Editorial Commentary

Exploring the capabilities and challenges of ChatGPT in disseminating urologic information

Amber McMahon[1], Cian L. Jacob[1], Vincent G. Bird[2], Bristol B. Whiles[1]^

[1]Department of Urology, University of Kansas Medical Center, Kansas City, KS, USA; [2]Department of Urology, University of Florida Medical Center, Gainesville, FL, USA

Correspondence to: Bristol B. Whiles, MD. Department of Urology, University of Kansas Medical Center, 4000 Cambridge Street, Mailstop#3016, Kansas City, KS 66160, USA. Email: bristolwhiles@gmail.com.

Comment on: Davis R, Eppler M, Ayo-Ajibola O, *et al.* Evaluating the Effectiveness of Artificial Intelligence-powered Large Language Models Application in Disseminating Appropriate and Readable Health Information in Urology. J Urol 2023;210:688-94.

Keywords: Artificial intelligence (AI); large language models (LLMs); urology; communication; ChatGPT

Submitted Oct 26, 2023. Accepted for publication Jan 24, 2024. Published online Mar 11 2024.
doi: 10.21037/tau-23-545
View this article at: https://dx.doi.org/10.21037/tau-23-545

With the current demand for instantly accessible information, the Internet is consistently the universal platform for knowledge dissemination. This has become particularly important in the realm of medical knowledge, and even more so with the rise of artificial intelligence (AI)-driven chatbots further ushering in a new era of language processing. With ChatGPT swiftly gaining prominence, it provides a mechanism for patients to easily ask medical questions.

This study by Davis *et al.* poised an interesting exploration into potential benefits and pitfalls of ChatGPT-3.5 in relation to disseminating appropriate and readable urologic healthcare information (1). Their team asked ChatGPT 18 questions they designed from Google trends specifically regarding urologic oncology, emergencies, and benign diseases. They investigated the quality, appropriateness, and readability of the responses.

Regarding response appropriateness, they adopted a unique approach by utilizing three different components to assess overall appropriateness: accuracy, comprehensiveness, and clarity. They reported an impressive response appropriateness rate at 77.8%. In comparison, other similar studies looking at response appropriateness to urologic questions found rates of 60% (2) and 52% (3). The authors note that of their three components, clarity scored higher more frequently at a significant level. This is interesting, because while ChatGPT's response can have good clarity, the information can still lack accuracy and comprehensiveness, potentially skewing their overall appropriateness score.

Davis *et al.*'s study, and the others referenced above, utilized ChatGPT-3.5; however, a newer version, GPT-4, has since been created. Since its release, a few studies have compared the responses of ChatGPT-3.5 to GPT-4 with all of them finding that GPT-4 outperformed ChatGPT-3.5 (4-6). In the study by Taloni *et al.* evaluating multiple-choice questions from the American Academy of Ophthalmology, GPT-4 accuracy was best at 82.4%, followed by humans at 75.7%, and then ChatGPT-3.5 at 65.9% (4). Another study compared the ability of GPT-4, ChatGPT-3.5 and Google Bard to accurately respond to myopia-related questions, and again GPT-4 was superior with 80.6% accurate answers compared to 61.3% for ChatGPT-3.5 and 54.8% for Google Bard (5). Finally, an additional study took a unique approach by comparing ChatGPT-3.5 to GPT-4 responses to neurosurgical questions to the responses by neurosurgeons of varying seniority. It found that responses of ChatGPT-3.5 were comparable to low-seniority surgeons

^ ORCID: 0000-0002-3525-7557.

© Translational Andrology and Urology. All rights reserved.

while the responses of GPT-4 were comparable to high-seniority surgeons (6). In addition to ChatGPT, there have been multiple other AI powered chatbots that have been introduced including but not limited to Claude, Bard, Bing AI and many others. A few studies have compared ChatGPT to these other chatbots. Again, GPT-4 often proves to reign supreme, but often the other chatbots evaluated tend to produce similar results on average (7,8). These findings underscore the exponential improvements in response quality with the new GPT-4 version of this large language model (LLM). Therefore, the results of the Davis *et al.* study may be improved if performed with the most recent version of ChatGPT. However, it is also important to consider that ChatGPT-3.5 is currently the free version of this LLM. Therefore, it is likely that the version utilized in this study would actually be used by the majority of users seeking answers to healthcare questions.

In addition, the questions for this study by Davis *et al.* were created to mimic a medical question from a layperson; however, questions were ultimately overseen by medical professionals. The way prompts are constructed significantly influences the responses generated by AI-powered chatbots. If a medical question is posed by a layperson rather than a medical professional, it could theoretically lead to a distinct difference in responses. In a study by Nguyen and colleagues, the use of prompt engineering techniques was associated with slightly higher accuracy and improved responses to open-ended prompts (9). These variations in user prompts may not only lead to different AI responses but could also contribute to increased mistrust of AI amongst healthcare professionals. Overcoming this mistrust is a major challenge for implementing AI in healthcare (10).

Another noteworthy finding is the readability level of the response. They assessed readability by using both Flesch Reading Ease and Flesch Kinkade Reading Grade Level scores. The average grade-level score was 13.5, indicating college-level reading. This correlates with other studies that evaluated the readability of urologic information by ChatGPT that also found the material to be college-level or greater (3,11).

A recurring challenge in evaluating the responses from AI language models in the context of medical advice is the choice of evaluation tools. This study employed a combination of Likert scales and components from QUEST and DISCERN tools, a strategy echoed in other studies that utilized tools such as the Brief DISCERN, DISCERN, newly created LIKERT scales, and other assessment variations (2,3,5,11-13). However, these tools were not specifically created or validated to evaluate medical information generated from AI tools or chatbots. Consequently, this issue complicates investigations when trying to compare and evaluate the results from different studies and highlights the need for a validated tool to evaluate AI chatbot responses to medical inquiries.

When it comes to creating a validated evaluation tool for assessing AI-generated responses for healthcare, there are many factors that must be considered. Foremost among these is ensuring the absolute accuracy of the information contained in the LLM response. While it can often appear logical and appropriate, the response validity is paramount, as inaccuracies could have serious consequences. It is equally imperative to evaluate the safety of the information provided. If the LLM provides an incorrect response, assessing whether it could potentially result in harm to the patient is crucial. Furthermore, AI chatbots should not only provide information but also actively encourage patients to seek further consultation from a qualified healthcare provider for additional explanation and clinical correlation. AI-generated responses have the potential ability to complement, but should not replace, medical advice from a qualified healthcare provider. These components would clearly distinguish a new evaluation tool and differentiate it from the DISCERN tool, one evaluation metric system currently utilized by some research groups. Unlike the DISCERN tool, the new AI specific tool should focus more on emphasizing the accuracy, appropriateness, and safety of the responses. A less significant weight could also be placed on the evaluation of references; although at least one question should evaluate and ensure provided resources are appropriate and functional, if additional reading is desired by the patient. This approach would ensure a more comprehensive assessment of AI-generated responses within medicine.

In conclusion, this study further reinforces existing data emphasizing the importance for continuous evaluation, development, and improvement of AI-driven language models, particularly in the context of medical information. In the realm of healthcare, any inappropriate response can potentially harm patients, making zero tolerance for inappropriate responses essential. Furthermore, responses that are overly complex and at a high reading level can lead to misinterpretation and limit the ability for use by the general population. Finally, to be able to completely evaluate and compare the responses to medical questions from a variety of LLM, a new tool needs to be created and validated. Therefore, individuals seeking medical advice

© Translational Andrology and Urology. All rights reserved.

through the current AI resources should exercise caution and ultimately consult with a healthcare provider to ensure appropriate responses and results.

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-23-545/prf

Conflicts of Interest: All authors have completed the ICMJE uniform disclosure form (available at https://tau.amegroups.com/article/view/10.21037/tau-23-545/coif). The 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.

Open Access Statement: This is an Open Access article distributed in accordance with the Creative Commons Attribution-NonCommercial-NoDerivs 4.0 International License (CC BY-NC-ND 4.0), which permits the non-commercial replication and distribution of the article with the strict proviso that no changes or edits are made and the original work is properly cited (including links to both the formal publication through the relevant DOI and the license). See: https://creativecommons.org/licenses/by-nc-nd/4.0/.

References

1. Davis R, Eppler M, Ayo-Ajibola O, et al. Evaluating the Effectiveness of Artificial Intelligence-powered Large Language Models Application in Disseminating Appropriate and Readable Health Information in Urology. J Urol 2023;210:688-94.
2. Whiles BB, Bird VG, Canales BK, et al. Caution! AI Bot Has Entered the Patient Chat: ChatGPT Has Limitations in Providing Accurate Urologic Healthcare Advice. Urology 2023;180:278-84.
3. 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;27:159-60.
4. Taloni A, Borselli M, Scarsi V, et al. Comparative performance of humans versus GPT-4.0 and GPT-3.5 in the self-assessment program of American Academy of Ophthalmology. Sci Rep 2023;13:18562.
5. 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.
6. Liu J, Zheng J, Cai X, et al. A descriptive study based on the comparison of ChatGPT and evidence-based neurosurgeons. iScience 2023;26:107590.
7. Cheong RCT, Pang KP, Unadkat S, et al. Performance of artificial intelligence chatbots in sleep medicine certification board exams: ChatGPT versus Google Bard. Eur Arch Otorhinolaryngol 2023. [Epub ahead of print]. doi: 10.1007/s00405-023-08381-3.
8. Lozić E, Štular B. Fluent but Not Factual: A Comparative Analysis of ChatGPT and Other AI Chatbots' Proficiency and Originality in Scientific Writing for Humanities. Future Internet 2023;15:336.
9. Nguyen D, Swanson D, Newbury A, et al. Evaluation of ChatGPT and Google Bard Using Prompt Engineering in Cancer Screening Algorithms. Acad Radiol 2023. [Epub ahead of print]. doi: 10.1016/j.acra.2023.11.002.
10. Sedaghat S. Future Potential Challenges of Using Large Language Models Like ChatGPT in Daily Medical Practice. J Am Coll Radiol 2024;21:344-5.
11. Pan A, Musheyev D, Bockelman D, et al. Assessment of Artificial Intelligence Chatbot Responses to Top Searched Queries About Cancer. JAMA Oncol 2023;9:1437-40.
12. Coskun B, Ocakoglu G, Yetemen M, et al. Can ChatGPT, an Artificial Intelligence Language Model, Provide Accurate and High-quality Patient Information on Prostate Cancer? Urology 2023;180:35-58.
13. Yeo YH, Samaan JS, Ng WH, et al. Assessing the performance of ChatGPT in answering questions regarding cirrhosis and hepatocellular carcinoma. Clin Mol Hepatol 2023;29:721-32.

Cite this article as: McMahon A, Jacob CL, Bird VG, Whiles BB. Exploring the capabilities and challenges of ChatGPT in disseminating urologic information. Transl Androl Urol 2024;13(3):470-472. doi: 10.21037/tau-23-545

© Translational Andrology and Urology. All rights reserved.