Artificial intelligence improves urologic oncology patient education and counseling

Yash B. Shah, MD, Anushka Ghosh, MD, Aaron Hochberg, MD, James R. Mark, MD, Costas D. Lallas, MD, Mihir S. Shah, MD
Department of Urology, Sidney Kimmel Medical College, Thomas Jefferson University, Philadelphia, Pennsylvania, USA

SHAH YB, GHOSH A, HOCHBERG A, MARK JR, LALLAS CD, SHAH MS. Artificial intelligence improves urologic oncology patient education and counseling. *Can J Urol* 2024;31(5):12013-12018.

Introduction: *Patients seek support from online resources when facing a troubling urologic cancer diagnosis. Physician-written resources exceed the recommended 6-8th grade reading level, creating confusion and driving patients towards unregulated online materials like AI chatbots. We aim to compare the readability and quality of patient education on ChatGPT against Epic and Urology Care Foundation (UCF).*

Materials and methods: *We analyzed prostate, bladder, and kidney cancer content from ChatGPT, Epic, and UCF. We further studied readability-adjusted responses using specific AI prompting (ChatGPT-a) and Epic material designated as Easy to Read. Blinded reviewers completed descriptive textual analysis, readability analysis via six validated formulas, and quality analysis via DISCERN, PEMAT, and Likert tools.*

Results: *Epic met the recommended grade level, while UCF and ChatGPT exceeded it (5.81 vs. 8.44 vs. 12.16, $p < 0.001$). ChatGPT text was longer with more complex wording ($p < 0.001$). Quality was fair for Epic, good for UCF, and excellent for ChatGPT (49.5 vs. 61.67 vs. 64.33). Actionability was overall poor but particularly lowest (37%) for Epic. On qualitative analysis, Epic lagged on all quality measures. When adjusted for user education level (ChatGPT-a and Epic Easy to Read), readability improved (7.50 and 3.53), but only ChatGPT-a retained high quality.*

Conclusions: *Online urologic oncology patient materials largely exceed the average American's literacy level and often lack real-world utility for patients. Our ChatGPT-a model indicates that AI technology can improve accessibility and usefulness. With development, a healthcare-specific AI program may help providers create content that is accessible and personalized to improve shared decision-making for urology patients.*

Key Words: ChatGPT, artificial intelligence, urologic oncology, patient education, health literacy, health systems research

Introduction

Due to the stressful nature of a cancer diagnosis and multitude of novel treatment options, patients increasingly complement their physician's counseling with online resources to better understand their care.[1,2] Generative artificial intelligence (AI) models like ChatGPT are user-friendly and mimic human conversation, offering a popular resource for self-education. However, physicians are rightfully concerned about content accuracy and potential effects on patient decision-making.[1,3,4]

Previous research shows AI chatbots effectively perform certain clinical tasks, including triaging

Accepted for publication September 2024

Address correspondence to Dr. Mihir S. Shah, Department of Urology, Thomas Jefferson University, 1025 Walnut St., Suite 1100, Philadelphia, PA 19107 USA

patients and supporting clinical decision-making,[5] interpreting pathologic, genomic, and radiologic data,[6] composing operative notes,[7] and providing a simplified patient summary of scientific literature.[8] There are key shortcomings including occasional misinformation[9-13] and poor accessibility for laypeople.[14,15] We previously demonstrated low readability when querying ChatGPT for men's sexual health education, although we found improvement with specific AI prompting.[14]

Because most physician-written online resources exceed recommended reading levels, AI has the potential to improve accessibility of online content. There has not been a comprehensive analysis of overall accessibility, quality, and actionability of cancer patient education written by AI. Optimal patient materials are not only accessible, but draw from scientifically-sound sources, provide comprehensive information, and help patients take meaningful action.[16,17]

Genitourinary cancer patients are commonly provided educational materials from Epic and Urology Care Foundation (UCF). To our knowledge, the current study is the first to compare these platforms with ChatGPT, aiming to understand the readability, quality, and actionability of AI-generated education.

Materials and methods

Data collection

UCF articles and Epic MyChart patient attachments covering kidney, bladder, and prostate cancer were identified. UCF subheadings were converted into questions to elicit responses from ChatGPT3.5. In total, 11 kidney, 39 bladder, and 29 prostate cancer questions were asked. Topics included physiology, symptoms, diagnostics, grading, staging, treatment, and prognosis of each cancer. All three platforms were queried on September 1, 2023.

To test the software's ability to adapt to the user's reading level, adjusted ChatGPT (ChatGPT-a) responses were obtained using the prompt "Explain it to me like I am in sixth grade" to match the 6-8th grade reading level as recommended by the National Institutes of Health and American Medical Association.[14] Similarly, an additional Epic Easy-to-Read attachment was available for prostate cancer and hence included in the analysis.

Readability

Word, complex word, sentence, and syllable count were evaluated for descriptive textual analysis. Readability analysis used six validated formulas, including Flesch-Kincaid Reading Ease Score (FRES), Flesch-Kincaid Grade Level (FKGL), Gunning-Fog Score (GFS), Simple Measure of Gobbledygook (SMOG), Coleman-Liau Index (CLI), and Automated Readability Index (ARI).[14] Scores were calculated and interpreted as previously described.[14]

Quality

Two blinded urologic oncologists scored all materials using the validated DISCERN, PEMAT understandability, and PEMAT actionability tools. A validated Likert scale (1 = Poor, 2 = Needs Improvement, 3 = Fair, 4 = Good, 5 = Excellent) was utilized for qualitative scoring of accuracy (Is the response evidence-based and medically accurate in comparison to AUA and NCCN guidelines?), comprehensiveness (Does the response provide sufficient information to fully inform patients about their diagnosis/treatment?), and understandability (Can the response be easily understood by average patients?).

Results

On descriptive textual analysis, ChatGPT content was lengthiest and contained the largest proportion of complex words across all three cancers, Table 1. Average word count for ChatGPT was 386.68; UCF was 17.6% shorter while Epic was 78.1% shorter. ChatGPT also had the largest proportion of complex words.

TABLE 1. **Descriptive textual analysis of patient education resources**

	ChatGPT	**Urology Care Foundation**	**Epic**	**ChatGPT-a**	**Epic Easy to Read[a]**
Word count	386.68	318.59	84.71	353.69	60.88
Complex word %	26.74	9.92	12.21	8.94	7.49
Words per sentence	17.20	16.73	10.42	15.37	8.91

[a]only available for prostate cancer queries

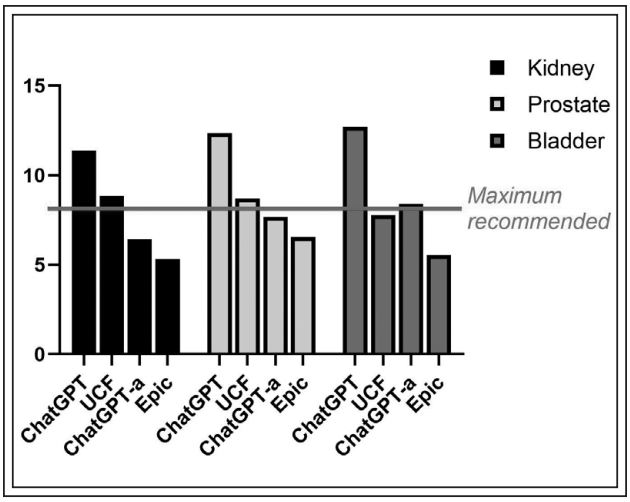

Figure 1. Readability by Flesch-Kincaid Grade Level across resources. Resources are recommended to be written at a maximum 8th grade reading level.

Readability was worst for ChatGPT across all three cancers, while only Epic content consistently met the national standard, Figure 1. When using FKGL, ChatGPT content was written at the 12.72 grade level for bladder cancer, 11.39 for kidney cancer, and 12.37 for prostate cancer. Conversely UCF content was written at 7.77, 8.85, and 8.71 grade levels, while Epic content was written at 5.56, 5.33, and 6.55 grade levels, respectively. This trend in readability held true across all six formulae, Table 2. With AI prompting for simpler information (ChatGPT-a), there was a notable drop in grade level required for adequate understanding.

Comparing ChatGPT-a to Epic Easy to Read, complex word usage fell to an average of 8.94% with ChatGPT-a and 7.49% with Epic Easy to Read (for prostate cancer). However, the Epic Easy to Read content lacked information on diagnosis and staging and qualitatively lacked relevant depth of information on topics such as treatment options.

In terms of quality, Epic lagged behind the other resources across all metrics. Across the cancers, DISCERN scores were fair for Epic at an average of 49.50, but good for UCF (61.67) and excellent for ChatGPT (64.33). PEMAT understandability was high for all three resources; conversely, PEMAT actionability was low for Epic at 37%, while it was fair for UCF and ChatGPT at 53% each. On qualitative analysis, accuracy, comprehensiveness, and quality were each lowest for Epic, while UCF and ChatGPT scored significantly higher.

When adjusted for user education level, ChatGPT-a demonstrated slightly decreased quality with DISCERN 56.17, PEMAT understandability 77%, and PEMAT actionability 53%. Accuracy also dropped to 3.67, comprehensiveness to 3.50, and understandability to 4.00. Epic Easy to Read saw notable drops in actionability and comprehensiveness.

Overall, ChatGPT offered more comprehensive and technical responses, with extensive detail in contrast to Epic – something which may be more useful for medical trainees or professionals. ChatGPT-a more commonly drew comparisons with everyday concepts, provided shorter answers, and used colloquial language. While these did not always meet the 6th grade level as requested by the prompt modifier, they did represent a sizeable improvement from original ChatGPT responses and met the higher 8th grade recommended

TABLE 2. **Readability analysis of patient education resources**

	ChatGPT	Urology Care Foundation	Epic	ChatGPT-a	Epic Easy to Read[a]
FKRE	38.53	64.31	71.94	66.44	85.68
FKGL	12.16	8.44	5.81	7.50	3.53
GFS	14.66	10.07	8.60	9.29	6.20
SMOG	10.95	7.33	6.55	6.77	4.66
CLI	15.56	11.03	11.08	10.01	7.74
ARI	12.27	8.41	4.44	7.76	1.88

[a]only available for prostate cancer queries
FKRE = Flesch Kincaid Reading Ease; FKGL = Flesch-Kincaid Grade Level; GFS = Gunning-Fog Score; SMOG = Simple Measure of Gobbledygook; CLI = Coleman-Liau Index; ARI = Automated Readability Index

benchmark. Epic used concise language and provided the shortest response. However, it scored low on quality and actionability, while ChatGPT-a appeared to perform highly on these metrics.

Discussion

Our study indicates poor readability for most online patient education materials, including those generated by AI, although Epic resources tended to be exempt from this trend. Nonetheless, while Epic resources were accessible, they lacked real utility for patients, as they were quite short and were not comprehensive or actionable. Accordingly, while Epic is a widely used electronic medical record across the United States, and its educational attachments are commonly the default resource that clinicians provide, our study indicates that these attachments are likely inadequate in assisting patients in thoroughly understanding and guiding their care.

In urologic oncology, where diagnoses are highly troubling and management options remain in flux, high quality patient education is important. Health literacy remains low in America, and poor patient understanding can harm adherence, autonomy, and ultimately outcomes.[18] Hence, readability and utility of online patient materials can significantly impact equity and public health. To improve outcomes in this field, we must better understand the content patients consume outside the clinic to ultimately counteract misinformation and fill gaps in understanding. Better understanding can also improve the design of future online materials including dynamic AI platforms.

Chatbots are commonly used in oncology and positively received by patients, yet studies are rare.[19,20] Upon PubMed search, we only located three studies investigating generative AI in oncology patient education. Given the limited and often conflicting conclusions, further research is needed to best understand the quality, accessibility, and limitations of contemporary AI. For instance, studies of breast implant-associated anaplastic large cell lymphoma and head and neck cancers found conflicting results on the quality of education from ChatGPT versus Google search, while the latter found that readability was poorer with ChatGPT.[21,22] Only one study investigated urologic cancers- finding that quality was moderate to high, understandability was moderate, and actionability was moderate to poor- on four AI chatbots. Quantitative readability was not calculated by the authors. Furthermore, there was no comparison to existing physician-written content, making it difficult to contextualize and appreciate the implications of these numerical findings for real-world application.[4]

We demonstrated that generative AI technology allows notable improvement in readability. Previous study has shown that while medical experts are widely aware of the poor readability in online content, they have been unsuccessful in addressing this flaw. This can be further shown in our study through the Epic Easy to Read prostate cancer resource, which is written by professionals intending to address the readability gap. This resource did include simpler, more accessible language, but it lost real-world utility with its short text that lacked accuracy, actionability, and comprehensiveness. Further, our qualitative Likert understandability ratings were unreliable and did not align with the quantitative readability formulae, showing that physician raters were not effective at interpreting a layman's ability to understand written text. Evidently, medical experts need assistance in improving readability within high-quality educational material.

Overall, though imperfect, ChatGPT-a might represent the most promising resource within our study, combining adjustable readability with high quality. Certainly, improvement is needed to ensure that the platform only draws from reliable evidence-based resources and is designed to provide the most actionable material for patients. Creation of a new generative AI resource that is specifically designed to provide patient education would be valuable. Such a tool can be specified to draw from limited, physician-designated medical literature, controlled to provide output that is highly actionable and specific, and instructed to incorporate more multimedia support. Cancer is heterogenous and requires decision-making based on innumerable variables including disease factors, comorbidities, and personal values. AI can gather patient data and provide personalized education, thoroughly discuss benefits and risks of diverse treatments, and exhaustively answer patient queries on-demand.[6,23]

Nonetheless, such a tool cannot fully replace physician communication. A physician's ability to communicate with empathy and humanity affects patient satisfaction, adherence to treatment plans, and even health outcomes. We posit that, with improvement, ChatGPT-like platforms can guide patients through in compiling relevant questions or establishing a baseline knowledge base ahead of appointments, ensuring that their conversations with physicians are focused and productive. Hence, AI allows physicians to dedicate more of their time to

personalized communication, where their expertise and empathy are most needed.

Additionally, there is promising research suggesting that AI can assist physicians in improving communication with patients online.[23] Given physicians receive numerous messages weekly, AI can alleviate some of physician burden by providing a draft of the response that the physician can edit, instead of starting from scratch. AI generated drafts are generally longer and appear more empathetic.[24] By integrating AI into messaging systems, doctors can maintain high-quality, personalized communication even when they are not in clinic, while alleviating some of the burdens associated with responding to large volumes of patient inquiries.

There are several limitations to our study. While ChatGPT can generate accurate data, it is prone to "hallucinations," where it sometimes produces responses that are factually incorrect or fabricated. "Hallucinations" occur responses are newly created based on patterns in data, rather than information directly drawn from reliable resources.[25] This yields stochasticity; repetition may produce new responses with slightly different quality or readability scores. However, we studied multiple cancers and asked 79 questions across a variety of topics to counteract this risk. Moreover, although the DISERN and PEMAT tools are widely validated, they were not originally designed for AI material. Accordingly, they may not perfectly represent the quality of these resources. For instance, ChatGPT is not traditionally designed to provide citations or visual media, which are criteria in these tools. Development of novel AI-specific metrics is needed.

Conclusions

Our study demonstrates continued flaws in existing physician-written educational materials within genitourinary cancer, including highly popular resources from Epic and Urology Care Foundation. Conversely, our findings indicate that generative AI can create accurate and accessible patient-facing content. With further development, AI may help physicians develop a new generation of useful, personalized content that helps patients understand their diagnoses and make management decisions in line with personal needs and values. □

References

1. Shah YB, Beiriger J, Mehta S, Cohen SD. Analysis of patient education materials on TikTok for erectile dysfunction treatment. *Int J Impot Res* Jul 7 2023 online ahead of print.
2. Cocci A, Pezzoli M, Lo Re M et al. Quality of information and appropriateness of ChatGPT outputs for urology patients. *Prostate Cancer Prostatic Dis* Epub Jul 29 2023.
3. Abramson M, Feiertag N, Javidi D, Babar M, Loeb S, Watts K. Accuracy of prostate cancer screening recommendations for high-risk populations on YouTube and TikTok. *BJUI Compass* 2023;4(2):206-213.
4. Musheyev D, Pan A, Loeb S, Kabarriti AE. How well do artificial intelligence chatbots respond to the top search queries about urological malignancies? *Eur Urol* 2024;85(1): 13-16.
5. 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.
6. Li Y, Gao W, Luan Z, Zhou Z, Li J. The impact of chat generative pre-trained transformer (ChatGPT) on oncology: application, expectations, and future prospects. *Cureus* 2023;15(11):e48670.
7. Waisberg E, Ong J, Masalkhi M et al. GPT-4 and ophthalmology operative notes. *Ann Biomed Eng* 2023;51(11):2353-2355.
8. Eppler MB, Ganjavi C, Knudsen JE 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(5):436-443.
9. 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.
10. Zhu L, Mou W, Chen R. Can the ChatGPT and other large language models with internet-connected database solve the questions and concerns of patient with prostate cancer and help democratize medical knowledge? *J Transl Med* 2023;21(1):269.
11. 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(4):409-415.
12. 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.
13. 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.
14. Shah YB, Ghosh A, Hochberg AR et al. Comparison of ChatGPT and traditional patient education materials for men's health. *Urol Pract* 2024;11(1):87-94.
15. 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* 2023;15(7):e42214.
16. Charnock D, Shepperd S, Needham G, Gann R. DISCERN: an instrument for judging the quality of written consumer health information on treatment choices. *J Epidemiol Community Health* 1999;53(2):105-111.
17. 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(3): 395-403.
18. Shah YB, Glatter R, Madad S. In layman's terms: the power and problem of science communication. *Disaster Med Public Health Prep* 2022:1-3.
19. Wang A, Qian Z, Briggs L, Cole AP, Reis LO, Trinh QD. The use of Chatbots in oncological care: a narrative review. *Int J Gen Med* 2023;16:1591-1602.

20. Gortz M, Baumgartner K, Schmid T et al. An artificial intelligence-based chatbot for prostate cancer education: Design and patient evaluation study. *Digit Health* 2023;9:20552076231173304.
21. Liu HY, Alessandri Bonetti M, De Lorenzi F, Gimbel ML, Nguyen VT, Egro FM. Consulting the digital doctor: Google versus ChatGPT as sources of information on breast implant-associated anaplastic large cell lymphoma and breast implant illness. *Aesthetic Plast Surg* 2024;48(4):590-607.
22. Wei K, Fritz C, Rajasekaran K. Answering head and neck cancer questions: an assessment of ChatGPT responses. *Am J Otolaryngol* 2024;45(1):104085.
23. Ayers JW, Poliak A, Dredze M et al. Comparing physician and artificial intelligence Chatbot responses to patient questions posted to a public social media forum. *JAMA Intern Med* 2023;183(6):589-596.
24. Tai-Seale M, Baxter SL, Vaida F et al. AI-generated draft replies integrated into health records and physicians' electronic communication. *JAMA Netw Open* 2024;7(4):e246565.
25. Teasdale A, Mills L, Costello R. Artificial intelligence-powered surgical consent: patient insights. *Cureus* 2024;16(8):e68134.