ORIGINAL ARTICLE

ChatGPT's competence in responding to urological emergencies

Mazhar Ortaç,[1] Rıfat Burak Ergül,[1] Hüseyin Burak Yazılı,[2] Muhammet Firat Özervarlı,[1] Şenol Tonyalı,[1] Omer Sarılar,[2] Faruk Özgör[2]

[1]Department of Urology, Istanbul Faculty of Medicine, Istanbul University, Istanbul-*Türkiye*
[2]Department of Urology, Haseki Training and Research Hospital, Istanbul-*Türkiye*

ABSTRACT

BACKGROUND: In recent years, artificial intelligence (AI) applications have been increasingly used as sources of medical information, alongside their applications in many other fields. This study is the first to evaluate ChatGPT's performance in addressing urological emergencies (UE).

METHODS: The study included frequently asked questions (FAQs) by the public regarding UE, as well as UE-related questions formulated based on the European Association of Urology (EAU) guidelines. The FAQs were selected from questions posed by patients to doctors and hospital accounts on social media platforms (Facebook, Instagram, and X) and on websites. All questions were presented to ChatGPT 4 (premium version) in English, and the responses were recorded. Two urologists assessed the quality of the responses using a Global Quality Score (GQS) on a scale of 1 to 5.

RESULTS: Of the 73 total FAQs, 53 (72.6%) received a GQS score of 5, while only two (2.7%) received a GQS score of 1. The questions with a GQS score of 1 pertained to priapism and urosepsis. The topic with the highest proportion of responses receiving a GQS score of 5 was urosepsis (82.3%), whereas the lowest scores were observed in questions related to renal trauma (66.7%) and postrenal acute kidney injury (66.7%). A total of 42 questions were formulated based on the EAU guidelines, of which 23 (54.8%) received a GQS score of 5 from the physicians. The mean GQS score for FAQs was 4.38±1.14, which was significantly higher (p=0.009) than the mean GQS score for EAU guideline-based questions (3.88±1.47).

CONCLUSION: This study demonstrated for the first time that nearly three out of four FAQs were answered accurately and satisfactorily by ChatGPT. However, the accuracy and proficiency of ChatGPT's responses significantly decreased when addressing guideline-based questions on UE.

Keywords: Artificial intelligence; ChatGPT; urological emergencies.

INTRODUCTION

Urological emergencies (UE) are defined as acute and unexpected pathologies of the urinary system and male genital organs that require immediate medical intervention. Delayed and/or inadequate management of UE can result in increased morbidity, organ loss, and mortality.[1] Bun et al.[2] analyzed data from a tertiary academic center over a five-year period and concluded that 357 of 15,834 patients admitted to the emergency department had urological disorders. In another study, Ndiaye et al.[3] reviewed patient charts from emergency admissions over three years, finding that 300 patients presented with urological complaints, averaging 8.3 admissions per month. Additionally, factors such as regional variations in patient reliance on emergency services, patients' social security status, and the adequacy of healthcare unit recording systems

Cite this article as: Ortaç M, Ergül RB, Yazılı HB, Özervarlı MF, Tonyalı Ş, Sarılar Ö, et al. ChatGPT's competence in responding to urological emergencies. Ulus Travma Acil Cerrahi Derg 2025;31:291-295.

Address for correspondence: Mazhar Ortaç
Department of Urology, Istanbul Faculty of Medicine, Istanbul University, Istanbul, Türkiye
E-mail: mazhar.ortac@istanbul.edu.tr

Ulus Travma Acil Cerrahi Derg 2025;31(3):291-295 DOI: 10.14744/tjtes.2024.03377
Submitted: 20.08.2024 Revised: 27.11.2024 Accepted: 30.12.2024 Published: 03.03.2025
OPEN ACCESS This is an open access article under the CC BY-NC license (http://creativecommons.org/licenses/by-nc/4.0/).

may influence the reported prevalence of urological emergencies. Today, many individuals rely on internet sources to gather information about their symptoms. Developing applications that provide assistance before individuals seek emergency care for urological complaints may help reduce emergency department congestion.[4] However, in recent years, the rapid adoption of artificial intelligence (AI) tools such as ChatGPT in healthcare has raised concerns about the reliability and accuracy of AI-generated medical information, particularly on social media platforms. The lack of a standardized control mechanism for evaluating the accuracy of shared content increases the risk of disseminating misleading information, which can negatively impact patient outcomes.

ChatGPT (OpenAI, California, USA) is a new artificial intelligence application that functions as a multilingual chatbot. [5] The competence and reliability of ChatGPT in addressing medical conditions is one of the most widely discussed topics today, and the advantages and limitations of its use in medicine are still under investigation. A study evaluating ChatGPT's responses to questions about pediatric urology found that ChatGPT provided satisfactory and accurate answers for nine out of ten public inquiries.[6] In contrast, Ozgor et al.[7] reported that ChatGPT had limited capacity to answer scientific questions about urological cancers, with an accuracy and proficiency rate of only 60% for such inquiries.

Although a limited number of studies have analyzed ChatGPT's performance in addressing urological diseases, to our knowledge, no study has evaluated its ability to answer questions related to UE. In this study, for the first time, we aimed to assess ChatGPT's performance in responding to UE-related inquiries.

MATERIALS AND METHODS

Frequently asked questions (FAQs) from the public regarding UE, as well as UE-related questions formulated based on the European Association of Urology (EAU) guidelines, were included in the study (Supplements 1 and 2). FAQs were selected from questions posed by patients to doctors and hospital accounts on social media platforms (Facebook, Instagram, and X) and various websites. The questions were categorized into the following subject groups: acute scrotum, priapism, postrenal acute kidney injury, urosepsis, renal trauma, and hematuria. Non-medical questions, redundant questions with identical meanings, grammatically inappropriate questions, and subjective personal inquiries (e.g., I have kidney stones, will I develop kidney failure?) were excluded from the study. Additionally, strongly recommended information from the relevant sections of the EAU 2023 guidelines on these topics was translated into question format for inclusion in the study.

All questions were presented to ChatGPT-4 (premium version) in English, and its responses were recorded. The answers were independently evaluated by two expert urologists with more than 10 years of experience in the field. The experts assessed the responses using the Global Quality Score (GQS) on a scale of 1 to 5.[8] For questions where both physicians assigned the same score, the common score was recorded. In cases where the scores differed, the median score was recorded as the final score. Since no patient data were used in this study, ethics committee approval was not required.

Global Quality Score

The GQS is a five-level scale developed by Langille et al. in 2010 to assess the quality of medical information materials. Level 1 represents the lowest quality (content with limited benefit to patients), while Level 5 represents the highest quality (very useful to patients and excellent quality).

Statistical Analysis

Statistical analysis was performed using the Statistical Package for the Social Sciences, version 27 (SPSS, IBM Corp., Armonk, NY, USA). The number of questions according to GQS score groups is presented as n (%). GQS scores for FAQs and EAU guideline-based questions were compared using the independent Student's t-test. Data were analyzed at a 95% confidence level, and a p value of less than 0.05 was considered statistically significant.

RESULTS

A flowchart depicting the FAQs included in the study is shown in Figure 1. Of the 122 questions evaluated, 49 were excluded from the study for not meeting the inclusion criteria.

GQS scores for FAQs across different categories are summarized in Table 1. Of the total 73 questions, 53 (72.6%) received a GQS score of 5, while only two questions (2.7%) received a GQS score of 1. The questions with a GQS score of 1 were related to priapism and urosepsis. The category

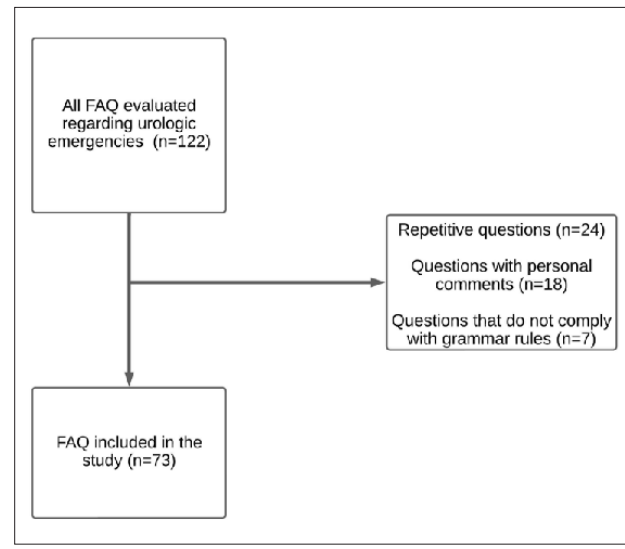

Figure 1. Study flowchart.

Table 1. Global Quality Scores for answers to frequently asked questions

GQS	1	2	3	4	5
Urologic Emergencies (n=73)	2 (2.7%)	7 (9.6%)	5 (6.9%)	6 (8.2%)	53 (72.6%)
Acute Scrotum (n=14)	-	2 (14.3%)	1 (7.1%)	1 (7.1%)	10 (71.5%)
Priapism (n=13)	1 (7.7%)	2 (15.4%)	-	2 (16.7%)	9 (69.2%)
Postrenal Acute Kidney Injury (n=12)	-	1 (8.3%)	1 (8.3%)	2 (16.7%)	8 (66.7%)
Urosepsis (n=17)	1 (5.9%)	-	2 (11.8%)	-	14 (82.3%)
Renal Trauma (n=9)	-	1 (11.1%)	-	2 (22.2%)	6 (66.7%)
Hematuria (n=8)	-	1 (12.5%)	1 (12.5%)	-	6 (75.0%)

GQS: Global Quality Score.

Table 2. Global Quality Scores for responses to questions based on the European Association of Urology (EAU) guideline recommendations

GQS	1	2	3	4	5
Urologic Emergencies (n=42)	4 (9.5%)	7 (16.6%)	2 (4.8%)	6 (14.3%)	23 (54.8%)

GQS: Global Quality Score; EAU: European Association of Urology.

with the highest proportion of responses receiving a GQS score of 5 was urosepsis (82.3%), while the lowest scores were observed in renal trauma (66.7%) and postrenal acute kidney injury (66.7%). For other topics, the percentage of questions receiving a GQS score of 5 was 71.5% for acute scrotum, 69.2% for priapism, and 75% for hematuria.

A total of 42 questions were created based on the EAU guidelines. Of these, 23 (54.8%) received a GQS score of 5 from the physicians. The distribution of other ratings was as follows: GQS score of 1: four questions (9.5%), GQS score of 2: seven questions (16.6%), GQS score of 3: two questions (4.8%), and GQS score of 4: six questions (14.3%) (Table 2). A comparison of mean GQS scores between FAQs and EAU guideline-based questions is provided in Table 3. The mean GQS score for FAQs was 4.38±1.14, which was statistically higher (p=0.009) than the mean GQS score for EAU guideline-based questions (3.88±1.47) (Table 3).

Table 3. Comparison of Global Quality Scores (GQS) for frequently asked questions (FAQs) and European Association of Urology (EAU) guideline questions

GQS Score	FAQ	EAU Guideline	p value
Urologic Emergencies	4.38±1.14	3.88±1.47	0.009

GQS: Global Quality Score; FAQ: Frequently Asked Questions; EAU: European Association of Urology.

DISCUSSION

AI is being increasingly integrated into various industries and aspects of daily life, and its application in medicine has become a highly intriguing topic for both the public and healthcare professionals. Previous studies have highlighted several potential benefits of AI in medicine, including reducing the time between symptom onset and diagnosis, enhancing the targeting of screening programs for specific populations, and improving patient adherence to treatment regimens. Additionally, the economic burden on the healthcare system can decrease, and redundant hospital admissions can be reduced with the use of AI.[9,10] However, some authors have expressed concerns regarding the adequacy and reliability of ChatGPT. Therefore, we conducted this study to analyze ChatGPT's responses to UE-related questions. Our findings revealed that ChatGPT provided satisfactory and accurate responses to three out of four FAQs about UE. However, this accuracy rate dropped to 54.76% when responding to guideline-based UE inquiries.

Internet-based applications have access to vast amounts of health information; however, many do not evaluate the accuracy and adequacy of the uploaded data. Betschart et al.[11] assessed the quality of YouTube videos about benign prostatic hyperplasia and concluded that these videos were generally of low quality, often containing misinformation and commercial bias. Similarly, Alsyouf et al.[12] analyzed urological cancer-related content on social media platforms and found that misleading information was significantly more prevalent than reliable data. In contrast, Alasker et al.[13] investigated

AI applications in answering prostate cancer-related questions and emphasized that ChatGPT provided understandable, accurate, and reliable responses about prostate cancer. Likewise, Cakır et al.[14] found that ChatGPT correctly and adequately answered 19 out of 20 inquiries regarding urinary system stone disease. In the present study, for the first time, we demonstrated that ChatGPT was capable of correctly and adequately answering more than seven out of ten FAQs about UE. The lack of a system for verifying the accuracy and reliability of shared content leads to the spread of misleading information on many social media platforms. ChatGPT can access numerous internet-based resources, including scientific articles, books, newspapers, etc., and we believe that its ability to scan this vast database contributes to the high accuracy and proficiency of its responses regarding UE.

Guidelines include specific recommendations for practitioners. In the process of developing guidelines, numerous academic studies, including reviews, meta-analyses, and original articles, are evaluated. The capability of ChatGPT to respond to inquiries involving complex scientific guidelines is debatable. Caglar et al.[6] reported a 93.6% accuracy rate when ChatGPT was used to answer pediatric urology guideline-based questions. In contrast, Ozgor et al.[7] found that the accuracy and proficiency of ChatGPT's responses significantly declined when answering urological cancer guideline-based inquiries compared to public inquiries. In this study, ChatGPT provided accurate and satisfactory responses to 54.76% of guideline-based UE questions. Additionally, the quality of ChatGPT's responses was significantly lower, as measured by the GQS, compared to its responses to public FAQs. Urological emergencies are less common than many other medical emergencies, and their heterogeneous nature may have contributed to ChatGPT's insufficiency in answering guideline-based questions about UE. The absence of a standardized system for validating AI-generated medical content has led to the widespread dissemination of inaccurate and potentially harmful information on social media platforms. We recommend integrating a multidisciplinary, physician-led validation mechanism into AI systems to enhance the accuracy and reliability of responses, particularly in applications intended for public use.

Although this study is the first to evaluate ChatGPT's ability to respond to inquiries about UE, it has some limitations. One limitation is the use of only two urologists to evaluate ChatGPT's responses. Urological emergencies often intersect with other medical specialties, such as nephrology, emergency medicine, and infectious diseases. Future studies could benefit from employing a multidisciplinary team to provide a more comprehensive evaluation of AI responses, particularly in scenarios requiring diverse expertise. Another concern is the lack of a standardized system to validate AI-generated medical content, which contributes to the spread of inaccurate and potentially harmful information on social media platforms. We recommend integrating a multidisciplinary, physician-led validation mechanism into AI systems to improve the accuracy and reliability of responses, particularly in applications intended for public use.

The study was conducted in English, which is the most commonly used language in academia and one of the most widely spoken languages in the world. However, many people access the internet in languages other than English. The accuracy of ChatGPT's responses regarding UE in languages other than English should be explored in future studies. Additionally, ChatGPT's knowledge is limited to the information available up to the date of the study, while a large amount of UE-related information continues to be uploaded to the internet daily. Furthermore, the GQS evaluation, while widely used for assessing the quality of medical information, inherently involves subjective judgments. To mitigate this, the evaluators adhered to predefined scoring criteria and reviewed responses independently. Nonetheless, differences in interpretation are unavoidable, underscoring the need for standardized evaluation frameworks in future research.

CONCLUSION

The present study demonstrated for the first time that nearly three out of four FAQs were accurately and satisfactorily answered by ChatGPT. In contrast, the accuracy and proficiency of ChatGPT's responses significantly decreased when answering guideline-based questions about UE. The findings of this study revealed that the use of ChatGPT in urology practice provides the public with information about UE. However, ChatGPT should be further updated and refined for professional healthcare applications related to UE before being implemented in urology practice.

Ethics Committee Approval: No patient data was used in our study and therefore ethics committee approval was not required.

Peer-review: Externally peer-reviewed.

Authorship Contributions: Concept: M.O., R.B.E.; Design: M.O., F.O.; Supervision: O.S., F.O.; Resource: R.B.E.,H.B.Y., S.T.; Materials: R.B.E., M.F.O.; Data Collection and/or Processing: H.B.Y., M.F.Ö., S.T.; Analysis and/or Interpretation: M.O.; Literature Review: M.O.; Writing: M.O.; Critical Review: O.S., F.O.

Conflict of Interest: None declared.

Financial Disclosure: The author declared that this study has received no financial support.

REFERENCES

1. Manjunath AS, Hofer MD. Urologic emergencies. Med Clin North Am 2018;102:373–85. [CrossRef]
2. Bun E, Edeh AJ, Umeji EI. Urological emergencies at ESUT Teaching Hospital, Parklane-Enugu: A five year review. Niger J Urol 2017;7:9–12.
3. Ndiaye M, Sow O, Sarr A, Ndiath A, Ondo CZ, Sine B, et al. Urological emergency in a university hospital setting: Epidemiological, diagnostic and therapeutic aspects. Int J Clin Urol 2020;4:51–4. [CrossRef]

4. Chen J, Wang Y. Social media use for health purposes: Systematic review. J Med Internet Res 2021;23:e17917. [CrossRef]
5. Zhou Z, Wang X, Li X, Liao L. Is ChatGPT an evidence-based doctor? Eur Urol 2023;84:355–6. [CrossRef]
6. 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–e5. [CrossRef]
7. Ozgor F, Caglar U, Halis A, Cakir H, Aksu UC, Ayranci A, et al. Urological cancers and ChatGPT: Assessing the quality of information and possible risks for patients. Clin Genitourin Cancer 2024;22:454–7.e4. [CrossRef]
8. Yuce A, Oto O, Vural A, Misir A. YouTube provides low-quality videos about talus osteochondral lesions and their arthroscopic treatment. Foot Ankle Surg 2023;29(5):441–5. [CrossRef]
9. Zhou Z. Evaluation of ChatGPT's capabilities in medical report generation. Cureus 2023;15:e37589. [CrossRef]
10. Xiao D, Meyers P, Upperman JS, Robinson JR. Revolutionizing healthcare with ChatGPT: An early exploration of an AI language model's impact on medicine at large and its role in pediatric surgery. J Pediatr Surg 2023;58:2410–5. [CrossRef]
11. Betschart P, Pratsinis M, Müllhaupt G, Rechner R, Herrmann TR, Gratzke C, et al. Information on surgical treatment of benign prostatic hyperplasia on YouTube is highly biased and misleading. BJU Int 2020;125:595–601. [CrossRef]
12. Alsyouf M, Stokes P, Hur D, Amasyali A, Ruckle H, Hu B. 'Fake News' in urology: Evaluating the accuracy of articles shared on social media in genitourinary malignancies. BJU Int 2019;124(4):701–6. [CrossRef]
13. Alasker A, Alsalamah S, Alshathri N, Almansour N, Alsalamah F, Alghafees M, et al. Performance of large language models (LLMs) in providing prostate cancer information. BMC Urol 2024;24(1):177. [CrossRef]
14. Cakir H, Caglar U, Yildiz O, Meric A, Ayranci A, Ozgor F. Evaluating the performance of ChatGPT in answering questions related to urolithiasis. Int Urol Nephrol 2024;56(1):17–21. [CrossRef]

ORİJİNAL ÇALIŞMA - *ÖZ*

Ürolojik acil durumlarda ChatGPT'nin yanıt yetkinliği

AMAÇ: Son yıllarda, yapay zekâ (AI) uygulamaları tıpta ve birçok diğer alanda bir bilgi kaynağı olarak kullanılmaktadır. Bu çalışma, ChatGPT'nin ürolojik aciller (ÜA) konusunda gösterdiği performansı değerlendiren ilk çalışmadır.
GEREÇ VE YÖNTEM: Çalışma, halk tarafından ürolojik acillerle ilgili sıkça sorulan soruları (SSS) ve Avrupa Üroloji Derneği (EAU) kılavuzlarını incelenerek oluşturulan ürolojik acillerle ilgili soruları içermektedir. SSS, sosyal medya (Facebook, Instagram ve X) veya doktor / hastane web sayfalarında halk tarafından sorulan sorular arasından seçilmiştir. Tüm sorular İngilizce olarak ChatGPT 4 (Premium versiyonu) ile sorulmuş ve cevaplar kaydedilmiştir. İki ürolog, yanıtları global kalite puanı (GQS) skalasına göre 1-5 puan arasında değerlendirmiştir.
BULGULAR: Toplam 73 yanıtın 53'ü (%72.6) 5 GQS puanına sahipti ve yalnızca 2 yanıt (%2.7) 1 GQS puanına sahipti. 1 GQS puanına sahip yanıtlar priapizm ve ürosepsis ile ilgiliydi. En yüksek GQS puanına (%82.3) sahip konu ürosepsis iken, en düşük puanlar renal travma (%66.7) ve postrenal akut böbrek 15 hasarı konularındaydı (%66.7). EAU kılavuzuna dayalı olarak oluşturulan soru sayısı 42 idi. Bu sorulara oluşturulan yanıtların 23'ü (%54.8) hekimlerden 5 GQS puanı aldı. SSS'ye yönelik yanıtlar için GQS ortalama puanı 4.38±1.14 idi ve bu, EAU kılavuzuna dayalı sorular için ortalama GQS puanından (3.88±1.47) istatistiksel olarak daha yüksekti (p=0.009).
SONUÇ: Bu çalışma, ilk kez ChatGPT'nin SSS'lerin yaklaşık dörtte üçünü doğru ve tatmin edici bir şekilde yanıtladığını göstermiştir. Buna karşılık, ÜA hakkında kılavuz temelli soruları yanıtlarken ChatGPT'nin doğruluğu ve yetkinliği önemli ölçüde azalmıştır.

Anahtar sözcükler: ChatGPT; ürolojik acil durumlar; yapay zeka.

Ulus Travma Acil Cerrahi Derg 2025;31(3):291-295 DOI: 10.14744/tjtes.2024.03377