DOI: 10.1002/bco2.359

ORIGINAL ARTICLE

BJUI COMPASS

Urology consultants versus large language models: Potentials and hazards for medical advice in urology

Johanna Eckrich[1] | Jörg Ellinger[1] | Alexander Cox[1] | Johannes Stein[1] | Manuel Ritter[1] | Andrew Blaikie[2] | Sebastian Kuhn[3] | Christoph Raphael Buhr[2,4]

[1]Department of Urology, University Hospital Bonn, Bonn, Germany

[2]School of Medicine, University of St Andrews, St Andrews, UK

[3]Institute of Digital Medicine Philipps-University Marburg and University Hospital of Giessen and Marburg, Marburg, Germany

[4]Department of Otorhinolaryngology, University Medical Center of the Johannes Gutenberg-University Mainz, Mainz, Germany

Correspondence
Christoph Raphael Buhr, Department of Otorhinolaryngology, University Medical Center of the Johannes Gutenberg-University Mainz, Langenbeckstraße 1, 55131 Mainz, Rhineland-Palatinate, Germany.
Email: buhrchri@uni-mainz.de

Abstract

Background: Current interest surrounding large language models (LLMs) will lead to an increase in their use for medical advice. Although LLMs offer huge potential, they also pose potential misinformation hazards.

Objective: This study evaluates three LLMs answering urology-themed clinical case-based questions by comparing the quality of answers to those provided by urology consultants.

Methods: Forty-five case-based questions were answered by consultants and LLMs (ChatGPT 3.5, ChatGPT 4, Bard). Answers were blindly rated using a six-step Likert scale by four consultants in the categories: 'medical adequacy', 'conciseness', 'coherence' and 'comprehensibility'. Possible misinformation hazards were identified; a modified Turing test was included, and the character count was matched.

Results: Higher ratings in every category were recorded for the consultants. LLMs' overall performance in language-focused categories (coherence and comprehensibility) was relatively high. Medical adequacy was significantly poorer compared with the consultants. Possible misinformation hazards were identified in 2.8% to 18.9% of answers generated by LLMs compared with <1% of consultant's answers. Poorer conciseness rates and a higher character count were provided by LLMs. Among individual LLMs, ChatGPT 4 performed best in medical accuracy ($p < 0.0001$) and coherence ($p = 0.001$), whereas Bard received the lowest scores. Generated responses were accurately associated with their source with 98% accuracy in LLMs and 99% with consultants.

Conclusions: The quality of consultant answers was superior to LLMs in all categories. High semantic scores for LLM answers were found; however, the lack of medical accuracy led to potential misinformation hazards from LLM 'consultations'. Further investigations are necessary for new generations.

KEYWORDS
artificial intelligence (AI), Bard, chatbots, ChatGPT, digital health, global health, large language models (LLMs), low- and middle-income countries (LMICs), telehealth, telemedicine, urology

[Correction added on 21 December 2024, after first online publication: The pagination of this article has been revised in this version.]

1 | INTRODUCTION

'I searched my symptoms on the internet…' is a frequently made statement by patients consulting medical advice in hospitals and private practice alike these days. This prevalent behaviour is a direct result of the ubiquitous availability of medical information from the internet. Although an informed patient is a desirable scenario, misinformation and wrongful preconceptions may hinder a trusting relationship with the attending healthcare professional.[1,2]

The current hype surrounding AI and large language models (LLMs) will also transform the process of medical self-information. Internet research formerly conducted through search engines like Google will transfer to open access services like Bard and ChatGPT (Chat-Generative Pre-trained Transformer), which are now being used for a broad range of tasks including medical information queries. Hence, LLM services will likely soon replace the traditional search engines as the primary source of medical information for lay people.[3–6]

LLMs are powered by a 'deep neural network architecture' and leverage principles of natural language processing. These models employ unsupervised learning, where they analyse extensive text data to learn linguistic patterns, grammar rules and contextual nuances. Deep neural networks, composed of multiple layers of interconnected nodes, process sequential data, such as sentences, by predicting the next word based on preceding words, developing a profound understanding of language structure. Hence, they excel in tasks like text generation and demonstrate remarkable linguistic abilities. LLMs are continuously trained on a vast amount of data including books, articles, websites and other textual content. The increasing amount of training data is just one factor contributing to continual model improvements. Architecture enhancements, fine-tuning techniques and improvements in training algorithms continue to progress and enhance the capabilities of these models.[7]

Beyond LLMs' vast database, their output of apparently human-like answers, as well as their multilingualism, makes them a seemingly competent contact point for medical consultation. However, using LLMs as a substitute for medical caregivers bears significant risks. The flawless semantic form may drive patients to avoid professional advice and lead to wrong diagnostic or therapeutic conclusions. These could result in severe misinterpretation of symptoms, hypochondria, increased anxiety and potentially harmful self-treatment or non-treatment.[1,8] The motivation behind internet research of medical symptoms is versatile and ranges from limited access to healthcare to the desire for reassurance or a second opinion. Furthermore, the search for deeper understanding and perceived external barriers to accessing information through traditional sources play a significant role. Additionally, factors like convenience, coverage and anonymity of medical internet research are relevant in this regard.[9]

Embarrassment, cultural ostracism and rejection of conventional norms emerge as pertinent determinants contributing to the avoidance of seeking counsel from medical practitioners, especially in medical domains such as urology. However, this delay leads to possibly adversarial postponement in the instigation of medical intervention.[10,11] This emphasizes the relevance of medical internet research, particularly in contexts encompassing potentially awkward or humiliating medical conditions.

Previous studies showcased by our working group showed the general ability of LLMs to answer medical case-based questions in the field of otorhinolaryngology correctly. Yet, the overall medical adequacy of the answers given was significantly inferior to those of specialists given the same questions.[3] As we believe in the high relevance of this topic, to further evaluate the capabilities and limitations of LLMs as providers of medical advice, we compared the answers given by three different LLMs for case-based medical questions to answers provided by specialists in the field of urology.

2 | METHODS

Urology study books, exemplary questions from urological journals and former exams were browsed for case-based questions. The selected questions were matched to clinical cases from the outpatient unit and the emergency centre, and corresponding questions were selected.[12,13]

After this process, 45 questions were selected resembling a broad range of urological pathologies. Subsequently, the questions were answered by four urology consultants (co-authors of this article) and three selected LLMs, respectively. LLMs ChatGPT 3.5 (free version of ChatGPT during trial), ChatGPT 4.0 (latest [paid] version ChatGPT) and Google Bard were utilized for this study because of their broad use and low barrier setup. On the other side, the consultants selected had at least 6 years of clinical experience in urology.

After the questions were answered, they were again randomized. A character count for every answer was determined and statistically compared.

After randomization, all answers given by the urology consultants and the LLMs respectively were rated by the homologue 4 consultants (their own questions excluded) using a 6-point Likert scale (1 = very poor, 6 = excellent). It must be noticed that answers provided by LLMs often include phrases that disclose a lack of qualification to answer medical questions or advise a medical consultation. To avoid possible bias and to allow a modified Turing test, these phrases were excluded before further evaluation.

Questions were rated for medical adequacy, conciseness, coherence and comprehensibility respectively in concordance with previous studies by our working group.[3]

Additionally, the hazardous potential of the answers provided was rated in a binary rating system (possible hazard: yes/no).

With the corresponding rating operandus, the consultants also assessed whether the answers were created by a urology consultant or by one of the LLMs respectively.

Gaussian distribution of ratings was evaluated after data acquisition utilizing the D'Agostino & Pearson test. After normality testing was performed, pairwise comparisons were realized with the non-parametric Mann–Whitney test or the Kruskal–Wallis test if more than two groups were compared. The statistical testing was

performed using GraphPad Prism, Version 10.0.3 (GraphPad Software, La Jolla, California, USA).

3 | RESULTS

As shown in Figure 1, evaluation of the character count showed a significantly reduced semantic output of the consultants compared with all three LLMs. Although the urology consultants utilized a Median of 359 characters per answer (Range 39–1135), ChatGPT 3.5 used 1926 (786–2762), ChatGPT 4 2600 (1572–3537) and Bard 1039 (256–2951) characters.

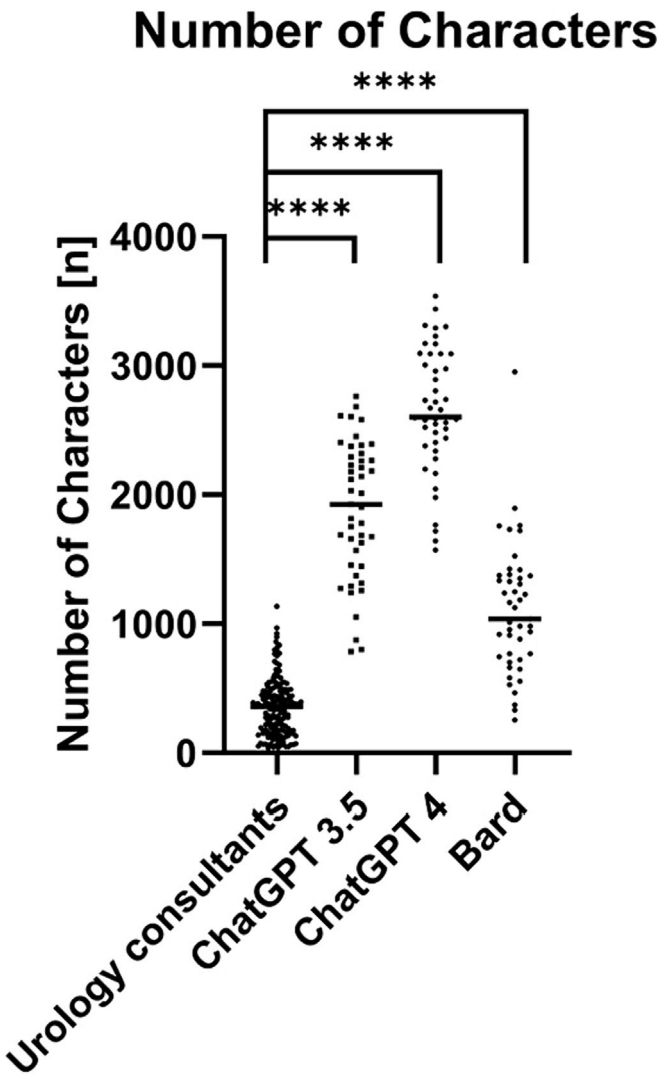

FIGURE 1 The number of characters per answer by urology consultants and large language models (LLMs; ChatGPT 3.5, Chat GT 4, Bard) for all evaluated categories. Data shown as a scatter dot blot with each point resembling an absolute value. Grey horizontal line = Median. The non-parametric Mann–Whitney test was used to compare the ratings for individual LLMs to the urology consultants (**** = $p < 0.0001$).

Even though the semantic quality of answers was rated comparatively good (Table 1), answers by the urology consultants were correctly assigned as the source of the origin for 99.01%, whereas the LLM was identified correctly in 98.00% of cases.

As shown in Figure 2, answers provided by urology consultants were rated superior to answers provided by the LLMs in every category.

A more detailed depiction of individual ratings and proportions is provided in Table 1.

Particularly, the ratings for medical adequacy and the conciseness of answers provided showed a relatively high qualitative discrepancy between urology consultants and the LLMs ($p < 0.0001$). Although differences in the rated categories still reached statistical significance, LLMs' performance was noticeably better in semantic evaluation criteria (coherence [$p = 0.0052$–$p < 0.0001$] and comprehensibility [$p = 0.023$–$p < 0.0001$]).

We noticed significant differences between the individual LLMs regarding the ratings for medical accuracy ($p < 0.0001$) as well as coherence ($p = 0.001$). In both categories, ChatGPT 4 was rated the most proficient, whereas Bard was rated with the lowest scores in both categories. Ratings for conciseness and comprehensibility, however, did not show any significant differences between all three LLMs.

To assess whether the answers provided could be the source of possible hazard, a binary rating system was included. Of the answers provided by urology consultants, 0.56% were rated as possibly hazardous for the patient, whereas answers provided by LLMs were rated as possibly hazardous in 2.78% for ChatGPT 4, 8.33% for ChatGPT 3.5 and 18.89% for answers provided by Bard. These findings are consistent with the distribution of ratings for medical adequacy for the individual LLMs.

Urology consultants were able to determine the source of the answers correctly in 99.01% for urology consultants and 98.00% for LLMs. For sample questions and answers, see Data S1.

4 | DISCUSSION

The potential of LLMs is a heavily discussed topic in today's society—and for a good reason! LLMs now offer the possibility of access to medical information in a convenient and understandable way. They are therefore very likely to be used by patients as a source of medical information. Hence, evaluation of their potential as well as their limitations is important. The quality of output as well as the accuracy of responses must ultimately be critically re-evaluated especially in the field of medical care.[14–16]

Our study therefore evaluated the performance of three commonly used LLMs for answering case-based questions in the field of urology. As expected, the LLMs' responses were of high semantic quality as underlined by the high-ranked overall comprehensibility and coherence (Table 1 and Figure 2). These findings support data recently published by Cocci et al. attesting a college graduate reading level for answers provided by ChatGPT as well as previously published findings

TABLE 1 Cumulative ratings for all categories evaluated by the urology consultants.

	Ratings (*n*)	Rating (Median; [Range])	Rating (Mean)	Rating [95% CI]	*p*-value[a]	*p*-value[b]
Medical adequacy						
Urology consultants	540	6 (1–6)	5.687	[6;6]		
ChatGPT 3.5	180	5 (1–6)	4.661	[5;5]	<0.0001	<0.0001
ChatGPT 4	180	5 (2–6)	5.200	[5;6]	<0.0001	
Bard	180	5 (1–6)	4.244	[4;5]	<0.0001	
Conciseness						
Urology consultants	540	6 (3–6)	5.893	[6;6]		
ChatGPT 3.5	180	5 (1–6)	4.450	[4;5]	<0.0001	n.s.
ChatGPT 4	180	5 (2–6)	4.444	[5;5]	<0.0001	
Bard	180	5 (1–6)	4.350	[4;5]	<0.0001	
Coherence						
Urology consultants	540	6 (4–6)	5.765	[6;6]		
ChatGPT 3.5	180	5 (4–6)	5.500	[6;6]	<0.0001	=0.001
ChatGPT 4	180	5 (2–6)	5.611	[6;6]	=0.0052	
Bard	180	5 (4–6)	5.289	[5;6]	<0.0001	
Comprehensibility						
Urology consultants	540	6 (4–6)	5.785	[6;6]		
ChatGPT 3.5	180	6 (4–6)	5.561	[6;6]	<0.0001	n.s.
ChatGPT 4	180	6 (4–6)	5.678	[6;6]	=0.0234	
Bard	180	6 (3–6)	5.561	[6;6]	<0.0001	

Note: Comparative statistics between the specific large language model (LLM) and the consultants were carried out using the Mann–Whitney test.
A comparative statistic evaluation between the three LLMs was carried out using the Kruskal–Wallis test.
[a]Compared to ratings of the urology consultants with the Mann–Whitney test.
[b]Comparison between the three different LLMs with the Kruskal–Wallis test.

by our working group in the field of otorhinolaryngology.[3,17] Nevertheless, even in semantic categories, the LLMs were still outperformed by the urology consultants as illustrated by their significantly higher ratings in the corresponding categories. Contrary to the relatively high comprehensibility and coherence of the answers provided, the LLMs showed a relative discrepancy regarding the conciseness of their answers compared with the urology consultants (Table 1/Figure 2). Taking these findings into consideration, it is important to acknowledge that the answers generated by LLMs were between 4 (Bard) to 9 times longer (ChatGPT 4) than the corresponding answers by the urology consultants (Figure 1). The lavish vocabulary, in contrast to the reduced conciseness, likely originates from the way that LLM chatbots are trained mimicking a 'human-like manner' by using a rather complex speech pattern instead of stenographic language. According to the OpenAI website, this phenomenon traces back to the feedback of the testers, who preferred 'longer answers that look more comprehensive' [openai.com].[18]

Even though the LLMs achieved a high semantic quality in our rating, the urology consultants were able to determine the source of the answers correctly in 99.01% for urology consultants and 98.00% for LLMs, respectively. These findings may contradict the excellent performance of today's LLMs in the modified Turing test but could be heavily biased by the fact that an expert is rating answers in the field of personal expertise as well as by the repetitive semantic structure and the significantly longer answers.[19]

The medical adequacy of the answers provided is ultimately by far the most relevant criterion of evaluation. In this category, all LLMs were highly significantly outperformed by the urology consultants (Table 1/Figure 2). Although a Median adequacy of 5.0 for all LLMs still deserves credit for an entity not specialized in medical care, the poor performance is highlighted by the percentage of possible hazards. The latter ranges from 2.78% for misinformation responses for ChatGPT 4 and 8.33% for ChatGPT 3.5 up to 18.89% for answers provided by Bard.

Medical adequacy in this current study was however still rated higher than corresponding ratings in studies performed by other working groups.[17,20,21] This difference may be accounted by the constant performance improvements of LLMs although the risk of misinformation still remains.

However, the potential of LLMs should not be ignored. In other specialties, LLMs have shown their potential to even outperform medical personnel as demonstrated by a recent study by Ayers et al.[22] The authors evaluated ChatGPTs' potential in answering patient questions in comparison with a licensed physician. To assess the quality, answers to questions posted to a public forum were answered by a physician and chatbot alike and subsequently evaluated by

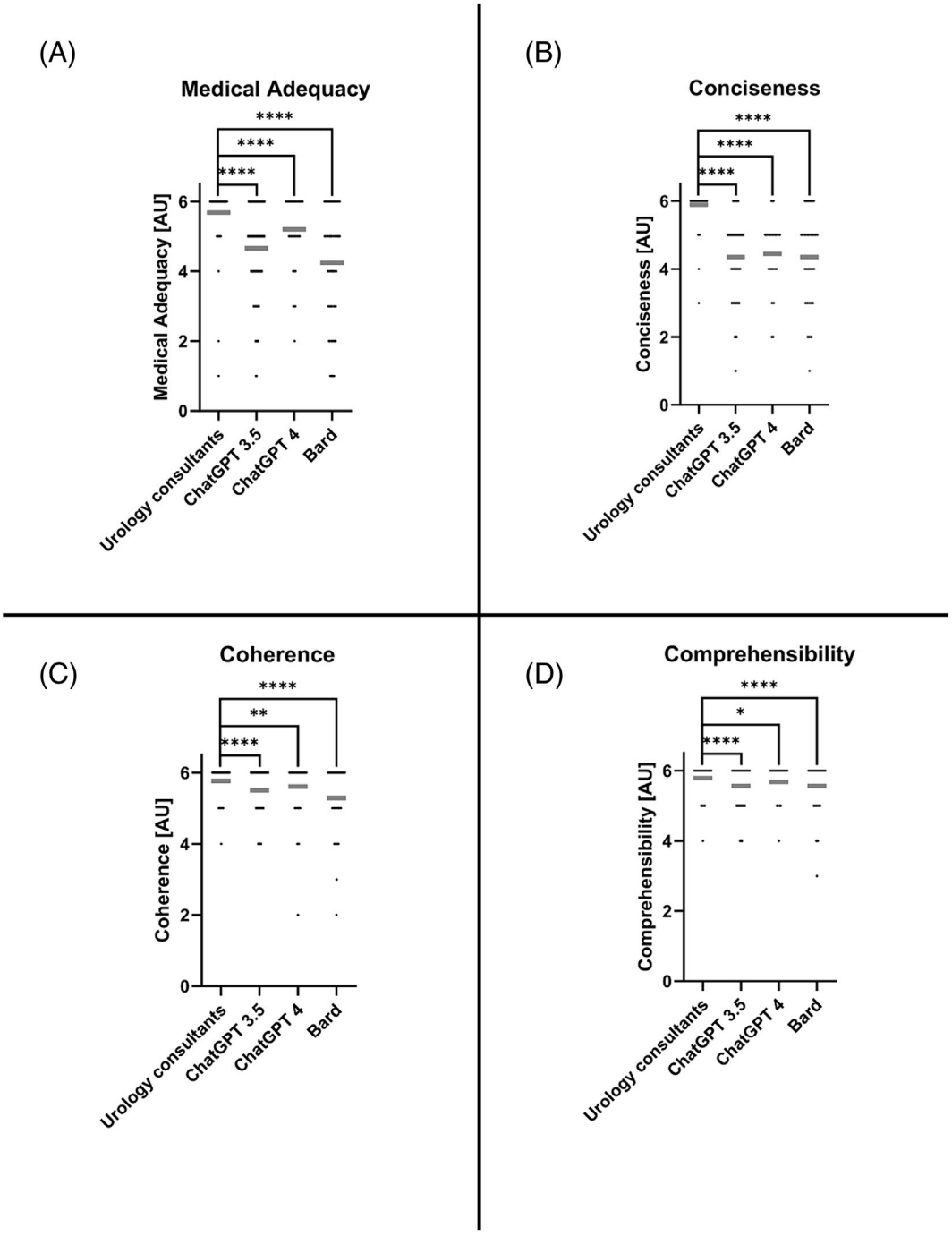

FIGURE 2 Comparison between urology consultants and large language model (LLMs; ChatGPT 3.5, Chat GPT 4, Bard) for all evaluated categories. Data shown as a scatter dot blot with each point resembling an absolute value. Grey horizontal line = Median. The non-parametric Mann–Whitney test was used to compare the ratings for individual LLMs to the urology consultants (**** = $p < 0.0001$; ** = $p < 0.01$; * = $p < 0.05$). Cumulative results of ratings for medical adequacy (A), conciseness (B), coherence (C) and comprehensibility (D).

healthcare professionals. Surprisingly, the trained personnel preferred chatbot responses to physician responses in 78.6% of the evaluations, even rating categories like empathy in favour of the LLM.

In our data, comparative analysis ratings for medical accuracy differed between the LLMs with ChatGPT4 being the most proficient of the three. Based on the limited sampling in our study, predictions on the evolution of medical accuracy of LLMs are hard to make. Yet, the significant increase in rating between ChatGPT 3.5 and ChatGPT 4 ($p < 0.0001$) may suggest improvement in that respect thereby contradicting recent findings by Zhu et al. and supporting findings by our own working group in the field of otorhinolaryngology.[3,23]

As LLMs provide an accessible and well-structured source of information, there are a variety of different use cases for LLMs in medical practice, ranging from providing additional information before

consulting a doctor to low-resource settings where a medical consultation in person is not available. Especially in urology, patients may find themselves in embarrassing situations, which is why they may want to avoid personal contact with the doctor. Other scenarios might occur after a diagnosis has already been made: the patient wants to gather more detailed information about their illness. In this manner, the LLM consultation can augment consultations with doctors and lead to empowerment of the patient for shared decision-making.[24]

Furthermore, LLMs have the potential to complement health care leading to more cost-efficient and timely delivery. Possible areas of applications include classification, organization and summarization of complex patient data, surveillance of complex medical co-founders or management of the increasing bureaucracy in the healthcare system.[25]

Apart from the influence on the doctor-patient relationship and economic effects, LLMs can improve global health issues in low- and middle-income countries, especially in areas with limited and untrained staff. As smartphones and internet access are often available in these settings, LLMs may provide useful access to medical advice for immediate management and triage.

However, before the actual impact of LLMs in medicine on a wider scale can be implemented, there are still concerns to manage. LLMs specially trained for medical purposes, such as Med-PaLM, will further improve the response to medical queries.[26] Moreover, LLMs with real-time access to the internet searching for up-to-date information and studies will take LLMs to the next level. Last but not least, special prompts will also optimize answers on medical questions. Here, our work can be helpful, as it reveals the inadequacies of the answers from a physician's perspective. Future work should analyse the needs and expectations of patients in more detail. Based on this information, further studies should build and evaluate LLMs with medical prompts on a larger scale in the future.

Whereas the lack of medical adequacy will likely improve when LLMs are specially trained for medical purposes, the regulation of LLMs handling highly sensitive patient and medical data might be more challenging and require strictly regulated and transparent standards.[27–29] Potential risks for patients' privacy are highlighted by current legislative initiatives such as the EU Artificial Intelligence Act.[30]

Currently, there are two main approaches for dealing with potential privacy risks. First, it is the users' responsibility to consider carefully which data they are passing to the LLMs. Therefore, data should only be entered pseudonymized; moreover, using vpn clients can help to make it more difficult to assign the data to specific patients. Second, commercial providers such as Aleph Alpha already recognized the need for privacy protection offering an AI infrastructure where the rights on personal data remain entirely with the user.[31] Unfortunately, these services have so far been exclusively reserved for commercial customers and are therefore only accessible for clinics, healthcare companies and authorities.

Obviously, our study has some limitations as only 45 case-based questions were used as input instead of patients passing their symptoms themselves to the LLM. However, the provided rating by clinically experienced doctors represents the gold standard of medical care as a benchmark. Further studies should include real patients and proof of the performance of LLMs in urology on a larger scale.

Although our data accentuate the potential of LLMs regarding linguistic performance, the limited medical adequacy and the higher risk of misinformation hazard emphasize the jeopardy associated with an unsupervised use of LLMs as a source of medical information. Hence, we sincerely believe that LLMs should be considered as an augmentative tool for providing as well as seeking healthcare and not an autarchic entity.

AUTHOR CONTRIBUTIONS

J Eckrich and CR Buhr conceived of the presented idea. J Eckrich browsed the textbooks and looked into the case based questions, which were then filtered and answered by J Ellinger, A Cox, J Eckrich and J Stein. CR Buhr entered all case based questions into the three LLMs, anonymized all answers. J Ellinger, A Cox, J Eckrich and J Stein rated the answers each by the other consultants and the LLMs. J Eckrich and CR Buhr then analyzed all answers and did the statistical evaluation. All authors discussed the results and contributed to the final manuscript.

ACKNOWLEDGEMENTS

Open Access funding enabled and organized by Projekt DEAL.

CONFLICT OF INTEREST STATEMENT

No third-party funding was utilized for the design of the study and collection, analysis and interpretation of data and in writing the manuscript. The authors declare no competing interests.

ETHICS STATEMENT

Written correspondence of March 3rd 2023 with the ethics committee of the regional medical association Rhineland-Palatinate determined no need for any specific ethical approval due to the use of anonymous text-based questions.

ORCID

Jörg Ellinger https://orcid.org/0000-0002-7526-0857
Alexander Cox https://orcid.org/0000-0001-7837-5740
Christoph Raphael Buhr https://orcid.org/0000-0002-9551-2310

REFERENCES

1. Jungmann SM, Brand S, Kolb J, Witthöft M. Do Dr. Google and health apps have (comparable) side effects? An experimental study. Clin Psychol Sci. 2020;8(2):306–17. https://doi.org/10.1177/2167702619894904
2. Cocco AM, Zordan R, Taylor DM, Weiland TJ, Dilley SJ, Kant J, et al. Dr Google in the ED: searching for online health information by adult emergency department patients. Med J Austr. 2018;209(8):342–7. https://doi.org/10.5694/mja17.00889
3. Buhr CR, Smith H, Huppertz T, Bahr-Hamm K, Matthias C, Blaikie A, et al. ChatGPT vs. consultants: a pilot study on answering otorhinolaryngology case-based questions. JMIR Med Educ (forthcoming). 2023;9:e49183. https://doi.org/10.2196/49183

4. Bommasani R, Hudson DA, Adeli E, Altman R, Arora S, von Arx S et al. On the opportunities and risks of foundation models. arXiv preprint arXiv:210807258. 2021.
5. Grant N, Metz C. A New Chat Bot Is a 'Code Red' for Google's Search Business. The New York Times, 2023.
6. Ayoub NF, Lee YJ, Grimm D, Divi V. Head-to-head comparison of ChatGPT versus Google search for medical knowledge acquisition. Otolaryngol–Head Neck Surgery. 2023. https://doi.org/10.1002/ohn.465
7. Radford A, Wu J, Child R, Luan D, Amodei D, Sutskever I. Language models are unsupervised multitask learners. OpenAI Blog. 2019;1:9.
8. McMullan RD, Berle D, Arnáez S, Starcevic V. The relationships between health anxiety, online health information seeking, and cyberchondria: systematic review and meta-analysis. J Affect Disord. 2019;245:270–8. https://doi.org/10.1016/j.jad.2018.11.037
9. Powell J, Inglis N, Ronnie J, Large S. The characteristics and motivations of online health information seekers: cross-sectional survey and qualitative interview study. J Med Internet Res. 2011;13(1):e20. https://doi.org/10.2196/jmir.1600
10. Skeppner E, Andersson SO, Johansson JE, Windahl T. Initial symptoms and delay in patients with penile carcinoma. Scand J Urol Nephrol. 2012;46(5):319–25. https://doi.org/10.3109/00365599.2012.677473
11. Ian Janes WC, Henley J, Andrews M, Organ M, Johnston P. A quality assurance review of penile cancer diagnostic delays and stage at presentation during the COVID-19 pandemic. Can Urol Assoc J. 2023;17(5):E134–40. https://doi.org/10.5489/cuaj.8143
12. Schmelz H-U, Leyh H. Facharztprüfung Urologie: 1000 kommentierte Prüfungsfragen. Georg Thieme Verlag; 2014.
13. Jung C, Tauber R. Urologie in Frage und Antwort. Elsevier, Urban&-FischerVerlag"; 2014.
14. 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):409–15. https://doi.org/10.1097/UPJ.0000000000000406
15. Sbaffi L, Rowley J. Trust and credibility in web-based health information: a review and agenda for future research. J Med Internet Res. 2017;19(6):e218. https://doi.org/10.2196/jmir.7579
16. Gilbert S, Harvey H, Melvin T, Vollebregt E, Wicks P. Large language model AI chatbots require approval as medical devices. Nat Med. 2023;29(10):2396–8. https://doi.org/10.1038/s41591-023-02412-6
17. 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;27(1):103–8. https://doi.org/10.1038/s41391-023-00705-y
18. OpenAI ChatGPT. OpenAI, 2021.
19. Turing AM. I.—Computing machinery and intelligence. Mind. 1950;LIX(236):433–60. https://doi.org/10.1093/mind/LIX.236.433
20. 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;180:278–84. https://doi.org/10.1016/j.urology.2023.07.010
21. Szczesniewski JJ, Tellez Fouz C, Ramos Alba A, Diaz Goizueta FJ, García Tello A, Llanes González L. ChatGPT and most frequent urological diseases: analysing the quality of information and potential risks for patients. World J Urol. 2023;41(11):3149–53. https://doi.org/10.1007/s00345-023-04563-0
22. Ayers JW, Poliak A, Dredze M, Leas EC, Zhu Z, Kelley JB, 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):589–96. https://doi.org/10.1001/jamainternmed.2023.1838
23. 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. https://doi.org/10.1186/s12967-023-04123-5
24. Daraz L, Morrow AS, Ponce OJ, Beuschel B, Farah MH, Katabi A, et al. Can patients trust online health information? A meta-narrative systematic review addressing the quality of health information on the internet. J Gen Intern Med. 2019;34(9):1884–91. https://doi.org/10.1007/s11606-019-05109-0
25. Thirunavukarasu AJ, Ting DSJ, Elangovan K, Gutierrez L, Tan TF, Ting DSW. Large language models in medicine. Nat Med. 2023;29(8):1930–40. https://doi.org/10.1038/s41591-023-02448-8
26. Tu T, Azizi S, Driess D, Schaekermann M, Amin M, Chang PC, et al. Towards generalist biomedical AI. N Engl J Med. 2024;1(3):1. https://doi.org/10.1056/AIoa2300138
27. Meskó B, Topol EJ. The imperative for regulatory oversight of large language models (or generative AI) in healthcare. npj Dig Med. 2023;6(1):120. https://doi.org/10.1038/s41746-023-00873-0
28. Minssen T, Vayena E, Cohen IG. The challenges for regulating medical use of ChatGPT and other Large language models. Jama. 2023;330(4):315–6. https://doi.org/10.1001/jama.2023.9651
29. Yaeger KA, Martini M, Yaniv G, Oermann EK, Costa AB. United States regulatory approval of medical devices and software applications enhanced by artificial intelligence. Health Policy Technol. 2019;8(2):192–7. https://doi.org/10.1016/j.hlpt.2019.05.006
30. European Commission. A European approach to artificial intelligence. https://digital-strategy.ec.europa.eu/en/policies/european-approach-artificial-intelligence
31. Aleph Alpha 2021 https://aleph-alpha.com

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Eckrich J, Ellinger J, Cox A, Stein J, Ritter M, Blaikie A, et al. Urology consultants versus large language models: Potentials and hazards for medical advice in urology. BJUI Compass. 2024;5(5):552–8. https://doi.org/10.1002/bco2.359