

The Scope and Limitations of Extant Research into ChatGPT as a Tool for Patient Education: Systematic Review

Reid Dale, Maggie Cheng, Katharine Casselman Pines, Maria Currie

Submitted to: Journal of Medical Internet Research
on: April 22, 2025

Disclaimer: © The authors. All rights reserved. This is a privileged document currently under peer-review/community review. Authors have provided JMIR Publications with an exclusive license to publish this preprint on its website for review purposes only. While the final peer-reviewed paper may be licensed under a CC BY license on publication, at this stage authors and publisher expressly prohibit redistribution of this draft paper other than for review purposes.

Table of Contents

Original Manuscript..... 5

Supplementary Files..... 31

 Figures 32

 Figure 1..... 33

The Scope and Limitations of Extant Research into ChatGPT as a Tool for Patient Education: Systematic Review

Reid Dale¹ PhD; Maggie Cheng¹ BA, BS; Katharine Casselman Pines¹ MPH; Maria Currie¹ MD, PhD

¹ Stanford Medicine Stanford US

Corresponding Author:

Maria Currie MD, PhD

Stanford Medicine
Center for Academic Medicine
453 Quarry Road, Room 267
Stanford
US

Abstract

Background: Chat Generative Pre-Trained Transformer (ChatGPT), a large language model (LLM) developed by OpenAI, has been extensively studied and embraced by medical researchers since its public release in November 2022. In addition to its benefits in generating summaries and predictive diagnostics, it has been proposed as a patient education tool. Existing research cites ChatGPT's potential to increase information availability and accessibility, improve the efficiency of clinical practice, and its high-quality responses to clinical questions as reasons to consider its adoption.

Objective: We assessed literature from PubMed on the quality and consistency of ChatGPT responses across medical specialties, evaluated the comprehensiveness of current research, and identified areas for future research.

Methods: We searched PubMed for published articles evaluating the consistency, reliability, and ethics of ChatGPT in patient education. Following the PRISMA guideline, we conducted a systematic literature review of the 567 retrieved records. After title, abstract, and full-text screens, 123 relevant records were included and synthesized in this review.

Results: We found a lack of consensus among ChatGPT studies. The model accuracy suffers from infrequent updates and generation of misleading information (hallucinations), and it lacks knowledge of current clinical guidelines. The consistency of the model falls short due to its sensitivity to prompt design and fine-tuning through user interaction, making ChatGPT research results almost impossible to peer review and validate. Relying on ChatGPT for clinical information risks spreading misinformation, disrupting trust in the medical system, and disobeying the principles of patient-centered care. This also shifts the focus of patient education from shared decision-making and information-building to information-giving, deviating from the objectives of Health Communication and Health Literacy outlined in Healthy People 2030.

Conclusions: We caution against the acceptance of ChatGPT as a patient education tool and encourage future efforts to incorporate community perspectives in AI research.

(JMIR Preprints 22/04/2025:76397)

DOI: <https://doi.org/10.2196/preprints.76397>

Preprint Settings

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

✓ **Please make my preprint PDF available to anyone at any time (recommended).**

Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.
Only make the preprint title and abstract visible.

No, I do not wish to publish my submitted manuscript as a preprint.

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

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

Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain visible to all users.

Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <http://www.jmir.org/preprint/76397>

No. Please do not make my accepted manuscript PDF available to anyone. I understand that if I later pay to participate in <https://www.jmir.org/preprint/76397>



Original Manuscript

Title:

The Scope and Limitations of Extant Research into ChatGPT as a Tool for Patient Education: Systematic Review

Reid Dale PhD^{1,2}, Maggie Cheng BA BS¹, Katharine Casselman Pines MPH¹, Maria Elizabeth Currie MD PhD^{1,2}

1. Department of Cardiothoracic Surgery, Stanford School of Medicine, Stanford, CA, USA
2. Stanford Cardiovascular Institute, Stanford, CA, USA

Abstract:

Background. Chat Generative Pre-Trained Transformer (ChatGPT), a large language model (LLM) developed by OpenAI, has been extensively studied and embraced by medical researchers since its public release in November 2022. In addition to its benefits in generating summaries and predictive diagnostics, it has been proposed as a patient education tool. Existing research cites ChatGPT's potential to increase information availability and accessibility, improve the efficiency of clinical practice, and its high-quality responses to clinical questions as reasons to consider its adoption.

Objective. We assessed literature from PubMed on the quality and consistency of ChatGPT responses across medical specialties, evaluated the comprehensiveness of current research, and identified areas for future research.

Methods. We searched PubMed for published articles evaluating the consistency, reliability, and ethics of ChatGPT in patient education. Following the PRISMA guideline, we conducted a systematic literature review of the 567 retrieved records. After title, abstract, and full-text screens, 123 relevant records were included and synthesized in this review.

Results. We found a lack of consensus among ChatGPT studies. The model accuracy suffers from infrequent updates and generation of misleading information (hallucinations), and it lacks knowledge of current clinical guidelines. The consistency of the model falls short due to its sensitivity to prompt design and fine-tuning through user interaction, making ChatGPT research results almost impossible to peer review and validate. Relying on ChatGPT for clinical information risks spreading misinformation, disrupting trust in the medical system, and

disobeying the principles of patient-centered care. This also shifts the focus of patient education from shared decision-making and information-building to information-giving, deviating from the objectives of Health Communication and Health Literacy outlined in Healthy People 2030.

Conclusions. We caution against the acceptance of ChatGPT as a patient education tool and encourage future efforts to incorporate community perspectives in AI research.

Keywords: Large Language Model; Health Literacy; Health Communication



Introduction

Chat Generative Pre-Trained Transformer (ChatGPT), initially developed by OpenAI in 2020, is a large language model (LLM) based on GPT-3.5. It was trained with a general information database and can recognize patterns in languages, extract information, and generate coherent and human-like texts.¹⁻⁴ GPT models have undergone several revisions, from GPT-1 in 2018 to GPT-4 in 2023, with progressively increased parameters and training dataset sizes. Although some studies noted significant improvement in accuracy as the model developed (especially comparing GPT 3.5 to GPT 4), others concluded that limitations in previous versions of GPT models still apply to the newest edition now.⁴⁻⁷

Although ChatGPT shows excellent performance in sourcing and summarizing information from its training data and could be useful in improving the quality of writing, it is fundamentally a language-based model that lacks expertise in specific subject areas.^{3,8-14} When applied to medical settings as a tool for individualized information sourcing or patient education, its inconsistency and lack of up-to-date medical knowledge raise technical and ethical concerns. Despite the presence of BioGPT, a model trained with biomedical research texts that presumably should be more accurate and rigorous in biomedical settings, the limitations of ChatGPT still apply.^{4,15} At the current stage, although artificial intelligence (AI) technologies have been incorporated into some medical settings such as designing radiology treatment, the incorporation of ChatGPT in offering clinical information to patients remains controversial.¹⁶

Much of the medical community and published literature support the utilization of ChatGPT in aiding clinical decision-making and as a valuable information source for patients despite recognizing the many flaws this technology carries, citing the model's near-instantaneous responses, accessibility, affordability, breadth of knowledge, capacity for a personalized experience, and minimization of bias with ongoing efforts to improve training dataset.^{4,12,16-41} However, opposite voices also exist, bringing to light the inconsistency, inaccuracy, privacy, and legal concerns regarding the incorporation of ChatGPT into clinical settings.^{2,7,42-58}

In this study, we reviewed literature from PubMed on the design, expandability, consistency, and accuracy of ChatGPT responses across medical specialties, evaluated current research, and identified areas for future research. Despite recognizing ChatGPT's merits in increasing efficiency and the accessibility of information, it is imperative to critically evaluate the potential concerns in embracing this change too quickly.

Methods

Search Strategy

This comprehensive review was conducted according to the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines. We searched PubMed for published articles (including preprints) on ChatGPT and its applications in providing medical information to patients. Keywords and search terms included “ChatGPT AND consistency”, “ChatGPT AND ethics”, “ChatGPT AND health literacy”, “ChatGPT AND patient education”, and “ChatGPT AND reliability.” Literature search concluded on January 7, 2024.

Inclusion and Exclusion Criteria

We included published articles (including preprints) on the use of ChatGPT in clinical information dissemination as well as ethics discussion of the model. Exclusion criteria included 1) Letter to Editors (article type), 2) Comments (article type), 3) Records discussing role of ChatGPT in professional/medical education, 4) ChatGPT in scientific writing and publishing, 5) ChatGPT-assisted writing, 6) ChatGPT in diagnostics and imaging, 7) ChatGPT in translation, 8) non-English records, and 9) other irrelevant records. We did not restrict the search and inclusion criteria only to records from the United States. Articles from other countries were included as long as the full text was available and accessible in English.

Results

Summary of Search Results and Record Screening Process

A total of 567 records were retrieved from PubMed. These represent a pooled result from all search terms: “ChatGPT AND consistency” (n = 97), “ChatGPT AND ethics” (n = 245), “ChatGPT AND health literacy” (n = 26), “ChatGPT AND patient education” (n = 64), and “ChatGPT AND reliability” (n = 136). 87 records were removed for duplication. A title screen (n = 481) was conducted next, and 274 articles that fitted the exclusion criteria were excluded. Article abstracts (n = 207) were retrieved and screened; 2 were removed for lack of access to the abstract, 14 were removed due to article type, and 56 were excluded for outside of the scope of this review. Lastly, we did a full-text article screen (n = 135); 7 articles were removed for lack of access, 3 were removed for being non-English records, and 1 were removed for being outside the scope of this review. At the end, 123 articles were included in this review. *Figure 1* shows this process in a PRISMA diagram.

ChatGPT Design

The development of ChatGPT and other ChatGPT-like AI chatbots have the potential to

revolutionize some facets of healthcare delivery. Although these models were not developed specifically for a healthcare context, such models could lead to more effective delegation of clinical responsibilities, increase accessibility to health information, and improve clinical decision-making.^{3,59} Beyond sourcing information, ChatGPT also summarizes and presents information concisely, saving readers time from visiting multiple different sites to extrapolate information themselves.^{60,61} It also can translate, making information available and accessible to patients of all language backgrounds, although the accuracy and reliability of ChatGPT in non-English contexts is yet to be fully evaluated.^{4,16} The readability of ChatGPT-generated content has also shown to be above the level that is appropriate to use in educational materials for the general public.^{22,50,61–69}

ChatGPT is a pre-trained model with the ability to fine-tune through user interactions. With delays in updating the training dataset and the resources needed for such training, ChatGPT's response lags behind current clinical guidelines and provides outdated information. GPT 3.5 was updated in January 2022, and GPT 4-Turbo was trained in April 2023.⁷⁰ Updates in clinical guidelines after those dates would be neglected in GPT responses, potentially leading to inaccurate information.⁷¹ ChatGPT is also subject to temperature control and top_p sampling, which determines the flexibility of the model in selecting the immediate next word as well as the set of phrases to use when generating responses. Lower temperature and top parameters give the model less flexibility in creatively generating responses and higher parameters give the model a higher degree of randomness and unpredictability. As both parameters are non-zero in GPT models, it is almost impossible to replicate an exact same response due to the irreducibly probabilistic nature of an LLM's response.^{72,73}

Discussion

Expandability and Privacy Concerns of ChatGPT

Existing studies identified several concerns regarding the usage of ChatGPT in a health education context.

First, the language model operates like a “black box”, not explaining its reasoning or judgments.^{16,74,75} From the user's perspective, we could only control the input but would not know what to expect as the output. The process that ChatGPT uses to generate responses is unclear, even to its developers from OpenAI.⁷⁶ This takes away the shared nature of decision-making that is supposed to happen in a clinical setting, deviating from the ideals of patient-centered medicine that the medical

communities strive to emphasize. Without information on how the model arrives at its recommendations, providers are ill-prepared should patients have any questions about ChatGPT's responses. Such an inability to properly explain or trace back the logic of AI may instill distrust between the patient and the provider, painting a picture of AI as more knowledgeable while the provider as inferior. Without information on how machine learning models work to generate a response, providers do not have adequate evidence to trust or question the validity of the conclusion, putting them in a dilemma. Their behaviors and clinical judgments may be affected by their willingness to minimize information discrepancy and maintain trust with patients, incentivizing them to side with information from the AI that they may not have full confidence in. In situations where seeking medical information from AI becomes a habit or norm, and patient is faced with conflicting information from the human physician and the AI, the patient would be unsure of how to proceed. Nov et al. gave early evidence of this trend, showing that participants could only correctly distinguish between AI vs. physician-generated responses 49-85.7% of the times, and they tended to trust ChatGPT responses when it comes to low-risk health questions.⁷⁷ When AI-generated information is used and acted upon in clinical settings, the decisions patients make may not be considered truly informed due to the unknown clinical reasoning process of ChatGPT.

Second, medico-legal considerations are widely mentioned in the published literature.^{35,47,57,60,78} To date, we lack regulations and accountability measures that govern the proper utilization of AI-based chatbots in medical settings.⁷⁹⁻⁸³ Developers are not required to publicly reveal when their models may fail.¹⁶ In the case of a medical incident, the liability of parties remains unclear.^{60,75} Therefore, when the liability risks are truly at stake, it is unclear whether developers would be as willing to release the AI chatbots to the general public as they do now.

Third, ChatGPT usage brings privacy concerns.^{50,57,74} Nasr et al. revealed possibilities to extract ChatGPT's training data, raising concerns about confidentiality.⁸⁴ With its ability to widely source and retain information, efforts should be made to limit the possibility of linking de-identified information to personally identifiable information to ensure confidentiality if ChatGPT were to be incorporated in clinical settings.¹¹

Consistency of ChatGPT Responses

With ChatGPT's sensitivity to prompt design and its constantly evolving nature, mixed results exist regarding the consistency of ChatGPT responses.

Through assessing ChatGPT's knowledge on five hepato-pancreatico-biliary-related conditions by inputting designed questions into the model three times, Walker et al. showed an

internal consistency of 100% of the algorithm.¹³ When two experts compared ChatGPT's answers with existing guidelines in the study, the interrater agreement showed complete agreement and that the information and diagnoses provided by ChatGPT are consistent with the national guidelines.¹³

However, Johnson et al. found statistically significant differences in the accuracy of ChatGPT responses to their question prompts when the same set of questions were asked to ChatGPT only a few days apart.⁸⁵ Johnson et al. explained such differences as ChatGPT's ability to learn and fine-tune through user interactions.⁸⁵ However, this can also be interpreted as a lack of consistency in ChatGPT responses. The inconsistency of ChatGPT was further supported by Salas et al. who found that ChatGPT provided different answers to the same prompt within a short time frame.⁸⁶

The finetuning of ChatGPT through user feedback makes it susceptible to manipulation. ChatGPT cannot necessarily distinguish between right or wrong information and lacks factual knowledge in some areas.^{87,88} When ChatGPT is corrected by the users, the corrections could potentially enter the training data of ChatGPT and affect model performance in the future.⁸⁷ If the user feedback were to be incorrect, the model could include inaccurate information or biased views in its training set, allowing the spread of misinformation iteratively.

A recent breakdown of the ChatGPT system raised concerns about the consistency of the model. Deviating from its usual behavior, users reported that the chatbot gave incorrect, irrelevant, and inappropriate answers to their questions and sometimes responded with the same phrase repeatedly without providing a comprehensible sentence. Although this system breakdown was quickly fixed by the developers, OpenAI could not explain why the model failed and admitted that the model behavior can be unpredictable.⁷⁶ Such unpredictability should not be tolerated if the model were to be used for health education or information collection, where each piece of information could inform one's health decision-making and affect trust in the medical system. With increasingly positive comments from the medical community on ChatGPT's use in clinics, patients may see those AI chatbots as increasingly reliable, therefore developing dependence on the use of such algorithms and foregoing the knowledge they may gain from interacting with human practitioners.^{30,31}

Moreover, ChatGPT is an LLM that detects patterns in language, and it is susceptible to prompt design variations. Jang and Lukasiewicz showed that both ChatGPT (based on GPT-3) and GPT-4 can be self-contradictory, generating different answers when the input questions convey the same meaning.⁸⁹ The same team also demonstrated that GPT-3 often violates logical consistency, and the frequency of such mistakes is not negligible. These limitations were observed in previous versions of GPT models like GPT-2 but remained uncorrected despite using a larger training dataset for subsequent GPT models.

Accuracy and Comprehensiveness of ChatGPT Responses

When it comes to the accuracy and comprehensiveness of answers, ChatGPT's responses vary by field. Many studies commented highly on the accuracy of ChatGPT in providing medical information.^{85,90-92} However, events of AI hallucination sometimes appear, providing information or sources that do not exist.^{8,13,35,46,87,93,94} The accuracy of ChatGPT reported by current literature range from 40-50% in some fields (e.g. hepatocellular carcinoma, retinal diseases, liver cancer) to greater than 90% in others (e.g. type 2 diabetes, thoracic surgery, robotic-assisted radical prostatectomy, otolaryngology).^{11,25,26,41,95-98} Assessing ChatGPT's knowledge across 17 specialties, Johnson et al. found that 8.3% of all answers generated were completely incorrect, and many more were acceptable but not comprehensive. Even for a new GPT model that was trained on biomedical texts, its accuracy was only 51%.⁸⁵ The consistency and accuracy of ChatGPT responses are expected to decrease with increased length of conversation, as ChatGPT sometimes struggles to recognize or maintain context in prolonged conversations.^{4,99}

Since ChatGPT is a pre-trained model with limited fine-tuning during use, its accuracy depends on the quality of the training data as well as the frequency of updates. Biases carried in training data will be represented, if not increased, in use.^{2,8,47,74,78,100} Alarming, Acerbi and Stubbersfield revealed that ChatGPT-3 showed bias towards content that is gender-stereotypical, negative, and threat-related, warning the public of the limitation of such model.¹⁰¹ Koranteng et al. found similar gender stereotypes in ChatGPT responses as well as its tendency to associate negative terms with African American names.¹⁰² Also, inadequate updates of the model lead to outdated and misleading information.^{1,71,103} Studying ChatGPT's response to misconceptions about vaccinations, Deiana et al. found various inaccuracies in responses from both ChatGPT GPT 3.5 and GPT 4, and the chatbot seemed to selectively disregard the benefits of vaccines.¹ Similarly, many other studies across specialties also pointed out that the information ChatGPT provides is not up to date and can be fabricated.^{71,91,104,105}

In addition to generating inaccurate answers to the user's questions, ChatGPT also frequently fails to provide accurate or accessible references, if at all. Eleven of the included articles pointed out that ChatGPT either did not provide references despite prompting or generated fabricated references that did not exist.^{93,105-114} This inability to cite its sources exist regardless of the information accuracy provided by ChatGPT. Alessandri-Bonetti et al. showed that among all requests, ChatGPT failed to provide references 33% of the time.¹⁰⁶ When references were provided, 36% were shown to be irrelevant or nonexistent.¹⁰⁶ McGowan et al. revealed that only 6% of the references provided by

ChatGPT in their study was accurate.¹⁰⁵ ChatGPT also has limited ability to assess the quality of such citations, potentially biasing towards sources that are present in higher quantity in its training data but not necessarily of high quality or accuracy.

Potential Impacts on Patient-Physician Relationship

Although only a few of the current studies clearly advice the incorporation of ChatGPT in clinical setting for information retrieval purposes, much of the literature supports the idea of using ChatGPT as a supplementary tool to aid health professionals in promoting patient education and health literacy.^{11,60,115–122} Whereas the efficiency and relative accuracy of ChatGPT maybe appealing, we should not overlook the challenges ChatGPT faces as well as the patient-provider relationship that is vital in building health literacy.^{123,124} Health literacy must not only advance people's knowledge and ability to obtain and understand health information, but also empower them to actively participate in the health decision-making process. This includes discerning the quality of information, communicating their concerns, asking questions, and self-manage their conditions.^{16,123–}

126

Putting the accuracy and consistency concerns about ChatGPT aside, merely providing information to patients is not an effective way to improve health literacy and can be dangerously biased. Just as the black-box nature of GPT models goes against the key principles of patient-centered medicine and informed consent, receiving health advice based on the limited patient-provided information to ChatGPT carries the same risks. Modern medicine emphasizes both evidence-based practice and patient-centered care. Recommending the utilization of ChatGPT for obtaining medical information may dangerously reduce people into numbers and data points and potentially ignore their beliefs and values.¹⁶ On the contrary, healthcare professionals play a key role in incorporating such considerations through the establishment of a trusting patient-provider relationship.¹⁶ Such a relationship should not be a hierarchical one of information-giving and receiving or one where providers work to correct recommendation inaccuracies from AI-based chatbots. Instead, it should be one of information- and relationship-building where the patient takes a proactive part to work with the healthcare provider and arrive at a final decision. ChatGPT also lacks incentives to improve its performance standards. Human practitioners have moral obligations and liability of malpractice and negligence to provide high standards of care, while this ability is not yet developed in ChatGPT.

Since ChatGPT is sensitive to prompt design and has the potential to provide irrelevant information, utilizing ChatGPT properly and reliably requires knowledge and skills to ask the right

questions and assess the quality of the information received that not everyone possesses.^{16,67,126–128} Although ChatGPT-based information is accessible, it can worsen the existing disparities around health literacy due to this initial knowledge threshold required to use the model. If not done correctly, ChatGPT may miss important information, give generic answers that are not specific to the patient's situation, or provide misleading or overly worrisome messages.^{1,129–131} As mentioned above, ChatGPT is subject to biases in its training data and memory, both are aspects the patient may not be aware of when obtaining or acting on the medical information provided by ChatGPT.

Although ChatGPT has an excellent ability to generate human-like and coherent texts, it cannot think like a medical professional or ask follow-up questions. On the contrary, medical professionals are crucial not only hearing the patients' experiences, but also offering clarification, actively engaging patients to ensure proper understanding, and providing emotional support.

Therefore, if ChatGPT or any AI-based sources were to be implemented in patient education in the future, close monitoring and guidance by medical professionals would be necessary to help patients develop the ability to identify reliable and authoritative sources and critically assess the quality of information online before they can be left on their own. These skills are especially important to develop with the increased democratization of knowledge and autonomy patients have in their own health education.^{1,85}

Gaps in ChatGPT Research

Primary studies that evaluate the consistency and quality of ChatGPT's responses vary in rigorousness. Some studies rate ChatGPT's responses on a Likert scale while others report the accuracy of ChatGPT responses in percentages. The lack of standardized metrics makes it hard to compare the accuracy of the model across specialties.

Mixed results exist in many areas of ChatGPT research. Some showed an internal consistency of 100% and commented highly on its accuracy while others found that even the most updated GPT model (GPT-4) frequently self-contradict, and ChatGPT only has an accuracy of around or less than 50%.^{11,13}

The sensitivity of the model to prompt design, its memory bias, the temperature control, and its constantly evolving nature make it almost impossible to peer review, reproduce, and validate study findings around ChatGPT. Contradicting results across studies can be justified as ChatGPT's ability to learn and evolve, while it may also be interpreted as a sign of inconsistency in the ChatGPT model.

Many other studies compare ChatGPT to Google searches or other AI-based chatbots (Google

Bard and BingAI), stating that ChatGPT's responses are comparable to or more valuable than the ones provided by Google and other AI chatbots.^{28,62,63,66,106,132–135} However, this comparison does not that ChatGPT should be used and trusted for medical information.

Furthermore, current ChatGPT research lacks a community perspective, using expert-designed or guideline-oriented questions to generate ChatGPT responses and seeking experts' views on incorporating this in clinical practice. Although healthcare professionals play a crucial role in clinical interactions, they are only part of the story. Patients' perspectives should also be gathered and incorporated through community outreach efforts. With ChatGPT's sensitivity to instructions and prompt design, the differences in questions asked between professionals and a layperson without much background in a specific topic area can make a significant difference in the quality of response they get.

Future Directions

The crucial community perspective is lacking in current ChatGPT research. Only two studies among all that are included in this review sought community's expertise on this issue.^{77,136} A crucial next step is to continue this effort, survey patients, and evaluate their experience, trust, and comfortability in utilizing AI-based models for medical information.¹³⁷ Although comparison between different versions of GPT models did not reach a consensus, there is early evidence suggesting a superior performance in the paid GPT 4 compared to the publicly accessible ChatGPT (GPT 3.5). This raises concerns from a health equity standpoint, and future research is needed to address this issue.

Inadequacies of ChatGPT are not unique to the model and can extend to other AI-based chatbots as well. With Nvidia and Hippocratic AI's recent collaboration on an empathetic AI agent that outperformed human nurses in offering clinical information and decisions, the speed and scope of such technological developments have been pushed to a new height, making accountability and ethical evaluations of such practices especially pressing and necessary.¹³⁸ Although there are some regulations and quality assessment tools in development, such progress is far from catching up with how fast AI-based chatbots has been spread and utilized in daily, academic, and medical settings.³⁶ Future research should focus on increasing patients' awareness of the limitations of AI-based chatbots while establishing regulations to ensure ethical and equitable use of such models.

Conclusions

We evaluated the potential of incorporating ChatGPT in clinical practice and revealed the

inconsistency, inaccuracy, and lack of accountability in the ChatGPT model. With significant delays in model training, events of hallucination, and a lack of transparency in its reasoning, ChatGPT is not yet a reliable source of health information. The purpose of patient education and increasing health literacy involves trust- and relationship-building between the patient and the provider, and it exists beyond what ChatGPT can provide. Therefore, we caution against the premature acceptance of ChatGPT as a patient education tool until its logical, legal, and accuracy concerns are addressed. We encourage future efforts to increase transparency of model development and reasoning and incorporate community perspectives in AI research.

Ethics approval and consent to participate

This study is a review of published literature and is exempt from human subjects research review by the Stanford University Institutional Review Board.

Availability of data and materials

Not applicable (this manuscript does not report data generation or analysis).

Conflicts of Interest

The authors declare that they have no competing interests.

Acknowledgement

Dr. Reid Dale is supported by Stanford University Department of Cardiothoracic Surgery [grant number 1252717-910-EAGHD].

Authors' contributions

Framework and design of the review: R. D. and M. E. C.; literature searches and review: M. C.; drafting the manuscript: M. C. and R. D.; revision of the manuscript: M. C., K. C. P., R. D., and M. E. C.; final approval of the manuscript for publication: M. C., K. C. P., R. D., and M. E. C.

List of Abbreviations

AI - Artificial Intelligence

ChatGPT - Chat Generative Pre-Trained Transformer

GPT – Generative Pre-Trained Transformers

LLM - Large Language Model

PRISMA - Preferred Reporting Items for Systematic Reviews and Meta-Analyses



References

1. Deiana G, Dettori M, Arghittu A, Azara A, Gabutti G, Castiglia P. Artificial Intelligence and Public Health: Evaluating ChatGPT Responses to Vaccination Myths and Misconceptions. *Vaccines*. 2023;11(7):1217. doi:10.3390/vaccines11071217
2. Nazir A, Wang Z. A comprehensive survey of ChatGPT: Advancements, applications, prospects, and challenges. *Meta-Radiol*. 2023;1(2):100022. doi:10.1016/j.metrad.2023.100022
3. Sng GGR, Tung JYM, Lim DYZ, Bee YM. Potential and Pitfalls of ChatGPT and Natural-Language Artificial Intelligence Models for Diabetes Education. *Diabetes Care*. 2023;46(5):e103-e105. doi:10.2337/dc23-0197
4. Ray PP. ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. *Internet Things Cyber-Phys Syst*. 2023;3:121-154. doi:10.1016/j.iotcps.2023.04.003
5. Frosolini A, Franz L, Benedetti S, et al. Assessing the accuracy of ChatGPT references in head and neck and ENT disciplines. *Eur Arch Otorhinolaryngol*. 2023;280(11):5129-5133. doi:10.1007/s00405-023-08205-4
6. Moshirfar M, Altaf AW, Stoakes IM, Tuttle JJ, Hoopes PC. Artificial Intelligence in Ophthalmology: A Comparative Analysis of GPT-3.5, GPT-4, and Human Expertise in Answering StatPearls Questions. *Cureus*. Published online June 22, 2023. doi:10.7759/cureus.40822
7. Wang G, Gao K, Liu Q, et al. Potential and Limitations of ChatGPT 3.5 and 4.0 as a Source of COVID-19 Information: Comprehensive Comparative Analysis of Generative and Authoritative Information. *J Med Internet Res*. 2023;25:e49771. doi:10.2196/49771
8. Cascella M, Montomoli J, Bellini V, Bignami E. Evaluating the Feasibility of ChatGPT in Healthcare: An Analysis of Multiple Clinical and Research Scenarios. *J Med Syst*. 2023;47(1):33. doi:10.1007/s10916-023-01925-4
9. Dossantos J, An J, Javan R. Eyes on AI: ChatGPT's Transformative Potential Impact on Ophthalmology. *Cureus*. Published online June 21, 2023. doi:10.7759/cureus.40765
10. Karabacak M, Margetis K. Embracing Large Language Models for Medical Applications: Opportunities and Challenges. *Cureus*. Published online May 21, 2023. doi:10.7759/cureus.39305
11. Liu J, Wang C, Liu S. Utility of ChatGPT in Clinical Practice. *J Med Internet Res*. 2023;25:e48568. doi:10.2196/48568
12. Seth I, Lim B, Xie Y, et al. Comparing the Efficacy of Large Language Models ChatGPT, BARD, and Bing AI in Providing Information on Rhinoplasty: An Observational Study. *Aesthetic Surg J Open Forum*. 2023;5:ojad084. doi:10.1093/asjof/ojad084

13. Walker HL, Ghani S, Kuemmerli C, et al. Reliability of Medical Information Provided by ChatGPT: Assessment Against Clinical Guidelines and Patient Information Quality Instrument. *J Med Internet Res*. 2023;25:e47479. doi:10.2196/47479
14. 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(3):721-732. doi:10.3350/cmh.2023.0089
15. Newton W. What is BioGPT and what does it mean for healthcare? Clinical Trials Arena. Published February 9, 2023. Accessed July 19, 2024. <https://www.clinicaltrialsarena.com/news/biogpt-healthcare/>
16. Bjerring JC, Busch J. Artificial Intelligence and Patient-Centered Decision-Making. *Philos Technol*. 2021;34(2):349-371. doi:10.1007/s13347-019-00391-6
17. Altamimi I, Altamimi A, Alhumimidi AS, Altamimi A, Temsah MH. Snakebite Advice and Counseling From Artificial Intelligence: An Acute Venomous Snakebite Consultation With ChatGPT. *Cureus*. Published online June 13, 2023. doi:10.7759/cureus.40351
18. Bernstein IA, Zhang Y (Victor), Govil D, et al. Comparison of Ophthalmologist and Large Language Model Chatbot Responses to Online Patient Eye Care Questions. *JAMA Netw Open*. 2023;6(8):e2330320. doi:10.1001/jamanetworkopen.2023.30320
19. Cankurtaran RE, Polat YH, Aydemir NG, Umay E, Yurekli OT. Reliability and Usefulness of ChatGPT for Inflammatory Bowel Diseases: An Analysis for Patients and Healthcare Professionals. *Cureus*. Published online October 9, 2023. doi:10.7759/cureus.46736
20. Caglar U, Yildiz O, Ozervarli MF, et al. Assessing the Performance of Chat Generative Pretrained Transformer (ChatGPT) in Answering Andrology-Related Questions. *Urol Res Pract*. 2023;49(6):365-369. doi:10.5152/tud.2023.23171
21. Draschl A, Hauer G, Fischerauer SF, et al. Are ChatGPT's Free-Text Responses on Periprosthetic Joint Infections of the Hip and Knee Reliable and Useful? *J Clin Med*. 2023;12(20):6655. doi:10.3390/jcm12206655
22. Duran GS, Yurdakurban E, Topsakal KG. The Quality of CLP-Related Information for Patients Provided by ChatGPT. *Cleft Palate Craniofacial J*. Published online December 21, 2023;10556656231222387. doi:10.1177/10556656231222387
23. Endo Y, Sasaki K, Moazzam Z, et al. Quality of ChatGPT Responses to Questions Related To Liver Transplantation. *J Gastrointest Surg*. 2023;27(8):1716-1719. doi:10.1007/s11605-023-05714-9
24. Franco D'Souza R, Amanullah S, Mathew M, Surapaneni KM. Appraising the performance of ChatGPT in psychiatry using 100 clinical case vignettes. *Asian J Psychiatry*. 2023;89:103770.

doi:10.1016/j.ajp.2023.103770

25. Gabriel J, Shafik L, Alanbuki A, Larner T. The utility of the ChatGPT artificial intelligence tool for patient education and enquiry in robotic radical prostatectomy. *Int Urol Nephrol*. 2023;55(11):2717-2732. doi:10.1007/s11255-023-03729-4
26. Hernandez CA, Vazquez Gonzalez AE, Polianovskaia A, et al. The Future of Patient Education: AI-Driven Guide for Type 2 Diabetes. *Cureus*. Published online November 16, 2023. doi:10.7759/cureus.48919
27. Huang SS, Song Q, Beiting KJ, et al. Fact Check: Assessing the Response of ChatGPT to Alzheimer's Disease Statements with Varying Degrees of Misinformation. Published online September 7, 2023. doi:10.1101/2023.09.04.23294917
28. Jeong T, Liu H, Alessandri-Bonetti M, Pandya S, Nguyen VT, Egro FM. Revolutionizing patient education: CHATGPT outperforms GOOGLE in answering patient queries on free flap reconstruction. *Microsurgery*. 2023;43(7):752-761. doi:10.1002/micr.31106
29. Köroğlu EY, Fakı S, Beştepe N, et al. A Novel Approach: Evaluating ChatGPT's Utility for the Management of Thyroid Nodules. *Cureus*. Published online October 24, 2023. doi:10.7759/cureus.47576
30. Krittanawong C, Virk HUH, Kaplin SL, Wang Z, Sharma S, Jneid H. Assessing the Potential of ChatGPT for Patient Education in Cardiac Catheterization Care. *JACC Cardiovasc Interv*. 2023;16(12):1551-1552. doi:10.1016/j.jcin.2023.04.042
31. Kuşcu O, Pamuk AE, Sütay Süslü N, Hosal S. Is ChatGPT accurate and reliable in answering questions regarding head and neck cancer? *Front Oncol*. 2023;13:1256459. doi:10.3389/fonc.2023.1256459
32. Mika AP, Martin JR, Engstrom SM, Polkowski GG, Wilson JM. Assessing ChatGPT Responses to Common Patient Questions Regarding Total Hip Arthroplasty. *J Bone Jt Surg*. 2023;105(19):1519-1526. doi:10.2106/JBJS.23.00209
33. Moazzam Z, Lima HA, Endo Y, Noria S, Needleman B, Pawlik TM. A Paradigm Shift: Online Artificial Intelligence Platforms as an Informational Resource in Bariatric Surgery. *Obes Surg*. 2023;33(8):2611-2614. doi:10.1007/s11695-023-06675-3
34. Pushpanathan K, Lim ZW, Er Yew SM, et al. Popular large language model chatbots' accuracy, comprehensiveness, and self-awareness in answering ocular symptom queries. *iScience*. 2023;26(11):108163. doi:10.1016/j.isci.2023.108163
35. Sallam M. ChatGPT Utility in Healthcare Education, Research, and Practice: Systematic Review on the Promising Perspectives and Valid Concerns. *Healthcare*. 2023;11(6):887.

doi:10.3390/healthcare11060887

36. Sallam M, Barakat M, Sallam M. Pilot Testing of a Tool to Standardize the Assessment of the Quality of Health Information Generated by Artificial Intelligence-Based Models. *Cureus*. Published online November 24, 2023. doi:10.7759/cureus.49373
37. Sharma S, Pajai S, Prasad R, et al. A Critical Review of ChatGPT as a Potential Substitute for Diabetes Educators. *Cureus*. Published online May 1, 2023. doi:10.7759/cureus.38380
38. Song H, Xia Y, Luo Z, et al. Evaluating the Performance of Different Large Language Models on Health Consultation and Patient Education in Urolithiasis. *J Med Syst*. 2023;47(1):125. doi:10.1007/s10916-023-02021-3
39. Rogasch JMM, Metzger G, Preisler M, et al. ChatGPT: Can You Prepare My Patients for [¹⁸ F]FDG PET/CT and Explain My Reports? *J Nucl Med*. 2023;64(12):1876-1879. doi:10.2967/jnumed.123.266114
40. Van Bulck L, Moons P. What if your patient switches from Dr. Google to Dr. ChatGPT? A vignette-based survey of the trustworthiness, value, and danger of ChatGPT-generated responses to health questions. *Eur J Cardiovasc Nurs*. 2024;23(1):95-98. doi:10.1093/eurjcn/zvad038
41. Zalzal HG, Abraham A, Cheng J, Shah RK. Can CHATGPT help patients answer their otolaryngology questions? *Laryngoscope Investig Otolaryngol*. 2024;9(1):e1193. doi:10.1002/liv.2.1193
42. Al-Dujaili Z, Omari S, Pillai J, Al Faraj A. Assessing the accuracy and consistency of ChatGPT in clinical pharmacy management: A preliminary analysis with clinical pharmacy experts worldwide. *Res Soc Adm Pharm*. 2023;19(12):1590-1594. doi:10.1016/j.sapharm.2023.08.012
43. Anastasio AT, Mills FB, Karavan MP, Adams SB. Evaluating the Quality and Usability of Artificial Intelligence-Generated Responses to Common Patient Questions in Foot and Ankle Surgery. *Foot Ankle Orthop*. 2023;8(4):24730114231209919. doi:10.1177/24730114231209919
44. Biswas S, Logan NS, Davies LN, Sheppard AL, Wolffsohn JS. Assessing the utility of ChatGPT as an artificial intelligence-based large language model for information to answer questions on myopia. *Ophthalmic Physiol Opt*. 2023;43(6):1562-1570. doi:10.1111/opo.13207
45. Bushuven S, Bentele M, Bentele S, et al. "ChatGPT, Can You Help Me Save My Child's Life?" - Diagnostic Accuracy and Supportive Capabilities to Lay Rescuers by ChatGPT in Prehospital Basic Life Support and Paediatric Advanced Life Support Cases – An In-silico Analysis. *J Med Syst*. 2023;47(1):123. doi:10.1007/s10916-023-02019-x
46. Branum C, Schiavenato M. Can ChatGPT Accurately Answer a PICOT Question? Assessing AI Response to a Clinical Question. *Nurse Educ*. 2023;48(5):231-233.

doi:10.1097/NNE.0000000000001436

47. Garg RK, Urs VL, Agrawal AA, Chaudhary SK, Paliwal V, Kar SK. Exploring the role of ChatGPT in patient care (diagnosis and treatment) and medical research: A systematic review. *Health Promot Perspect.* 2023;13(3):183-191. doi:10.34172/hpp.2023.22
48. Jazi AHD, Mahjoubi M, Shahabi S, et al. Bariatric Evaluation Through AI: a Survey of Expert Opinions Versus ChatGPT-4 (BETA-SEOV). *Obes Surg.* 2023;33(12):3971-3980. doi:10.1007/s11695-023-06903-w
49. McCarthy CJ, Berkowitz S, Ramalingam V, Ahmed M. Evaluation of an Artificial Intelligence Chatbot for Delivery of IR Patient Education Material: A Comparison with Societal Website Content. *J Vasc Interv Radiol.* 2023;34(10):1760-1768.e32. doi:10.1016/j.jvir.2023.05.037
50. Mishra A, Begley SL, Chen A, et al. Exploring the Intersection of Artificial Intelligence and Neurosurgery: Let us be Cautious With ChatGPT. *Neurosurgery.* 2023;93(6):1366-1373. doi:10.1227/neu.0000000000002598
51. Momenaei B, Wakabayashi T, Shahlaee A, et al. Appropriateness and Readability of ChatGPT-4-Generated Responses for Surgical Treatment of Retinal Diseases. *Ophthalmol Retina.* 2023;7(10):862-868. doi:10.1016/j.oret.2023.05.022
52. Nastasi AJ, Courtright KR, Halpern SD, Weissman GE. A vignette-based evaluation of ChatGPT's ability to provide appropriate and equitable medical advice across care contexts. *Sci Rep.* 2023;13(1):17885. doi:10.1038/s41598-023-45223-y
53. Sütçüoğlu BM, Güler M. Appropriateness of premature ovarian insufficiency recommendations provided by ChatGPT. *Menopause.* 2023;30(10):1033-1037. doi:10.1097/GME.0000000000002246
54. Stephens LD, Jacobs JW, Adkins BD, Booth GS. Battle of the (Chat)Bots: Comparing Large Language Models to Practice Guidelines for Transfusion-Associated Graft-Versus-Host Disease Prevention. *Transfus Med Rev.* 2023;37(3):150753. doi:10.1016/j.tmr.2023.150753
55. Stroop A, Stroop T, Zawy Alsofy S, et al. Large language models: Are artificial intelligence-based chatbots a reliable source of patient information for spinal surgery? *Eur Spine J.* Published online October 11, 2023. doi:10.1007/s00586-023-07975-z
56. Temsah MH, Aljamaan F, Malki KH, et al. ChatGPT and the Future of Digital Health: A Study on Healthcare Workers' Perceptions and Expectations. *Healthcare.* 2023;11(13):1812. doi:10.3390/healthcare11131812
57. Wang C, Liu S, Yang H, Guo J, Wu Y, Liu J. Ethical Considerations of Using ChatGPT in Health Care. *J Med Internet Res.* 2023;25:e48009. doi:10.2196/48009

58. Uz C, Umay E. “Dr CHATGPT ”: Is it a reliable and useful source for common rheumatic diseases? *Int J Rheum Dis*. 2023;26(7):1343-1349. doi:10.1111/1756-185X.14749
59. Krittanawong C, Rodriguez M, Kaplin S, Tang WHW. Assessing the potential of ChatGPT for patient education in the cardiology clinic. *Prog Cardiovasc Dis*. 2023;81:109-110. doi:10.1016/j.pcad.2023.10.002
60. Dave T, Athaluri SA, Singh S. ChatGPT in medicine: an overview of its applications, advantages, limitations, future prospects, and ethical considerations. *Front Artif Intell*. 2023;6:1169595. doi:10.3389/frai.2023.1169595
61. Mondal H, Mondal S, Podder I. Using ChatGPT for writing articles for patients’ education for dermatological diseases: A pilot study. *Indian Dermatol Online J*. 2023;14(4):482. doi:10.4103/idoj.idoj_72_23
62. Ayoub NF, Lee Y, Grimm D, Divi V. Head-to-Head Comparison of ChatGPT Versus Google Search for Medical Knowledge Acquisition. *Otolaryngol Neck Surg*. 2024;170(6):1484-1491. doi:10.1002/ohn.465
63. Bellinger JR, De La Chapa JS, Kwak MW, Ramos GA, Morrison D, Kesser BW. BPPV Information on Google Versus AI (ChatGPT). *Otolaryngol Neck Surg*. 2024;170(6):1504-1511. doi:10.1002/ohn.506
64. Golan R, Ripps SJ, Reddy R, et al. ChatGPT’s Ability to Assess Quality and Readability of Online Medical Information: Evidence From a Cross-Sectional Study. *Cureus*. Published online July 20, 2023. doi:10.7759/cureus.42214
65. Haidar O, Jaques A, McCaughran PW, Metcalfe MJ. AI-Generated Information for Vascular Patients: Assessing the Standard of Procedure-Specific Information Provided by the ChatGPT AI-Language Model. *Cureus*. Published online November 30, 2023. doi:10.7759/cureus.49764
66. Hristidis V, Ruggiano N, Brown EL, Ganta SRR, Stewart S. ChatGPT vs Google for Queries Related to Dementia and Other Cognitive Decline: Comparison of Results. *J Med Internet Res*. 2023;25:e48966. doi:10.2196/48966
67. Spallek S, Birrell L, Kershaw S, Devine EK, Thornton L. Can we use ChatGPT for Mental Health and Substance Use Education? Examining Its Quality and Potential Harms. *JMIR Med Educ*. 2023;9:e51243. doi:10.2196/51243
68. Ulusoy I, Yilmaz M, Kivrak A. How Efficient Is ChatGPT in Accessing Accurate and Quality Health-Related Information? *Cureus*. Published online October 7, 2023. doi:10.7759/cureus.46662
69. Yurdakurban E, Topsakal KG, Duran GS. A comparative analysis of AI-based chatbots: Assessing data quality in orthognathic surgery related patient information. *J Stomatol Oral*

Maxillofac Surg. 2024;125(5):101757. doi:10.1016/j.jormas.2023.101757

70. Whitney L. ChatGPT is no longer as clueless about recent events | ZDNET. ZDNET. Published November 7, 2023. Accessed July 19, 2024. <https://www.zdnet.com/article/chatgpt-is-no-longer-as-clueless-about-recent-events/>

71. Grünebaum A, Chervenak J, Pollet SL, Katz A, Chervenak FA. The exciting potential for ChatGPT in obstetrics and gynecology. *Am J Obstet Gynecol.* 2023;228(6):696-705. doi:10.1016/j.ajog.2023.03.009

72. Peeperkorn M, Kouwenhoven T, Brown D, Jordanous A. Is Temperature the Creativity Parameter of Large Language Models? Published online May 1, 2024. Accessed July 19, 2024. <http://arxiv.org/abs/2405.00492>

73. Watkins R. Guidance for researchers and peer-reviewers on the ethical use of Large Language Models (LLMs) in scientific research workflows. *AI Ethics.* Published online May 16, 2023:s43681-023-00294-00295. doi:10.1007/s43681-023-00294-5

74. Jeyaraman M, Balaji S, Jeyaraman N, Yadav S. Unraveling the Ethical Enigma: Artificial Intelligence in Healthcare. *Cureus.* Published online August 10, 2023. doi:10.7759/cureus.43262

75. Victor G, Bélisle-Pipon JC, Ravitsky V. Generative AI, Specific Moral Values: A Closer Look at ChatGPT's New Ethical Implications for Medical AI. *Am J Bioeth.* 2023;23(10):65-68. doi:10.1080/15265161.2023.2250311

76. Tran TH. OpenAI's ChatGPT Went Completely Off the Rails for Hours. *The Daily Beast.* <https://www.thedailybeast.com/openais-chatgpt-went-completely-off-the-rails-for-hours>. Published February 21, 2024. Accessed February 21, 2024.

77. Nov O, Singh N, Mann D. Putting ChatGPT's Medical Advice to the (Turing) Test: Survey Study. *JMIR Med Educ.* 2023;9:e46939. doi:10.2196/46939

78. Hung YC, Chaker SC, Sigel M, Saad M, Slater ED. Comparison of Patient Education Materials Generated by Chat Generative Pre-Trained Transformer Versus Experts: An Innovative Way to Increase Readability of Patient Education Materials. *Ann Plast Surg.* 2023;91(4):409-412. doi:10.1097/SAP.0000000000003634

79. Athavale A, Baier J, Ross E, Fukaya E. The potential of chatbots in chronic venous disease patient management. *JVS-Vasc Insights.* 2023;1:100019. doi:10.1016/j.jvsvi.2023.100019

80. Klang E, Sourosh A, Nadkarni GN, Sharif K, Lahat A. Evaluating the role of ChatGPT in gastroenterology: a comprehensive systematic review of applications, benefits, and limitations. *Ther Adv Gastroenterol.* 2023;16:17562848231218618. doi:10.1177/17562848231218618

81. Minssen T, Vayena E, Cohen IG. The Challenges for Regulating Medical Use of ChatGPT

and Other Large Language Models. *JAMA*. 2023;330(4):315. doi:10.1001/jama.2023.9651

82. Wilhelm TI, Roos J, Kaczmarczyk R. Large Language Models for Therapy Recommendations Across 3 Clinical Specialties: Comparative Study. *J Med Internet Res*. 2023;25:e49324. doi:10.2196/49324

83. Yun JY, Kim DJ, Lee N, Kim EK. A comprehensive evaluation of ChatGPT consultation quality for augmentation mammoplasty: A comparative analysis between plastic surgeons and laypersons. *Int J Med Inf*. 2023;179:105219. doi:10.1016/j.ijmedinf.2023.105219

84. Nasr M, Carlini N, Hayase J, et al. Scalable Extraction of Training Data from (Production) Language Models. Published online 2023. doi:10.48550/ARXIV.2311.17035

85. Johnson D, Goodman R, Patrinely J, et al. Assessing the Accuracy and Reliability of AI-Generated Medical Responses: An Evaluation of the Chat-GPT Model. Published online February 28, 2023. doi:10.21203/rs.3.rs-2566942/v1

86. Salas A, Rivero-Calle I, Martín-Torres F. Chatting with ChatGPT to learn about safety of COVID-19 vaccines – A perspective. *Hum Vaccines Immunother*. 19(2):2235200. doi:10.1080/21645515.2023.2235200

87. Meyrowitsch DW, Jensen AK, Sørensen JB, Varga TV. AI chatbots and (mis)information in public health: impact on vulnerable communities. *Front Public Health*. 2023;11:1226776. doi:10.3389/fpubh.2023.1226776

88. Zhuo TY, Huang Y, Chen C, Xing Z. Red teaming ChatGPT via Jailbreaking: Bias, Robustness, Reliability and Toxicity. Published online May 29, 2023. Accessed July 19, 2024. <http://arxiv.org/abs/2301.12867>

89. Jang ME, Lukasiewicz T. Consistency Analysis of ChatGPT. Published online 2023. doi:10.48550/ARXIV.2303.06273

90. Ferro Desideri L, Roth J, Zinkernagel M, Anguita R. Application and accuracy of artificial intelligence-derived large language models in patients with age related macular degeneration. *Int J Retina Vit*. 2023;9(1):71. doi:10.1186/s40942-023-00511-7

91. Goodman RS, Patrinely JR, Stone CA, et al. Accuracy and Reliability of Chatbot Responses to Physician Questions. *JAMA Netw Open*. 2023;6(10):e2336483. doi:10.1001/jamanetworkopen.2023.36483

92. Tanaka OM, Gasparello GG, Hartmann GC, Casagrande FA, Pithon MM. Assessing the reliability of ChatGPT: a content analysis of self-generated and self-answered questions on clear aligners, TADs and digital imaging. *Dent Press J Orthod*. 2023;28(5):e2323183. doi:10.1590/2177-6709.28.5.e2323183.oar

93. Sharun K, Banu SA, Pawde AM, et al. ChatGPT and artificial hallucinations in stem cell research: assessing the accuracy of generated references – a preliminary study. *Ann Med Surg.* 2023;85(10):5275-5278. doi:10.1097/MS9.0000000000001228
94. Sridi C, Brigui S. The use of ChatGPT in occupational medicine: opportunities and threats. *Ann Occup Environ Med.* 2023;35(1):e42. doi:10.35371/aoem.2023.35.e42
95. Potapenko I, Boberg-Ans LC, Stormly Hansen M, Klefter ON, Van Dijk EHC, Subhi Y. Artificial intelligence-based chatbot patient information on common retinal diseases using CHATGPT. *Acta Ophthalmol (Copenh).* 2023;101(7):829-831. doi:10.1111/aos.15661
96. Cao JJ, Kwon DH, Ghaziani TT, et al. Accuracy of Information Provided by ChatGPT Regarding Liver Cancer Surveillance and Diagnosis. *Am J Roentgenol.* 2023;221(4):556-559. doi:10.2214/AJR.23.29493
97. Sun H, Zhang K, Lan W, et al. An AI Dietitian for Type 2 Diabetes Mellitus Management Based on Large Language and Image Recognition Models: Preclinical Concept Validation Study. *J Med Internet Res.* 2023;25:e51300. doi:10.2196/51300
98. Shao C ye, Li H, Liu X long, et al. Appropriateness and Comprehensiveness of Using ChatGPT for Perioperative Patient Education in Thoracic Surgery in Different Language Contexts: Survey Study. *Interact J Med Res.* 2023;12:e46900. doi:10.2196/46900
99. Whiles BB, Bird VG, Canales BK, DiBianco JM, Terry RS. Caution! AI Bot Has Entered the Patient Chat: ChatGPT Has Limitations in Providing Accurate Urologic Healthcare Advice. *Urology.* 2023;180:278-284. doi:10.1016/j.urology.2023.07.010
100. Wang G, Liu Q, Chen G, et al. AI's deep dive into complex pediatric inguinal hernia issues: a challenge to traditional guidelines? *Hernia.* 2023;27(6):1587-1599. doi:10.1007/s10029-023-02900-1
101. Acerbi A, Stubbersfield JM. Large language models show human-like content biases in transmission chain experiments. *Proc Natl Acad Sci.* 2023;120(44):e2313790120. doi:10.1073/pnas.2313790120
102. Koranteng E, Rao A, Flores E, et al. Empathy and Equity: Key Considerations for Large Language Model Adoption in Health Care. *JMIR Med Educ.* 2023;9:e51199. doi:10.2196/51199
103. Haze T, Kawano R, Takase H, Suzuki S, Hirawa N, Tamura K. Influence on the accuracy in ChatGPT: Differences in the amount of information per medical field. *Int J Med Inf.* 2023;180:105283. doi:10.1016/j.ijmedinf.2023.105283
104. Bhattacharyya M, Miller VM, Bhattacharyya D, Miller LE. High Rates of Fabricated and Inaccurate References in ChatGPT-Generated Medical Content. *Cureus.* 2023;15(5):e39238.

doi:10.7759/cureus.39238

105. McGowan A, Gui Y, Dobbs M, et al. ChatGPT and Bard exhibit spontaneous citation fabrication during psychiatry literature search. *Psychiatry Res.* 2023;326:115334. doi:10.1016/j.psychres.2023.115334

106. Alessandri-Bonetti M, Liu HY, Palmesano M, Nguyen VT, Egro FM. Online patient education in body contouring: A comparison between Google and ChatGPT. *J Plast Reconstr Aesthet Surg.* 2023;87:390-402. doi:10.1016/j.bjps.2023.10.091

107. Chervenak J, Lieman H, Blanco-Breindel M, Jindal S. The promise and peril of using a large language model to obtain clinical information: ChatGPT performs strongly as a fertility counseling tool with limitations. *Fertil Steril.* 2023;120(3):575-583. doi:10.1016/j.fertnstert.2023.05.151

108. Copeland-Halperin LR, O'Brien L, Copeland M. Evaluation of Artificial Intelligence-generated Responses to Common Plastic Surgery Questions. *Plast Reconstr Surg - Glob Open.* 2023;11(8):e5226. doi:10.1097/GOX.0000000000005226

109. Crook BS, Park CN, Hurley ET, Richard MJ, Pidgeon TS. Evaluation of Online Artificial Intelligence-Generated Information on Common Hand Procedures. *J Hand Surg.* 2023;48(11):1122-1127. doi:10.1016/j.jhsa.2023.08.003

110. Randhawa J, Khan A. A Conversation With ChatGPT About the Usage of Lithium in Pregnancy for Bipolar Disorder. *Cureus.* Published online October 5, 2023. doi:10.7759/cureus.46548

111. Sevgi UT, Erol G, Doğruel Y, Sönmez OF, Tubbs RS, Güngör A. The role of an open artificial intelligence platform in modern neurosurgical education: a preliminary study. *Neurosurg Rev.* 2023;46(1):86. doi:10.1007/s10143-023-01998-2

112. Suppadungsuk S, Thongprayoon C, Krisanapan P, et al. Examining the Validity of ChatGPT in Identifying Relevant Nephrology Literature: Findings and Implications. *J Clin Med.* 2023;12(17):5550. doi:10.3390/jcm12175550

113. Vaira LA, Lechien JR, Abbate V, et al. Accuracy of ChatGPT-Generated Information on Head and Neck and Oromaxillofacial Surgery: A Multicenter Collaborative Analysis. *Otolaryngol Neck Surg.* 2024;170(6):1492-1503. doi:10.1002/ohn.489

114. Warren E, Hurley ET, Park CN, et al. Evaluation of information from artificial intelligence on rotator cuff repair surgery. *JSES Int.* 2024;8(1):53-57. doi:10.1016/j.jseint.2023.09.009

115. Abbasi J, Hsuen Y. How AI Assistants Could Help Answer Patients' Messages—and Potentially Improve Their Outcomes. *JAMA.* 2024;331(2):95. doi:10.1001/jama.2023.22555

116. Alten A, Gündeş E, Tuncer E, Kozanoğlu E, Akalın BE, Emekli U. Integrating artificial

intelligence in orthognathic surgery: A case study of ChatGPT's role in enhancing physician-patient consultations for dentofacial deformities. *J Plast Reconstr Aesthet Surg.* 2023;87:405-407. doi:10.1016/j.bjps.2023.10.097

117. Holmes J, Liu Z, Zhang L, et al. Evaluating large language models on a highly-specialized topic, radiation oncology physics. *Front Oncol.* 2023;13:1219326. doi:10.3389/fonc.2023.1219326

118. Jin Y, Liu H, Zhao B, Pan W. ChatGPT and mycosis– a new weapon in the knowledge battlefield. *BMC Infect Dis.* 2023;23(1):731. doi:10.1186/s12879-023-08724-9

119. Marcantonio TL, Nielsen KE, Haikalis M, et al. Hey ChatGPT, Let's Talk About Sexual Consent. *J Sex Res.* Published online September 14, 2023:1-12. doi:10.1080/00224499.2023.2254772

120. Wang A, Kim E, Oleru O, Seyidova N, Taub PJ. Artificial Intelligence in Plastic Surgery: ChatGPT as a Tool to Address Disparities in Health Literacy. *Plast Reconstr Surg.* 2024;153(6):1232e-1234e. doi:10.1097/PRS.00000000000011202

121. Wójcik S, Rulkiewicz A, Pruszczyk P, Lisik W, Poboży M, Domienik-Karłowicz J. Beyond ChatGPT: What does GPT-4 add to healthcare? The dawn of a new era. *Cardiol J.* Published online October 12, 2023:VM/OJS/J/97515. doi:10.5603/cj.97515

122. 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(6):1790-1796. doi:10.1007/s11695-023-06603-5

123. Health Communication - Healthy People 2030 | health.gov. Accessed July 19, 2024. <https://health.gov/healthypeople/objectives-and-data/browse-objectives/health-communication>

124. Health Literacy in Healthy People 2030 - Healthy People 2030 | health.gov. Accessed July 19, 2024. <https://health.gov/healthypeople/priority-areas/health-literacy-healthy-people-2030>

125. Wittink H, Oosterhaven J. Patient education and health literacy. *Musculoskelet Sci Pract.* 2018;38:120-127. doi:10.1016/j.msksp.2018.06.004

126. Schulz PJ, Nakamoto K. Patient behavior and the benefits of artificial intelligence: The perils of “dangerous” literacy and illusory patient empowerment. *Patient Educ Couns.* 2013;92(2):223-228. doi:10.1016/j.pec.2013.05.002

127. Bilika P, Stefanouli V, Strimpakos N, Kapreli EV. Clinical reasoning using ChatGPT: Is it beyond credibility for physiotherapists use? *Physiother Theory Pract.* Published online December 11, 2023:1-20. doi:10.1080/09593985.2023.2291656

128. Lin Z. Why and how to embrace AI such as ChatGPT in your academic life. *R Soc Open Sci.* 2023;10(8):230658. doi:10.1098/rsos.230658

129. Hillmann HAK, Angelini E, Karfoul N, Feickert S, Mueller-Leisse J, Duncker D. Accuracy and comprehensibility of chat-based artificial intelligence for patient information on atrial fibrillation and cardiac implantable electronic devices. *Europace*. 2023;26(1):euad369. doi:10.1093/europace/euad369
130. Kuang YR, Zou MX, Niu HQ, Zheng BY, Zhang TL, Zheng BW. ChatGPT encounters multiple opportunities and challenges in neurosurgery. *Int J Surg*. 2023;109(10):2886-2891. doi:10.1097/JS9.0000000000000571
131. Lautrup AD, Hyrup T, Schneider-Kamp A, Dahl M, Lindholt JS, Schneider-Kamp P. Heart-to-heart with ChatGPT: the impact of patients consulting AI for cardiovascular health advice. *Open Heart*. 2023;10(2):e002455. doi:10.1136/openhrt-2023-002455
132. Ayoub NF, Lee YJ, Grimm D, Balakrishnan K. Comparison Between ChatGPT and Google Search as Sources of Postoperative Patient Instructions. *JAMA Otolaryngol Neck Surg*. 2023;149(6):556. doi:10.1001/jamaoto.2023.0704
133. Lim ZW, Pushpanathan K, Yew SME, et al. Benchmarking large language models' performances for myopia care: a comparative analysis of ChatGPT-3.5, ChatGPT-4.0, and Google Bard. *eBioMedicine*. 2023;95:104770. doi:10.1016/j.ebiom.2023.104770
134. Mu X, Lim B, Seth I, et al. Comparison of large language models in management advice for melanoma: Google's AI BARD, BingAI and ChatGPT. *Skin Health Dis*. 2024;4(1):e313. doi:10.1002/ski2.313
135. Rahsepar AA, Tavakoli N, Kim GHJ, Hassani C, Abtin F, Bedayat A. How AI Responds to Common Lung Cancer Questions: ChatGPT versus Google Bard. *Radiology*. 2023;307(5):e230922. doi:10.1148/radiol.230922
136. Shahsavari Y, Choudhury A. User Intentions to Use ChatGPT for Self-Diagnosis and Health-Related Purposes: Cross-sectional Survey Study. *JMIR Hum Factors*. 2023;10:e47564. doi:10.2196/47564
137. Pillai M, Griffin AC, Kronk CA, McCall T. Toward Community-Based Natural Language Processing (CBNLP): Cocreating With Communities. *J Med Internet Res*. 2023;25:e48498. doi:10.2196/48498
138. Dumas B. Nvidia announces AI-powered health care "agents" that outperform nurses — and cost \$9 an hour. *FOXBusiness*. Published March 21, 2024. Accessed July 19, 2024. <https://www.foxbusiness.com/technology/nvidia-announces-ai-powered-health-care-agents-outperform-nurses-cost-9-hour>

Supplementary Files

Figures

PRISMA 2020 flow diagram.

