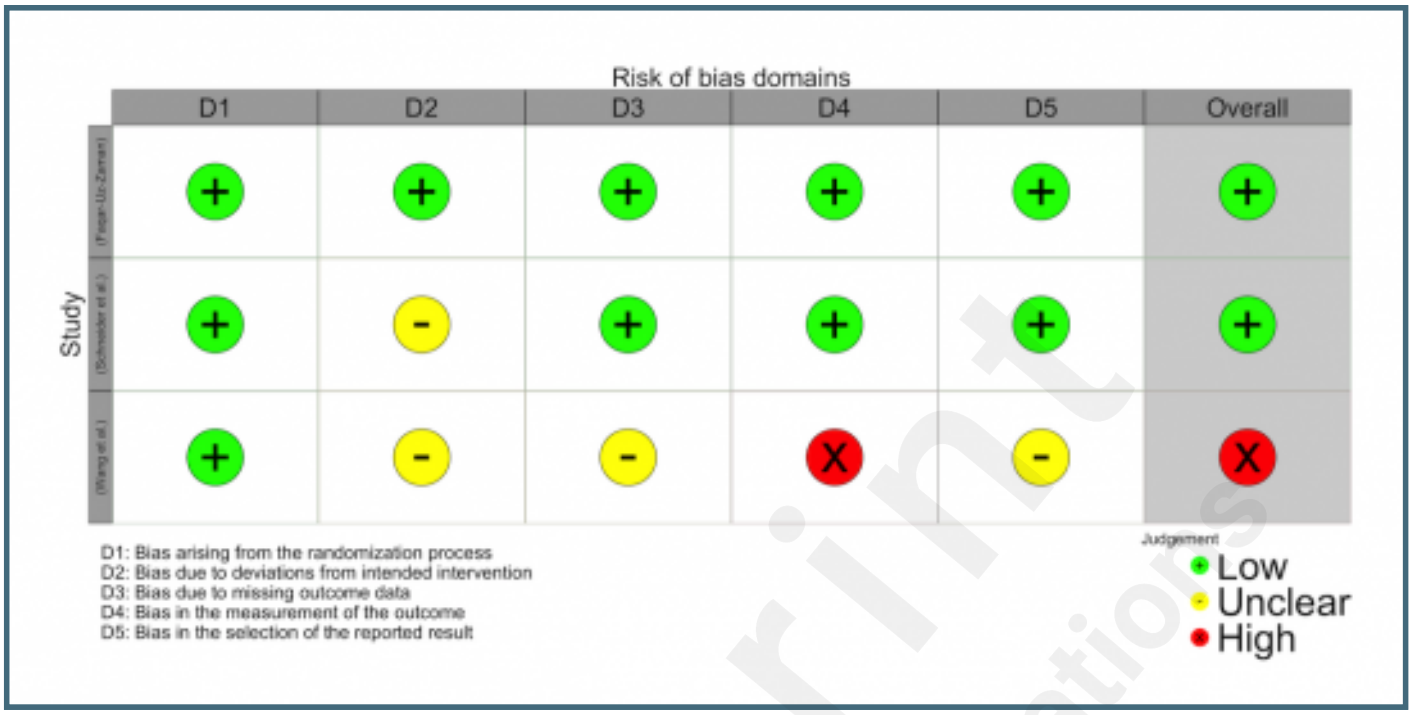
World map showing the number of articles published in each country.



Fulfilment of STROBE criteria and categorisation.



Risk of bias domains (RoB-tool) for RCTs.



Multimedia Appendixes

Search strategies conducted, overview of studies, quality assessment of included studies.
URL: http://asset.jmir.pub/assets/a7d0832aacde62442b851f0c97c8c5f5.pdf

CONSORT (or other) checklists

PRISMA Checklist.
URL: http://asset.jmir.pub/assets/ee93e7e0a96500dc6c642356bbca58ce.pdf