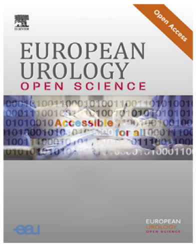
available at www.sciencedirect.com
journal homepage: www.eu-openscience.europeanurology.com



European Association of Urology

Brief Correspondence

Artificial Intelligence: Ready To Pass the European Board Examinations in Urology?

Benoît Mesnard[a], *Aurélie Schirmann*[a], *Julien Branchereau*[b], *Ophélie Perrot*[a], *Guy Bogaert*[c], *Yann Neuzillet*[a], *Thierry Lebret*[a], *François-Xavier Madec*[a,*]

Article info

Article history:
Accepted January 12, 2024

Associate Editor:
Silvia Proietti

Keywords:
Artificial intelligence
Urology
Urology degree
Clinical reasoning

Abstract

The role of artificial intelligence (AI) in the medical domain is increasing on an annual basis. AI allows instant access to the latest scientific data in urological surgery, facilitating a level of theoretical knowledge that previously required several years of practice and training. To evaluate the capability of AI to provide robust data in a specialized domain, we submitted the in-service assessment of the European Board of Urology to three different AI tools: ChatGPT 3.5, ChatGPT 4.0, and Bard. The assessment consists of 100 single-answer questions with four multiple-choice options. We compared the responses of 736 participants to the AI responses. The average score for the 736 participants was 67.20. ChatGPT 3.5 scored 59 points, ranking in 570th place. ChatGPT 4.0 scored 80 points, ranking 80th, just on the border of the top 10%. Google Bard scored 68 points, ranking 340th. Our study demonstrates that AI systems have the capability to participate in a urological examination and achieve satisfactory results. However, a critical perspective must be maintained, as current AI systems are not infallible. Finally, the role of AI in the acquisition of knowledge and the dissemination of information remains to be delineated.

Patient summary: We submitted questions from the European Diploma in Urological Surgery to three artificial intelligence (AI) systems. Our findings reveal that AI tools show remarkable performance in assessments of urological surgical knowledge. However, certain limitations were also observed.

The role of artificial intelligence (AI) in the medical domain is increasing on an annual basis. The most advanced AI applications pertain to medical screening and diagnostics, with current relevance in radiotherapy [1], radiology [2], histopathology, and ophthalmology [3]. AI is progressively being assigned a more significant role in surgical management, particularly in urology. Its potential applications include diagnosis, outcome prediction, treatment planning, and assessment of operative techniques, primarily in oncourology but also in other urological specialties [4,5]. AI can also play a role in urological learning and education. AI can review lots of data and make us think differently about how we teach new doctors and update more experienced doctors. AI provides quick access to the latest

https://doi.org/10.1016/j.euros.2024.01.002

research in urological surgery, making it easier to find out about things that used to take years to learn. AI tools could really change the way in which we learn.

Fellow of the European Board of Urology (FEBU) is the title awarded after a candidate has passed the EBU examinations and confirms a high level of knowledge in urology. The title can only be obtained after several years of both theoretical and practical training. Before applying for the FEBU designation, each candidate has the opportunity to prepare and assess their knowledge via an in-service assessment. Final-year residents and certified urologists in an EBU country are eligible as FEBU candidates. In-service assessment is available on the EBU website (https://www.ebu.com/examination/in-service-assessment/important-information/).

To evaluate the ability of AI tools to pass this examination, we submitted questions from the EBU in-service assessment to three different AI tools: ChatGPT 3.5, ChatGPT 4.0, and Bard. ChatGPT has been developed by OpenAI and relies on a generative pretrained transformer architecture that uses machine learning algorithms to comprehend and generate text naturally. The model is trained on an extensive corpus of textual data, enabling it to respond to a wide array of queries. With its ability to process and generate language, ChatGPT is used in various applications such as customer support and virtual assistants. The most recent update to the program was in 2021; therefore, its knowledge base is limited to events and information up to that year, constraining its ability to address recent developments. ChatGPT was initially released in July 2020 and was made available for free use, in contrast to ChatGPT 4.0, which was launched in March 2023 and is a subscription-based service offering more precise and quicker answers owing to a larger neural network architecture and a more expansive data set. Bard has been developed by Google (released in July 2023) and is a direct competitor to ChatGPT, offering similar functionalities. One of the main distinctions lies in the continuous updating model of Bard, in contrast to that of ChatGPT. As of the current date, Bard is available for free as an experimental version to individuals with a Google account.

We were able to obtain the 100 questions that constitute the EBU in-service assessment. These are single-answer questions with four options, covering the entire spectrum of urological surgery in terms of pathophysiology, prevalence, diagnosis, and treatment. We were also able to collect responses from 736 European urologists who participated in the in-service assessment as FEBU candidates. The AI tools were queried in the following manner: "I will pose questions to you concerning urological surgery; you will need to select the correct answer each time", and a series of 100 questions was then subsequently presented.

The average score for the 736 participants was 67.20, with standard deviation of 10.37. The 25th percentile was 60 and the 75th percentile was 75. ChatGPT 3.5 scored 59 points, ranking in 570th place, within the bottom 25%. ChatGPT 4.0 scored 80 points, ranking 80th, just on the border of the top 10%. Google Bard scored 68 points, ranking 340th and falling in the upper half of the distribution. The data are plotted in Figure 1.

Our results underline the ability of AI to answer specialized urological questions, with one tool even capable of ranking among the top participants. The results demonstrate the ability of AI to retrieve accurate knowledge across diverse domains in an unparalleled time frame, thereby suggesting the potential to transform our methods for learning and practice. However, the three AI tools were not infallible during the tests, with some exhibiting a significant error rate. This can be attributed to two phenomena. First, the AI models are trained on databases that, although exhaustive, are not certified or validated in the peer-reviewed literature. Second, the architectural limitations of the AI tools preclude the incorporation of multiple criteria in generating responses. Specifically, analysis of the results and incorrect answers indicated suboptimal performance by the AI tools for patient-adapted clinical management and the severity of clinical situations. Nonetheless, the improvement in score from ChatGPT 3.5 to ChatGPT 4.0 suggests that these algorithms are undergoing constant evolution with rapid progression. In the current state, the use of data generated by AI requires a vigilant and critical approach.

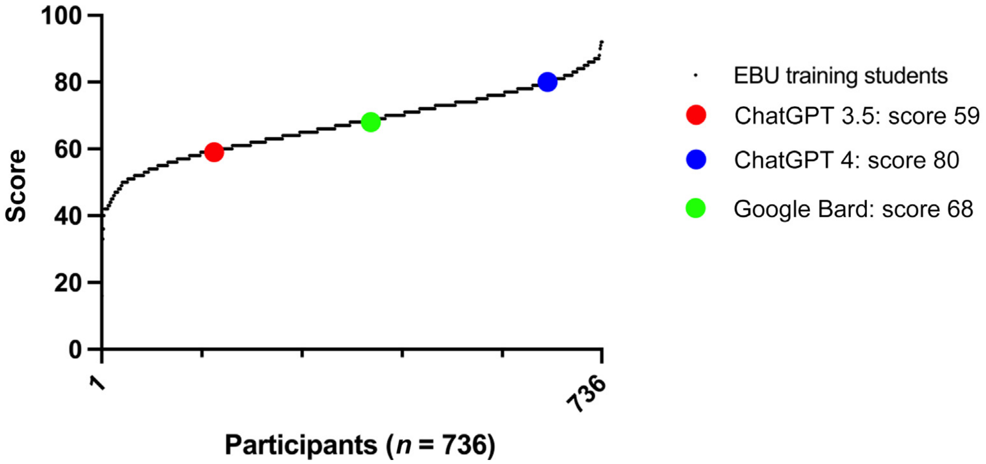


Fig. 1 – Comparative evaluation of the performance of 736 urology trainees and artificial intelligence tools (ChatGPT 3.5, ChatGPT 4.0, Bard) on the European Board of Urology (EBU) in-service assessment.

Table 1 – Studies on artificial intelligence performance in urology education

Study	Year	Question source	Questions evaluated	Accuracy	ChatGPT Version tested
Huynh [6]	2023	AUA 2022 SASP	150	26.7% for open-ended questions and 28.2% for MC questions	3.5
Deebel [7]	2023	AUA 2021–2022 SASP	268	42.3% for questions from 2022 and 53.8% for questions from 2021	3

AUA = American Urological Association; MC = multiple choice; SASP = self-assessment study program.

Other authors have assessed the role of AI in urological surgery. Thirteen studies on this subject were found in the literature. These studies evaluated the ability of AI systems to answer questions regarding clinical cases across general urology, oncology, and pediatrics. Table 1 summarizes two studies that, like ours, assessed how well AI tools answer examination questions in the field of urology [6,7]. The majority of the results attest to the reliability of AI tools in urology, aligning with our study conclusions while emphasizing the need for vigilance when interpreting the results.

In conclusion, our study findings demonstrate that AI can accurately respond to numerous questions across a wide range of themes in urology. Although the use of these tools is not recommended for clinical practice as of yet, they hold the potential to assist physicians in making informed decisions and improving the overall quality and efficiency of health care delivery. Further deliberations are warranted regarding regulation of the use of these chatbots to guide the development of AI in an ethical and responsible manner.

Author contributions: Benoît Mesnard had full access to all the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis.

Study concept and design: Mesnard, Bogaert, Madec.
Acquisition of data: Mesnard, Perrot.
Analysis and interpretation of data: Mesnard, Schirmann, Branchereau, Perrot, Neuzillet, Lebret, Madec.
Drafting of the manuscript: Mesnard, Madec.
Critical revision of the manuscript for important intellectual content: Branchereau, Perrot, Bogaert, Neuzillet, Lebret, Madec.
Statistical analysis: Mesnard, Madec.
Obtaining funding: None.
Administrative, technical, or material support: None.
Supervision: None.
Other: None.

Financial disclosures: Benoît Mesnard certifies that all conflicts of interest, including specific financial interests and relationships and affiliations relevant to the subject matter or materials discussed in the manuscript (eg, employment/affiliation, grants or funding, consultancies, honoraria, stock ownership or options, expert testimony, royalties, or patents filed, received, or pending), are the following: None.

Funding/Support and role of the sponsor: None.

Data sharing statement: All data generated or analyzed during this study are included in this article. Further enquiries can be directed to the corresponding author.

Acknowledgments: The authors thank the committee of the European Board of Urology for making the data available.

Ethics considerations: The generation of the data presented in this article necessitated use of artificial intelligence.

References

[1] Liang X, Bibault JE, Leroy T, et al. Automated contour propagation of the prostate from pCT to CBCT images via deep unsupervised learning. Med Phys 2021;48:1764–70. https://doi.org/10.1002/mp.14755.

[2] Ardila D, Kiraly AP, Bharadwaj S, et al. End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. Nat Med 2019;25:954–61. https://doi.org/10.1038/s41591-019-0447-x.

[3] Gulshan V, Peng L, Coram M, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. JAMA 2016;316:2402–10. https://doi.org/10.1001/jama.2016.17216.

[4] Hameed BMZ, Dhavileswarapu AVL, Raza SZ, et al. Artificial intelligence and its impact on urological diseases and management: a comprehensive review of the literature. J Clin Med 2021;10:1864. https://doi.org/10.3390/jcm10091864.

[5] Hameed BMZ, Shah M, Naik N, et al. The ascent of artificial intelligence in endourology: a systematic review over the last 2 decades. Curr Urol Rep 2021;22:53. https://doi.org/10.1007/s11934-021-01069-3.

[6] 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. https://doi.org/10.1097/UPJ.0000000000000406.

[7] Deebel NA, Terlecki R. ChatGPT Performance on the American Urological Association self-assessment study program and the potential influence of artificial intelligence in urologic training. Urology 2023;177:29–33. https://doi.org/10.1016/j.urology.2023.05.010.

[a] *Urology Department, Hôpital Foch, Suresnes, France*
[b] *Urology Department, Nantes University Hospital, Nantes, France*
[c] *Urology Department, University of Leuven, Leuven, Belgium*

* Corresponding author. Urology Department, Hôpital Foch, Suresnes, France. Tel. +33 6 95063200.
E-mail address: f.madec@hopital-foch.com (F.-X. Madec).