

COMMENT OPEN



Comment on: Assessing ChatGPT's ability to answer questions pertaining to erectile dysfunction

Jacob S. Hershenhouse^{1,2,3} and Giovanni E. Cacciamani^{1,2,3}✉

© The Author(s) 2024

IJIR: Your Sexual Medicine Journal; <https://doi.org/10.1038/s41443-023-00821-2>

Due to the sensitive and sexual nature of the condition, many men opt to consult the internet first when seeking to understand Erectile Dysfunction (ED), anonymously researching the subject and any potential solutions online [1] prior to enlisting their primary care physician or urologist for advice and treatment. Now with the rapid release of publicly accessible, consumer-facing Generative Artificial Intelligence (GAI) like ChatGPT, studying the efficacy of patient-facing GAI is crucial as the technology evolves. We have read with interest the recent study by Razdan et al., which raises important questions about the capabilities and limitations of large language models like ChatGPT applied to answering common inquiries about ED [2]. A number of significant conclusions can be drawn from their results.

It is crucial to recognize that GPT models are not specifically trained in medical knowledge, unlike specialized systems such as Medpalm2 [3]. Despite GPT's impressive ability to respond to medical inquiries, there is an inherent risk of inaccuracies when addressing public questions. A significant concern is the manner in which these chatbots deliver responses. The "human-like" format and style of their outputs can mislead patients into over-trusting this "AI oracle" without skepticism. This is particularly perilous because, unlike Google searches where users actively select from numerous links or webpages, the sources behind AI-powered chatbot responses are obscured, and the output is singular. This raises an essential issue: as physicians, we must engage actively in the development and evaluation of these AI-chatbots, rather than passively accept them or become involved only at the final stages. In the business of medicine, which fundamentally relies on trust, the accuracy and validation of the knowledge used by GPTs in responding to patient inquiries must be meticulously scrutinized and validated [4].

We found the results of the authors' readability analysis particularly telling and generalizable. Prior studies have shown that ChatGPT produces (when not otherwise prompted) standard outputs of post-secondary grade level when employed for various medical use cases [5, 6]. We agree with the authors that the reading level and terminology used by ChatGPT exceeds the health literacy level of many ED patients. It is best practice to not exceed a Flesch-Kincaid Grade Level readability score above grade level 8 on medical documentation intended for patient understanding, as is the case for informed consent forms [7]. It is the role of the physician to educate the patient about finding verifiable, understandable information online for their conditions [8]. Simply put, the more

patients truly comprehend, the better. To improve the readability of generated outputs, and therefore the level of understanding of the online patient population, we believe that prompting the chatbot to simplify its explanations and use more colloquial terms could make ChatGPT's responses more accessible when the responses are intended specifically to be patient-facing.

We also found their interpretation of ChatGPT's "empathy" to merit further investigation. Men's health conditions necessitate empathetic answers from healthcare professionals and source material, which extends inherently to online resources. ChatGPT's ability to demonstrate empathy and provide couples counseling reveals the technology's potential to aid in reducing the stigma around ED and encourage open communication on the subject before one comes into clinic. This ability to incorporate supportive language suggests this technology could be useful in co-piloting, alongside healthcare professionals, online information seeking particularly for men's health diseases.

In response to the subjective interpretation of response quality, we caution against overgeneralizing ChatGPT's lack of nuance regarding treatment modalities as proof it is incapable or untrustworthy in a medical context. OpenAI's usage policies warn against this potential use case for the software: "You should never use our models to provide diagnostic or treatment services for serious medical conditions [9]." Despite this warning, the software does still attempt to provide this information, though its knowledge base is restricted to before Q4 2021, which leads to an expected dip in performance when assessed by up-to-date, expert providers [10–12]. These results underscore the need for greater fine-tuning of these models if they are to fit this use case in patient education and we join other researchers in calling for greater regulation of this software in all medical applications [13–15]. Medical researchers conducting exploratory research into other medical use cases should heed the usage policies for the software under examination as well as ethical concerns surrounding the technology as well. With this in mind, it begs the question: why is there no standardized method to ascertain the quality of chatbot generated responses to medical questions? In an effort to allow researchers to compare and replicate studies such as this one, we propose the following 5-item quantitative analysis for assessing response quality: Accuracy, Completeness, Clarity, Readability, and Understandability and empathy (see Fig. 1). Without assessing each of these essential reflections of quality, the validity of conclusions based in an incomplete research methodology in this domain could be called into question.

¹USC Institute of Urology and Catherine and Joseph Aresty Department of Urology, Keck School of Medicine, University of Southern California, Los Angeles, CA, USA. ²Artificial Intelligence Center, USC Institute of Urology, University of Southern California, Los Angeles, CA, USA. ³These authors contributed equally: Jacob S. Hershenhouse, Giovanni E. Cacciamani. ✉email: giovanni.cacciamani@med.usc.edu

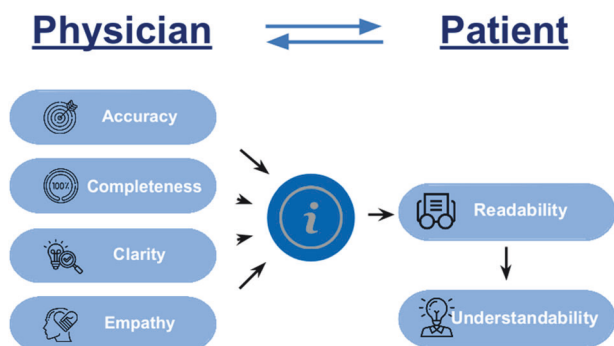


Fig. 1 Keyfeatures for assessing AI-powered medical chatbots response quality.

While an imperfect source of ED information for patients, ChatGPT demonstrates the promise of AI to supplement human clinical expertise and disseminate medical information if employed correctly. Yet, the information provided, even when generated by machine, should be always supported by clinical evidence and reference. With appropriate oversight and updating as the technology evolves, large language models could meaningfully expand access to consumer health education on sexual health. More interdisciplinary research is needed, but our shared goal remains empowering patients with accurate, empathetic ED knowledge.

REFERENCES

- Russo GI, di Mauro M, Cocci A, Cacciamani G, Cimino S, Serefoglu EC, et al. Consulting "Dr Google" for sexual dysfunction: a contemporary worldwide trend analysis. *Int J Impot Res.* 2020;32:455–61.
- Razdan S, Siegal AR, Brewer Y, Slijovich M, Valenzuela RJ. Assessing ChatGPT's ability to answer questions pertaining to erectile dysfunction: can our patients trust it? *Int J Impot Res.* 2023. <https://doi.org/10.1038/s41443-023-00797-z>.
- Singhal K, Azizi S, Tu T, Mahdavi SS, Wei J, Chung HW, et al. Large language models encode clinical knowledge. *Nature.* 2023;620:172–80.
- Huo B, Cacciamani GE, Collins GS, McKechnie T, Lee Y, Guyatt G. Reporting standards for the use of large language model-linked chatbots for health advice. *Nat Med.* 2023;29:2988.
- Momenaei B, Wakabayashi T, Shahlaee A, Durrani AF, Pandit SA, Wang K, et al. Appropriateness and readability of ChatGPT-4-generated responses for surgical treatment of retinal diseases. *Ophthalmol Retin.* 2023;7:862–8.
- Pan A, Musheyev D, Bockelman D, Loeb S, Kabarriti AE. Assessment of Artificial Intelligence Chatbot responses to top searched queries about cancer. *JAMA Oncol.* 2023;9:1437–40.
- Pandiya A. Readability and comprehensibility of informed consent forms for clinical trials. *Perspect Clin Res.* 2010;1:98–100.
- Cocci A, Pezzoli M, Lo Re M, Russo GI, Asmundo MG, Fode M, et al. Quality of information and appropriateness of ChatGPT outputs for urology patients. *Prostate Cancer Prostatic Dis.* 2023. <https://doi.org/10.1038/s41391-023-00754-3>
- Usage policies: OpenAI; 2023 [updated March 23, 2023]. Available from: <https://openai.com/policies/usage-policies>.

- Davis R, Eppler M, Ayo-Ajibola O, Loh-Doyle JC, Nabhani J, Samplaski M, 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.
- Eppler MB, Ganjavi C, Knudsen JE, Davis RJ, Ayo-Ajibola O, Desai A, et al. Bridging the gap between urological research and patient understanding: the role of large language models in automated generation of Layperson's summaries. *Urol Pract.* 2023;10:436–43.
- Huynh LM, Bonebrake BT, Schultis K, Quach A, Deibert CM. New artificial intelligence ChatGPT performs poorly on the 2022 Self-assessment study program for urology. *Urol Pract.* 2023;10:409–15.
- Baumgartner C, Baumgartner D. A regulatory challenge for natural language processing (NLP)-based tools such as ChatGPT to be legally used for healthcare decisions. where are we now? *Clin Transl Med.* 2023;13:e1362.
- Cacciamani GE, Chen A, Gill IS, Hung AJ. Artificial intelligence and urology: ethical considerations for urologists and patients. *Nat Rev Urol.* 2023. <https://doi.org/10.1038/s41585-023-00796-1>.
- Meskó B, Topol EJ. The imperative for regulatory oversight of large language models (or generative AI) in healthcare. *NPJ Digit Med.* 2023;6:120.

FUNDING

Open access funding provided by SCELC, Statewide California Electronic Library Consortium.

COMPETING INTERESTS

The authors declare no competing interests.

ADDITIONAL INFORMATION

Correspondence and requests for materials should be addressed to Giovanni E. Cacciamani.

Reprints and permission information is available at <http://www.nature.com/reprints>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2024