Ther Adv Gastroenterol

2023, Vol. 16: 1–10

DOI: 10.1177/
17562848231218618

Evaluating the role of ChatGPT in gastroenterology: a comprehensive systematic review of applications, benefits, and limitations

Eyal Klang, Ali Sourosh, Girish N. Nadkarni, Kassem Sharif and Adi Lahat

Abstract

Background: The integration of artificial intelligence (AI) into healthcare has opened new avenues for enhancing patient care and clinical research. In gastroenterology, the potential of AI tools, specifically large language models like ChatGPT, is being explored to understand their utility and effectiveness.

Objectives: The primary goal of this systematic review is to assess the various applications, ascertain the benefits, and identify the limitations of utilizing ChatGPT within the realm of gastroenterology.

Design: Through a systematic approach, this review aggregates findings from multiple studies to evaluate the impact of ChatGPT on the field.

Data sources and methods: The review was based on a detailed literature search of the PubMed database, targeting research that delves into the use of ChatGPT for gastroenterological purposes. It incorporated six selected studies, which were meticulously evaluated for quality using the Joanna Briggs Institute critical appraisal instruments. The data were then synthesized narratively to encapsulate the roles, advantages, and drawbacks of ChatGPT in gastroenterology.

Results: The investigation unearthed various roles of ChatGPT, including its use in patient education, diagnostic self-assessment, disease management, and the formulation of research queries. Notable benefits were its capability to provide pertinent recommendations, enhance communication between patients and physicians, and prompt valuable research inquiries. Nonetheless, it encountered obstacles in decoding intricate medical questions, yielded inconsistent responses at times, and exhibited limitations in generating novel content. The review also considered ethical implications.

Conclusion: ChatGPT has demonstrated significant potential in the field of gastroenterology, especially in facilitating patient–physician interactions and managing diseases. Despite these advancements, the review underscores the necessity for ongoing refinement, customization, and ethical regulation of AI tools. These findings serve to enrich the dialog concerning AI's role in healthcare, with a specific focus on ChatGPT's application in gastroenterology.

Plain language summary

Checking how ChatGPT works in gastroenterology: a detailed look at its uses, advantages, and challenges

Goal We looked at how ChatGPT, a computer program, is used in the study Gastroenterology. We wanted to understand what's good about it, what's challenging, and how it can help doctors and patients. How We Did It We searched for articles about ChatGPT in Gastroenterology on PubMed. We found six suitable articles and checked their quality using the Joanna Briggs Institute (JBI) critical appraisal tools. Then, we put all the information

Correspondence to:
Adi Lahat
Department of Gastroenterology, Sheba Medical Center at Tel Hashomer, Ramat Gan, 52621 Affiliated with Tel Aviv Medical School, Tel Aviv University, Tel Aviv, Israel
zokadi@gmail.com

Eyal Klang
Division of Data-Driven and Digital Medicine (D3M), Icahn School of Medicine at Mount Sinai, New York, NY, USA

The Charles Bronfman Institute of Personalized Medicine, Icahn School of Medicine at Mount Sinai, New York, NY, USA

ARC Innovation Center, Sheba Medical Center at Tel Hashomer Affiliated with Tel Aviv Medical School, Tel Aviv University, Tel Aviv, Israel

Ali Sourosh
Girish N. Nadkarni
Division of Data-Driven and Digital Medicine (D3M), Icahn School of Medicine at Mount Sinai, New York, NY, USA

The Charles Bronfman Institute of Personalized Medicine, Icahn School of Medicine at Mount Sinai, New York, NY, USA

Kassem Sharif
Department of Gastroenterology, Sheba Medical Center at Tel Hashomer Affiliated with Tel Aviv Medical School, Tel Aviv University, Tel Aviv, Israel

together to get a clear picture. What We Found Doctors and researchers use ChatGPT in many ways. Some use it to teach patients about their health, while others use it to help patients check their symptoms or manage their conditions. It can even help come up with research questions. The good things about ChatGPT are that it gives helpful advice, makes talking between doctors and patients easier, and helps come up with research topics. But, sometimes it doesn't understand hard medical questions, gives different answers for the same question, or lacks new ideas. There are also concerns about using it the right way. What This Means ChatGPT can be a helpful tool in Gastroenterology, especially when talking with patients and managing their health. But, there are challenges that need to be fixed. Our review helps people understand how ChatGPT can be used in health care, especially in the field of Gastroenterology.

Keywords: artificial intelligence, ChatGPT, gastroenterology, large language models, review

Received: 3 October 2023; revised manuscript accepted: 16 November 2023.

Introduction

Artificial intelligence (AI) and large language models (LLMs), such as ChatGPT developed by OpenAI, have been increasingly recognized for their potential to revolutionize various sectors, including healthcare.[1,2] In medicine, and more specifically in gastroenterology, these models have shown promise as supportive tools for clinicians, enhancing patient care and improving healthcare delivery.[3] However, while the potential benefits are substantial, the application of AI in healthcare is not without its challenges and limitations.[4,5]

ChatGPT, a conversational AI system based on the generative pre-trained transformer (GPT) architecture, has demonstrated impressive capabilities in various gastroenterological applications. These include answering common patient questions,[6] taking part in self-assessment tests,[7] and even identifying research priorities.[8] Despite these promising applications, the performance of ChatGPT in the medical domain has been inconsistent, with concerns raised about its accuracy and efficacy.[9]

A recent review article[10] noted that GPT-4 could be beneficial for patient–physician communication, patient education, and continuous patient care, potentially mitigating factors related to physicians' burnout. However, the authors highlighted key limitations and ethical considerations of this AI technology, including patient confidentiality and data security, algorithmic bias, inconsistent and inaccurate responses, plagiarism concerns, compliance with data privacy regulations, and the irreplaceable role of human judgment.

This review aims to provide an evaluation of the role of ChatGPT in gastroenterology, drawing from existing literature. By analyzing studies on the application of ChatGPT in patient communication, medical education, disease management, and research prioritization, we aim to provide a perspective on the potential and challenges of this tool in gastroenterology.

Methods

Study selection

For this systematic review, we included studies that examined the application of ChatGPT in gastroenterology. We excluded studies that focused on other AI models or other areas of healthcare.

Search strategy

We conducted a comprehensive literature search using the PubMed database. The search strategy incorporated a combination of Medical Subject Headings (MeSH) terms and keywords related to 'ChatGPT', and 'Gastroenterology'. The search was limited to articles published in English. Reference lists of included studies and relevant reviews were also manually searched to identify any additional studies.

Data extraction

Two independent reviewers extracted data from the included studies using a standardized data extraction form. Discrepancies were resolved through discussion or consultation with a third reviewer. The extracted information included: study design, sample size, application of ChatGPT (e.g. patient education, self-assessment, continuous care), main findings, and limitations.

Quality assessment

The quality of the included studies was assessed using the Joanna Briggs Institute (JBI) critical appraisal tools, appropriate to each study design. These tools assess the methodological quality of a study and the extent to which a study has addressed the possibility of bias in its design, conduct, and analysis. Studies were categorized as high, moderate, or low quality based on their scores. According to the JBI guidelines, it is recommended that critical appraisal be undertaken by at least two independent reviewers to minimize potential bias. In our study, the quality assessment using the modified JBI critical appraisal tools was conducted with the agreement of three authors (AL, EK, and KS) to ensure the objectivity and robustness of the evaluation.

Data synthesis

We conducted a narrative synthesis of the findings from the included studies. Due to the anticipated heterogeneity in study designs and outcomes, a meta-analysis was not planned. Instead, we focused on summarizing the applications, benefits, and limitations of ChatGPT in gastroenterology as reported in the studies, and on identifying areas for future research.

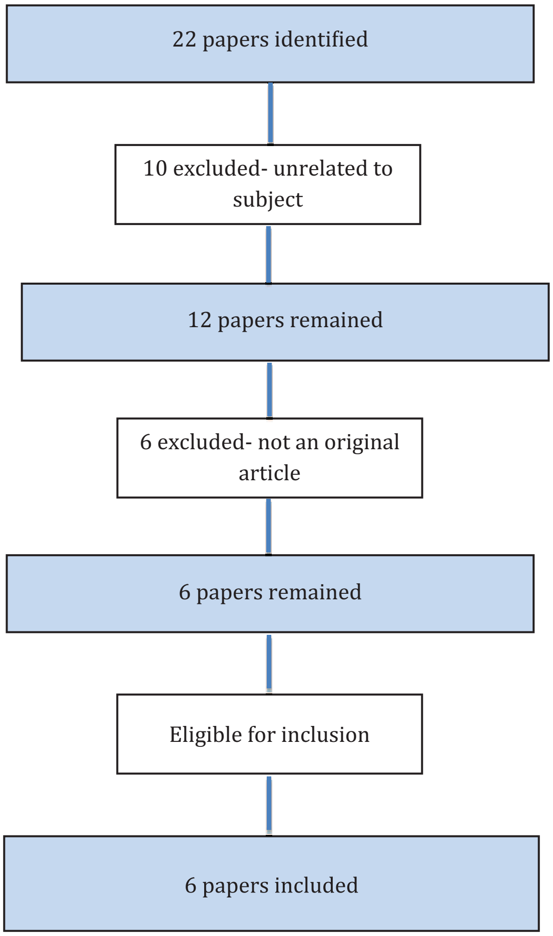

Figure 1. Flowchart delineating the selection procedure of the studies included in the review.

Results

The systematic review included six studies that evaluated the application, benefits, and limitations of ChatGPT in the field of gastroenterology. The studies were diverse in their objectives and methodologies, and they covered various aspects of gastroenterology, including patient education, self-assessment, patient–physician communication, disease management, and research question generation. The flowchart delineating the selection procedure of the studies included is depicted in Figure 1.

Table 1 summarizes the characteristics of studies included in the review.

Table 2 summarizes the main findings, and the identified benefits and limitations of ChatGPT in gastroenterology as presented in the studies included.

Table 3 summarizes the quality assessment performed according to JBI critical appraisal tools.

ChatGPT as a tool for patients

Two studies examined the efficacy of ChatGPT as a tool for patients, mainly in answering common patient questions. The first study by Lee *et al.*,[6] 'ChatGPT answers common patient questions about colonoscopy' found that ChatGPT provided answers similar in quality to non-AI answers but the text similarity was low. The AI-generated answers were written at higher grade reading levels, exceeding the recommended eighth-grade threshold.

Table 1. Characteristics of studies included in the review.

Paper title	Objective	Methodology	Participants	GPT type	Study type
ChatGPT answers common patient questions about colonoscopy[6]	Investigate the efficacy of ChatGPT in answering common questions about colonoscopy	Presented ChatGPT with eight common questions about colonoscopy, compared answers to hospital webpage answers	Four gastroenterologists	GPT-3.5	Comparative analysis
Chat generative pre-trained transformer fails the multiple-choice American College of Gastroenterology Self-Assessment Test[7]	Evaluate the performance of ChatGPT on gastroenterology self-assessment tests	Input the exact questions from the tests into ChatGPT-3 and ChatGPT-4		GPT-3.5, GPT-4	Performance evaluation
Evaluating the utility of a large language model in answering common patients' gastrointestinal health-related questions: are we there yet?[9]	Evaluate the performance of ChatGPT in answering patients' gastrointestinal health-related questions	ChatGPT was asked 110 real-life questions, responses were rated by three experienced gastroenterologists	Three experienced gastroenterologists	GPT-3.5	Performance evaluation
Evaluation of the potential utility of an artificial intelligence ChatBot in GERD management[11]	Evaluate the utility of ChatGPT in the management of GERD	Submitted 23 GERD management prompts to ChatGPT, responses were rated by three gastroenterologists and eight patients	Three gastroenterologists, eight patients	GPT-3.5	Performance evaluation
Evaluating the use of large language model in identifying top research questions in gastroenterology[8]	Evaluate the potential of ChatGPT in identifying research priorities in gastroenterology	ChatGPT was asked to generate five research questions for each of the four GI topics. Responses were evaluated by a panel of experienced gastroenterologists	A panel of experienced gastroenterologists	GPT-3.5	Performance evaluation
Harnessing language models for streamlined post-colonoscopy patient management: a novel approach[12]	Evaluate the potential of ChatGPT in providing post-colonoscopy guideline-based recommendations	Twenty clinical scenarios were presented to ChatGPT in the form of structured reports and free-text notes. Two senior gastroenterologists evaluated the responses generated by ChatGPT	Two senior gastroenterologists	GPT-3.5	Performance evaluation

GERD, gastroesophageal reflux disease; GI, gastrointestinal; GPT, generative pre-trained transformer.

Table 2. Key findings and limitations of ChatGPT as identified by studies included.

Paper title	Key findings	Limitations
ChatGPT answers common patient questions about colonoscopy[6]	ChatGPT provided answers similar in quality to non-AI answers, but text similarity was low. AI answers were written at higher grade reading levels	ChatGPT-generated medical information is not based on clinical evidence, the potential for manipulation through prompt engineering, concerns about implicit bias
Chat generative pre-trained transformer fails the multiple-choice American College of Gastroenterology Self-Assessment Test[7]	ChatGPT did not pass the gastroenterology self-assessment test	ChatGPT was not specifically trained on medical literature, models lacked up-to-date information, relied on freely available information for training
Evaluating the utility of a large language model in answering common patients' gastrointestinal health-related questions: are we there yet?[9]	ChatGPT could provide accurate and clear answers in some cases, but not in others	ChatGPT's understanding of complex medical questions is insufficient, accuracy, and helpfulness of responses varied significantly

(Continued)

Table 2. (Continued)

Paper title	Key findings	Limitations
Evaluation of the potential utility of an artificial intelligence ChatBot in GERD management[11]	ChatGPT provided appropriate and specific recommendations for GERD management in 91.3% of cases	Inconsistencies in responses to the same prompt, some potential PPI risks were stated as fact, some responses provided limited specific guidance
Evaluating the use of large language model in identifying top research questions in gastroenterology[8]	ChatGPT was able to generate relevant, clear, and moderately specific research questions, but struggled with originality	A small panel of experts, subjective nature of the ratings, focus on specific subfields of gastroenterology, lack of experiments with different prompts
Harnessing language models for streamlined post-colonoscopy patient management: a novel approach[12]	ChatGPT has the potential to assist healthcare providers in making well-informed decisions and improving adherence to post-colonoscopy surveillance guidelines	A limited group of specialists, subjective evaluation criteria, examining only 20 clinical post-colonoscopy scenarios

AI, artificial intelligence; GERD, gastroesophageal reflux disease; PPI, proton pump inhibitor.

Table 3. Quality assessment of studies included in the review using modified JBI critical appraisal tools.

Critical appraisal questions	Lahat *et al.*[9]	Lee *et al.*[6]	Suchman *et al.*[7]	Henson *et al.*[11]	Lahat *et al.*[8]	Gorelik *et al.*[12]
1. Was the sample frame appropriate to address the target population?	Yes	Yes	Yes	Yes	Yes	Yes
2. Were study participants sampled in an appropriate way?	Yes	Yes	Yes	Yes	Unclear	Yes
3. Was the sample size adequate?	Yes	Yes	Yes	No	Unclear	No
4. Were the study subjects and the setting described in detail?	Yes	Yes	Yes	Yes	Yes	Yes
5. Was the data analysis conducted with sufficient coverage of the identified sample?	Yes	Yes	Yes	Yes	Yes	Yes
6. Were valid methods used for the identification of the condition?	Yes	Yes	Yes	Yes	Unclear	Yes
7. Was the condition measured in a standard, reliable way for all participants?	Yes	Yes	Yes	Yes	Yes	Yes
8. Was there appropriate statistical analysis?	Yes	Yes	Yes	Yes	Yes	Yes
9. Was the response rate adequate, and if not, was the low response rate managed appropriately?	Unclear	No	No	Unclear	No	Yes
10. Were the findings robust to sensitivity analyses?	Unclear	Unclear	Unclear	Unclear	Unclear	Unclear
11. Are the results generalizable to the target population?	Yes	Unclear	No	Yes	Yes	Yes
12. Do the conclusions and recommendations follow from the results?	Yes	Yes	Yes	Yes	Yes	Yes
13. Can the results be applied to your local population or the population you are interested in?	Yes	Yes	No	Yes	Yes	Yes
14. Were all important outcomes considered?	Yes	Yes	Yes	Yes	Yes	Yes
15. Are the benefits worth the harms and costs?	Yes	Yes	No	No	Yes	Yes

Criteria	Lee *et al.*[6]	Suchman *et al.*[7]	Lahat *et al.*[9]	Henson *et al.*[11]	Lahat *et al.*[8]	Gorelik *et al.*[12]
Clear aims and objectives	Yes	Yes	Yes	Yes	Yes	Yes
Methodology	Yes	Yes	Yes	Yes	Yes	Yes
Data collection	Yes	Yes	Yes	Yes	Yes	Yes
Data analysis	Yes	Yes	Yes	Yes	Yes	Yes
Ethical considerations	Yes	Yes	Not mentioned	Not mentioned	Yes	Unclear
Results	Yes	Yes	Yes	Yes	Yes	Yes
Discussion and conclusion	Yes	Yes	Yes	Yes	Yes	Yes
Source of funding	Unclear	Unclear	Unclear	Unclear	Unclear	Unclear

JBI, Joanna Briggs Institute.

In another study,[9] our group evaluated the utility of ChatGPT in answering a variety of clinical questions addressing a wide range of topics, including common symptoms, diagnostic tests, and treatments for various gastrointestinal conditions. The study revealed that ChatGPT could provide accurate and clear answers in some cases, but not in others, indicating the need for further development. Notably, both studies only examined GPT-3.5, an older and less capable ChatGPT model that is free to access.

ChatGPT as a tool for physicians

As a tool for physicians, ChatGPT was evaluated in several aspects: clinical reasoning, knowledge, and education.

In the clinical field, ChatGPT was evaluated in two different domains; management of Gastroesophageal Reflux Disease (GERD),[11] and optimizing post-colonoscopy management.[12]

In the study 'Evaluation of the potential utility of an artificial intelligence ChatBot in GERD management',[11] evaluated the utility of ChatGPT in the management of GERD. The authors did not specify which ChatGPT model they investigated for this study. The results showed that ChatGPT provided appropriate and specific recommendations for GERD management in 91.3% of cases, with 29.0% considered completely appropriate and 62.3% mostly appropriate. However, inconsistencies were noted in responses to the same prompt, and some potential proton pump inhibitor (PPI) risks were stated as facts. Notably, patients from diverse educational backgrounds universally regarded the responses as comprehensible and beneficial. Furthermore, all respondents expressed their inclination to consider the tool as a valuable resource for obtaining medical information, highlighting the superior utility of the response format compared to that of a conventional search engine.

In the second domain, Gorelik *et al.*[12] evaluated the responses generated by ChatGPT to 20 clinical scenarios presented to ChatGPT in the form of structured reports and free-text notes. The responses generated by ChatGPT were assessed focusing on measuring adherence to guidelines, accuracy of responses, and inter-rater agreement between the gastroenterologists. Results demonstrated a 90% adherence to guidelines and an 85% accuracy in responses.

In the subject of knowledge and education, Suchman *et al.*[7] evaluated the performance of ChatGPT on the American College of Gastroenterology Self-Assessment Tests.

Surprisingly, ChatGPT was not able to pass multiple versions of the exams, indicating its limitations with gastroenterology subspeciality level question and answer tasks. Notably, in this study, the authors examined and compared both versions of ChatGPT (versions GPT-3.5 and GPT-4), with no actual difference in the results achieved: ChatGPT-3.5 scored 65.1% and GPT-4 scored 62.4% (passing grade was 70%).

ChatGPT as a tool for researchers

In assessing ChatGPT as a tool for researchers, our group evaluated the use of a ChatGPT for highlighting research priorities and identifying open and meaningful top research questions in gastroenterology.[8] The research questions outputted by ChatGPT achieved high ratings for relevance and clarity as well as an average rating for specificity but performed poorly in terms of originality.

Studies heterogeneity

When diving into the complexities of research involving AI models, understanding the underlying methods is pivotal. Several factors can introduce variability in the outcomes of such studies, especially when dealing with models like ChatGPT. Factors contributing to this heterogeneity include the following:

1. The absence of specific model version details (e.g. 3.5 *versus* 4.0 or the exact date of each model). Different versions may exhibit varied performances.
2. Lack of clarity on the method of question submission – whether they were sent collectively or in individual chat sessions for each prompt and response. The chosen method can significantly influence the replies.
3. Non-disclosure of the prompts utilized in the research, hindering the reproducibility of the findings.
4. Uncertainty over whether the study repeated the same prompts to gauge the consistency in the model's responses. The criteria for prompt selection remain vague. Even minor alterations in the wording can drastically change outcomes, a phenomenon particularly evident with GPT 3.5. Enhancing the quality of prompts can substantially refine many of these investigations.

Benefits and limitations of ChatGPT in gastroenterology

From the included studies, several benefits of ChatGPT in gastroenterology were identified. These include the ability of ChatGPT to provide appropriate and specific recommendations, aid in patient–physician communication, patient education, and continuous patient care, and generate relevant and clear research questions. However, limitations were also noted, including ChatGPT's insufficient understanding of complex medical questions, inconsistencies in responses, some potential PPI risks being stated as fact, some responses providing limited specific guidance, and struggles with originality. Ethical considerations were also raised, such as confidentiality and data security, stereotypes, bias and inaccuracy, plagiarism concerns, compliance with data privacy regulations, and the irreplaceable role of human judgment.

Discussion

The results of this systematic review provide a comprehensive evaluation of the role of ChatGPT, an LLM, in the field of gastroenterology. The included studies highlight the potential of ChatGPT in various applications, including patient education, self-assessment, patient–physician communication, disease management, and research question generation. However, they also underscore several limitations and ethical considerations that warrant further exploration and careful regulation.

In the era of digital health, it is essential to critically evaluate emerging technologies and their potential impact on healthcare delivery. This review contributes to this ongoing discourse, offering a focused examination of ChatGPT in the context of gastroenterology.

Our review focused on ChatGPT, as it stands out as the most popular LLM chat tool, for patients and physicians alike. Therefore, it is important to assess its performance on tasks relevant to both groups.

We believe that the insights gleaned from this review will be valuable not only to practitioners and researchers in gastroenterology, but also to policymakers, AI developers, and the broader healthcare community as we navigate the integration of AI into healthcare.[4,5]

The studies evaluating the efficacy of ChatGPT in answering common patient questions[6,9] reveal a mixed picture. While ChatGPT demonstrated the ability to generate credible medical information, its performance was inconsistent. Some responses were accurate and clear, while others were not, indicating an insufficient understanding of complex medical information. Moreover, the AI-generated answers were written at significantly higher grade reading levels than recommended, potentially limiting their accessibility to patients with lower literacy levels. Notably, this point can easily be fixed with prompt adjustments at a system level. These findings echo the cautionary note in the editorial 'Will ChatGPT transform healthcare?',[4] and highlight the need for further development and fine-tuning of ChatGPT to ensure its reliability and accessibility in patient education.

In the context of gastroenterology board exam-style medical reasoning, ChatGPT did not achieve a passing score using the methods in the recently published study. This indicates its limitations as an educational tool in its current form.[7] Notably, this study examined and compared the performance of both versions of the chatbot – the free version (GPT-3.5) and the advanced version (ChatGPT-4). It is noteworthy that the advanced version did not demonstrate an advantage over the free version. On the contrary, it lagged by three points. This finding emphasizes the need for continuous updates and the development of fine-tuned models specifically geared toward medical education, as suggested in the study, or the use of additional LLM augmentation methods like database linkage. Given the dynamic nature of medical knowledge, AI tools used in medical education need to be capable of providing accurate and updated information and following new information and guidelines that become available.

ChatGPT shows promise in enhancing patient–physician communication and continuous patient care.[10] Its ability to take patients' medical history, present the information in a concise, structured format, and continuously learn and improve based on the responses it receives could potentially improve healthcare outcomes. Moreover, by taking on tasks such as patient education and medical history taking, ChatGPT could help reduce physician burnout. However, the ethical considerations and limitations of AI, including confidentiality and data security, stereotypes, bias and inaccuracy, plagiarism concerns, compliance with data privacy regulations, and the irreplaceable role of human judgment, need to be addressed. AI technologies like ChatGPT should complement, and not replace, the human elements of empathy and professional judgment.

In the management of GERD, ChatGPT provided appropriate and specific recommendations in the majority of cases.[11] However, inconsistencies in responses to the same prompt and some potential PPI risks being stated as fact were identified as limitations. These findings highlight the need for rigorous clinical oversight in the use of ChatGPT in disease management.

ChatGPT's ability to generate relevant, clear, and moderately specific research questions is noteworthy.[8] However, it struggled with originality, suggesting the need for further work to improve the novelty of the generated research questions.

While it is not disclosed what was the specific training data used to train ChatGPT, it is likely that similar to other LLMs, it was trained on vast information derived from the internet and other open-access sources. However, ChatGPT is not innately attuned to medical nuances.[6,9] This may explain its inconsistency in providing clear and accurate information on gastroenterological issues.

In addition, ChatGPT's information might not always be up-to-date. Especially, regarding recent medical research and guidelines.[7] Its training data depends on available open-source information at the time of training. Thus, it might not be aware of newer studies or guidelines unless it is retrained on newer data.

A significant challenge is ChatGPT's language complexity.[6] This complexity is not an inherent flaw but a byproduct of the data it was trained on. However, for patient interactions, it is beneficial that the information is delivered at an accessible reading level.

Bias and the potential for manipulation through prompt engineering[6] arise because the model reflects the data it was trained on. If biases exist in those datasets, they will also be present in the model's outputs. This can be hazardous in medical applications where impartiality is essential.

The inconsistencies in responses to similar prompts[11] may be a byproduct of the model's vast training data. This is because the model may attempt to generate varied responses. However, in a clinical context, consistency is vital. Thus, this unpredictable behavior is a clear limitation.

A limitation demonstrated in the field of gastroenterology, particularly in medical reasoning and board exams, is the lack of domain-specific training for ChatGPT.[7] A model specifically fine-tuned on gastroenterological data could potentially outperform the general ChatGPT.

Notably, this could be related to how the task was presented to the model and a limitation of ChatGPT in particular. Perhaps with prompt engineering and increasing the temperature (creativity) of the model, more original responses could have been created.

This finding aligns with the review published by Sharma and Parasa,[3] which emphasizes the need for careful implementation and regulation of AI tools in healthcare.

ChatGPT proved its capability in handling various scenarios and descriptions effectively, providing concise patient letters in post-colonoscopy management by offering guideline-based recommendations.[12] These findings suggest that ChatGPT has the potential to assist healthcare providers in streamlining post-colonoscopy decision-making and improving adherence to post-colonoscopy surveillance guidelines.

This review has some limitations. First, the number of studies included in the review was relatively small, limiting the generalizability of the findings. However, as far as we know, it summarizes the current literature on the topic of ChatGPT in the field of gastroenterology. Second, the included studies were diverse in their objectives and methodologies, making it challenging to make quantitative analyses.

Finally, given the fast-paced advancements in AI technology, the conclusions of this review might soon be obsolete. Notably, the majority of the studies assessed the free version of ChatGPT.[3,5] However, recent research suggests that the improved version, ChatGPT-4, performs better in the medical field.[13]

In conclusion, the review of ChatGPT's application in gastroenterology reveals mixed outcomes. While showing promise as a tool for physicians (e.g. in GERD management and post-colonoscopy adherence to guidelines), it struggled with inconsistencies in patient education and failed in self-assessment tests. However, most data were created using the free version ChatGPT,[3,5] while the improved version (GPT-4) may achieve better results. Our findings emphasize the potential of ChatGPT and also underline clear limitations and the need for further refinement and ethical scrutiny.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

All authors consented publication.

Author contributions

Eyal Klang: Conceptualization; Data curation; Formal analysis; Investigation; Methodology; Software; Supervision; Validation; Visualization; Writing – original draft; Writing – review & editing.

Ali Sourosh: Investigation; Software; Visualization; Writing – original draft; Writing – review & editing.

Girish N. Nadkarni: Investigation; Supervision; Writing – original draft; Writing – review & editing.

Kassem Sharif: Investigation; Resources; Validation; Visualization; Writing – original draft; Writing – review & editing.

Adi Lahat: Conceptualization; Data curation; Formal analysis; Investigation; Methodology; Project administration; Resources; Supervision; Validation; Writing – original draft; Writing – review & editing.

Acknowledgements

None.

Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

Competing interests

The authors declare that there is no conflict of interest.

Availability of data and materials

Not applicable.

ORCID iD

Adi Lahat https://orcid.org/0000-0003-1513-7280

References

1. 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.
2. Tustumi F, Andreollo NA and Aguilar-Nascimento JE. Future of the language models in healthcare: the role of ChatGPT. *Arq Bras Cir Dig* 2023; 36: e1727.
3. Sharma P and Parasa S. ChatGPT and large language models in gastroenterology. *Nat Rev Gastroenterol Hepatol* 2023; 20: 481–482.
4. Will ChatGPT transform healthcare? *Nat Med* 2023; 29: 505–506.
5. Lahat A and Klang E. Can advanced technologies help address the global increase in demand for specialized medical care and improve telehealth services? *J Telemed Telecare* 2023: 1357633X231155520. DOI: 10.1177/1357633X231155520.
6. Lee TC, Staller K, Botoman V, *et al.* ChatGPT answers common patient questions about colonoscopy. *Gastroenterology* 2023; 165: 509–511.e7.
7. Suchman K, Garg S and Trindade AJ. Chat generative pretrained transformer fails the Multiple-Choice American College of Gastroenterology Self-assessment Test. *Am J Gastroenterol*. Epub ahead of print June 2023. DOI: 10.14309/ajg.0000000000002320.
8. Lahat A, Shachar E, Avidan B, *et al.* Evaluating the use of large language model in identifying top research questions in gastroenterology. *Sci Rep* 2023; 13: 4164.
9. Lahat A, Shachar E, Avidan B, *et al.* Evaluating the utility of a large language model in answering common patients' gastrointestinal health-related questions: are we there yet? *Diagnostics* (*Basel*) 2023; 13: 1950.
10. Nashwan AJ, Abujaber AA and Choudry H. Embracing the future of physician–patient communication: GPT-4 in gastroenterology. *Gastroenterol Endosc* 2023; 1: 132–135.
11. Henson JB, Glissen Brown JR, Lee JP, *et al.* Evaluation of the potential utility of an artificial intelligence ChatBot in GERD management. *Am J Gastroenterol*. Epub ahead of print July 2023. DOI: 10.14309/ajg.0000000000002397.
12. Gorelik Y, Ghersin I, Maza I, *et al.* Harnessing language models for streamlined post-colonoscopy patient management: a novel approach. *Gastrointest Endosc*. Epub ahead of print June 2023. DOI: 10.1016/j.gie.2023.06.025.
13. Nori H, King N, McKinney SM, *et al.* Capabilities of GPT-4 on medical challenge problems. *arXiv:2303.13375*, 2023.

Visit Sage journals online journals.sagepub.com/home/tag

Sage journals