Editorial Commentary

Evaluating the efficacy of artificial intelligence chatbots in urological health: insights for urologists on patient interactions with large language models

Benjamin D. Simon[1,2], **David G. Gelikman**[1], **Baris Turkbey**[1]

[1]Molecular Imaging Branch, National Cancer Institute, National Institutes of Health, Bethesda, MD, USA; [2]Institute of Biomedical Engineering, Department Engineering Science, University of Oxford, Oxford, UK

Correspondence to: Baris Turkbey, MD. Molecular Imaging Branch, National Cancer Institute, National Institutes of Health, 9000 Rockville Pike, Bethesda, MD 20892, USA. Email: turkbeyi@mail.nih.gov.

Comment on: Musheyev D, Pan A, Loeb S, *et al.* How Well Do Artificial Intelligence Chatbots Respond to the Top Search Queries About Urological Malignancies? Eur Urol 2024;85:13-6.

Keywords: Artificial intelligence (AI); chatbot; urology

Submitted Dec 08, 2023. Accepted for publication Mar 17, 2024. Published online May 07, 2024.
doi: 10.21037/tau-23-635
View this article at: https://dx.doi.org/10.21037/tau-23-635

Introduction

As OpenAI's ChatGPT and other large language models (LLMs) receive increasing attention and slowly become integrated into our everyday lives, we are witnessing the emergence of transformative technology. Comparable to the emergence and evolution of the first computers, we are experiencing the first LLMs to become available to the public. As computers quickly evolved from information storage systems to interactive systems encompassing fields such as banking and social media, it is difficult to predict how far artificial intelligence (AI) will extend. With the recent adoption of AI by researchers, students, and physicians, questions regarding security, accountability, and authorship arise, especially in medicine (1,2). Tools such as ChatGPT are in their early stages and are at risk of generating inaccurate and potentially non-existent information (3), leaving users at risk of fraud or error if relied upon too heavily.

As many rush into using AI tools broadly and prematurely, a recent article by Musheyev *et al.* takes a deliberate approach to evaluating AI chatbot responses to common urological malignancy queries (4). Using the top 5 Google Trends queries related to 4 common cancers, chatbot responses were characterized using selected metrics. This editorial delves into their findings, critiques their methods, and explores the importance of evaluating emerging LLMs, offering urologists insight into impact of chatbots on patient care. Musheyev *et al.* take a key step in the critical evaluation of LLMs, aiming to assess the capabilities and limitations of ChatGPT, among other LLMs, in answering queries in the urology oncology space. While their analysis is pioneering, it begs the question: what is an appropriate way to evaluate a dynamic and quickly evolving LLM such as ChatGPT? Further, how will this tool impact the accessibility and utility of information in the field of urology?

Study summary

Musheyev *et al.* evaluated the performance of four AI chatbots (ChatGPT v3.5, Perplexity, Chat Sonic, and Microsoft Bing AI) in providing information on common search queries regarding various urological malignancies (prostate, bladder, kidney, and testicular cancer). Each chatbot received the top 5 Google Trends phrases for each cancer as exact inputs with the following settings: default for ChatGPT v3.5, concise for Perplexity, Google Integrated concise for Chat Sonic, and balanced results for Microsoft Bing AI (with memory cleared for all between each input). The study used evaluation metrics such as DISCERN (a tool for evaluating the quality of written healthcare information) (5), the Patient Education Materials

Table 1 Description of metrics as used by Musheyev *et al.*

Metric	Scale description	Range	Interpretation
DISCERN	5-point scale for quality assessment	1	Low quality
		2–3	Moderate quality
		4–5	High quality
PEMAT-P	Percentage scores for understandability and actionability	0–59%	Poor
		60–79%	Fair
		80–100%	Good to excellent
Likert scale	5-point scale to gauge misinformation	1	No misinformation
		2–4	Some misinformation
		5	High misinformation
Flesch-Kincaid readability test	Grade level of text readability	0–5	Very easy
		6–8	Easy/standard
		9–12	Fairly difficult
		13–16	Difficult
		17+	Very difficult

PEMAT-P, Patient Education Materials Assessment Tool for Printable Materials.

Assessment Tool for Printable Materials (PEMAT-P) (6), a five-point Likert scale for misinformation (7), and the Flesch-Kincaid readability test to evaluate the outputs (see *Table 1* for scales) (8).

Their primary findings indicate that AI chatbots generally provide responses with moderate to high information quality with moderate understandability and low actionability. This corresponds to a median DISCERN score of 4 out of 5, median PEMAT-P understandability score of 66.7%, and median PEMAT-P actionability score of 40%, respectively. A median Likert score of 1 for misinformation across all models indicated high factual accuracy in responses; however, the readability level was determined to be fairly difficult, exceeding the recommended level for disseminating consumer health information. The study found that LLM responses for prostate cancer queries had the lowest median scores for understandability and actionability, though, across all four cancers, the quality of information was high and free from misinformation. In comparing the models themselves, ChatGPT had the lowest median DISCERN score, at least partly because it did not cite any sources.

While the study suggests that AI chatbots provide accurate information and do not necessarily spread misinformation, their performance could be improved by making responses more approachable to the average audience, possibly including visuals or clear, actionable instructions. The study acknowledges its limitations appropriately, such as the potential mismatch between study input phrasing and real user queries due to the use of top Google searches as inputs. The authors suggest future studies to evaluate if alternative responses from AI chatbots would impact the evaluation scores. In summary, the authors acknowledge that AI may present a more accurate source of medical information when it comes to urological malignancies compared to other online platforms, but there is significant room for improvement in accessibility.

Analysis of findings and critical evaluation

This study draws much-needed attention to how individuals use AI LLMs to gain medical knowledge, whether it be for their own health or to help someone they care for. While the authors' approach of using carefully selected, published metrics for the evaluation of AI responses to controlled queries minimizes bias and maximizes fairness, it also prevents the dynamic nature of these models from being explored. LLMs like ChatGPT are not simple querying systems akin to Google or Bing. These are complex pattern-matching, language "experts" that respond dynamically to feedback. They are interactive, and therefore should be evaluated as such.

Despite the dynamic nature of these tools, Musheyev *et al.* seemed to take a conceptually similar approach to most other current evaluative literature in urology. A study by Cocci *et al.* compared ChatGPT's ability to provide accurate, high-quality information for urology cases to expert urologists and used metrics such as DISCERN and Flesch-Kincaid readability scores (9). They found that the median DISCERN metric of 15 corresponded to poor quality, and the Flesch-Kincaid score of 15.8 corresponded to college graduate reading level. This discrepancy in DISCERN scores for quality between this study and Musheyev *et al.* calls into question the subjectivity of both studies and the DISCERN metric itself. Musheyev *et al.* seem to be assessing quality using DISCERN from the viewpoint of the average patient or consumer, while Cocci *et al.* seem to compare this to the quality of expert urologists. While tools like DISCERN may assist in quantifying subjective analysis, the assessment of quality remains largely dependent on the initial expectations of the person completing the questionnaire.

A study published outside of the urology space by Biswas *et al.* used a more overtly subjective ranking, asking experts to use a Likert scale to assess AI chatbot quality of information about myopia (10). While unrelated to urology, their decision to use a Likert scale to assess quality (while Musheyev *et al.* use the same metric to assess misinformation) draws attention to the many methods used to leverage and evaluate healthcare information even when selecting from the same few quantitative metrics.

Evaluating dynamical LLMs

Although it is important to rely on quantitative metrics and standardized queries for each language model, a dynamic tool such as ChatGPT requires a more adaptable evaluation. Equally vital as adhering to the scientific method is the relevance to the research question. If the goal is to understand how everyday users might employ ChatGPT to understand urologic malignancies, a study allowing open interaction with the chatbot and having users rate their understanding and the usefulness of the information could yield more insightful results. Additionally, observing the follow-up questions asked by users after the initial response might be more telling of the chatbot's effectiveness than the initial answer itself. For instance, if, as Musheyev *et al.* suggest, average users find language model responses to common queries overly complex, it would be informative to observe how they navigate this challenge. A user cannot tell a typical search engine "I'm confused" and ask for clarification but can with an interactive LLM.

Consider if ChatGPT was aware that it was being evaluated on specific metrics and received a user response such as "I don't know what to do with this information" after providing an answer. This situation highlights a limitation in Musheyev *et al.*'s methodology. Their approach overlooks the interactive capability of AI models to clarify complicated medical terms, which could otherwise result in poor understandability or require a college-level comprehension. Unlike static information sources like Google, TikTok, or YouTube, ChatGPT can directly and efficiently explain intricate medical jargon upon request. Therefore, the key to enhancing the poor understandability of LLMs might lie not in changing the models themselves, but in educating users on how to interact with these chatbots more effectively.

Recommendations

Considering the insights published by Musheyev *et al.* and other recent studies, several recommendations emerge for the future application and evaluation of LLMs such as ChatGPT in healthcare.

Dynamic evaluative studies

Future studies should emphasize the dynamic interactions between users and AI chatbots. Instead of relying solely on predetermined queries, researchers should document real-time user interactions and subsequent questions to better understand when and how users seek clarification. This approach may better capture the interactive nature of these tools to provide insight into their practical impact.

User education

These studies indicate the importance of educating potential users and the public about the capabilities and limitations of AI chatbots, particularly regarding medical information. Users need to understand how to interact with these models most effectively, distinguishing them from familiar resource platforms such as Google or YouTube.

Model improvement

AI chatbots should provide clear, actionable advice when needed. This may involve integrating visual aid into outputs (this is beginning to be incorporated into ChatGPT, and a recent release also incorporates the inclusion of images as an input) (11). Developers should continue to improve these models to enhance accuracy and clarity, using research like that of Musheyev *et al.* as a guide.

Ethical and responsible use

While not the primary focus of this editorial, the ethical and responsible use of AI, especially in sensitive areas like healthcare, remains paramount. This involves clear communication about the limitations of AI and advising users to consult healthcare professionals prior to making crucial health decisions. It is unclear what the potential impact of chatbots may be on the patient-physician relationship, but it remains clear that they will continue to impact what information is available to patients and how patients seek information. Although there is promise that AI may improve accuracy of certain healthcare decision making processes, there is also concern that AI may exacerbate existing biases. For the purposes of their study, Musheyev *et al.*, wiped the memory of each chatbot, yet in practice chatbots can have a rich memory of user history potentially impacting the information and phrasing of its responses. How this impacts bias, health disparities, and more is yet to be established.

Conclusions and closing remarks

The investigation of LLMs in healthcare, as demonstrated by Musheyev *et al.*, is a significant step in grasping the potential and limitations of emerging AI technologies and formulating methods to evaluate them. Their study emphasizes the accuracy and potential of AI chatbots in delivering medical information, while also underscoring the need to improve the understandability and actionability of the information provided. Given the interactive nature of LLMs like ChatGPT, evaluations should be user-focused, assessing not only the AI's initial responses but also its capacity for engaging in meaningful dialogue and responding to user feedback. More comprehensive approaches to understanding quality, actionability, and understandability of AI should be explored in urology. As we enter a new era of technology and healthcare, a cautious yet inquisitive stance towards AI integration is essential. Following the example of Musheyev *et al.*, these tools must be evaluated carefully and systematically, considering their rapid evolution. The development of public-facing AI is in its infancy, and its future will be shaped by collaborative efforts among users, researchers, and medical professionals alike.

Acknowledgments

Funding: None.

Footnote

Provenance and Peer Review: This article was commissioned by the editorial office, *Translational Andrology and Urology*. The article has undergone external peer review.

Peer Review File: Available at https://tau.amegroups.com/article/view/10.21037/tau-23-635/prf

Conflicts of Interest: All authors have completed the ICMJE uniform disclosure form (available at https://tau.amegroups.com/article/view/10.21037/tau-23-635/coif). B.T. reports that he receives support as part of a CRADA between NIH and NVIDIA and between NIH and Philips Healthcare; he also receives royalties from NIH, and is party to patents or potential patents relevant to the subject matter. The other authors have no conflicts of interest to declare.

Ethical Statement: The authors are accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

References

1. 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.

2. Solomon DH, Allen KD, Katz P, et al. ChatGPT, et al… Artificial Intelligence, Authorship, and Medical Publishing. Arthritis Rheumatol 2023;75:867-8.
3. Sallam M. ChatGPT Utility in Healthcare Education, Research, and Practice: Systematic Review on the Promising Perspectives and Valid Concerns. Healthcare (Basel) 2023;11:887.
4. Musheyev D, Pan A, Loeb S, et al. How Well Do Artificial Intelligence Chatbots Respond to the Top Search Queries About Urological Malignancies? Eur Urol 2024;85:13-6.
5. Charnock D, Shepperd S, Needham G, et al. DISCERN: an instrument for judging the quality of written consumer health information on treatment choices. J Epidemiol Community Health 1999;53:105-11.
6. Shoemaker SJ, Wolf MS, Brach C. Development of the Patient Education Materials Assessment Tool (PEMAT): a new measure of understandability and actionability for print and audiovisual patient information. Patient Educ Couns 2014;96:395-403.
7. Loeb S, Sengupta S, Butaney M, et al. Dissemination of Misinformative and Biased Information about Prostate Cancer on YouTube. Eur Urol 2019;75:564-7.
8. Ley P, Florio T. The use of readability formulas in health care. Psychol Health Med 1996;1:7-28.
9. Cocci A, Pezzoli M, Lo Re M, et al. Quality of information and appropriateness of ChatGPT outputs for urology patients. Prostate Cancer Prostatic Dis 2024;27:103-8.
10. Biswas S, Logan NS, Davies LN, et al. Assessing the utility of ChatGPT as an artificial intelligence-based large language model for information to answer questions on myopia. Ophthalmic Physiol Opt 2023;43:1562-70.
11. Santana LADM, Floresta LG, Alves ÊVM, et al. Can GPT-4 be a viable alternative for discussing complex cases in digital oral radiology? A critical analysis. EXCLI J 2023;22:749-51.

Cite this article as: Simon BD, Gelikman DG, Turkbey B. Evaluating the efficacy of artificial intelligence chatbots in urological health: insights for urologists on patient interactions with large language models. Transl Androl Urol 2024;13(5):879-883. doi: 10.21037/tau-23-635