Are Questionnaire Forms Necessary in the Age of Artificial Intelligence? A Comparative Study of AI Tools in LUTS Assessment

Fatih Gökalp * , Eser Ördek , Ali Borekoglu , Sadık Görür

Posted Date: 26 February 2026

doi: 10.20944/preprints202602.1633.v1

Keywords: artificial intelligence; prostate; lower urinary tract symptoms; diagnosis; questionnaire

 doi:10.20944/preprints202602.1633.v1

Article

Are Questionnaire Forms Necessary in the Age of Artificial Intelligence? A Comparative Study of AI Tools in LUTS Assessment

Fatih Gökalp [1], Eser Ördek [1], Ali Borekoglu [2] and Sadık Görür [1]

[1] Hatay Mustafa Kemal University, Faculty of Medicine, Department of Urology, Hatay, Turkey

[2] Mersin State Hospital, Department of Urology

* Correspondence: fatihgokalp85@gmail.com

Highlight

AI tools have exciting potential as a assistance in urology practice. The primary point is these tools can accurately evaluate patients' history and calculate the questionnaire forms. These are the first steps of these tools, that will start to evaluate patients in the future.

Abstract

Background: This study aimed to evaluate the accuracy and clinical applicability of various artificial intelligence (AI) tools (ChatGPT vs. Gemini vs. Grok) in interpreting patient history, calculating the International Prostate Symptom Score (IPSS).**Methods:** Between November 2024 and January 2025, 32 male patients aged 45–70 years presenting with LUTS to a urology outpatient clinic were enrolled. Patients with a history of cancer, elevated PSA, or previous urological intervention were excluded. Each patient's history was documented using standard IPSS questions in an open-ended format, which was then inputted into four AI tools. Each AI tool calculated the IPSS score, assessed symptom severity, and suggested treatment options. **Results:** The median age was 62.5 years, with a median PSA of 1.08 ng/dL and the median prostate volume of 38 cc. All AI tools calculated IPSS scores with statistically comparable accuracy to patient-reported scores ($p < 0.005$). Grok provided specific treatment recommendations for all patients (100%), followed by ChatGPT (78.1%). Grok also demonstrated the highest agreement with physician treatment plans ($\kappa = 0.709$, $p < 0.001$), while ChatGPT and Gemini showed lower compatibility ($p > 0.05$). Supplementary analysis revealed Grok's unique approach in calculating PSA density and identifying potential malignancy risk.**Conclusions:** AI tools can accurately interpret patient history and assess LUTS severity. Grok showed the highest compatibility with urologist recommendations, while ChatGPT offered the most nuanced therapeutic suggestions. These findings suggest that AI tools may soon become valuable clinical decision support systems.

Keywords: artificial intelligence; prostate; lower urinary tract symptoms; diagnosis; questionnaire

Introduction

Lower urinary tract symptoms (LUTS) are a group of symptoms related to urination, including leakage, frequency, or a weak stream. LUTS are divided into two main branches: storage and voiding symptoms. In particular, the worsening of voiding symptoms, with or without comorbidities in the community, is a significant concern in clinical practice. These symptoms cause distress and reduce the quality of life [1]. The guidelines recommended validated questionnaire forms to evaluate the patients [2]. The IPSS and a 10-minute self-administered questionnaire can now be used to objectively evaluate the subjective burden of a patient's LUTS [3]. The IPSS became a standard tool in the assessment of LUTS in men over 30 years [4]. This form is crucial for assessing the patient and how well they are responding to treatment. However, these forms can sometimes be forgotten, or in the

 doi:10.20944/preprints202602.1633.v1

changing digital world, it may be difficult to access these forms in print, and perhaps even become unnecessary in the future.

Artificial intelligence (AI) has the potential to save costs while improving patient diagnosis and treatment [5]. Because AI is still in its early stages of development, its potential benefits in medicine are undervalued. Nevertheless, this instrument remains an essential aid. The era of artificial intelligence is just beginning to blossom. A novel AI tool accelerated the development of language models and natural language processing (NLP) applications. These tools make people's lives easier in every aspect. The reliance of AI on pre-existing datasets is important [6].

Throughout history, urology has adapted very rapidly to new technologies. AI tools have rapidly been adopted in urology, being used for clinical decisions, medical documentation, or patient education [6]. The current literature also demonstrated that AI tools provide accurate medical records and enhance patients' outcomes [7]. ChatGPT has shown its usefulness in medical recording and patient discharge summaries. Urologists can employ ChatGPT to input patient clinical history and test results to obtain optimal recommendations for patient management, ultimately enhancing medical care [8]. Additionally, AI tools offer accurate and reproducible responses to BPH-related questions [9]. Current studies are examining the reliability and accuracy of AI tools in providing information about BPH or prostate cancer [9,10]. Nonetheless, literature contains a limited number of studies that help in diagnosis, including those focused on anamnesis or questionnaire forms.

In this study, we aimed to evaluate the AI tools to calculate the IPSS form from the patients' history and their treatment options for LUTS.

Materials and Methods

Patients with lower urinary tract symptoms who applied to the Mersin State Hospital urology outpatient clinic between November 2024 and January 2025 were included in the study to compare the AI's IPSS answers. Due to the potential for heterogeneity, the study excluded people younger than 45 and older than 70 who had a history of cancer, had a high PSA, or had undergone urological intervention in the past. The study was conducted in accordance with the Declaration of Helsinki, and approved by the Hatay Mustafa Kemal University Ethics Committee (protocol code #32 and May 2025). Informed consent was obtained from all subjects involved in the study.

In order to help artificial intelligence analyze the data, patients were asked open-ended questions through the international prostate symptom score questions which listed below and answers were recorded like "Never, Rarely, Sometimes, Half the time, Often, Always". During the patient's history-taking, the inquiry was posed as follows: "In the past month, how frequently have you experienced the sensation of incomplete bladder emptying after urination?" ; "In the previous month, how frequently have you needed to urinate again within two hours of completing urination?"; "In the previous month, how frequently have you had interruptions during urination, stopping and starting multiple times?"; "In the preceding month, how frequently have you experienced difficulty in delaying urination?"; "In the past month, how frequently have you had a diminished urine stream?"; "In the past month, how frequently have you experienced the need to exert effort to initiate urination?"; "During the preceding month, how frequently did you generally awaken each night to urinate from the moment you retired to bed until you arose in the morning?"; (this inquiry was posed to assess quality of life)."; "If you were to spend the rest of your life with your urinary condition just the way it is now, how would you feel about that?" and answers were recorded in the patients' history. The PSA levels, prostate volume calculated during the ultrasonography, and uroflowmetry parameters of the individuals were also recorded. All patients filled out IPSS forms, and their responses also recorded to data. In order to calculate the patient's IPSS score by AI tools, the history—which included the previously mentioned questions—was first given to the AI tools. The response of AI tools obtained and recorded.

After the artificial intelligence tools calculated the IPSS score, the patient's PSA value, prostate volume, and uroflowmetry data were provided to these tools; recommendations regarding the patient's appropriate treatment method were requested, and the responses were noted. Questions

 doi:10.20944/preprints202602.1633.v1

and answers were recorded by accessing the official addresses for artificial intelligence tools and we used Gemini* 2.0 from https://gemini.google.com/app; ChatGPT* 4.0 from https://chatgpt.com/?model=auto, and Copilot*; https://copilot.microsoft.com/chats, and Grok* 3.0 beta version; https://grok.com/?referrer=website.

Statistical Analysis

All statistical analyses were conducted using IBM SPSS Statistics for Windows, Version 26.0 (IBM Corp., Armonk, NY, USA). Descriptive statistics were used to summarize patient characteristics, including median and interquartile range (IQR) for continuous variables such as age, PSA level, and prostate volume. To compare the IPSS scores calculated by patients and each AI tool (ChatGPT, Gemini, Copilot, and Grok), the Wilcoxon signed-rank test was used due to the non-parametric distribution of the data. Inter-rater agreement for disease severity categorization (mild, moderate, severe) between AI tools and patient responses was assessed using Cohen's kappa (κ) coefficient. A κ value >0.60 was considered substantial agreement, and values >0.80 indicated almost perfect agreement. The frequency of specific treatment recommendations generated by each AI tool was compared using the Chi-square test. To evaluate the consistency between AI-recommended treatments and those provided by the treating urologist, Cohen's kappa test was again applied. A p-value of less than 0.05 was considered statistically significant for all analyses.

Results

A total of 32 patients were included in the study. The median age of the patients was 62.50 years. The median PSA level and prostate volume were 1.08 ng/dl (0.86-1.60) and 38.00 cc (30.00-45.00) (Table 1). All four AI tools performed similar results when calculating patient IPSS scores. Although Copilot calculations were marginally higher, there was no statistically significant difference between the IPSS scores of patients and those calculated from AI tools ($p<0.005$ for each tool). All AI tools correctly calculated and indicated the severity of the disease. ($p<0.05$ for each AI tool when compared to patients) (Table 2).

Table 1. General data.

	Median (IQR)
Age	62.50 (59.00-66.00)
PSA level	1.08 (0.86-1.60)
Prostate volume	38.00 (30.00-45.00)
IPSS pateints	12.00 (11.00-15.00)
Gemini calculated IPSS	13.00 (10.00-15.00)
ChatGPT calculated IPSS	12.50 (10.00-14.00)
Copilot calculated IPSS	15.00 (11.00-20.00)
Grok calculated IPSS	11.00 (10.00-14.00)

Table 2. Calculated severity of AI tools.

	Mild	**Moderate**	**Severe**	**P value**
Patients' severity	6.0 (18.8%)	23.0 (71.9%)	3.0 (9.4%)	
Gemini calculated severity	5.0 (15.6%)	23.0 (71.9%)	4.0 (12.5%)	0.029 *
ChatGPT calculated severity	6.0 (18.8%)	24.0 (75.0%)	2.0 (6.2%)	0.018&
Copilot calculated severity	5.0 (15.6%)	18.0 (56.3%)	9.0 (28.1%)	0.038#
Grok calculated severity	7.0 (21.9%)	23.0 (71.9%)	2.0 (6.3%)	0.044^

* kappa = 0.294; & kappa = 0.330; # kappa = 0.248; ^kappa = 0.284.

 doi:10.20944/preprints202602.1633.v1

In response to a question concerning the AI tools' recommendations for treating LUTS, Grok suggested specific treatments for every patient (100.0%), while ChatGPT produced the second-highest percentage of specific treatments (78.1%) (Table 3). When the compatibility of specific treatment with the treatment given by the patient's physician was evaluated, Grok made recommendations with a very high level of compatibility (kappa = 0.709, p<0.001). Furthermore, neither ChatGPT's compatibility nor Gemini's specific treatment recommendations were statistically significantly compatible (kappa = 0.010, p = 0.909 and kappa = 0.031, p = 0.485, respectively). (Table 4).

Table 3. Specific treatment advised from AI tools.

	Absent	Present
Gemini treatment advice	21.0 (65.6%)	11.0 (34.4%)
ChatGPT treatment advice	7.0 (2.19%)	25.0 (78.1%)
Copilot treatment advice	31.0 (96.9%)	1.0 (3.1%)
Grok treatment advice	0.0 (0.0%)	32.0 (100.0%)

Table 4. Comparison of specific treatment from AI tools to patients received.

		Gemini		ChatGPT		Co-pilot		Grok	
		Value	P value	Value	P value	Value	P value	Value	P value
Interval by Interval	**Pearson's R**	0.003	0.988	-0.361	0.042	-0.27	0.134	0.727	<0.001
Ordinal by Ordinal	**Spearman Correlation**	0.05	0.785	-0.304	0.091	-0.255	0.159	0.72	**<0.001**
Measure of Agreement	**Kappa**	0.031	0.485	0.010	0.909	-0.019	0.219	**0.709**	0
N of Valid Cases		32		32		32		32	

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

c. Based on normal approximation.

Supplementary Result

The assessment of artificial intelligence technologies' responses yielded significantly varied results. When Gemini was initially being considered, Gemini's patients stood out in terms of evaluation in light of the data. Gemini emphasized that patients faced a risk of retention if their prostate volume was above 100 cc and their uroflowmetry results were inadequate. Once more, Gemini stated that individuals with lower urinary symptoms and a normal prostate volume and good uroflowmetry might not have benign prostate obstruction but rather an overactive bladder. ChatGPT distinguished itself in providing therapy advice for patients. ChatGPT suggested combination therapy as the primary treatment for patients with enlarged prostate size. ChatGPT also advocated anticholinergics alongside alpha blocker medication for patients with moderate prostate volume and good uroflowmetry results. Additionally, ChatGPT suggested desmopressin alongside lifestyle modifications for patients exhibiting severe nocturia symptoms, a low IPSS score, and normal uroflowmetry results. Grok meticulously assesses patients, comparable to Gemini, and offers possible and preliminary diagnoses, while also suggesting more precise therapies, similar to ChatGPT. Additionally, unlike other artificial intelligence tools, it requests additional information about the patient by emphasizing the points that will be missing during the diagnosis phase and recommends

 doi:10.20944/preprints202602.1633.v1

treatment by prioritizing this. Grok evaluated patients in terms of prostate cancer by calculating PSAD in every patient without exception. As an example, when a patient with a mild-moderate IPSS score was given a PSA value of 5, it directly calculated PSAD and recommended a urology consultation first, then an MRI with free PSA due to the patient's risk of cancer due to the patient's PSAD being above 0.15 and stated that cancer exclusion was prioritized.

Discussion

The AI tools became popular and an indispensable part of life. They have begun to hold a critical place in health care, as in all professions. Our study demonstrates that AI tools can properly evaluate patients' history and calculate the questionnaire forms. According to our research, these instruments will usher in a new era in medical practice. One of the most important tools for healthcare providers will be AI.

Health outcomes and medical care could be greatly enhanced by artificial intelligence (AI) systems. Therefore, it is essential to make sure that the concepts of explainability and trust are used to guide the development of clinical AI. A crucial initial step in assessing these attributes is comparing the medical expertise of AI to that of skilled human practitioners. Medical education could be another application for AI tools. An early study showed that ChatGPT was given open-ended and multiple-choice questions from Step 1, Step 2 CK, and Step 3 of the US Medical Licensing Examination (USMLE) to see how well it could learn difficult medical and clinical information. ChatGPT had the lowest accuracy in Step 1 of the USMLE, which is regarded as the most challenging test [11]. Consistent with these findings, a different study evaluated ChatGPT's responses to urological emergencies and demonstrated that ChatGPT offered correct and satisfactory answers. However, when answering urological emergency questions based on guidelines, ChatGPT's accuracy rate fell to 54.76%. In later studies, artificial intelligence tools predict many post-surgical outcomes [12,13] and additionally help the clinicians' preoperative planning [14]. Urologists could use ChatGPT for various tasks, such as information extraction, text classification, summarization, and creation. Other uses for ChatGPT include assistance with administrative tasks integrated into contemporary clinical procedures. AI tools have been used to provide patients with individualized resources, simplify medical concepts, answer commonly asked questions regarding postoperative care, and create operation reports and discharge summaries based on the data supplied [14]. Urologists around the world were using AI tools in their clinical practice and research. Similar to these results, our study supported that AI tools will summarize patients' history, calculate questionnaire forms, and suggest accurate treatment in the future. Based on these results, we believe that future artificial intelligence technologies will be included into systems, significantly assisting clinicians in diagnostic processes by automatically providing questionnaire responses during patient history intake.

Implementing AI tools in medicine demonstrates significant potential for transforming traditional research, education, and diagnosis. Current literature showed that there are some differences in these AI tools [15–17]. A comparative investigation of ChatGPT and Gemini across multiple medical disciplines demonstrates notable differences in their accuracy and response length performance. ChatGPT routinely exhibits superior accuracy rates relative to Gemini across the majority of specializations. This tendency is apparent in specialties such as clinical diagnosis (90% vs. 80%), neuroradiology (100% vs. 86.21%), hematology (63% vs. 44%), physiology (79% vs. 53%), and neurodegenerative disorders (84% vs. 76%). ChatGPT gave more accurate information in various medical fields when compared to Gemini [15–18]. The author noted that Gemini exhibited a prolonged response time. They stated that the extended response length of Gemini indicates its potential to provide more detailed and comprehensive information, advantageous in situations requiring thorough explanations [19]. Another research evaluating and comparing the accuracy, conciseness, and readability of responses from OpenAI ChatGPT-4 and Google Bard regarding prostate cancer revealed that ChatGPT-4 generated more accurate answers than Bard (2.95 ± 0.671 vs 2.73 ± 0.732, p=0.027). Nonetheless, Bard's responses revealed superior readability compared to ChatGPT-4 (2.79 ± 0.408 vs 2.94 ± 0.243, p=0.003) [20]. Similar to these results, our study demonstrated

 doi:10.20944/preprints202602.1633.v1

that ChatGPT gave more accurate information about patients' treatment; however, Gemini provided additional information about patients' diagnoses.

We must acknowledge that artificial intelligence tools serve as assistants rather than substitutes for healthcare personnel. Erkan et al. study assessed the reliability of the information provided by artificial intelligence chatbots (ChatGPT vs. Gemini vs. Copilot) about urogenital cancer treatments. They demonstrated that ChatGPT and Gemini gave moderate-quality information, while Copilot's answers were very low [21]. Furthermore, the PEMAT understandability scores for each chatbot were notably low, while the PEMAT actionability levels were moderate alone for Gemini. A recent study evaluated the responses of three chatbots (ChatGPT, Gemini, and Copilot) regarding the following pathologies and their therapies as offered by AI: prostate cancer, kidney cancer, bladder cancer, urinary lithiasis, and benign prostatic hypertrophy (BPH). The responses showed that Copilot obtained the highest scores. Nonetheless, it was seen on the appropriateness scale that their responses were not the most suitable. Additionally, Gemini achieved the highest scores in surgical treatment, followed by ChatGPT, with Copilot ranking last [22]. Controversial to literature, our study indicated that Copilot calculates IPSS score and severity with high accuracy, at least as much as other AI tools. However, ChatGPT and Gemini listed detailed information for treatment options or diagnosis, whereas Copilot listed only standard information.

Alongside the advancement of artificial intelligence tools, there is a corresponding rise in their quantity. The differences between artificial intelligence tools have been investigated by studies as new artificial intelligence tools emerge [19,22,23]. With the emergence of the artificial intelligence tool Grok, researchers aim to investigate its effectiveness in the medical field. An early study that compared five AI tools answers about kidney stones showed that Grok demonstrated the highest Flesch–Kincaid Reading Ease score (55.6 ± 7.1) and the lowest Flesch–Kincaid Grade Level (10.0 ± 1.1) ratings ($p = 0.001$), whereas Claude outperformed the other chatbots in its DISCERN scores (47.6 ± 1.2) ($p = 0.001$). PEMAT understandability was the lowest in GPT-4 (53.2 ± 2.0), and actionability was the highest in Claude (61.8 ± 3.5) ($p = 0.001$). The authors concluded that ChatGPT had the most complex language structure of the five chatbots, making it the most difficult to read and comprehend, whereas Grok was the simplest [23]. Similar to these results, our study indicated that Grok recommended the treatment closest to the urologist's treatment option.

The responses generated by artificial intelligence tools remain unsatisfactory. A further issue is that the responses of artificial intelligence systems may vary. Nonetheless, it demonstrates that they can be highly beneficial in interrogative formats and that patients provide ratings comparable to their responses. The thorough responses provided by AI technologies for patients instill optimism over their potential advantages in anamnesis in the near future. Artificial intelligence can summarize patient data for the physician by conducting a preliminary assessment and offering initial information and recommendations based on the patient's database. It can sequentially show the stages by analyzing the patient's data, thereby minimizing the aspects that physicians may overlook due to their demanding schedules.

This study has several limitations that should be acknowledged. First, the sample size was relatively small, with only 32 patients included, which may limit the generalizability of the findings. Larger-scale studies are needed to validate the performance and consistency of AI tools across diverse patient populations and clinical settings. Another limitation is that AI tools were provided with structured and complete patient histories; this may not reflect the incomplete or ambiguous nature of real-life patient inputs. The AI's accuracy in processing free-text data or responding to inconsistent information was not fully evaluated. Another important limitation is the nature of AI tools that are rapidly evolving and their performance changing over time. Therefore, the findings of this study may not reflect future versions of the same tools.

Conclusion

The AI tools show tremendous promise in assessing lower urinary tract symptoms (LUTS), calculating IPSS scores, and recommending treatment options with accuracy similar to that of

 doi:10.20944/preprints202602.1633.v1

healthcare professionals. All AI tools precisely assessed symptom severity and showed equivalent efficacy in IPSS calculation. These findings highlight the increasing potential of AI tools to aid in history-taking, clinical evaluation, and treatment planning, contingent upon their utilization being directed and overseen by healthcare professionals. Further research, including bigger cohorts and real-time clinical integration, is necessary to evaluate these tools and create criteria for their safe and successful application in routine practice.

Author Contributions: Conceptualization FG, EO, AB; Data curation FG, AB; Formal analysis FG; Methodology FG, AB; Supervision SG; Writing – original draft FG, EO; Writing – review & editing FG, EO, AB, SG.

Funding: This research received no external funding.

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki, and approved by the Hatay Mustafa Kemal University Ethics Committee (protocol code #32 and May 2025). Informed consent was obtained from all subjects involved in the study.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study. Written informed consent has been obtained from the patient(s) to publish this paper.

Data Availability Statement: The data that support the findings of this study are available from the corresponding author upon reasonable request.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviation

AI	*Artificial intelligence*
IPSS	*International Prostate Symptom Score*
LUTSÇ	*Lower urinary tract symptoms*
PSA	*Prostate sepesific antigen*
NLP	*Natural learning processing*
BPH	*benign prostate hyperplasia*

References

1. Martin SA, Haren MT, Marshall VR, Lange K, Wittert GA, Members of the Florey Adelaide Male Ageing Study. Prevalence and factors associated with uncomplicated storage and voiding lower urinary tract symptoms in community-dwelling Australian men. *World J Urol*. 2011;29(2):179-184. doi:10.1007/s00345-010-0605-8
2. Bosch JLHR, Abrams P, Cotterill N. Etiology, patient assessment and predicting outcome from therapy. *Male Low Urin Tract Symptoms*. Published online January 1, 2013:37-133.
3. Barry MJ, Fowler FJ, O'Leary MP, et al. The American Urological Association symptom index for benign prostatic hyperplasia. The Measurement Committee of the American Urological Association. *J Urol*. 1992;148(5):1549-1557; discussion 1564. doi:10.1016/s0022-5347(17)36966-5
4. Yao MW, Green JSA. How international is the International Prostate Symptom Score? A literature review of validated translations of the IPSS, the most widely used self-administered patient questionnaire for male lower urinary tract symptoms. *Low Urin Tract Symptoms*. 2022;14(2):92-101. doi:10.1111/luts.12415
5. Alexa R, Kranz J, Kuppe C, Hayat S, Hoffmann M, Saar M. [Artificial intelligence in urology-opportunities and possibilities]. *Urol Heidelb Ger*. 2023;62(4):383-388. doi:10.1007/s00120-023-02026-3
6. Tzelves L, Kapriniotis K, Feretzakis G, et al. ChatGPT in Clinical Medicine, Urology and Academia: A Review. *Arch Esp Urol*. 2024;77(7):708-717. doi:10.56434/j.arch.esp.urol.20247707.99
7. Cocci A, Pezzoli M, Lo Re M, et al. Quality of information and appropriateness of ChatGPT outputs for urology patients. *Prostate Cancer Prostatic Dis*. 2024;27(1):103-108. doi:10.1038/s41391-023-00705-y
8. Juhi A, Pipil N, Santra S, Mondal S, Behera JK, Mondal H. The Capability of ChatGPT in Predicting and Explaining Common Drug-Drug Interactions. *Cureus*. 2023;15(3):e36272. doi:10.7759/cureus.36272

 doi:10.20944/preprints202602.1633.v1

9. Zhang Y, Dong Y, Mei Z, et al. Performance of large language models on benign prostatic hyperplasia frequently asked questions. *The Prostate*. 2024;84(9):807-813. doi:10.1002/pros.24699
10. Alasker A, Alsalamah S, Alshathri N, et al. Performance of large language models (LLMs) in providing prostate cancer information. *BMC Urol*. 2024;24(1):177. doi:10.1186/s12894-024-01570-0
11. Kung TH, Cheatham M, Medenilla A, et al. Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. *PLOS Digit Health*. 2023;2(2):e0000198. doi:10.1371/journal.pdig.0000198
12. Noh SH, Cho PG, Kim KN, Kim SH, Shin DA. Artificial Intelligence for Neurosurgery : Current State and Future Directions. *J Korean Neurosurg Soc*. 2023;66(2):113-120. doi:10.3340/jkns.2022.0130
13. Park J, Bonde PN. Machine Learning in Cardiac Surgery: Predicting Mortality and Readmission. *ASAIO J Am Soc Artif Intern Organs 1992*. 2022;68(12):1490-1500. doi:10.1097/MAT.0000000000001696
14. Iqbal J, Jahangir K, Mashkoor Y, et al. The future of artificial intelligence in neurosurgery: A narrative review. *Surg Neurol Int*. 2022;13:536. doi:10.25259/SNI_877_2022
15. Patil NS, Huang RS, van der Pol CB, Larocque N. Comparative Performance of ChatGPT and Bard in a Text-Based Radiology Knowledge Assessment. *Can Assoc Radiol J J Assoc Can Radiol*. 2024;75(2):344-350. doi:10.1177/08465371231193716
16. Kumari A, Kumari A, Singh A, et al. Large Language Models in Hematology Case Solving: A Comparative Study of ChatGPT-3.5, Google Bard, and Microsoft Bing. *Cureus*. 15(8):e43861. doi:10.7759/cureus.43861
17. Dhanvijay AKD, Pinjar MJ, Dhokane N, Sorte SR, Kumari A, Mondal H. Performance of Large Language Models (ChatGPT, Bing Search, and Google Bard) in Solving Case Vignettes in Physiology. *Cureus*. 2023;15(8):e42972. doi:10.7759/cureus.42972
18. Muhialdeen AS, Mohammed SA, Ahmed NHA, et al. Artificial Intelligence in Medicine: A Comparative Study of ChatGPT and Google Bard in Clinical Diagnostics. *Barw Med J*. Published online November 6, 2023. doi:10.58742/pry94q89
19. Fattah FH, Salih AM, Salih AM, et al. Comparative analysis of ChatGPT and Gemini (Bard) in medical inquiry: a scoping review. *Front Digit Health*. 2025;7:1482712. doi:10.3389/fdgth.2025.1482712
20. Belge Bilgin G, Bilgin C, Childs DS, et al. Performance of ChatGPT-4 and Bard chatbots in responding to common patient questions on prostate cancer 177Lu-PSMA-617 therapy. *Front Oncol*. 2024;14:1386718. doi:10.3389/fonc.2024.1386718
21. Erkan A, Koc A, Barali D, et al. Can Patients With Urogenital Cancer Rely on Artificial Intelligence Chatbots for Treatment Decisions? *Clin Genitourin Cancer*. 2024;22(6):102206. doi:10.1016/j.clgc.2024.102206
22. Szczesniewski JJ, Ramos Alba A, Rodríguez Castro PM, Lorenzo Gómez MF, Sainz González J, Llanes González L. Quality of information about urologic pathology in English and Spanish from ChatGPT, BARD, and Copilot. *Actas Urol Esp*. 2024;48(5):398-403. doi:10.1016/j.acuroe.2024.02.009
23. Şahin MF, Topkaç EC, Doğan Ç, et al. Still Using Only ChatGPT? The Comparison of Five Different Artificial Intelligence Chatbots' Answers to the Most Common Questions About Kidney Stones. *J Endourol*. 2024;38(11):1172-1177. doi:10.1089/end.2024.0474