JOURNAL OF CLINICAL TRIALS AND EXPERIMENTAL INVESTIGATIONS JCTEI

DOI: https://doi.org/10.5281/zenodo.6686078

ORIGINAL ARTICLE

Evaluating the accuracy of ChatGPT addressing urological questions: A pilot study

Suleyman Sagir

Islahiye State Hospital, Urology Clinic, Gaziantep Turkey.

Abstract

Objective: This research aimed to assess the accuracy of the ChatGPT 3.5 model in providing information related to various urological diseases.

Materials and methods: Eighty questions regarding urological diseases were presented to ChatGPT in December 2022. Responses were recorded and subsequently cross-referenced with the European Urology Association (EUA) guidelines to determine their correctness. Diseases were categorized into subgroups: Urolithiasis, Bladder cancer, Urethroplasty, Renal cancer, and Andrology. Accuracy percentages were calculated for each disease subgroup and the total dataset.

Results: For Urolithiasis, out of 25 responses, 10 (40%) were true and 15 (60%) were false. Bladder cancer had an even distribution, with 50% of the responses (10 out of 20) being true and the remaining 50% being false. Renal cancer showed a higher proportion of true responses, with 14 out of 22 responses (approximately 63.6%) being true and 8 (approximately 36.4%) being false. In the case of Urethroplasty, out of 25 responses, 13 (52%) were true while 12 (48%) were false.

Conclusions: ChatGPT showcased varying degrees of accuracy across different urological disease subgroups. While it demonstrates potential utility as a supportive tool for urological questions, the observed accuracy levels highlight the need for cautious interpretation.

Keywords: Artificial intelligence, ChatGPT, urology.

OPEN ACCESS

Correspondence:

Suleyman Sagir, MD
Suleyman Sagir, Islahiye State
Hospital, Gaziantep Turkey.

e-mail:
dr.sagiroglu414@gmail.com

Specialty section:
This article was submitted
to Urology section

Received: 15 December 2022
Revised: 27 December 2022
Accepted: 28 December 2022
Published: 30 December 2022

Published by The JCTEI.

Introduction

The Chat Generative Pre-trained Transformer (ChatGPT) exemplifies a pinnacle of natural language processing, fuelled by a rich foundation of expansive data and the robust capabilities of artificial intelligence (1). Since its initial public release in November 2022, ChatGPT has captivated widespread attention owing to its remarkable ability to formulate human-esque responses during textual conversations. The launch of ChatGPT was underpinned by GPT-3.5, a foundational large language model, providing a solid base for its sophisticated conversational aptitude (2).

The emergence and development of Artificial Intelligence (AI) across multiple disciplines have transformed our comprehension and application of knowledge (3,4). The incorporation of AI into medicine, especially within the sphere of pediatric surgery, provides a novel perspective through which to investigate, interpret, and address intricate surgical challenges. While the AI language models GPT-3.5, crafted by OpenAI, have showcased potential across various domains, including medicine, an exhaustive exploration into their ability to provide precise responses to specialized, field-related inquiries is still pending comprehensive scrutiny.

The advances in machine learning, particularly with AI models like ChatGPT, have heralded a new age in information technology and human-computer interaction (5). Its intricate algorithms, supported by countless layers of neural networks, epitomize the forefront of technological progress. This exceptional prowess in natural language processing has enabled ChatGPT to comprehend and generate contextually relevant information, even in fields as specialized as quantum physics or ancient history.

While AI has been infiltrating various sectors, its application in the medical arena is of particular interest. The confluence of AI and medical practice promises not only improvements in diagnostic precision but also the potential for tailoring treatment plans based on individual patient profiles (6). Here, the utility of GPT-3.5, and by extension ChatGPT, lies in its ability to analyze vast datasets, ranging from medical journals to patient histories, thus providing healthcare professionals with real-time insights.

However, with every technological leap, there arise legitimate concerns and challenges. The utilization of AI in fields that demand high degrees of accuracy, like pediatric surgery, poses inherent risks. Reliance on AI-generated insights, without adequate vetting, might result in suboptimal or, in the worst cases, harmful outcomes (7). Hence, while AI models like ChatGPT offer unprecedented capabilities, their integration into critical domains mandates rigorous testing, validation, and a clear understanding of their limitations. Nevertheless, the very existence of such sophisticated AI tools paves the way for revolutionary breakthroughs in numerous disciplines. As we continue to harness and refine the potential of ChatGPT and similar models, the horizon of possibilities only seems to expand, promising a future where the synergy of human intelligence and artificial assistance reaches unparalleled heights (8).

In this study, it was aimed to investigate the accuracy level of the information provided by ChatGPT 3.5 in various diseases of urology.

Materials and methods

This study was conducted in December 2022 using the OpenAI ChatGPT program via the web. Various questions concerning urological diseases were asked to ChatGPT. The answers received from ChatGPT were recorded. The accuracy percentages of the received answers were calculated.

In our study, a total of 80 questions were asked to ChatGPT. The obtained answers were checked using the European Urology Association (EUA) guidelines. Correct and incorrect answers were noted individually according to the diseases and analyses were performed. These diseases consisted of the following subgroups: Urolithiasis, Bladder cancer, Urethroplasty, Renal cancer, and Andrology.

Once all answers were evaluated and categorized, a statistical analysis was performed. Accuracy percentages were calculated for each disease subgroup and for the overall dataset. The intent was to discern patterns or particular strengths and weaknesses in ChatGPT's responses.

Results

In our study assessing various diseases, the following results were obtained based on the given responses: For Urolithiasis, out of 25 responses, 10 (40%) were true and 15 (60%) were false. Bladder cancer had an even distribution, with 50% of the responses (10 out of 20) being true and the remaining 50% being false. Renal cancer showed a higher proportion of true responses, with 14 out of 22 responses (approximately 63.6%) being true and 8 (approximately 36.4%) being false.

In the case of Urethroplasty, out of 25 responses, 13 (52%) were true while 12 (48%) were false. Andrology had 11 out of 20 responses (55%) being true and 9 (45%) being false. In total, out of 112 responses across all diseases, 58 (approximately 51.8%) were true and 54 (approximately 48.2%) were false (Table 1 and Figure 1).

Table 1: Questions and answers

Questions	Answers		True	False	
Diseases	**n**	**n**	**%**	**n**	**%**
Urolithiasis	25	10	40	15	60
Bladder cancer	20	10	50	10	50
Renal cancer	22	14	63.6	8	36
Urethroplasty	25	13	52	12	48
Andrology	20	11	55	9	45
Total	112	58	51.8	54	48

Discussion

To the best of our knowledge, this study is the first study on urological issues using ChatGPT. In our research, ChatGPT was posed with 80 questions. The responses were then verified against the European Urology Association (EUA) standards. Each answer, right or wrong, was cataloged based on the specific disease and subsequent evaluations were made. The diseases were categorized into these subgroups: Urolithiasis, Bladder cancer, Urethroplasty, Renal cancer, and Andrology.

The rapid advancement in artificial intelligence, particularly in natural language processing, has led to the development of tools like OpenAI's ChatGPT. Our investigation aimed to evaluate the accuracy of this tool in answering questions related to urological diseases, a sector that has seen limited AI application to date (1,3,8).

Our findings highlight that ChatGPT's performance in answering urological questions varied across different disease subgroups. For diseases like Renal cancer, the model displayed a reasonably high accuracy rate of approximately 63.6%. In contrast, its accuracy for Urolithiasis was considerably lower at 40%.

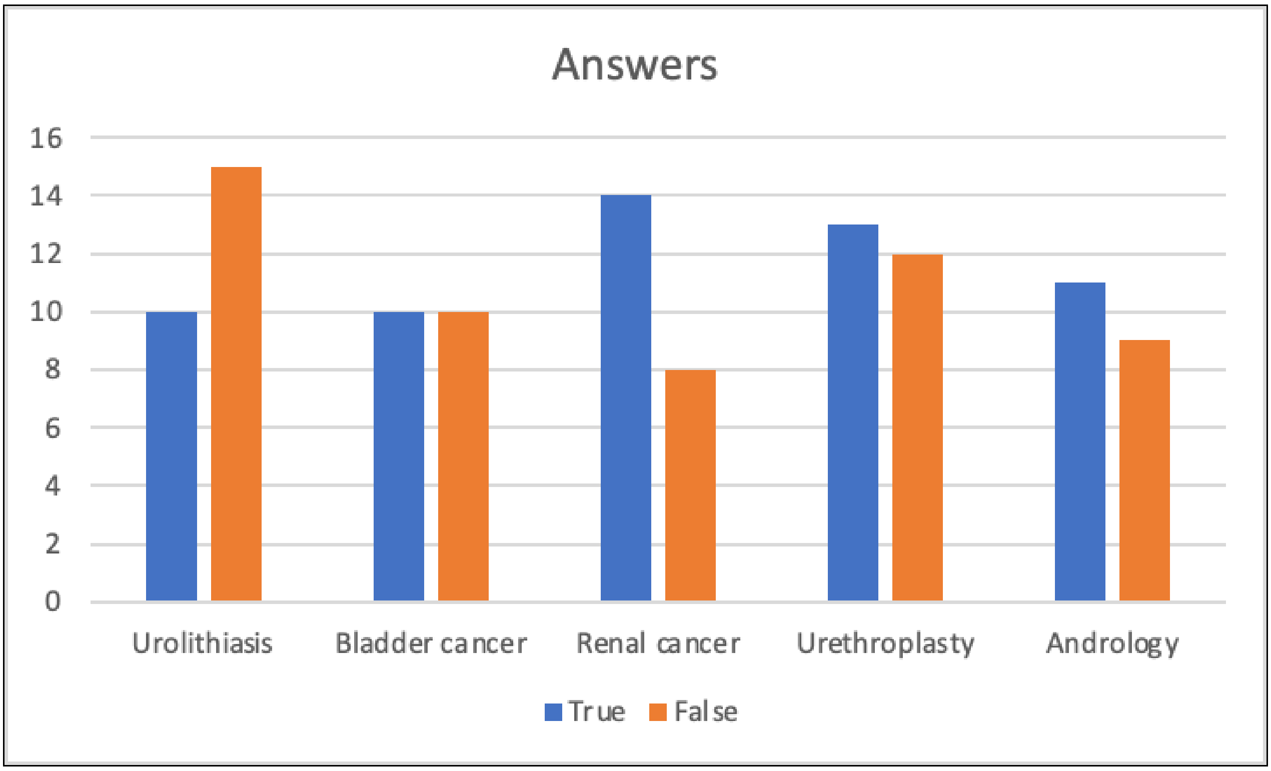

Figure 1: Questions and answer of the ChatGPT software

In comparison to literature, the utilization of AI in the medical field, especially in specialized domains like urology, remains in its nascent stages (7-9). A few studies have assessed the potential of AI in medical diagnosis and decision-making, but to our knowledge, ours is among the first to gauge the efficacy of ChatGPT in the realm of urology.

The even distribution of correct and incorrect answers for Bladder cancer, with an exact 50% accuracy, is intriguing. It raises questions about the model's training data for this particular disease or potential ambiguities in the EUA guidelines concerning it.

Our study has some limitations. While we utilized the EUA guidelines as a reference, it's important to note that medical knowledge is continuously evolving. New findings and advancements might not be immediately reflected in the guidelines or the AI model's training data. Moreover, the framing and specificity of the questions posed to ChatGPT can influence the accuracy of its responses.

Conclusions

While ChatGPT exhibits promise as a supplementary tool for urological queries, its current accuracy levels necessitate careful interpretation of its responses. Future iterations of the model, updated with the latest medical data, might prove to be more reliable. Until then, relying solely on AI for medical decisions, without human oversight, remains premature.

Conflict of interest

The authors report no conflict of interest.

Funding source

No funding was required.

Ethical approval

This is a computer-based study. No need for ethical approval

Acknowledgments

No

Peer-review

Externally. Evaluated by independent reviewers working in at least two different institutions appointed by the field editor.

Data availability

The datasets generated during and/or analyzed during the current study are available from the corresponding author on reasonable request.

Contributions

Research concept and design: SS

Data analysis and interpretation: SS

Collection and/or assembly of data: SS

Writing the article: SS

Critical revision of the article: SS

Final approval of the article: SS

References

1. Stokel-Walker C. AI bot ChatGPT writes smart essays - should professors worry? Nature. 2022 Dec 9. doi: 10.1038/d41586-022-04397-7. Epub ahead of print. PMID: 36494443.
2. OpenAI. Introducing ChatGPT. Accessed from: https://openai.com/ blog/chatgpt, Accessed Dec 2, 2022.
3. Graham F. Daily briefing: Will ChatGPT kill the essay assignment? Nature. 2022 Dec 12. doi: 10.1038/d41586-022-04437-2. Epub ahead of print. PMID: 36517680.
4. Yang DB, Smith AD, Smith EJ, Naik A, Janbahan M, Thompson CM, et al. The State of Machine Learning in Outcomes Prediction of Transsphenoidal Surgery: A Systematic Review. J Neurol Surg B Skull Base. 2022;84(6):548-59.
5. Weng SF, Reps J, Kai J, Garibaldi JM, Qureshi N. Can machine-learning improve cardiovascular risk prediction using routine clinical data? PLoS One. 2017;12(4):e0174944.
6. Mashraqi AM, Allehyani B. Current trends on the application of artificial intelligence in medical sciences. Bioinformation. 2022;18(11):1050-61.
7. Gallagher AG, Ritter EM, Champion H, Higgins G, Fried MP, Moses G, et al. Virtual reality simulation for the operating room: proficiency-based training as a paradigm shift in surgical skills training. Ann Surg. 2005;241(2):364-72.
8. Jin Q, Dhingra B, Liu Z, Cohen WW, Lu X. PubMedQA: a dataset for biomedical research question answering. arXiv doi: 10.48550/ arXiv.1909.06146. Preprint posted online on September 13, 2019
9. Ha LA, Yaneva V. Automatic question answering for medical MCQs: can it go further than information retrieval? Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019); RANLP 2019; September 2-4, 2019; Varna, Bulgaria. 2019. pp. 418-2.
10. Jin D, Pan E, Oufattole N, Weng WH, Fang H, Szolovits P. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. Applied Sciences. 2021;11(14):6421.
11. Kuleshov V, Ding J, Vo C, Hancock B, Ratner A, Li Y, et al. A machine-compiled database of genome-wide association studies. Nat Commun. 2019;10(1):3341.