Systematic Review/Meta-analysis

ANNALS OF MEDICINE & SURGERY

OPEN

Exploring ChatGPT in clinical inquiry: a scoping review of characteristics, applications, challenges, and evaluation

Shahabeddin Abhari, PhD[a], Yasna Afshari, BSc[b,c], Farhad Fatehi, MD PhD[d], Hosna Salmani, PhD[e,*], Ali Garavand, PhD[f], Dmytro Chumachenko, PhD[g], Somayyeh Zakerabasali, PhD[h], Plinio P. Morita, PEng, PhD[a,i,j,k,l]

Introduction: Recent advancements in generative AI, exemplified by ChatGPT, hold promise for healthcare applications such as decision-making support, education, and patient engagement. However, rigorous evaluation is crucial to ensure reliability and safety in clinical contexts. This scoping review explores ChatGPT's role in clinical inquiry, focusing on its characteristics, applications, challenges, and evaluation.
Methods: This review, conducted in 2023, followed PRISMA-ScR guidelines (Supplemental Digital Content 1, http://links.lww.com/MS9/A636). Searches were performed across PubMed, Scopus, IEEE, Web of Science, Cochrane, and Google Scholar using relevant keywords. The review explored ChatGPT's effectiveness in various medical domains, evaluation methods, target users, and comparisons with other AI models. Data synthesis and analysis incorporated both quantitative and qualitative approaches.
Results: Analysis of 41 academic studies highlights ChatGPT's potential in medical education, patient care, and decision support, though performance varies by medical specialty and linguistic context. GPT-3.5, frequently referenced in 26 studies, demonstrated adaptability across diverse scenarios. Challenges include limited access to official answer keys and inconsistent performance, underscoring the need for ongoing refinement. Evaluation methods, including expert comparisons and statistical analyses, provided significant insights into ChatGPT's efficacy. The identification of target users, such as medical educators and nonexpert clinicians, illustrates its broad applicability.
Conclusion: ChatGPT shows significant potential in enhancing clinical practice and medical education. Nevertheless, continuous refinement is essential for its successful integration into healthcare, aiming to improve patient care outcomes, and address the evolving needs of the medical community.

Keywords: artificial intelligence, ChatGPT, clinical questions, generative AI, healthcare

Introduction

In recent years, the field of artificial intelligence (AI), particularly focusing on generative AI, has experienced a remarkable surge in both research and real-world applications[1,2]. This surge has become even more pronounced in recent months, highlighting the rapidly evolving landscape of AI technologies[3]. An example of this remarkable progress is ChatGPT, an AI model that has attracted considerable attention due to its exceptional ability to generate human-like text. This upswing in AI, as seen through models like ChatGPT, represents a broader trend where AI technologies are becoming increasingly indispensable in various domains[4]. These applications range from natural language processing and content generation to sectors as diverse as

[a]School of Public Health Sciences, University of Waterloo, Waterloo, Ontario, Canada, [b]Department of Radiology and Nuclear Medicine, Erasmus MC University Medical Center Rotterdam, Rotterdam, [c]Department of Epidemiology, Erasmus MC University Medical Center Rotterdam, Rotterdam, The Netherlands, [d]Business School, The University of Queensland, Brisbane, Australia, [e]Department of Health Information Management, School of Health Management and Information Sciences, Iran University of Medical Sciences, Tehran, Iran, [f]Department of Health Information Technology, School of Allied Medical Sciences, Lorestan University of Medical Sciences, Khorramabad, Iran, [g]Department of Mathematical Modeling and Artificial Intelligence, National Aerospace University 'Kharkiv Aviation Institute', Kharkiv, Ukraine, [h]Department of Health Information Management, Clinical Education Research Center, Health Human Resources Research Center, School of Health Management and Information Sciences, Shiraz University of Medical Sciences, Shiraz, Iran, [i]Department of Systems Design Engineering, University of Waterloo, [j]Research Institute for Aging, University of Waterloo, Waterloo, Ontario, Canada, [k]Centre for Digital Therapeutics, Techna Institute, University Health Network, Toronto and [l]Dalla Lana School of Public Health, Institute of Health Policy, Management, and Evaluation, University of Toronto, Toronto, Ontario, Canada

Sponsorships or competing interests that may be relevant to content are disclosed at the end of this article.

*Corresponding author. Address: Department of Health Information Management, School of Health Management and Information Sciences, Iran University of Medical Sciences, Tehran, Iran. Tel.: +0218 879 4301. E-mail: salmani.h@iums.ac.ir (H. Salmani).

Annals of Medicine & Surgery (2024) 86:7094–7104

Received 19 August 2024; Accepted 25 October 2024

Supplemental Digital Content is available for this article. Direct URL citations are provided in the HTML and PDF versions of this article on the journal's website, www.lww.com/annals-of-medicine-and-surgery.

Published online 8 November 2024

http://dx.doi.org/10.1097/MS9.0000000000002716

healthcare and customer support[5–7]. The swift and multifaceted growth of AI, particularly generative AI, continues to be a central point of discussion in contemporary technological discourse, shaping the way we interact with and apply AI in various aspects of our lives[4].

The applications of generative AI, with a notable focus on ChatGPT, within the healthcare domain, are profoundly influencing a wide spectrum of healthcare-related interactions and processes. ChatGPT, as an advanced language model, embodies a diverse range of applications, with a particular emphasis on its potential utility in the clinical domain and its capacity to respond to clinical questions posed by various stakeholders, including healthcare providers, clinicians, medical sciences students, and patients[8,9]. Within this framework, ChatGPT emerges as a multifaceted resource, holding great promise for supporting numerous critical healthcare functions[5]. Healthcare providers and clinicians can harness ChatGPT's capabilities to aid in screening, diagnosis, treatment planning, medication management, and rehabilitation planning, potentially optimizing patient care and clinical decision-making[6,10,11]. Medical sciences students stand to benefit from ChatGPT as an educational tool, enabling them to receive guidance, references, and clarifications on intricate medical concepts and practices, thereby advancing their knowledge base[6,12,13]. Furthermore, patients can find ChatGPT a valuable ally in enhancing health literacy and patient education, affording them comprehensible and reliable healthcare information and increasing more informed and empowered involvement in their healthcare journey[7,14].

The significance of rigorously evaluating ChatGPT's performance in responding to clinical questions cannot be overstated. Such evaluations are pivotal in ensuring the reliability, accuracy, and safety of AI-driven interactions within the healthcare domain[15–17]. Clinical questions often encompass critical aspects of medical diagnosis, treatment planning, and patient care, making the quality of responses a matter of paramount importance[18,19]. An inadequately evaluated AI model may inadvertently provide misinformation, potentially compromising patient safety, and healthcare decision-making. Therefore, comprehensive assessments of ChatGPT's responses are essential to ascertain the model's ability to furnish accurate medical information, provide sound diagnostic insights, and provide appropriate treatment recommendations. Furthermore, these evaluations help uncover potential biases, ethical concerns, and limitations, providing invaluable perspectives into responsible AI implementation in healthcare. Robust evaluations serve not only as a means of quality assurance but also as a cornerstone for enhancing trust among healthcare providers, clinicians, and patients, ultimately strengthening the foundation of AI integration in clinical settings[20–22].

During this timeframe, numerous studies have assessed ChatGPT's performance in responding to clinical questions and reported their findings. However, there has been a notable gap in synthesizing and consolidating this disparate literature into a comprehensive overview. Thus, the aim of this study was to conduct a systematic mapping and synthesis of existing literature concerning the utilization of ChatGPT, or similar conversational AI models, in addressing clinical queries across diverse healthcare domains. Our objective was to elucidate the characteristics, applications, challenges, and evaluation methods associated with ChatGPT's role in clinical inquiry. By examining the scope of research, methodologies, clinical settings explored, and outcomes assessed, we aimed to provide insights for future research directions, clinical implementation strategies, and ethical considerations in leveraging ChatGPT for clinical decision support.

HIGHLIGHTS

- Explores the potential of ChatGPT in medical education, patient care, and clinical decision support.
- Conducted a scoping review of 41 academic studies following PRISMA-ScR guidelines.
- Found GPT 3.5 to be the most frequently studied version in diverse medical scenarios.
- Included comparisons with expert opinions and statistical analyses to assess ChatGPT's performance.
- Noted performance differences across medical specialties and linguistic contexts.
- Highlighted issues such as limited access to official answer keys and the need for ongoing refinement.
- Emphasized the importance of continued research, ethical considerations, and regulatory frameworks for AI integration in healthcare.

Methods

In November 2023, we conducted a comprehensive scoping review, systematically searching scientific databases, including PubMed, Scopus, IEEE, Web of Science, Google Scholar, and Cochrane. Our search was guided by relevant keywords, and our article selection process adhered to the guidelines delineated in the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA, Supplemental Digital Content 1, http://links.lww.com/MS9/A636) extension for Scoping Reviews (PRISMA-ScR, Supplemental Digital Content 1, http://links.lww.com/MS9/A636). Following the application of predefined inclusion and exclusion criteria, a meticulous curation process resulted in the inclusion of a total of 41 articles. Subsequently, we meticulously summarized and reported the collated data.

Information sources

The primary electronic search commenced on 1st August and concluded on 15th August 2023. Searches were conducted in scientific databases, including PubMed, Scopus, IEEE, Web of Science, Google Scholar, and Cochrane. Given the interdisciplinary nature of our research, we expanded our search to include databases and knowledge repositories in the fields of health sciences and engineering.

Search strategy

The combination of related keywords is detailed in Table 1. All search steps were executed by the PRISMA-ScR flowchart (Supplemental Digital Content 1, http://links.lww.com/MS9/A636).

Selection criteria

i. Use of ChatGPT.
ii. Application within the clinical domain (medical education, diagnosis, treatment, or rehabilitation).
iii. Conducted evaluation and reported results quantitatively.

Table 1

Search strategy

Search strategy	
Databases: PubMed, Scopus, IEEE, Web of Science, Google Scholar, and Cochrane	
Limits: Language (only resources with English(, Species (studies on human)	
Date: 1 All literature to 20 November 2023	
Search strategy: #1 AND #2 AND #3	
#1	'ChatGPT'
#2	'Evaluation' OR 'Performance'
#3	'health*' OR 'medical' OR 'clinical'

iv. Journal and conference article.
v. All types of papers (original, letter, viewpoint, case report, short communication, etc.), except for review articles.

Selection process

The article selection process adhered to PRISMA-ScR guidelines (Supplemental Digital Content 1, http://links.lww.com/MS9/A636), resulting in the inclusion of 41 papers. Two authors conducted all steps in the selection and quality evaluation of the papers. Cases of disagreement were referred to a third reviewer for final decisions.

Data extraction

The data extraction was done using a data extraction form which was designed based on the objectives of the study. The searches were independently conducted by two authors to mitigate potential bias. In disagreement cases, the search results were reviewed by a third party who resolved any discrepancies. Searches were limited to published articles in English, without any specific time limitation.

The data extraction process encompassed several categories, including Title and Author, Year of publication, country, type of publication, study medical domain (primary care specialty and types), the user asking the questions (physicians, nurses, students, pharmacists, etc.), Aim of the study, Evaluation Methods, Performance measure metrics (accuracy, F1, Coppa, etc.), Gold standards applied, Summary of results, Conclusion, Limitations/challenges of the study, and version of ChatGPT (see Appendix 1, Supplemental Digital Content 2, http://links.lww.com/MS9/A636).

Data analysis

Both quantitative and qualitative methods were used for data analysis. We used descriptive methods to analyze quantitative variables and then conducted a thematic analysis on data from three qualitative variables. For quantitative assessment, data extracted were imported into SPSS v26 (IBM), where basic descriptive statistics were computed to outline information about publication year, country of origin, publication type, medical domain studied, target audience, evaluation methodology, gold standard, and the version of ChatGPT utilized. For qualitative examination, thematic analysis was employed to scrutinize three qualitative variables: challenges and limitations, evaluation methods, and study conclusions. Data from each variable were individually imported into NVivo v14 (QSR International), and thematic analysis was conducted by the framework delineated by Thomas and Harden[23] encompassing three stages: initial free coding of primary study findings, subsequent organization of these codes into cohesive descriptive themes, and final development of analytical themes.

Results

In our scoping review, we identified 1002 articles, out of which 41 academic papers were included (Fig. 1). Through a meticulous analysis of existing literature, the review shows the model's promising role in medical education, patient care, and decision support, while also identifying challenges such as access to official answer keys, representativeness of question sources, and variations in performance across different medical specialities and linguistic contexts. The diverse array of questionnaire sources and thorough examination of Generative pretrained transformer (GPT) versions provide a comprehensive overview of the landscape of research and development in natural language processing applied to healthcare. In general, the findings contribute valuable perspectives to the evolving discourse surrounding AI-driven technologies in healthcare, informing future research directions and guiding the integration of ChatGPT into clinical practice and medical education.

Characteristics of study and literature distribution

According to Figure 2, the United States emerged as the primary contributor to publications, (36.6%, $n=15$), while Turkey, Korea, Israel, UK, and Singapore each contributed (2.4%, $n=1$).

Study medical domain

The field of Health and Medical Education accounted for 12.19% ($n=5$) of the total[24–28]. Other areas included Ophthalmology (9.7%, $n=4$)[29–32] and Clinical and Medical Informatics (7.3%, $n=3$),[10,33,34] and various other fields (4.8%, $n=2$ and 2.4% $n=1$). This comprehensive coverage of medical domains ensured a thorough evaluation of ChatGPT's performance across a wide range of clinical scenarios and diagnostic challenges. By assessing its capabilities in various specialities, the scoping review provided valuable perspectives into the model's potential applications in clinical practice, medical education, and research (see Table 2).

Challenges and limitations

The scoping review of the evaluation of ChatGPT's performance in responding to clinical questions identified nine challenges and limitations. According to Table 3, the most frequently cited challenges and limitations of GPT were 'Subjectivity and Bias' constituting $n=9$ (15%)[20,21,35–41]. 'Sample Size Limitations',[35,39,40,42–45] 'Representation and Generalizability of Data',[21,34,37,39,40,44,46] 'Outdated Information and Training Data',[24,40,47–51] 'Variability in Responses and Interpretations'[27,38,40,43,49–51], and 'Lack of Validation or Verification Mechanisms'[21,29,37,40,50–52] were the second most frequently mentioned $n=7$, representing 11.66% of the occurrences. The third theme was 'Limitations in Clinical Reasoning and Judgment' with six studies included (10%)[30,38,43,46,48,53]. In general, these challenges and limitations emphasized the need for continued refinement and improvement in utilizing ChatGPT for clinical decision-making purposes.

Evaluation methods

The evaluation of ChatGPT's performance encompassed a wide array of methodologies, reflecting a rigorous academic approach aimed at assessing its capabilities across diverse medical domains. These methodologies can be classified into several categories.

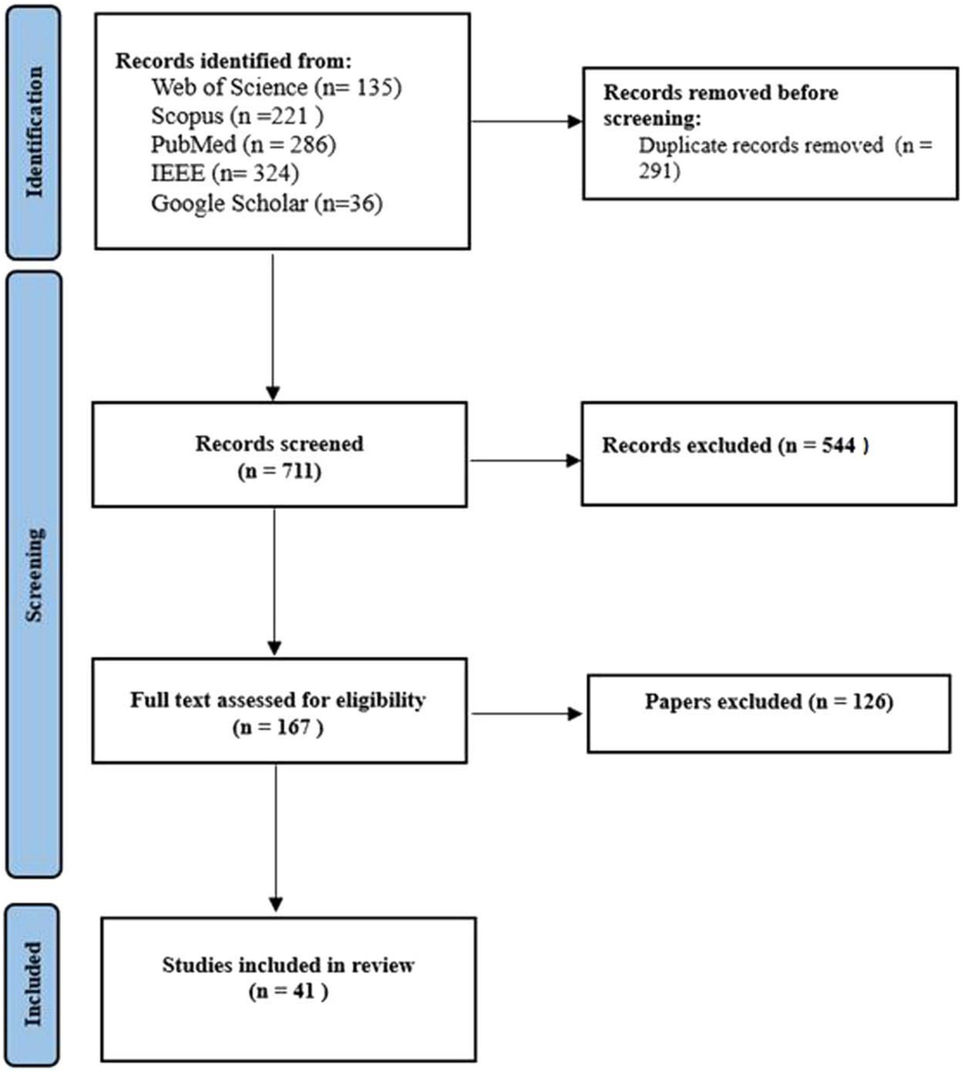
Identification
Records identified from:
Web of Science (n= 135)
Scopus (n =221)
PubMed (n = 286)
IEEE (n= 324)
Google Scholar (n=36)
Records removed before screening:
Duplicate records removed (n = 291)
Screening
Records screened (n = 711)
Records excluded (n = 544)
Full text assessed for eligibility (n = 167)
Papers excluded (n = 126)
Included
Studies included in review (n = 41)

Figure 1. Scoping literature review procedure based on PRISMA-ScR.

Firstly, there was a significant emphasis on comparing ChatGPT's performance with expert opinion and physicians' decision -making[37,38,40,49]. This involved benchmarking against human experts, evaluating responses against established guidelines, and assessing its ability to handle real-world clinical scenarios based on published vignettes. Secondly, standardized exam questions played a crucial role, with evaluations conducted using multiple-choice questions, theoretical questions, and diagnostic reasoning tasks across various medical specialties[28,39,43,52,54,55]. Statistical analysis was also employed extensively to analyze evaluation data and results[27,35]. Additionally, academic evaluations were conducted in academic settings, providing further insights into ChatGPT's performance[56]. Communication and decision support abilities were assessed through scenarios and suggestions[50,57]. Other methodologies included grading responses by specialists, investigating the impact of health literacy, evaluating emotional awareness, and assessing performance across specific medical subfields[36,42,46,47,51,58].

Another crucial aspect was the creation and assessment of datasets used to train and test ChatGPT[21]. Efforts were made to generate diverse datasets to ensure comprehensive evaluation and gage the complexity of the data utilized. Moreover, quality assessments were conducted using global scales and guidelines to measure the overall accuracy and adherence of ChatGPT's responses to established medical standards[29,42,51].

Furthermore, ChatGPT's performance was evaluated through comparison with historical cohorts of human experts, providing insights into its advancements compared to past practices[58]. Additionally, the impact of language and cultural nuances was explored by evaluating ChatGPT's performance in tasks related to Chinese medical knowledge[58].

Moreover, specific medical specialties, such as ophthalmology and dermatology, underwent detailed evaluations to assess ChatGPT's performance in these domains[29,44,47]. Assessment methodologies also included analyzing ChatGPT's responses to specialty-specific examination questions, such as those from the Dermatology Specialty Certificate Examination and pharmacist licensing exams[44,57]. Furthermore, the ability to generate clinical letters, provide radiology screening reports, and suggest appropriate clinical decision support alerts were evaluated to assess

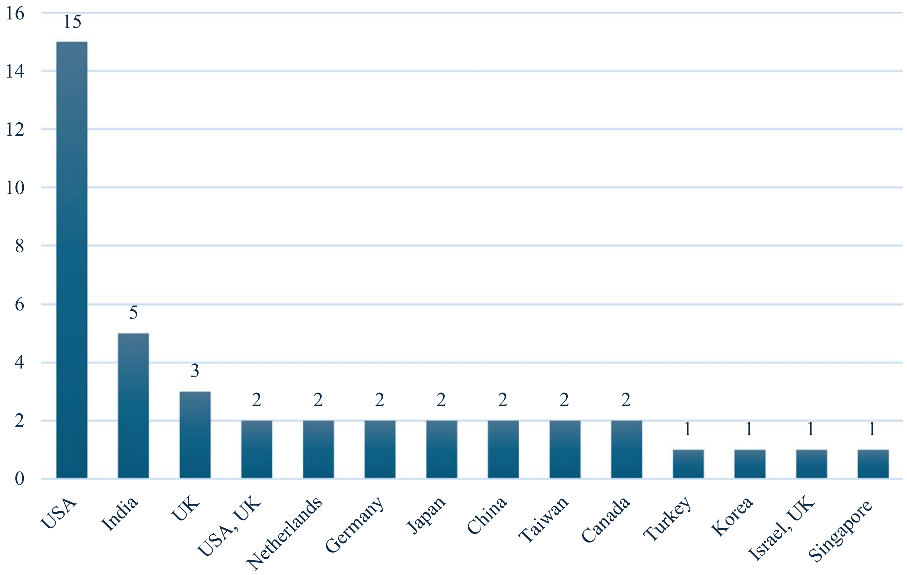

Figure 2. Distribution of documents by year.

ChatGPT's practical utility in clinical settings[39,50,59]. These evaluation methods, combined with the previously mentioned approaches, provided a comprehensive understanding of ChatGPT's performance and capabilities in the medical domain.

Questionnaire sources

This study presents a comprehensive analysis of the diverse array of medical questions and scenarios collected from various sources and contexts. These encompass clinical consultations, educational curricula, standardized medical exams, and training resources. In this section, we categorize the data into three main themes to provide a structured overview of the sources and contexts from which these medical inquiries were gathered.

Clinical consultations and physician queries: This category encompasses a wide range of medical questions and scenarios sourced directly from clinical practice and professional inquiries. It includes data from clinical consultations at a tertiary hospital[37], questions generated by physicians across multiple specialties[21], and patient queries curated from various sources such as bariatric surgery support groups and medical encyclopedias[46].

Educational and examination materials: This category involves questions derived from educational curricula and standardized medical exams. It includes questions based on competency-based medical education (CBME) curriculum for microbiology[35], licensing exams in different countries[37,41,48,52,57], specialty board exams like ophthalmology and dermatology[29,31,32,44], and assessments for general practitioners[39].

Training and assessment resources: This category encompasses datasets and resources utilized for training machine learning models and assessing medical knowledge. It includes datasets of annotated discharge summaries and screening reports for machine learning training[55,60], self-assessment exams for neurosurgery and general medical knowledge[25,39,61,62], as well as clinical scenarios and quizzes for educational purposes[45,51,59].

These categories provide a structured overview of the diverse sources and contexts from which medical questions and scenarios were gathered, providing valuable perspectives for analysis and learning. Moreover, the questionnaire sources encompassed a wide range of medical assessments, including biochemistry question papers, clinical vignettes from renowned medical manuals, and specialty board examination questions. These resources provided diverse clinical scenarios and diagnostic challenges, allowing for a robust evaluation of ChatGPT's performance across different medical disciplines and contexts. In summary, the questionnaire sources in this study were thoughtfully selected to provide a rigorous and academically sound evaluation of ChatGPT's performance in medical question answering. The diverse array of sources

Table 2

Summary of the retrieved publications' characteristics

Study medical domain	Frequency ($n = 41$) (Percent (100%)
Health and medical education	5 (12.20)
Ophthalmology	4 (9.76)
Clinical and medical informatics	3 (7.32)
Surgery	2 (4.88)
Biochemistry	2 (4.88)
Cardiology	2 (4.88)
Dermatology	2 (4.88)
General medicine	2 (4.88)
Microbiology	2 (4.88)
Multiple specialties	2 (4.88)
Neurosurgery	2 (4.88)
Urology	2 (4.88)
Radiology	2 (4.88)
Clinical workflow	1 (2.44)
ENT	1 (2.44)
Family medicine	1 (2.44)
Gastroenterology and hepatology	1 (2.44)
Obstetrics and gynecology	1 (2.44)
Pharmacy	1 (2.44)
Physiology	1 (2.44)
Psychology	1 (2.44)
Oncology	1 (2.44)

Table 3

Summary of results from thematic analysis of challenges and limitations of GPT

Number	Key emergent themes	No. of publications defining this theme	Percentages	Explanation
1.	Subjectivity and bias	9	15	Numerous statements discuss the subjective nature of evaluations, scoring, or biases introduced by human adjudication
2.	Sample size limitations	7	11.66	Several statements highlight the modest sample size used in their studies, indicating potential limitations in generalizability
3.	Representation and generalizability of data	7	11.66	Many statements mention limitations in the representativeness of the dataset or questions used, which could impact the model's performance across different scenarios or medical specialities
4.	Outdated information and training data	7	11.66	Multiple statements raise concerns about the model being trained on outdated data or its inability to incorporate new information beyond a certain date
5.	Variability in responses and interpretations	7	11.66	Several statements mention variations in responses based on prompt wording, user interactions, or differences in interpretation
6.	Lack of validation or verification mechanisms	7	11.66	Some statements express concerns about the absence of validation mechanisms or the inability to verify the accuracy of responses
7.	Limitations in clinical reasoning and judgment	6	10	Various statements discuss limitations related to the model's clinical reasoning, judgment, or ability to provide accurate recommendations
8.	Language and cultural variations	5	8.33	Several statements mention the influence of language differences, regional variations, or cultural contexts on response accuracy and validity
9.	Inability to process visual information:	5	8.33	Many statements highlight the model's inability to process images, graphs, or clinical photographs, which could limit its performance in certain medical specialities

ensured the breadth and depth of evaluation, contributing to the credibility and validity of the study's findings.

ChatGPT target users

Our study targeted a diverse array of stakeholders within the medical community, aiming to address the needs of different user groups. Medical education providers ($n=3$)[31,45,57], medical students ($n=1$)[35], nonexpert clinicians and healthcare providers ($n=1$)[48], and skin cancer patients ($n=1$)[59] were among the target users identified in the scoping review.

Medical students can benefit from enhanced learning experiences and clinical reasoning skills, while nonexpert clinicians and healthcare providers can leverage ChatGPT as a supplementary tool for clinical decision-making. Medical education providers stand to gain insights into integrating AI technologies like ChatGPT into their curriculum, increasing interactive, and engaging learning environments. Moreover, skin cancer patients can access clear and concise information about their condition and treatment options, empowering them to make informed decisions about their healthcare journey. This comprehensive approach to addressing the needs of diverse user groups highlights the potential of ChatGPT to revolutionize medical education, patient care, and knowledge dissemination within the medical community.

Test and comparison of GPT with other models

In evaluating ChatGPT's performance, the scoping review included tests and comparisons with other systems or benchmarks to assess its effectiveness and identify areas for improvement. The analysis revealed instances where ChatGPT was tested or compared against other models or approaches, demonstrating a significant focus on evaluating its performance in various scenarios.

Specifically, ChatGPT was tested in 26 studies[20,21,25–31,34,35,37–39,41–43,46–48,52,53,56,57,59,63] and medical contexts, indicating the model's versatility and potential applications across different domains. Additionally, there were 15 studies[24,32,36,40,44,45,49–51,54,55,58,60–62] where GPT was directly compared with other models or approaches, highlighting the importance of comparative analyses to assess GPT's effectiveness relative to existing solutions. Interestingly, in one instance, GPT was mentioned in both testing and comparison contexts, showcasing the multifaceted nature of its evaluation (Fig. 3).

These tests and comparisons provided valuable perspectives into ChatGPT's strengths and weaknesses, informing future research and development efforts aimed at enhancing its performance and utility in healthcare settings. By assessing its performance against other systems or benchmarks, the scoping review contributed to advancing the understanding of ChatGPT's capabilities and limitations in addressing clinical questions.

Gold standards

In evaluating ChatGPT's accuracy and reliability, the scoping review analyzed the use of gold standards in assessing its performance. Gold standards refer to established criteria or benchmarks used to validate the accuracy and reliability of AI models' responses.

The analysis of the use of Gold Standards in GPT revealed two studies where studies indicated the use of gold standards to assess ChatGPT's performance, highlighting the importance of rigorous evaluation methods[55,58]. However, a significant proportion of studies ($n=39$) did not utilize gold standards, suggesting variations in evaluation practices across different studies[20,21,24–32,34–57,59–63]. (see Table 3).

Despite the challenges associated with establishing gold standards in AI model evaluation, their incorporation is crucial for ensuring the validity and reliability of assessment outcomes. By adhering to established criteria or benchmarks, researchers can enhance the credibility and reproducibility of their findings, thereby advancing the field of AI in healthcare.

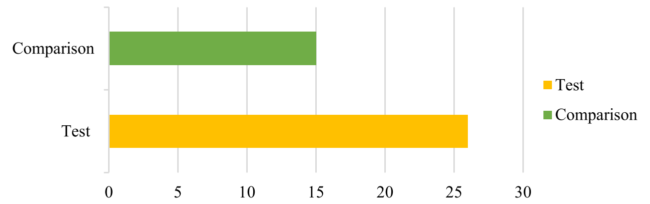

Figure 3. The distribution of test and comparison of GPT with other models.

GPT versions

In assessing the performance of ChatGPT, the scoping review examined different versions of the GPT model to identify trends in usage and prevalence. The analysis revealed varying frequencies of mentions for different GPT versions, with GPT 3.5 emerging as the most frequently mentioned version, garnering a frequency count of 30[20,21,25,26,28–31,34–42,45,46,48–53,56–59,63]. This indicates a significant focus on evaluating the performance of GPT 3.5 within the context of responding to clinical questions. Following closely, the combination of GPT 3.5 and GPT 4 was mentioned seven times[32,43,44,54,60–62], suggesting a potential interest in comparing the capabilities of these two versions.

Interestingly, the combination of GPT 3 and GPT 3.5 appeared twice[24,55], indicating that researchers may have explored the performance of earlier versions alongside more recent iterations. Additionally, individual mentions of GPT 4[47] and GPT 3[27] indicate the relevance of these versions in specific evaluation contexts, albeit with lower frequencies (Fig. 4).

Study conclusions

This study provides an overview of the conclusions extracted from studies evaluating ChatGPT's performance, reliability, educational utility, limitations, future directions, ethical considerations, potential applications, and comparisons with other models. Through a thematic analysis of these studies, this review aims to elucidate the current understanding of ChatGPT's role and impact in healthcare and to identify areas for further research and development.

Performance evaluation of ChatGPT

Studies have evaluated ChatGPT's performance in various medical contexts, including board exams, specialty exams, and clinical decision-making scenarios. Some studies report promising results, while others highlight limitations and areas for improvement[20,21,26,27,30,31,35–39,41,43,44,52,53,55,59,63].

Accuracy and reliability

While ChatGPT demonstrates high accuracy and completeness in generating responses, there are concerns regarding reliability, especially in complex or ambiguous scenarios. Some studies emphasize the need for verification by human experts to ensure the correctness of information provided by ChatGPT[21,25,26,35,37–39,43,47–49,52,54,56,59,62,63].

Educational tools and clinical support

Authors acknowledge the potential of ChatGPT as a supplementary tool for medical education, board exam preparation, and clinical decision-making. ChatGPT's ability to provide accurate information can aid both learners and practitioners in accessing relevant medical knowledge[24,32,34,41,43,47,52,54,56–58,63].

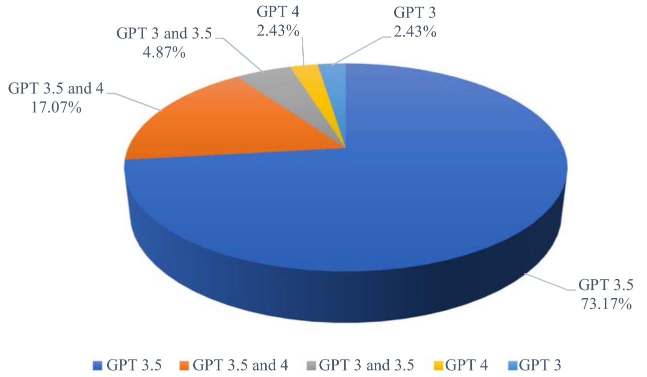

Figure 4. The distribution of the GPT Versions.

Limitations and challenges

Several studies highlight the limitations of ChatGPT, including its inability to handle complex or ambiguous questions, the potential for generating incorrect responses, and the need for continual improvement in training and development. Challenges such as reliance on human verification, ethical considerations, and potential biases are also discussed[31,34,44,49,50,54,56,58,63].

Future directions and recommendations

The authors suggest avenues for future research and development to enhance ChatGPT's performance, address limitations, and ensure safe and effective integration into healthcare practice. Recommendations include multidisciplinary collaboration, ongoing evaluation, regulation, and user education[20,31,40,43,47,52,54,57].

Ethical and regulatory considerations

Discussions on the ethical implications of using ChatGPT in healthcare, including concerns about patient safety, privacy, algorithm transparency, and responsible use. Recommendations are made for regulatory frameworks and guidelines to govern the development and deployment of AI in clinical settings[26,31,34,36,49,59].

Potential applications and benefits

Despite challenges and limitations, authors recognize the potential benefits of ChatGPT in improving access to medical information, supporting learning, enhancing clinical decision-making, and addressing healthcare workforce shortages[20,24,26,29,32,36,43,50,52].

Performance comparison and model evolution

Some studies compare the performance of different versions of ChatGPT or evaluate its performance against other models or human performance. These comparisons provide insights into the evolution of language models and their potential applications in healthcare[24,26,32,43,60,62].

This thematic analysis provides a comprehensive overview of the conclusions drawn by various authors regarding ChatGPT's performance, limitations, potential applications, and future directions in healthcare. Each theme reflects key aspects of ChatGPT's role and impact in medical practice and education, along with considerations for its safe and effective use.

Discussion

The scoping review conducted in this study provides a comprehensive evaluation of ChatGPT's performance in responding to clinical questions across various medical domains and user groups within the healthcare sector. This analysis indicates the multifaceted capabilities and limitations of ChatGPT in supporting critical healthcare functions. Moreover, it provides a comprehensive analysis of various studies assessing ChatGPT's efficacy across diverse medical domains within the healthcare sector. Our scoping review conducted in this study aims to enrich our understanding of ChatGPT's applicability in healthcare and educational settings across various medical domains, as evidenced by studies conducted by Sarink *et al.* (2023)[49], Johnson *et al.* (2023)[21], Samaan *et al.* (2023)[46], Das *et al.* (2023)[35], Yeo *et al.* (2023)[40], Mahat *et al.* (2023)[56], and others.

Sarink *et al.*[49] (2023) investigated ChatGPT's performance in 40 clinical consultations at a tertiary hospital in the Netherlands, focusing on medical microbiology, thereby aligning with our aim of assessing its utility in real-world clinical scenarios. Similarly, Johnson *et al.* (2023) evaluated ChatGPT's accuracy and comprehensiveness across 17 medical specialities in the USA, providing valuable perspectives into its performance in diverse clinical settings. Their findings contribute significantly to our understanding of ChatGPT's applicability across various medical contexts[21]. In addition, Samaan *et al.*[46] (2023) examined ChatGPT's accuracy and reproducibility in answering patient questions about bariatric surgery in the USA and UK, aligning with our goal of assessing its effectiveness in patient communication and education. Yeo *et al.*[40] (2023) investigated ChatGPT's accuracy in providing knowledge, management, and emotional support for cirrhosis and hepatocellular carcinoma, clarifying its potential role in supporting patients with complex medical conditions. Furthermore, Mahat *et al.* (2023) assessed ChatGPT's performance in solving biochemistry questions in India, further emphasizing its role in healthcare domains. These studies contribute to our understanding of ChatGPT's performance and potential applications in various medical domains and educational settings, highlighting both its strengths and areas for improvement[56].

This study indicates the importance of continuous improvement and rigorous evaluation of AI models like ChatGPT to ensure their safe and effective integration into clinical practice. As AI technologies become increasingly indispensable in healthcare, it is crucial to address any challenges and limitations identified during the evaluation process. By doing so, healthcare organizations can enhance the reliability, accuracy, and safety of AI-driven interactions, ultimately improving patient care outcomes and enhancing trust among healthcare providers, clinicians, and patients. Among these challenges is the lack of access to official answer keys, which raises concerns about the reliability and accuracy of ChatGPT's responses. Additionally, issues regarding the representativeness of question sources and the opacity of the dataset used for training highlight the need for transparency and rigor in AI model development and evaluation. Moreover, our review indicates the potential for confabulation or 'hallucination' in responses, emphasizing the importance of stringent validation of AI-generated outputs. Despite these challenges, ChatGPT's diverse applications within the healthcare landscape, from medical education to patient empowerment, are acknowledged.

Studies have been conducted in different countries, investigating its effectiveness in diverse medical specialities and educational contexts. In recent years, there has been a surge in research aimed at evaluating the performance and applicability of ChatGPT across various domains within the medical field. Rao *et al.*[53] (2023) evaluated ChatGPT's capacity for ongoing clinical decision support in the USA, focusing on standardized clinical vignettes. Moreover, Huang *et al.* (2023)[43] benchmarked ChatGPT's performance in radiation oncology in Germany, providing valuable perspectives into its utility in specialized medical fields. Their study aligns with our aim of exploring ChatGPT's effectiveness across diverse medical domains.

Balel (2023)[42] assessed the usability of information generated by ChatGPT in oral and maxillofacial surgery in Turkey. Similarly, Kaneda *et al.* (2023)[37] examined the potential utilization of ChatGPT in the Japanese clinical setting by investigating the Japanese National Medical Licensing Examination. Zhu *et al.* (2023)[51] evaluated the ability of ChatGPT to provide correct and useful information on common problems related to prostate cancer in China. These studies contribute to our understanding of ChatGPT's performance and potential applications in various medical domains and educational settings. They highlight both the strengths and limitations of ChatGPT in addressing complex medical queries and providing significant insights for future development and implementation.

While other studies focus on specific areas like biochemistry university examinations or microbiology knowledge questions, the scoping review provides a comprehensive overview of ChatGPT's performance across diverse medical domains. This holistic approach enables healthcare providers and educators to gain insights into ChatGPT's potential applications and tailor its usage to specific clinical or educational needs. Das *et al.* (2023)[35] evaluated ChatGPT's performance in answering microbiology-related questions in India, clarifying its utility in different cultural and medical contexts. Moreover, Mahat *et al.* (2023)[56] assessed ChatGPT's performance in solving biochemistry questions in India, highlighting its utility in medical education and knowledge dissemination. Their study is aligned with our aim of exploring ChatGPT's effectiveness in educational contexts. This study provides valuable perspectives into ChatGPT's cross-cultural applicability. For instance, ChatGPT can be utilized as an educational tool for medical sciences students, aiding in their learning process and enhancing their understanding of complex medical concepts. Moreover, medical students, nonspecialist clinicians, medical educators, and patients can potentially benefit from ChatGPT's capabilities, albeit with the recognition of the variability in its performance across different medical domains. Therefore, continual refinement and improvement are imperative to ensure the reliability and accuracy of ChatGPT in diverse clinical scenarios.

The analysis reveals a clear dominance of GPT 3.5, indicating its strong foothold in natural language processing tasks and text generation within the healthcare domain. However, the inclusion of other versions like GPT 3 and GPT 4 suggests their relevance in specialized applications or research domains. This nuanced understanding of GPT model usage patterns provides valuable perspectives into the evolving dynamics surrounding the adoption and utilization of different GPT versions for distinct purposes within the healthcare sector.

Aligned with its overarching objective, this review meticulously examines and synthesizes findings from multiple studies to provide significant insights into ChatGPT's utility, constraints, and associated challenges in addressing clinical queries posed by stakeholders across the healthcare continuum. This approach contrasts with prior research, which often focused on specific medical specialties or user groups, thus providing a more holistic understanding of ChatGPT's performance.

The review illuminates various facets of ChatGPT's performance, highlighting both commendable attributes and areas necessitating improvement. While ChatGPT demonstrates promise in supporting healthcare functions such as screening, diagnosis, treatment planning, medication management, and patient education, it faces challenges that hinder its seamless integration into clinical practice. The future of the study suggests several avenues for advancing our understanding of ChatGPT's role in healthcare. Future research should prioritize longitudinal studies to assess ChatGPT's performance over extended periods, personalized responses tailored to individual patient characteristics, improved transparency and interpretability of ChatGPT's output, and integration with existing healthcare systems to streamline its use in clinical practice. By addressing these areas, future research can enhance ChatGPT's utility as a valuable tool for supporting clinicians and improving patient care.

Implication

This study provides several significant benefits. Firstly, it provides a comprehensive analysis of existing research on ChatGPT's efficacy across diverse medical domains and user demographics within the healthcare sector. By synthesizing insights from various studies, our review provides a significant understanding of ChatGPT's utility, constraints, and associated challenges in addressing clinical queries. Additionally, our study highlights both commendable attributes and areas necessitating improvement, thus guiding future research and development efforts. Furthermore, our scoping review contributes to advancing the field of AI in healthcare by clarifying the potential applications of ChatGPT in clinical practice and medical education. In summary, our study provides valuable perspectives that can inform healthcare professionals, researchers, and policymakers about the role of ChatGPT in modern healthcare delivery.

Limitations

One limitation of this study lies in the search strategy and information sources employed. Despite conducting searches in various scientific databases and repositories, including PubMed, Scopus, IEEE, Web of Science, Google Scholar, and Cochrane, there is a possibility that relevant studies may have been missed. Additionally, our search was limited to papers published in English and studies conducted on human subjects only, which may have excluded valuable research published in other languages or focusing on animal models. Furthermore, the search timeframe from 1st August to 15th August 2023, imposes a constraint that may have excluded relevant studies published before or after this period. These limitations could potentially impact the comprehensiveness of our review and the generalizability of our findings.

Conclusion

In conclusion, this study conducted a scoping review to explore ChatGPT in clinical inquiry, with an emphasis on its characteristics, applications, challenges, and evaluation. Through this review, we have highlighted ChatGPT's promising role in medical education, patient care, and decision support, while also identifying areas for improvement, such as reliability, accuracy, and cross-cultural applicability. By addressing these challenges and leveraging ChatGPT's strengths, we can optimize its integration into clinical practice, ultimately enhancing patient care outcomes. The study results contribute to the ongoing discourse surrounding AI-driven technologies in healthcare, driving further research and development in this dynamic field.

Ethical approval

Not applicable.

Consent

Not applicable.

Source of funding

Not applicable.

Author contribution

S.A., Y.A., and H.S.: conceived the idea of the manuscript and made substantial contributions to this scoping review, conducted the literature search, and conducted study selection and data extraction; S.A., Y.A., F.F., A.G., and H.S.: conceptualized and operationalized the object of interest; S.A. and H.S.: conducted the data analysis and synthesis and were involved in the drafting of the manuscript; S.A., Y.A., F.F., A.G., H.S., D.C., S.Z., and P.M.: revised the manuscript for important intellectual content; P.M., S.A., and F.F.: supervised the project and allocated resources accordingly. All authors read and approved the final manuscript.

Conflicts of interest disclosure

The authors declare no conflict of interest.

Research registration unique identifying number (UIN)

Our article is a scoping review, and since it is not a systematic review, it cannot be registered. PROSPERO does not accept scoping reviews, literature reviews, or mapping reviews. Additionally, this study was conducted without any funding, so it is not feasible for us to pay registration fees on paid websites. We hope that we can publish this work in your journal. Thank you for your consideration.

Guarantor

Hosna Salmani, Department of Health Information Management, School of Health Management and Information Sciences, Iran University of Medical Sciences, Tehran, Iran. E-mail: salmani.h@iums.ac.ir.

Data availability statement

The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request.

Provenance and peer review

Not commissioned, externally peer-reviewed.

Acknowledgements

Not applicable.

References

[1] Feuerriegel S, Hartmann J, Janiesch C, *et al.* Generative AI. Bus Inform Syst Eng 2024;66:111–26.
[2] Jo A. The promise and peril of generative AI. Nature 2023;614:214–6.
[3] Euchner J. Generative AI. Res Technol Manag 2023;66:71–4.
[4] Fui-Hoon Nah F, Zheng R, Cai J, *et al.* Generative AI and ChatGPT: Applications, challenges, and AI-human collaboration. Taylor & Francis; 2023:277–304.
[5] Cascella M, Montomoli J, Bellini V, *et al.* Evaluating the feasibility of ChatGPT in healthcare: an analysis of multiple clinical and research scenarios. J Med Syst 2023;47:33.
[6] Clusmann J, Kolbinger FR, Muti HS, *et al.* The future landscape of large language models in medicine. Communicat Med 2023;3:141.
[7] Morita PP, Abhari S, Kaur J, *et al.* Applying ChatGPT in public health: a SWOT and PESTLE analysis. Front Public Health 2023;11:1225861.
[8] Javaid M, Haleem A, Singh RP. ChatGPT for healthcare services: an emerging stage for an innovative perspective. BenchCouncil Trans Benchmarks Standards Eval 2023;3:100105.
[9] Li J, Dada A, Puladi B, *et al.* ChatGPT in healthcare: a taxonomy and systematic review. Comput Methods Programs Biomed 2024;245:108013.
[10] Liu J, Wang C, Liu S. Utility of ChatGPT in clinical practice. J Med Internet Res 2023;25:e48568.
[11] Sandmann S, Riepenhausen S, Plagwitz L, *et al.* Systematic analysis of ChatGPT, Google search and Llama 2 for clinical decision support tasks. Nat Commun 2024;15:2050.
[12] Khan RA, Jawaid M, Khan AR, *et al.* ChatGPT-Reshaping medical education and clinical management. Pak J Med Sci 2023;39:605.
[13] Lee H. The rise of ChatGPT: exploring its potential in medical education. Anatom Sci Educat 2024;17:926–31.
[14] Sallam M, Salim NA, Al-Tammemi AB, *et al.* ChatGPT output regarding compulsory vaccination and COVID-19 vaccine conspiracy: a descriptive study at the outset of a paradigm shift in online search for information. Cureus J Med Sci 2023;15:e35029.
[15] Meo SA, Al-Masri AA, Alotaibi M, *et al.* ChatGPT knowledge evaluation in basic and clinical medical sciences: multiple choice question examination-based performance. Healthcare 2023;11:2046.
[16] Oztermeli AD, Oztermeli A. ChatGPT performance in the medical specialty exam: an observational study. Medicine 2023;102:e34673.
[17] Ali H, Patel P, Obaitan I, *et al.* Evaluating the performance of ChatGPT in responding to questions about endoscopic procedures for patients. iGIE 2023;2:553–9.
[18] Branum C, Schiavenato M. Can ChatGPT accurately answer a PICOT question? Assessing AI response to a clinical question. Nurse Educ 2023;48:231–3.
[19] Del Fiol G, Workman TE, Gorman PN. Clinical questions raised by clinicians at the point of care: a systematic review. JAMA Intern Med 2014;174:710–8.
[20] Ghosh A, Bir A. Evaluating ChatGPT's ability to solve higher-order questions on the competency-based medical education curriculum in medical biochemistry. Cureus 2023;15:e37023.
[21] Johnson D, Goodman R, Patrinely J, *et al.* Assessing the accuracy and reliability of AI-generated medical responses: an evaluation of the ChatGPT model. Res Sq 2023:rs.3.rs-2566942.
[22] Zhou Z. Evaluation of ChatGPT's capabilities in medical report generation. Cureus 2023;15:e37589.
[23] Thomas J, Harden A. Methods for the thematic synthesis of qualitative research in systematic reviews. BMC Med Res Methodol 2008;8:1–10.
[24] Gilson A, Safranek CW, Huang T, *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 2023;9:e45312.
[25] Huynh LM, Bonebrake BT, Schultis K, *et al.* New artificial intelligence ChatGPT performs poorly on the 2022 self-assessment study program for urology. Urol Pract 2023;10:409–15.
[26] Sharma M, Sharma S. Transforming maritime health with ChatGPT-powered healthcare services for mariners. Ann Biomed Eng 2023;51:1123–5.
[27] Strong E, DiGiammarino A, Weng Y, *et al.* Performance of ChatGPT on free-response, clinical reasoning exams. medRxiv 2023;183:1028–30.
[28] Benoit J. ChatGPT for Clinical Vignette Generation, Revision, and Evaluation2023.
[29] Ali MJ. ChatGPT and lacrimal drainage disorders: performance and scope of improvement. Ophthalmic Plast Reconstr Surg 2023;39:221–5.
[30] Antaki F, Touma S, Milad D, *et al.* Evaluating the performance of ChatGPT in ophthalmology: an analysis of its successes and shortcomings. Ophthalmol Sci 2023;3:100324.
[31] Mihalache A, Popovic MM, Muni RH. Performance of an artificial intelligence Chatbot in ophthalmic knowledge assessment. JAMA Ophthalmol 2023;141:589–97.
[32] Teebagy S, Colwell L, Wood E, *et al.* Improved performance of ChatGPT-4 on the OKAP exam: a comparative study with ChatGPT-3.5. medRxiv 2023;15:e184–7.
[33] Huh S. Are ChatGPT's knowledge and interpretation ability comparable to those of medical students in Korea for taking a parasitology examination?: a descriptive study. J Educ Eval Health Prof 2023;20:1.
[34] Kumah-Crystal Y, Mankowitz S, Embi P, *et al.* ChatGPT and the clinical informatics board examination: the end of unproctored maintenance of certification? J Am Med Inform Assoc 2023;30:1558–60.
[35] Das D, Kumar N, Longjam LA, *et al.* Assessing the capability of ChatGPT in answering first- and second-order knowledge questions on microbiology as per competency-based medical education curriculum. Cureus 2023;15:e36034.
[36] Elyoseph Z, Hadar-Shoval D, Asraf K, *et al.* ChatGPT outperforms humans in emotional awareness evaluations. Front Psychol 2023;14: 1199058.
[37] Kaneda Y, Tanimoto T, Ozaki A, *et al.* Can ChatGPT Pass the 2023 Japanese National Medical Licensing Examination?2023.
[38] Rao A, Pang M, Kim J, *et al.* Assessing the utility of ChatGPT throughout the entire clinical workflow. medRxiv 2023;25:e48659.
[39] Thirunavukarasu AJ, Hassan R, Mahmood S, *et al.* Trialling a large language model (ChatGPT) in general practice with the applied knowledge test: observational study demonstrating opportunities and limitations in primary care. JMIR Med Educ 2023;9:e46599.
[40] Yeo YH, Samaan JS, Ng WH, *et al.* Assessing the performance of ChatGPT in answering questions regarding cirrhosis and hepatocellular carcinoma. Clin Mol Hepatol 2023;29:721–32.
[41] Yu P, Fang C, Liu X, *et al.* Performance of ChatGPT on the Chinese postgraduate examination for clinical medicine: survey study. JMIR Med Educ 2024;10:e48514.
[42] Balel Y. Can ChatGPT be used in oral and maxillofacial surgery? J Stomatol Oral Maxillofac Surg 2023;124:101471.
[43] Huang Y, Gomaa A, Semrau S, *et al.* Benchmarking ChatGPT-4 on a radiation oncology in-training exam and Red Journal Gray Zone cases: potentials and challenges for AI-assisted medical education and decision making in radiation oncology. Front Oncol 2023;13:1265024.
[44] Passby L, Tso S, Wernham A. Performance of ChatGPT on dermatology specialty certificate examination multiple choice questions. Clin Exp Dermatol 2023;48:585–90.
[45] Harskamp RE, De Clercq Clercq L. Performance of ChatGPT as an AI-assisted decision support tool in medicine: a proof-of-concept study for interpreting symptoms and management of common cardiac conditions (AMSTELHEART-2). Acta Cardiol 2024;79:358–66.
[46] Samaan JS, Yeo YH, Rajeev N, *et al.* Assessing the accuracy of responses by the language model ChatGPT to questions regarding bariatric surgery. Obes Surg 2023;33:1790–6.
[47] Hoch CC, Wollenberg B, Lüers JC, *et al.* ChatGPT's quiz skills in different otolaryngology subspecialties: an analysis of 2576 single-choice and multiple-choice board certification preparation questions. Eur Arch Otorhinolaryngol 2023;280:4271–8.
[48] Kusunose K, Kashima S, Sata M. Evaluation of the accuracy of ChatGPT in answering clinical questions on the japanese society of hypertension guidelines. Circ J 2023;87:1030–3.

[49] Sarink MJ, Bakker IL, Anas AA, *et al.* A study on the performance of ChatGPT in infectious diseases clinical consultation. Clin Microbiol Infect 2023;29:1088–9.

[50] Liu S, Wright AP, Patterson BL, *et al.* Using AI-generated suggestions from ChatGPT to optimize clinical decision support. J Am Med Inform Assoc 2023;30:1237–45.

[51] Zhu L, Mou W, Chen R. Can the ChatGPT and other large language models with internet-connected database solve the questions and concerns of patient with prostate cancer and help democratize medical knowledge? J Transl Med 2023;21:269.

[52] Weng TL, Chen TJ. ChatGPT failed Taiwan's family medicine board exam. J Chin Med Assoc 2023;86:865.

[53] Rao A, Kim J, Kamineni M, *et al.* Evaluating ChatGPT as an adjunct for radiologic decision-making. medRxiv 2023;20:990–7.

[54] Oh N, Choi GS, Lee WY. ChatGPT goes to the operating room: evaluating GPT-4 performance and its potential in surgical education and training in the era of large language models. Ann Surg Treat Res 2023; 104:269–73.

[55] Hu Y, Chen Q, Du J, *et al.* Improving large language models for clinical named entity recognition via prompt engineering. J Am Med Inform AssocJAMIA 2024;31:1812–20.

[56] Mahat RK, Jantikar AM, Rathore V, *et al.* Assessing the performance of ChatGPT to solve biochemistry question papers of university examination. Adv Physiol Educ 2023;47:528–9.

[57] Wang YM, Shen HW, Chen TJ. Performance of ChatGPT on the pharmacist licensing examination in Taiwan. J Chin Med Assoc 2023;86: 653–8.

[58] Li SW, Kemp MW, Logan SJS, *et al.* ChatGPT outscored human candidates in a virtual objective structured clinical examination in obstetrics and gynecology. Am J Obstet Gynecol 2023;229:172.e1–12.

[59] Ali SR, Dobbs TD, Hutchings HA, *et al.* Using ChatGPT to write patient clinic letters. The Lancet Digital Health 2023;5:e179–81.

[60] Lyu Q, Tan J, Zapadka ME, *et al.* Translating radiology reports into plain language using ChatGPT and GPT-4 with prompt learning: results, limitations, and potential. Vis Comput Ind Biomed Art 2023;6:9.

[61] Ali R, Tang OY, Connolly ID, *et al.* Performance of ChatGPT, GPT-4, and Google Bard on a neurosurgery oral boards preparation question bank. Neurosurgery 2023;93:1090–8.

[62] Ali R, Tang OY, Connolly ID, *et al.* Performance of ChatGPT and GPT-4 on neurosurgery written board examinations. Neurosurgery 2023;93: 1353–65.

[63] Subramani M, Jaleel I, Krishna Mohan S. Evaluating the performance of ChatGPT in medical physiology university examination of phase I MBBS. Adv Physiol Educ 2023;47:270–1.