Original Article - Artificial Intelligence in Urology

Investig Clin Urol 2025;66:188-193.
https://doi.org/10.4111/icu.20250040
pISSN 2466-0493 • eISSN 2466-054X

INVESTIGATIVE AND CLINICAL UROLOGY
ICUROLOGY

Assessment of artificial intelligence performance in answering questions on onabotulinum toxin and sacral neuromodulation

Ibrahim Hacibey[1], **Ahmet Halis**[2]

[1]*Department of Urology, Basaksehir Çam and Sakura City Hospital, Istanbul,* [2]*Department of Urology, Yedikule Chest Diseases and Chest Surgery Training and Research Hospital, Istanbul, Türkiye*

Purpose: This study aimed to evaluate the performance of three artificial intelligence (AI) models—ChatGPT, Gemini, and Copilot—in addressing clinically relevant questions about onabotulinum toxin and sacral neuromodulation (SNM) for the management of overactive bladder (OAB).

Materials and Methods: A set of 30 questions covering mechanisms of action, indications, contraindications, procedural details, efficacy, and safety profiles was posed to each AI model. Responses were assessed by a panel of four urology specialists using predefined criteria: accuracy, completeness, clarity, and consistency. A multi-dimensional scoring framework evaluated the performance across five dimensions: factual accuracy, relevance, clarity/coherence, structure, and utility. Responses were scored on a 4-point Likert scale, and statistical analyses were conducted using one-way ANOVA to compare model performance.

Results: ChatGPT achieved the highest mean score (3.98/4) across all dimensions, with statistically significant differences compared to Gemini (3.20/4) and Copilot (2.60/4) (p=0.001 for all dimensions). ChatGPT excelled particularly in clinical application, procedure, and safety categories, consistently delivering accurate and comprehensive answers. No statistically significant differences were found between Gemini and Copilot in most categories.

Conclusions: ChatGPT demonstrated superior performance in generating accurate, complete, and clinically relevant responses for OAB management, highlighting its potential as a reliable tool for both healthcare professionals and patients. However, the variability observed in Gemini and Copilot underscores the need for further refinement of these models. Future studies should explore real-world integration of AI models into clinical workflows to enhance patient care and decision-making.

Keywords: Artificial intelligence; Onabotulinum toxin; Sacral neuromodulation

INTRODUCTION

The integration of artificial intelligence (AI) in healthcare has witnessed significant advancements, particularly with the development of large language models (LLMs) such as ChatGPT, Copilot, and Gemini. These models, powered by natural language processing, are increasingly being utilized to provide medical information, assist in clinical decision-making, and address patient inquiries. Among these applications, the management of overactive bladder (OAB) repre-

Received: January 26, 2025 • **Revised:** February 24, 2025 • **Accepted:** March 5, 2025 • **Published online:** April 10, 2025
Corresponding Author: Ibrahim Hacibey https://orcid.org/0000-0002-2212-5504
Department of Urology, Basaksehir Çam and Sakura City Hospital, Başakşehir Mahallesi G-434 Caddesi No: 2L, Başakşehir, Istanbul 34480, Türkiye
TEL: +90-543-728-7787, E-mail: drihacibey@gmail.com

sents an area of growing interest due to its high prevalence and the complexity of its treatment pathways. OAB, characterized by symptoms of urgency, frequency, and nocturia, often necessitates advanced therapeutic interventions such as onabotulinum toxin injections and sacral neuromodulation (SNM) when first-line treatments fail [1-4].

While onabotulinum toxin and SNM are well-established, their management involves nuanced decision-making informed by clinical guidelines and individualized patient factors. Patients and clinicians alike increasingly turn to AI tools for accessible, evidence-based answers about these treatments. Despite this growing reliance, the accuracy, completeness, and consistency of AI-generated responses on these topics remain underexplored. Evaluations of AI systems like ChatGPT, Copilot, and Gemini have demonstrated variable performance across medical specialties, with studies noting high accuracy for general questions but significant limitations when responses required guideline adherence or addressed complex clinical scenarios [5-7].

This study aims to assess the performance of ChatGPT, Copilot, and Gemini in addressing questions related to onabotulinum toxin and SNM for OAB management. Specifically, it evaluates the models' ability to generate accurate, complete, and guideline-consistent responses to a predefined set of clinically relevant questions. By doing so, the study seeks to illuminate the potential of AI tools in supporting clinical practice while identifying areas for improvement in their application to specialized fields.

MATERIALS AND METHODS

This study employed a descriptive, evaluative approach to assess the performance of ChatGPT version 4.0 (OpenAI), Copilot (GitHub), and Gemini in answering questions related to onabotulinum toxin and SNM in OAB management. A comprehensive set of 30 clinically relevant questions was curated, focusing on the mechanisms of action, indications, contraindications, procedural details, efficacy, and safety profiles of onabotulinum toxin and SNM. Questions were sourced from clinical guidelines, frequently asked questions on healthcare platforms, and inputs from urology experts specializing in OAB management. Since patient data were not used in the study, ethics committee approval was not required.

Each question was individually posed to ChatGPT, Copilot, and Gemini in separate sessions to prevent contextual learning from previous interactions. The models were queried using their default parameters and prompts requesting concise, evidence-based answers with references where applicable. Responses were recorded verbatim and documented for subsequent analysis.

The responses were evaluated by a panel of four urology specialists based on predefined metrics: accuracy, completeness, clarity, and consistency. Accuracy assessed the correctness of information relative to established clinical guidelines. Completeness evaluated whether responses covered all key aspects of the query. Clarity examined the readability and coherence of answers for a clinician audience, while consistency measured the uniformity of responses when similar questions were posed. Each metric was scored on a 4-point Likert scale, with the following definitions: 1=poor (significant inaccuracies or deficiencies), 2=fair (minor inaccuracies or incomplete coverage), 3=good (mostly accurate with minor omissions), and 4=excellent (fully accurate and comprehensive). The mean scores across the four evaluators were calculated for each question. To benchmark the performance of ChatGPT, Copilot, and Gemini, gold-standard answers were developed for each question based on current guidelines and expert consensus. The AI-generated responses were compared to these gold standards to identify discrepancies and gaps.

No patient data were used in this study, and all evaluations were based on publicly available guidelines and hypothetical scenarios. Ethical approval was deemed unnecessary as per institutional policies.

To assess the quality of responses generated by ChatGPT, Gemini, and Copilot, a multi-dimensional scoring framework was designed. This framework evaluated responses based on five distinct dimensions: factual accuracy, relevance, clarity/coherence, structure, and utility. Each dimension was rated individually on a 1-to-5 scale (1=very poor, 5=excellent). Responses to the same set of questions were scored across all three models for each dimension independently. The mean scores for each dimension were calculated and visualized using radar charts, which effectively illustrated the differences in performance. This method allowed for a thorough and detailed comparison of the response quality across the models, both holistically and within specific dimensions.

Statistical analysis was performed using IBM SPSS version 27 (IBM Corp.). Normality assessment was checked with the Shapiro–Wilk test. Scores of frequently asked questions (FAQ) subcategories are presented as percentages. Scores were compared between ChatGPT, Gemini, and Copilot using the one-way ANOVA test. Data were analyzed at 95% confidence level, and a p-value less than 0.05 was considered statistically significant.

Table 1. Mean AI scores for different topics

Topic	ChatGPT	Gemini	Copilot	p-value
FAQs (n=30)	3.98[a]	3.20[b]	2.60[b]	0.001*
General understanding	4.00[a]	3.30[b]	2.60[c]	0.001*
Clinical application	4.00[a]	3.30[b]	2.40[b]	0.001*
Procedure and administration	3.95[a]	2.90[b]	2.15[b]	0.001*
Safety and complications	4.00[a]	3.20[b]	2.80[b]	0.001*
Patient outcomes	3.98[a]	2.80[b]	2.65[b]	0.001*
Comparative insights	3.92[a]	3.53[b]	2.80[c]	0.001*

Lower-case letters are used to identify the group that makes the difference. The same letters (such as a-a) indicate that there is no difference, different letters (such as a-b) indicate that there is a difference.
AI, artificial intelligence; FAQs, frequently asked questions.
*p<0.05.

RESULTS

The analysis evaluated the performance of ChatGPT, Gemini, and Copilot in addressing questions related to onabotulinum toxin and SNM across various thematic categories. Reviewer scores were averaged to assess the comparative accuracy and reliability of the AI models.

ChatGPT demonstrated the highest overall performance, with a mean score of 3.98 out of 4 across all topics, indicating consistent accuracy and reliability in addressing the questions. ChatGPT's scores were statistically significantly higher than those of the other two groups; however, no statistically significant difference was observed between Gemini (mean score, 3.20) and Copilot (mean score, 2.60) (Table 1, Fig. 1).

ChatGPT achieved a perfect score of 4.00 out of 4 in general understanding, outperforming Gemini 3.30 and Copilot 2.60. This highlights its strength in providing accurate and comprehensive responses to foundational inquiries. In clinical applications, ChatGPT maintained its lead with a mean score of 4.00, while Gemini scored 3.30 and Copilot lagged with a score of 2.40. In procedure and administration, ChatGPT achieved another perfect mean score of 3.95, with Gemini scoring 2.90 and Copilot 2.15. In safety and complications, ChatGPT achieved another perfect mean score of 4.00, with Gemini scoring 3.20 and Copilot 2.80 out of 4. For patient outcome, ChatGPT maintained its lead with a mean score of 3.98, while Gemini scored 2.80 and Copilot lagged with a score of 2.65 out of 4. For comparative insights, ChatGPT again led with a mean score of 3.92, followed by Gemini 3.53 and Copilot 2.80 (Table 1).

Lower-case letters are used to identify the group that makes the difference. The same letters (such as a-a) indicate that there is no difference, different letters (such as a-b) indicate that there is a difference.

Fig. 2 (referencing the radar chart) illustrates the comparison of ChatGPT, Copilot, and Gemini across five critical dimensions: factual accuracy, relevance, clarity/coherence, structure, and utility. ChatGPT demonstrated superior performance across all dimensions, with statistically significant differences compared to the other models (p<0.001 for each dimension) (Table 2).

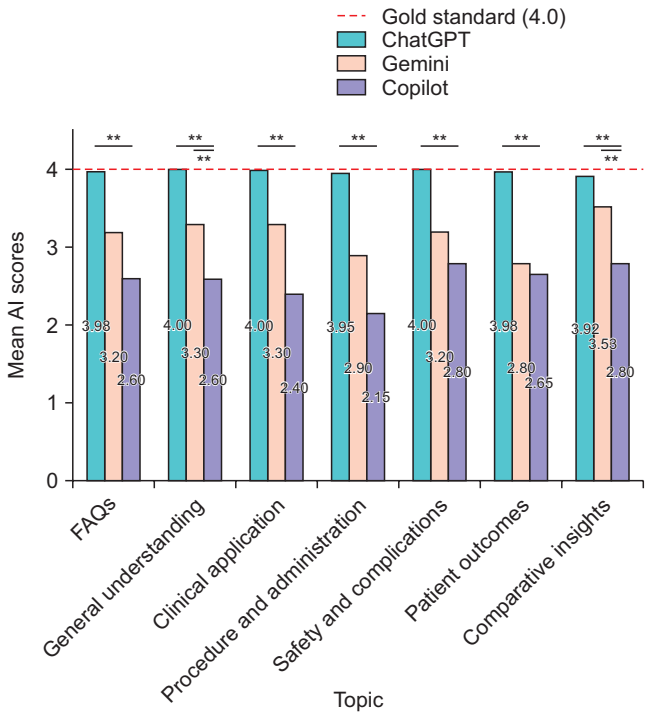

Fig. 1. Mean artificial intelligence (AI) scores for different topics. FAQs, frequently asked questions. **p<0.01.

The consistency of ChatGPT's scores across multiple reviewers highlights its reliability in generating responses. Variability was more pronounced for Gemini and Copilot, indicating occasional inconsistencies in their outputs.

Lower-case letters are used to identify the group that makes the difference. The same letters (such as a-a) indicate

https://doi.org/10.4111/icu.20250040

that there is no difference, different letters (such as a-b or b-c) indicate that there is a statistically difference.

DISCUSSION

The findings align with prior research on AI performance in medical contexts. ChatGPT's high scores across most topics are consistent with studies demonstrating its superior ability to generate accurate, contextually relevant responses in urology and other medical specialties. For example, Caglar et al. [8] reported ChatGPT's accuracy rates exceeding 90% in pediatric urology, emphasizing its potential as a valuable tool for medical education and patient information dissemination.

Similarly, Carlson et al. [9] highlighted ChatGPT's superior performance compared to other AI models in addressing vasectomy-related questions, further corroborating its dominance in AI-powered healthcare applications. The current study builds on this evidence, underscoring ChatGPT's consistent excellence in addressing questions about onabotulinum toxin and SNM [9]. However, the use of such tools requires careful consideration, as it can introduce both opportunities and challenges. On one hand, they provided invaluable insights and streamlined tasks. On the other, it is crucial to be mindful of over-reliance on automated systems, as they may not fully capture the depth of human expertise required in certain contexts. As AI technologies continue to evolve, their integration should always be aligned with domain-specific knowledge and human judgment to ensure balanced, high-quality outcomes. In this study, we employed three AI language models—Gemini, Copilot, and ChatGPT—to support our analysis. Each model contributed unique insights due to differences in their underlying architectures, training datasets, and access to medical guidelines. While this multi-model approach allowed us to cross-verify findings and enhance overall robustness, it also introduced subtle variations in the output. These differences highlight the importance of critically evaluating AI-generated data and underscore the need for further standardization to ensure consistency and reliability in clinical research.

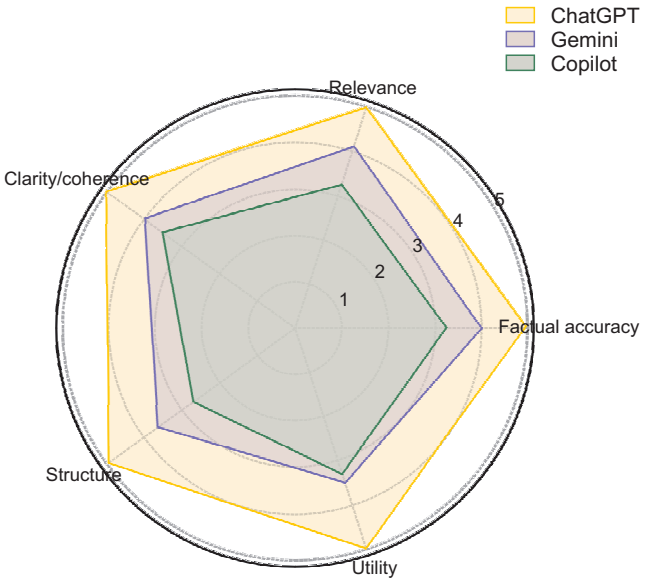

Fig. 2. Performance comparison across dimensions.

The superior performance of ChatGPT across all categories can be attributed to its advanced natural language processing capabilities and comprehensive training dataset. However, Gemini and Copilot exhibited notable weaknesses, particularly in topics requiring procedural knowledge and comparative insights. These discrepancies may stem from differences in training datasets or model architecture. Specifically, Gemini and Copilot struggled the most in safety-related and procedural questions, likely due to the lack of direct access to updated clinical guidelines and structured medical databases. Previous studies have demonstrated that LLMs trained with direct access to peer-reviewed literature and structured clinical guidelines tend to outperform those

Table 2. Summary of AI model performance

Subcategories of response/AI model	ChatGPT	Gemini	Copilot	p-value
Factual accuracy	4.80[a]	4.00[b]	3.20[c]	<0.001*
Relevance	4.90[a]	4.20[b]	3.30[c]	<0.001*
Clarity/coherence	4.90[a]	4.00[b]	3.70[b]	<0.001*
Structure	4.70[a]	3.80[b]	2.80[c]	<0.001*
Utility	4.90[a]	3.70[b]	3.50[b]	<0.001*

Lower-case letters are used to identify the group that makes the difference. The same letters (such as a-a) indicate that there is no difference, different letters (such as a-b) indicate that there is a difference.
AI, artificial intelligence.
*p<0.05.

relying on general web-based corpora [10].

Recent literature further supports the transformative impact of AI in urologic oncology and clinical decision-making. Pak et al. [11] provided a comprehensive review of AI applications in urologic cancers, demonstrating the effectiveness of machine learning and deep learning models in prostate, bladder, and kidney cancer management. Their findings highlight how AI-assisted tools can enhance diagnostic accuracy, predict treatment outcomes, and support clinical decision-making. These insights align with our study, which suggests that ChatGPT, the highest-performing LLM in our assessment, holds significant potential in medical education and clinical support [11].

Additionally, the study by Park et al. [12] evaluated the real-world application of IBM Watson for Oncology (WFO) in managing urologic malignancies. Their research found a 92.7% concordance rate between WFO recommendations and decisions made by a multidisciplinary tumor board, emphasizing the clinical applicability of AI-based decision support tools. However, their study also revealed that certain AI-generated treatment recommendations were not feasible due to regional healthcare policies, underscoring the limitations of AI in real-world medical practice. This is particularly relevant to our findings, as we observed that while ChatGPT performed well in delivering medical knowledge, Gemini and Copilot struggled in areas requiring precise adherence to clinical guidelines, likely due to differences in their training data and access to structured medical resources [12].

Despite ChatGPT's superior performance in accuracy and completeness, its application in clinical decision-making must be approached with caution. AI-generated responses may still contain misinformation or hallucinations, which can have serious clinical implications. Thus, AI outputs should always be cross-verified with clinical guidelines and expert consultation. The responsible integration of AI in clinical workflows should involve human oversight, ensuring that responses align with current medical standards. Regulatory bodies may need to establish guidelines for the appropriate use of AI in healthcare, particularly in areas requiring high diagnostic and procedural accuracy.

Incorporating AI into healthcare, including urology practice, presents significant regulatory and ethical challenges, particularly regarding risk classification and AI-generated misinformation. Baldassarre and Padovan [13] emphasize the necessity of a structured regulatory framework and continuous monitoring to mitigate these risks, particularly within the European AI Act. In urology, AI models could be practically integrated to enhance diagnostic accuracy, surgical planning, and patient management by assisting in image interpretation, risk stratification, and predictive analytics. However, ensuring patient safety and adherence to medical guidelines requires a balanced approach, where AI serves as a supportive tool rather than a replacement for clinical expertise. While AI has the potential to streamline workflows and improve decision-making, unchecked reliance may lead to misdiagnoses, ethical dilemmas, and data privacy concerns. Therefore, developing clear governance strategies and maintaining human oversight are crucial for maximizing AI's benefits while minimizing its risks in urology and broader clinical practice [13].

While AI models such as ChatGPT demonstrate significant potential in delivering medical information, their integration into clinical decision-making requires careful consideration of both risks and practical limitations. Recent studies have highlighted key challenges associated with AI-generated medical content, particularly concerning misinformation, algorithmic bias, and the ethical implications of AI-driven healthcare recommendations [14,15].

The results suggest that ChatGPT can serve as a reliable tool for healthcare professionals and patients seeking information on onabotulinum toxin and SNM. Its ability to consistently deliver accurate and contextually appropriate responses positions it as a valuable resource for enhancing patient education and supporting clinical decision-making. However, the variability observed in Gemini and Copilot underscores the need for cautious use of AI tools in clinical settings. Further refinement and validation are required to ensure their reliability.

Specifically, we suggest conducting real-world validation studies where AI-generated responses are systematically compared against clinical decisions made by urology specialists in real-world scenarios. This will help assess AI's reliability, applicability, and potential limitations in clinical workflows. Additionally, we propose prospective clinical studies integrating AI models into urological practice, where AI-assisted decision-making is compared with standard clinical judgment in areas such as patient counseling, treatment selection, and postoperative management. These studies will provide insight into whether AI can effectively complement clinical decision-making while ensuring patient safety and adherence to established guidelines.

CONCLUSIONS

This study highlights ChatGPT's superior performance in addressing questions related to onabotulinum toxin and SNM. While Gemini and Copilot show potential, their limitations emphasize the importance of continuous development

https://doi.org/10.4111/icu.20250040

and rigorous validation of AI models for medical applications. ChatGPT's robust performance reaffirms its potential as a valuable tool in the evolving landscape of AI-powered healthcare.

CONFLICTS OF INTEREST

The authors have nothing to disclose.

FUNDING

None.

ACKNOWLEDGMENTS

The authors wish to express their gratitude to Dr. Ufuk Caglar, Dr. Enes Pay, and Dr. Faruk Ay for their assistance in providing an objective evaluation of the responses generated by the artificial intelligence models.

AUTHORS' CONTRIBUTIONS

Research conception and design: Ibrahim Hacibey and Ahmet Halis. Data acquisition: Ibrahim Hacibey. Statistical analysis: Ibrahim Hacibey. Data analysis and interpretation: Ibrahim Hacibey and Ahmet Halis. Drafting of the manuscript: Ibrahim Hacibey and Ahmet Halis. Critical revision of the manuscript: Ibrahim Hacibey and Ahmet Halis. Administrative, technical, or material support: Ibrahim Hacibey and Ahmet Halis. Supervision: Ibrahim Hacibey. Approval of the final manuscript: all authors.

SUPPLEMENTARY MATERIAL

Supplementary material can be found via https://doi.org/10.4111/icu.20250040.

REFERENCES

1. Cameron AP, Chung DE, Dielubanza EJ, Enemchukwu E, Ginsberg DA, Helfand BT, et al. The AUA/SUFU guideline on the diagnosis and treatment of idiopathic overactive bladder. J Urol 2024;212:11-20.
2. European Association of Urology (EAU). Non-neurogenic Female LUTS [Internet]. EAU [cited 2025 Jan 12]. Available from: https://uroweb.org/guidelines/non-neurogenic-female-luts.
3. Gormley EA, Lightner DJ, Burgio KL, Chai TC, Clemens JQ, Culkin DJ, et al.; American Urological Association; Society of Urodynamics, Female Pelvic Medicine & Urogenital Reconstruction. Diagnosis and treatment of overactive bladder (non-neurogenic) in adults: AUA/SUFU guideline. J Urol 2012;188(6 Suppl):2455-63.
4. Gajewski JB, Schurch B, Hamid R, Averbeck M, Sakakibara R, Agrò EF, et al. An International Continence Society (ICS) report on the terminology for adult neurogenic lower urinary tract dysfunction (ANLUTD). Neurourol Urodyn 2018;37:1152-61.
5. ChatGPT version 4.0 [Internet]. OpenAI [cited 2025 Jan 2]. Available from: https://chatgpt.com/
6. Copilot [Internet]. GitHub [cited 2025 Jan 3]. Available from: https://copilot.microsoft.com/
7. Gemini [Internet]. Google DeepMind [cited 2025 Jan 3]. Available from: https://gemini.google.com/
8. Caglar U, Yildiz O, Meric A, Ayranci A, Gelmis M, Sarilar O, et al. Evaluating the performance of ChatGPT in answering questions related to pediatric urology. J Pediatr Urol 2024;20:26.e1-5.
9. Carlson JA, Cheng RZ, Lange A, Nagalakshmi N, Rabets J, Shah T, et al. Accuracy and readability of artificial intelligence chatbot responses to vasectomy-related questions: public beware. Cureus 2024;16:e67996.
10. Pressman SM, Borna S, Gomez-Cabello CA, Haider SA, Haider CR, Forte AJ. Clinical and surgical applications of large language models: a systematic review. J Clin Med 2024;13:3041.
11. Pak S, Park SG, Park J, Cho ST, Lee YG, Ahn H. Applications of artificial intelligence in urologic oncology. Investig Clin Urol 2024;65:202-16.
12. Park T, Gu P, Kim CH, Kim KT, Chung KJ, Kim TB, et al. Artificial intelligence in urologic oncology: the actual clinical practice results of IBM Watson for Oncology in South Korea. Prostate Int 2023;11:218-21.
13. Baldassarre A, Padovan M. Regulatory and ethical considerations on artificial intelligence for occupational medicine. Med Lav 2024;115:e2024013.
14. Williamson SM, Prybutok V. The era of artificial intelligence deception: unraveling the complexities of false realities and emerging threats of misinformation. Information 2024;15:299.
15. Okonji OR, Yunusov K, Gordon B. Applications of generative AI in healthcare: algorithmic, ethical, legal and societal considerations. arXiv:2406.10632 [Preprint]. 2024 [cited 2025 Jan 19]. Available from: https://doi.org/10.48550/arXiv.2406.10632