ORIGINAL RESEARCH ORİJİNAL ARAŞTIRMA

DOI: 10.5336/urology.2024-107872

Researching the Reliability of ChatGPT-4, 4 Plus and Google Gemini in Responding to Questions About the Management of Lower Urinary Tract Trauma: Cross-Sectional Study

Alt İdrar Yolu Travmasının Yönetimi ile İlgili Sorulara Yanıt Vermede ChatGPT-4, 4 Plus ve Google Gemini'nin Güvenilirliğinin Araştırılması: Kesitsel Çalışma

Yunus ÇOLAKOĞLU[a], Ali AYTEN[b]

[a]Bahçelievler Memorial Hospital, Clinic of Urology, İstanbul, Türkiye
[b]Gaziosmanpaşa Training and Research Hospital, Clinic of Urology, İstanbul, Türkiye

ABSTRACT Objective: The emergence of artificial intelligence chatbots in recent years has dramatically changed the landscape of education and access to knowledge. The role of these large language models (LLM) within medicine is still relatively unclear due to their novelty. The question of whether such systems can replace the knowledge and experience of trained doctors is highly debated. Lower urinary tract trauma is a very complex emergency to manage. First responding physicians often need consultation. In this study, we aimed to investigate whether these three LLMs are reliable enough in management of lower urinary tract trauma. **Material and Methods:** The recommendation tables of the bladder, urethra and genital trauma management section of the European Association of Urology 2024 guidelines were analyzed. Those with strong and weak recommendation levels were translated into a questionnaire. Questions were asked in English via ChatGPT-4, ChatGPT-4 Plus and Google Gemini. Answers were evaluated by two surgeons experienced in urogenital trauma and subjected to statistical analysis by calculating the mean score. **Results:** In total, 27 questions were included in the study. While ChatGPT-4 and 4 Plus were more successful than Google Gemini in urethral trauma management, Google Gemini and ChatGPT-4 Plus were statistically better in approaching bladder trauma ($p<0.001$). In total, ChatGPT-4 and Google Gemini gave 81.4% correct and sufficient answers to the questions, while this rate was 88.8% in ChatGPT-4 Plus ($p=0.618$). **Conclusion:** Our study confirms that ChatGPT-4, 4 Plus and Google Gemini are a reliable resource in lower urinary tract trauma management. In future, it may become a platform to help healthcare professionals in determining the managing trauma.

Keywords: Lower urinary tract traumas; ChatGPT; Google Gemini

ÖZET Amaç: Son yıllarda yapay zekâ sohbet robotlarının ortaya çıkması, eğitim ve bilgiye erişim manzarasını önemli ölçüde değiştirdi. Bu büyük dil modellerinin [large language models (LLM)] tıbbın içindeki rolü, yenilikleri nedeniyle hâlâ nispeten belirsizdir. Bu tür sistemlerin, eğitimli doktorların geniş bilgi ve deneyiminin yerini alıp alamayacağı sorusu oldukça tartışılmaktadır. Alt üriner sistem travmaları yönetimi oldukça zor ve karmaşık bir acil durumdur. Hastaya ilk müdahale eden hekimler genellikle konsültasyona ihtiyaç duymaktadırlar. Bu çalışmada, bu üç LLM'nin alt üriner sistem travmalarının yönetimi konusunda yeterince güvenilir mi sorusunu araştırmayı amaçladık. **Gereç ve Yöntemler:** Avrupa Üroloji Derneği 2024 kılavuzunun mesane, üretra ve genital travma yönetimi bölümünün öneri tabloları analiz edildi. Güçlü ve zayıf öneri düzeyine sahip olanlar soru formuna çevrildi. Sorular İngilizce olarak ChatGPT-4, ChatGPT-4 Plus ve Google Gemini üzerinden soruldu. Cevaplar ürogenital travma konusunda deneyimli iki cerrah tarafından değerlendirildi ve ortalama puan hesaplanarak istatistiksel analize tabi tutuldu. **Bulgular:** Toplamda 27 soru çalışmaya dâhil edildi. Üretral travma yönetimi konusunda ChatGPT-4 ve 4 Plus, Google Gemini'ye göre daha başarılıyken, mesane travmasına yaklaşım konusunda Google Gemini ve ChatGPT-4 Plus istatistiksel olarak daha iyiydi ($p<0,001$). Totalde ChatGPT-4 ve Google Gemini sorulara %81,4 oranında doğru ve yeterli cevap verirken, ChatGPT-4 Plus'da bu oran %88,8 idi ($p=0,618$). **Sonuç:** Çalışmamız ChatGPT-4, 4 Plus ve Google Gemini'nin alt üriner sistem travma yönetiminde güvenilir bir kaynak olduğunu doğrulamaktadır. İlerleyen yıllarda acil tedavi planının belirlenmesi ve travmanın yönetiminde sağlık profesyonellerine yardımcı bir platform hâline gelebilir.

Anahtar Kelimeler: Alt üriner sistem travmaları; ChatGPT; Google Gemini

TO CITE THIS ARTICLE:
Çolakoğlu Y, Ayten A. Researching the reliability of ChatGPT-4, 4 plus and Google Gemini in responding to questions about the management of lower urinary tract trauma: Cross-sectional study. J Reconstr Urol. 2024;14(3):102-6.

Correspondence: Yunus ÇOLAKOĞLU
Bahçelievler Memorial Hospital, Clinic of Urology, İstanbul, Türkiye
E-mail: dr.yunusc@gmail.com

Peer review under responsibility of Journal of Reconstructive Urology.

Received: 24 Dec 2024 ***Received in revised form:*** 25 Dec 2024 ***Accepted:*** 27 Dec 2024 ***Available online:*** 30 Dec 2024

In recent years, the widespread use of artificial intelligence (AI) chatbots has changed the way we access education and information. Chatbots have been shown to be quite capable of synthesizing information from different sources, effectively sorting through multiple data sources to produce relevant and accurate answers. Due to the novelty of these large language models (LLM), there are still questions about their use in the medical field. Whether such systems can replace the vast knowledge and experience of trained doctors is a highly debatable question.[1]

The first AI program was developed in 1956 as a Dartmouth summer research project. Since that first description, the field has rapidly evolved and expanded. AI is now widely available and becoming more mainstream with the emergence of universally accessible programs. One of the widely recognized AI tools, Chat Generative Pre-Trained Transformer (ChatGPT, Openai, USA), has achieved quite popularity, reaching over one million users in the first five days of its launch. With the launch and growing popularity of ChatGPT, other competitive AI models were launched in response. On March 21, 2023, Google (Google Deepmind, California, USA) launched its first competitor called Google Bard (now Gemini).[2]

The outstanding success of these three LLMs in the field of health has been demonstrated in many studies.[3,4] Healthcare professionals also need information from time to time and are increasingly consulting AI programs. Lower urinary tract trauma is a very difficult and complex emergency to manage. First responding physicians often require consultation.[5] In this study, we aimed to investigate whether these three LLMs are reliable enough for the management of lower urinary tract trauma.

MATERIAL AND METHODS

The recommendation tables of the bladder, urethra and genital trauma management section of the European Association of Urology 2024 guidelines were analyzed. Those with strong and weak recommendation levels were translated into a questionnaire. Questions were asked in English via ChatGPT-4, ChatGPT-4 Plus and Google Gemini, and the answers were evaluated by two surgeons experienced in urogenital trauma. The different scores given to the answers were averaged and subjected to statistical analysis. Responses were scored 1-4 by the reviewers.

4: Correct and sufficient answer (no further information to add)

3: Correct answer but insufficient (more detailed explanation required)

2: Accurate and misleading information in one

1: Wrong or irrelevant answer

These three versions were then subjected to statistical analysis. Since real patient data were not used in the study, ethics committee approval was not required.

STATISTICAL ANALYSIS

Data analysis was performed using IBM SPSS Statistics for Windows, version 28.0 (Armonk, NY, USA). Categorical data are presented as n (%) and analyzed using Pearson's chi-square test to determine differences in accuracy between groups and subtopics. Before applying ANOVA, the assumptions of normality and homogeneity of variances were tested using the Shapiro-Wilk test and Levene's test, respectively. Since both assumptions were met ($p>0.05$), a one-way ANOVA was performed to compare group performance across multiple topics. Statistical significance was set at $p<0.05$.

RESULTS

A total of 27 questions were included in the study. The questions and the grading of the answers are shown in Table 1. LLMs did not give a score of 2 or 1 to any question. ChatGPT-4 and 4 Plus were better than Google Gemini in urethral trauma management, whereas Google Gemini and ChatGPT-4 Plus were statistically better in bladder trauma management ($p<0.001$). In total, ChatGPT-4 and Google Gemini gave 81.4% correct and sufficient answers to the questions, while this rate was 88.8% in ChatGPT-4 Plus. Statistical analysis showed no statistical difference between the responses of ChatGPT-4, ChatGPT-4 Plus and Google Gemini ($p=0.618$) Table 2.

TABLE 1: Scoring of ChatGPT-4, ChatGPT-4 Plus and Google Gemini asked questions and answers.

	ChatGPT-4	ChatGPT-4 Plus	Google Gemini
Urethral trauma management			
1. What to do to prevent traumatic catheterization?	4	4	4
2. How male urethral injuries should be evaluated at diagnosis?	4	4	4
3. How should female urethral injuries be evaluated in diagnosis?	4	4	4
4. How iatrogenic anterior urethral injuries should be treated?	4	4	4
5. How should partial blunt anterior urethral injuries be treated?	4	4	4
6. How complete blunt anterior urethral injuries should be treated?	4	4	4
7. In hemodynamically unstable patients, pelvic fracture-related urethral injuries (PFUI) how to treat in the beginning?	4	4	3
8. Should early endoscopic realignment be performed in male pelvic fracture-related urethral injuries (PFUI)?	3	3	3
9. Should endoscopic treatments be repeated after failed realignment for male PFUI?	4	4	4
10. How should partial posterior urethral injuries be treated?	4	4	4
11. Should urgent urethroplasty (<48 hours) be performed in male pelvic fracture associated urethral injuries (PFUI)?	3	3	3
12. In which case can early urethroplasty (two days to six weeks) be performed for male PFUIs with complete separation?	3	4	3
14. How should male PFUIs with complete posterior separation be treated?	4	4	4
Bladder injury management			
15. Which diagnostic tool should be used for bladder injury in the presence of macroscopic hematuria and pelvic fracture?	4	4	4
16. Which diagnostic method should be used in case of suspicion of iatrogenic bladder injury in the postoperative period?	4	4	4
17. How should cystography be performed?	4	4	4
18. What should be done to exclude bladder injury during retropubic sub-urethral sling procedures?	4	4	4
19. What should be the treatment approach in uncomplicated extraperitoneal bladder injuries after blunt trauma?	3	4	4
20. How should blunt extraperitoneal bladder injuries be treated when there is bladder neck involvement and/or associated injuries requiring surgical intervention?	3	3	3
21. How should blunt intraperitoneal injuries be treated?	4	4	4
22. How should small, uncomplicated, intraperitoneal bladder injuries that occur during endoscopic procedures be managed?	4	4	4
23. What should be done to assess bladder wall healing after repair of a complex injury or if there are risk factors for wound healing?	4	4	4
Genital trauma management			
24. Should urethral injury be evaluated in penile fracture cases?	4	4	4
25. What should be done in the diagnosis of testicular trauma?	4	4	4
26. What are the tissues that should be sutured in the surgical treatment of penile fractures?	4	4	4
27. What should be the treatment approach in cases of testicular trauma with testicular rupture and suspicious USG findings?	4	4	4

TABLE 2: Percentage of complete responses to Questions from ChatGPT-4, ChatGPT-4 Plus and Google Gemini by topics.

	ChatGPT-4	ChatGPT-4 Plus	Google Gemini	p value
Urethral trauma management	78.5%*	85.7%*	71.4%	<0.001
Bladder injury management	77.7%	88.8%+	88.8%+	<0.001
Genital trauma management	100%	100%	100%	0.998
Average	81.4%	88.8%	81.4%	0.618

p<0.05 from ANOVA indicates significant differences; *Different from Google Gemini, +Different from ChatGPT-4. This table has been prepared based on the numbered questions in Table 1.

DISCUSSION

The use of AI is increasing day by day in medicine and other fields. In recent research, AI tools have scored above the median score for the Medical College Admission Test and achieved passing scores on medical board exams, including cardiology and neurosurgery, as well as the United States Medical Licensing Examination step 1 exam.[6] LLMs answer questions with information based on previously published articles and books. This suggests that LLMs access quality, accurate information more frequently than other social media platforms.[7] In this study, we evaluated the extent to which these three popular AI applications are a reliable source for the approach to lower urinary tract trauma.

Lower urinary tract trauma may require urgent surgical intervention or may be approached conservatively in the presence of life-threatening additional organ injuries.[8] There is no clear consensus about urethral traumas in particular, which causes complexity in patient management.[9] It is thought that AI robots, which are currently being renewed day by day, can provide consultancy services to physicians who first examine and intervene in difficult situations.[10] In this respect, in the current study, we have shown that ChatGPT-4, 4 plus and Google Gemini are reliable sources.

In their published study, Mankowski et al. tested how ChatGPT could be used in kidney transplantation by comparing it with human participants. Nephrology residents and nephrology fellowship program directors were asked 12 multiple-choice questions about kidney transplantation on the American Society of Nephrology fellowship exam using ChatGPT versions 3.5, 4, 4 Visual (4 V). The study found that the 4V version performed as well as nephrology residents and fellowship program directors.[11] Cakir et al. reached an excellent accuracy of over 90% on the questions they posed to ChatGPT with female urology.[12] Another study evaluated ChatGPT-4 and Gemini on their performance in an ophthalmology exam and found that ChatGPT-4 was more successful.[13] Baturu et al. evaluated ChatGPT 3.5, ChatGPT-4 and BARD answers to frequently asked questions about erectile dysfunction. There was no significant difference between ChatGPT 3.5 and ChatGPT-4 in terms of answer quality, but both outperformed BARD.[14] In our study, although there were differences under sub-headings, there was no significant difference between the success of the three LLM versions. All three programs achieved a valid rating by performing above 80%. The ability of AI software to access different literature sources and its capacity to continuously improve itself are among the important factors in the high rate of correct responds.

Additionally, reproducibility is an issue to be considered in AI-supported programs. A high reproducibility rate of over 90% was observed in the answers to questions on andrology. In addition, the answers were in an easy-to-understand language.[15] One of the limitations of our study is that this repeatability was not evaluated. Another limitation is that a single surgeon scored the answers. On the other hand, to our knowledge, this is the first study to evaluate the reliability of three popular LLMs in lower urinary tract trauma.

CONCLUSION

Our study confirms that ChatGPT-4, 4 Plus and Google Gemini are a reliable resource in lower urinary tract trauma management. In the following years, it may become a platform to help healthcare professionals in determining the emergency treatment plan and managing trauma.

Source of Finance

During this study, no financial or spiritual support was received neither from any pharmaceutical company that has a direct connection with the research subject, nor from a company that provides or produces medical instruments and materials which may negatively affect the evaluation process of this study.

Conflict of Interest

No conflicts of interest between the authors and / or family members of the scientific and medical committee members or members of the potential conflicts of interest, counseling, expertise, working conditions, share holding and similar situations in any firm.

Authorship Contributions

Idea/Concept: *Yunus Çolakoğlu;* ***Design:*** *Yunus Çolakoğlu, Ali Ayten;* ***Control/Supervision:*** *Yunus Çolakoğlu;* ***Data Collection and/or Processing:*** *Yunus Çolakoğlu, Ali Ayten;* ***Analysis and/or Interpretation:*** *Yunus Çolakoğlu, Ali Ayten;* ***Literature Review:*** *Yunus Çolakoğlu, Ali Ayten;* ***Writing the Article:*** *Yunus Çolakoğlu, Ali Ayten;* ***Critical Review:*** *Yunus Çolakoğlu;* ***References and Fundings:*** *Yunus Çolakoğlu, Ali Ayten;* ***Materials:*** *Yunus Çolakoğlu, Ali Ayten.*

REFERENCES

1. Sezgin E. Artificial intelligence in healthcare: complementing, not replacing, doctors and healthcare providers. Digit Health. 2023;9:20552076231186520. [Crossref] [PubMed] [PMC]
2. Terwilliger E, Bcharah G, Bcharah H, Bcharah E, Richardson C, Scheffler P. Advancing medical education: performance of generative artificial intelligence models on otolaryngology board preparation questions with image analysis insights. Cureus. 2024;16(7):e64204. [Crossref] [PubMed] [PMC]
3. Caglar U, Yildiz O, Meric A, Ayranci A, Yusuf R, Sarilar O, et al. Evaluating the performance of ChatGPT in answering questions related to benign prostate hyperplasia and prostate cancer. Minerva Urol Nephrol. 2023;75(6):729-33. [Crossref] [PubMed]
4. Secinaro S, Calandra D, Secinaro A, Muthurangu V, Biancone P. The role of artificial intelligence in healthcare: a structured literature review. BMC Med Inform Decis Mak. 2021;21(1):125. [Crossref] [PubMed] [PMC]
5. Abedi AR, Hosseini J, Hojjati SA, Alinejad Khorram A, Nikmaram R, Fakhar F. Efficacy and complications of mitrofanoff continent urinary diversion in adults with complex urethral strictures: a single-center experience. Urol J. 2024;21(6):404-9. [PubMed]
6. Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepaño C, et al. Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. PLOS Digit Health. 2023;2(2):e0000198. [Crossref] [PubMed] [PMC]
7. Antaki F, Touma S, Milad D, El-Khoury J, Duval R. Evaluating the performance of ChatGPT in ophthalmology: an analysis of its successes and shortcomings. Ophthalmol Sci. 2023;3(4):100324. [Crossref] [PubMed] [PMC]
8. Mahat Y, Leong JY, Chung PH. A contemporary review of adult bladder trauma. J Inj Violence Res. 2019;11(2):101-6. [Crossref] [PubMed] [PMC]
9. Song BJ, Wang Q, Feng W, Zhu DJ, Zhang XJ. Comparison of pediatric pelvic fractures and associated injuries caused by different types of road traffic accidents. Chin J Traumatol. 2024;27(6):372-9. [Crossref] [PubMed] [PMC]
10. Echefu G, Batalik L, Lukan A, Shah R, Nain P, Guha A, et al. The digital revolution in medicine: applications in cardio-oncology. Curr Treat Options Cardiovasc Med. 2025;27(1). [Crossref] [PubMed]
11. Mankowski MA, Jaffe IS, Xu J, Bae S, Oermann EK, Aphinyanaphongs Y, et al. ChatGPT solving complex kidney transplant cases: a comparative study with human respondents. Clin Transplant. 2024;38(10):e15466. [Crossref] [PubMed] [PMC]
12. Cakir H, Caglar U, Halis A, Sarilar O, Yazili HB, Ozgor F. Assessing the knowledge of ChatGPT in answering questions regarding female urology. Urol J. 2024;21(6):410-4. [PubMed]
13. Gill GS, Tsai J, Moxam J, Sanghvi HA, Gupta S. Comparison of Gemini Advanced and ChatGPT 4.0's Performances on the Ophthalmology Resident Ophthalmic Knowledge Assessment Program (OKAP) Examination Review Question Banks. Cureus. 2024;16(9):e69612. [Crossref]
14. Baturu M, Solakhan M, Kazaz TG, Bayrak O. Frequently asked questions on erectile dysfunction: evaluating artificial intelligence answers with expert mentorship. Int J Impot Res. 2024. [Crossref] [PubMed]
15. Caglar U, Yildiz O, Ozervarli MF, Aydin R, Sarilar O, Ozgor F, et al. Assessing the performance of chat generative pretrained transformer (ChatGPT) in answering andrology-related questions. Urol Res Pract. 2023;49(6):365-9. [PubMed] [PMC]