JOURNAL OF
CONTEMPORARY MEDICINE

DOI:10.16899/jcm.1741131
J Contemp Med 2025;15(5):216-220

Original Article / Orijinal Araştırma

Artificial Intelligence in Pediatric Urology: Accuracy and Consistency of ChatGPT's Responses on Hypospadias

Pediatrik Ürolojide Yapay Zeka: ChatGPT'nin Hipospadias Konusundaki Yanıtlarının Doğruluğu ve Tutarlılığı

Emre Kandemir[1], Mehmet Sarikaya[2]

[1]Department of Urology, Faculty of Medicine, Karamanoğlu Mehmetbey University, Karaman, Türkiye
[2]Department of Pediatric Surgery, Faculty of Medicine, Selcuk University, Konya, Türkiye

Abstract

Aim: This study aimed to evaluate the accuracy and reproducibility of ChatGPT (GPT-4-turbo) responses to frequently asked questions regarding hypospadias, a common congenital urological condition. As artificial intelligence (AI) becomes increasingly integrated into patient education, its reliability in delivering sensitive and clinically relevant information warrants empirical investigation.

Material and Method: Frequently asked questions about hypospadias were compiled from pediatric urology association websites, public health portals, and social media platforms. Questions were classified into five categories: general information, diagnosis, treatment, follow-up, and guideline-based recommendations. After excluding duplicate, vague, or subjective questions, 97 unique items were entered into ChatGPT. Two independent pediatric urologists rated the answers on a four-point scale (1 = completely correct, 4 = completely incorrect), and responses were repeated on separate devices to assess reproducibility.

Results: Of the 97 responses, 87.6% were graded as completely correct, 7.2% as correct but insufficient, 4.1% as partially misleading, and 1.0% as completely incorrect. The highest rate of accurate answers was observed in the diagnosis and follow-up categories (90.0%), while treatment-related questions showed slightly lower accuracy (86.7%). Guideline-based questions were answered correctly in 87.5% of cases. Overall reproducibility across all categories was 91.7%, with the highest consistency in diagnostic responses.

Conclusions: ChatGPT demonstrated high accuracy and reproducibility in answering patient-centered questions related to hypospadias, particularly in diagnosis and general information domains. However, variability in treatment-related content and limitations in referencing highlight the importance of cautious interpretation. While AI may serve as a supplementary educational tool in pediatric urology, clinical oversight remains essential to ensure safe and reliable information dissemination.

Keywords: Hypospadias, artificial intelligence, ChatGPT, pediatric urology

Öz

Amaç: Bu çalışma, yaygın bir konjenital ürolojik durum olan hipospadias ile ilgili sık sorulan sorulara ChatGPT (GPT-4-turbo) yanıtlarının doğruluğunu ve tekrarlanabilirliğini değerlendirmeyi amaçlamıştır. Yapay zeka (AI) hasta eğitimine giderek daha fazla entegre hale geldikçe, hassas ve klinik olarak ilgili bilgileri sağlamadaki güvenilirliği ampirik araştırmayı gerektirmektedir.

Gereç ve Yöntem: Hipospadias hakkında sıkça sorulan sorular, pediatrik üroloji derneği web sitelerinden, halk sağlığı portallarından ve sosyal medya platformlarından derlenmiştir. Sorular beş kategoride sınıflandırıldı: genel bilgi, tanı, tedavi, takip ve kılavuza dayalı öneriler. Mükerrer, belirsiz veya öznel sorular elendikten sonra 97 benzersiz soru ChatGPT'ye girilmiştir. İki bağımsız pediatrik ürolog yanıtları dört puanlık bir ölçekte (1 = tamamen doğru, 4 = tamamen yanlış) değerlendirdi ve yanıtlar tekrarlanabilirliği değerlendirmek için ayrı cihazlarda tekrarlandı.

Bulgular: 97 yanıtın %87,6'sı tamamen doğru, %7,2'si doğru ancak yetersiz, %4,1'i kısmen yanıltıcı ve %1,0'ı tamamen yanlış olarak derecelendirildi. En yüksek doğru cevap oranı tanı ve takip kategorilerinde gözlenirken (%90,0), tedavi ile ilgili sorular biraz daha düşük doğruluk oranı göstermiştir (%86,7). Kılavuza dayalı sorular vakaların %87,5'inde doğru yanıtlanmıştır. Tüm kategorilerdeki genel tekrarlanabilirlik %91,7 olup, en yüksek tutarlılık tanısal yanıtlarda görülmüştür.

Sonuç: ChatGPT, özellikle tanı ve genel bilgi alanlarında olmak üzere hipospadias ile ilgili hasta merkezli soruları yanıtlamada yüksek doğruluk ve tekrarlanabilirlik göstermiştir. Bununla birlikte, tedaviyle ilgili içerikteki değişkenlik ve referans vermedeki sınırlamalar dikkatli yorumlamanın önemini vurgulamaktadır. Yapay zeka pediatrik ürolojide tamamlayıcı bir eğitim aracı olarak hizmet edebilirken, güvenli ve güvenilir bilgi yayılımını sağlamak için klinik gözetim gerekli olmaya devam etmektedir.

Anahtar Kelimeler: Hipospadias, yapay zeka, ChatGPT, pediyatrik üroloji

Corresponding *(İletişim)*: Mehmet Sarıkaya, Department of Pediatric Surgery, Faculty of Medicine, Selcuk University, Konya, Türkiye
E-mail *(E-posta)*: drmehmetsarikaya@hotmail.com
Received *(Geliş Tarihi)*: 13.07.2025 **Accepted *(Kabul Tarihi)*:** 02.09.2025

INTRODUCTION

Hypospadias, a common congenital anomaly characterized by an ectopic location of the urethral meatus along the ventral aspect of the penis, presents a spectrum of surgical and psychosocial challenges. Despite advances in surgical techniques and postoperative care, variability in treatment strategies and outcomes persists.[1] As such, caregivers of affected children frequently seek supplementary medical information beyond clinical consultations, often turning to digital platforms for accessible guidance.[2]

In parallel with the growing reliance on artificial intelligence (AI) technologies in healthcare, natural language processing models such as ChatGPT have gained significant attention as potential tools for patient education. These systems generate human-like responses to user inquiries, offering real-time and conversational access to a vast body of knowledge. However, the reliability and consistency of AI-generated medical content, particularly for sensitive topics like congenital urological anomalies, remain uncertain and warrant empirical investigation.[3]

Hypospadias is a relatively common congenital anomaly, with an estimated incidence of approximately 1 in every 250 live male births.[4] Numerous surgical techniques have been described for its correction, reflecting the heterogeneity and complexity of the condition. Postoperative recurrence rates vary depending on the severity of the hypospadias but may approach 30% in more severe cases.[5] Given these challenges, it is not uncommon for parents to seek more detailed information regarding treatment options. While traditional search engines were once the primary source of such information, they are increasingly being replaced by artificial intelligence–powered platforms, which offer more interactive and dynamic access to health-related knowledge.

This study aims to evaluate the accuracy and reproducibility of ChatGPT's responses to a broad set of frequently asked questions related to hypospadias. By systematically assessing the quality of information provided by the model across multiple domains—general knowledge, diagnosis, treatment, follow-up, and guideline-based recommendations—we seek to determine the model's utility as a potential adjunct in the dissemination of pediatric urological information.

MATERIAL AND METHOD

Ethic Statement

Since this study involved the analysis of responses generated by an artificial intelligence model (ChatGPT) to publicly available and anonymized questions, and did not include any human participants, patient data, or identifiable personal information, ethical approval was not required in accordance with institutional and international research ethics guidelines. Since no patient data were used in the study, informed consent was not required.

Commonly asked questions regarding hypospadias were collected from the official websites of pediatric urology associations, university hospitals, and public health portals. Additionally, patient and caregiver questions posted on social media platforms (such as YouTube and Facebook) were reviewed. Evidence-based recommendations with strong recommendation levels from the European Association of Urology (EAU) 2024 Guidelines on Pediatric Urology and Hypospadias Management were used as reference points.

Questions that were repetitive, grammatically unclear, or that invited overly subjective interpretation were excluded from the analysis. All included questions were curated by two board-certified pediatric urologists, each with a minimum of five years of independent clinical experience in hypospadias surgery. Questions were categorized into four domains: general information, diagnosis, treatment options (including surgical techniques), and postoperative care/outcomes.

Each question was entered sequentially into ChatGPT (GPT-4-turbo, OpenAI) and the responses were recorded. All questions were submitted to ChatGPT (GPT-4-turbo, OpenAI) between March 1 and March 15, 2025, using separate browser sessions. Responses were independently evaluated by two pediatric urologists using a 4-point rating scale:

1 = Completely correct
2 = Correct but insufficient
3 = Contains both correct and misleading information
4 = Completely incorrect

When both evaluators assigned the same score to a response, the shared score was recorded. Each question was asked twice using separate devices to assess response reproducibility. Reproducibility was defined as consistency in the score category and informational depth of the two responses. If a repeated answer received a different score or varied significantly in content detail, it was considered non-reproducible. Discrepancies between evaluators or inconsistencies in reproducibility were reviewed and resolved by a third senior pediatric urologist. Questions based on guideline-derived strong recommendations were also evaluated for alignment with the original guideline content.

Statistical Analysis

All data analyses were performed using IBM SPSS Statistics for Windows, Version 25.0 (IBM Corp., Armonk, NY, USA). Questions related to hypospadias were evaluated within their respective categories. Scores assigned to ChatGPT responses were expressed as percentages and presented descriptively.

RESULTS

The flowchart of the question selection process for inclusion in the study is presented in **Figure 1**. Of the 160 hipospadias-related questions initially assessed for eligibility, 28 duplicate questions, 20 grammatically incorrect or ambiguous questions, and 15 non-clinical or subjectively interpretable questions were excluded. Consequently, a total of 97 questions were included in the final analysis as shown in the flow diagram in **Figure 1**.

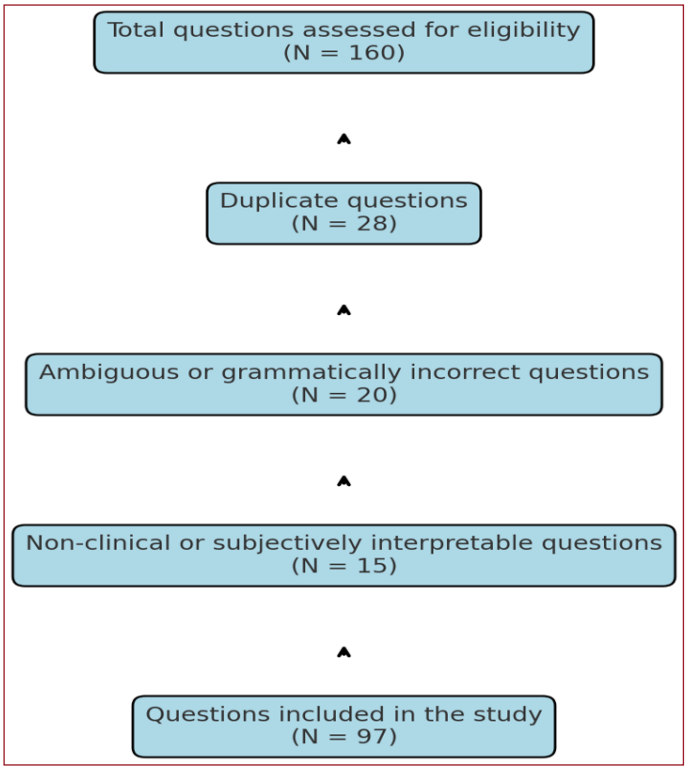

Figure 1. Flow diagram of questions evaluated and included in the analysis

The answers provided by ChatGPT were graded by two independent, board-certified pediatric urologists with at least five years of clinical experience, using a predefined four-point accuracy scale. The overall distribution of response quality is summarized in **Table 1**. Of the 97 total questions, 87.6% were graded as completely correct (Grade 1), 7.2% as correct but insufficient (Grade 2), 4.1% as partially misleading (Grade 3), and 1.0% as completely incorrect (Grade 4). Among the subcategories, the highest proportion of Grade 1 responses was observed in the diagnosis (90.0%) and prevention/follow-up (90.0%) categories. The lowest Grade 1 rate was noted in the treatment-related questions (86.7%). For questions derived from established guideline recommendations, 35 out of 40 (87.5%) were answered completely correctly.

Table 1. Grade of Responses by ChatGPT to Questions Related to Hipospadias

Category	Grade 1 (Completely correct)	Grade 2 (Correct but insufficient)	Grade 3 (Partially misleading)	Grade 4 (Completely incorrect)
All questions (N=97)	85 (87.6%)	7 (7.2%)	4 (4.1%)	1 (1.0%)
General information (N=25)	22 (88.0%)	2 (8.0%)	1 (4.0%)	0
Diagnosis (N=20)	18 (90.0%)	1 (5.0%)	1 (5.0%)	0
Treatment (N=30)	26 (86.7%)	2 (6.7%)	1 (3.3%)	1 (3.3%)
Prevention/ Follow-up (N=10)	9 (90.0%)	1 (10.0%)	0	0
Guideline-based (N=40)	35 (87.5%)	3 (7.5%)	1 (2.5%)	1 (2.5%)

To assess the reproducibility of ChatGPT responses, each question was submitted twice on different computers. The similarity rates of responses by category are presented in **Figure 2**. The overall reproducibility rate was 91.7%. The highest reproducibility was observed in the diagnosis category (95.0%), followed by prevention/follow-up (93.0%), general information (91.0%), and treatment (89.5%). For guideline-based questions, a similarity rate of 88.0% was recorded.

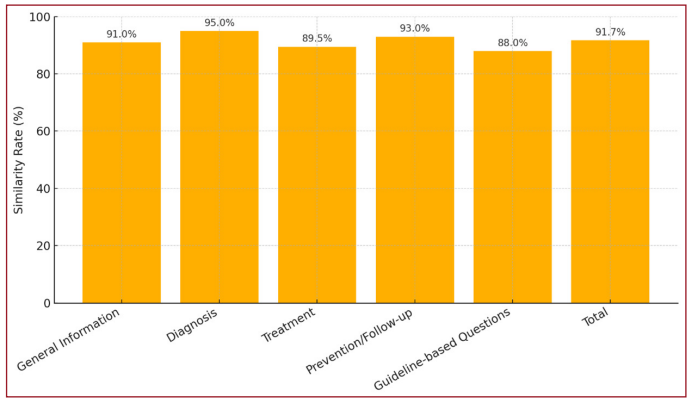

Figure 2. Analysis of the similarity in answers to questions

DISCUSSION

This study aimed to evaluate the factual accuracy and reproducibility of ChatGPT's responses to patient-oriented questions concerning hypospadias. The high proportion of responses graded as completely correct (87.6%) highlights ChatGPT's potential as an accessible and informative digital tool in the field of pediatric urology. Particularly in the domains of diagnosis and prevention/follow-up, where accuracy rates exceeded 90%, the model demonstrated a strong capacity to deliver clinically appropriate content. This finding aligns with the increasing integration of AI tools in medical communication and patient education.

Interestingly, treatment-related questions exhibited a slightly lower rate of fully accurate answers (86.7%) and a higher proportion of both partially misleading and incorrect responses. This discrepancy may stem from the inherent complexity and variability of surgical approaches to hypospadias, where individual patient factors often guide therapeutic decision-making. Moreover, nuances in current surgical techniques and postoperative care protocols—often updated in subspecialty literature—may not be fully captured in ChatGPT's existing training data.

The rapid integration of artificial intelligence into daily life is expected to significantly impact the trust-based relationship between patients and physicians. In a 2023 study conducted by Ayers et al., 195 patient questions sourced from Reddit's r/AskDocs forum were analyzed.[6] The responses generated by human physicians and those provided by ChatGPT were compared. Remarkably, participants preferred the AI-generated responses in 78.6% of cases, citing superior

quality and empathy. While AI tools have only recently been introduced into clinical communication, early evidence suggests that they may become valuable adjuncts in supporting medical advice and information dissemination.

The reproducibility analysis revealed an overall similarity rate of 91.7% across repeated responses, suggesting a commendable level of internal consistency. However, the observed variability in guideline-based and treatment responses implies that even minor shifts in prompt phrasing or model context can influence the level of detail or emphasis in generated outputs. This level of consistency is promising but also reveals the importance of prompt phrasing and context when interpreting AI-generated responses.

Compared to traditional health information platforms, ChatGPT provides an interactive and user-friendly interface. However, unlike static websites maintained by urological societies, its responses are dynamically generated and not peer-reviewed. This raises important considerations about the platform's reliability in real-time patient education.[7,8] A recent comparative study by Sarikaya et al. similarly demonstrated that AI applications including ChatGPT can achieve high compliance with pediatric urology guidelines when evaluated systematically, reinforcing the clinical utility of such platforms in structured contexts.[9] Given the emotionally charged nature of congenital anomalies such as hypospadias often diagnosed in infancy and requiring parental consent for surgical correction—ensuring the quality and consistency of educational content becomes paramount.[10]

This study has several strengths, including the comprehensive coverage of question types and the structured grading system applied by experienced urologists. Nevertheless, its limitations must also be acknowledged. All questions were evaluated in English using a single version of ChatGPT, and real-time updates to the model could alter its future performance. Furthermore, the model's inability to cite verifiable sources remains a notable constraint in clinical contexts.[11] Moreover, the inability of ChatGPT to cite verifiable, peer-reviewed sources limits its credibility in clinical decision-making

Despite the remarkable progress in artificial intelligence, it cannot yet be considered entirely accurate or flawless. While AI tools tend to provide clear and precise answers to basic medical questions, they often generate inconsistent or contradictory responses when addressing more complex or nuanced inquiries. This observation is supported by a study conducted by Shiferaw et al., in which ChatGPT was evaluated based on its responses to various types of medical questions.[12] The model answered "what" questions correctly in approximately 67% of cases; however, responses to "why" and "how" questions frequently contained inconsistencies and errors. These findings suggest that AI still lacks reliability in contexts requiring detailed and specific information. Therefore, it remains essential that treatment decisions and follow-up strategies be guided by qualified healthcare professionals.

It would be inaccurate to assume that artificial intelligence will be utilized effectively only by patients in the coming years. Even in complex domains such as surgical intervention, particularly in the context of hypospadias, AI is poised to play a significant role in supporting clinical decision-making. A study conducted by Abbas et al. investigated the use of artificial intelligence in the preoperative assessment of hypospadias by applying deep learning algorithms to 2D clinical images.[13] The model accurately identified five key POST (Plate Objective Scoring Tool) landmarks on the urethral plate with a precision rate of 99.5%. This study demonstrates that AI can enhance the consistency and objectivity of preoperative evaluations, thereby improving the overall quality of surgical planning in pediatric urology.

Despite these limitations, our findings suggest that ChatGPT may serve as a supplementary tool for disseminating basic medical information about hypospadias, especially when its use is supervised by qualified clinicians to ensure accuracy and appropriateness. Future developments in language models should prioritize integration with up-to-date clinical databases and transparent source attribution to enhance credibility and trust in medical applications.[14]

CONCLUSION

The findings of this study suggest that ChatGPT can deliver generally accurate and consistent information in response to questions related to hypospadias. Its high rate of completely correct responses, particularly in domains such as diagnosis, underscores the model's potential as a supplementary educational resource. However, variability in treatment-related answers and limitations in reproducibility highlight the need for cautious interpretation of AI-generated content in clinical contexts.

While ChatGPT holds promise in enhancing digital health literacy, especially for non-specialist users, its integration into patient care should be approached judiciously. Ongoing refinement of AI models, including linkage to updated clinical guidelines and the incorporation of verifiable references, is essential for improving reliability. Until such advancements are achieved, clinician oversight remains indispensable when AI tools are used in the dissemination of sensitive or nuanced medical information.

ETHICAL DECLARATIONS

Ethics Committee Approval: Since this study involved the analysis of responses generated by an artificial intelligence model (ChatGPT) to publicly available and anonymized questions, and did not include any human participants, patient data, or identifiable personal information, ethical approval was not required in accordance with institutional and international research ethics guidelines.

Informed Consent: Since no patient data were used in the study, informed consent was not required.

Referee Evaluation Process: Externally peer-reviewed.
Conflict of Interest Statement: The authors have no conflicts of interest to declare.
Financial Disclosure: The author declared that this study has received no financial support.
Author Contributions: All of the authors declare that they have all participated in the design, execution, and analysis of the paper, and that they have approved the final version.

REFERENCES

1. Gabrielson AT, Galansky L, Shneyderman M, Cohen AJ. The impact of hypogonadism on surgical outcomes following primary urethroplasty: analysis of a large multi-institutional database. Urology. 2024;185:116-23.
2. Wang F, Casalino LP, Khullar D. Deep learning in medicine-promise, progress, and challenges. JAMA Intern Med. 2019;179(3):293-4.
3. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat Med. 2019;25(1):44-56.
4. Baskin LS, Ebbers MB. Hypospadias: anatomy, etiology, and technique. J Pediatr Surg. 2006;41(3):463-72.
5. Spinoit AF, Poelaert F, Van Praet C, Groen LA, Van Laecke E, Hoebeke P. Grade of hypospadias is the only factor predicting for re-intervention after primary hypospadias repair: a multivariate analysis from a cohort of 474 patients. J Pediatr Urol. 2015;11(2):70.e1-70.e706.
6. Ayers JW, Poliak A, Dredze M, et al. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. JAMA internal medicine. 2023;183(6):589-96.
7. Betschart P, Pratsinis M, Müllhaupt G, et al. Information on surgical treatment of benign prostatic hyperplasia on YouTube is highly biased and misleading. BJU Int. 2020;125(4):595-601.
8. 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):701-706.
9. Sarikaya M, Ozcan Siki F, Ciftci I. Use of artificial intelligence in vesicoureteral reflux disease: a comparative study of guideline compliance. J Clin Med. 2025;14(7):2378.
10. Nguyen DD, Trinh QD, Cole AP, et al. Impact of health literacy on shared decision making for prostate-specific antigen screening in the United States. Cancer. 2021;127(2):249-56.
11. Checcucci E, Verri P, Amparore D, et al. Generative Pre-training Transformer Chat (ChatGPT) in the scientific community: the train has left the station. Minerva Urol Nephrol. 2023;75(2):131-3.
12. Shiferaw MW, Zheng T, Winter A, Mike LA, Chan LN. Assessing the accuracy and quality of artificial intelligence (AI) chatbot-generated responses in making patient-specific drug-therapy and healthcare-related decisions. BMC Med Inform Decis Mak. 2024;24(1):404.
13. Abbas TO, AbdelMoniem M, Khalil IA, Abrar Hossain MS, Chowdhury MEH. Deep learning based automated quantification of urethral plate characteristics using the plate objective scoring tool (POST). J Pediatr Urol. 2023;19(4):373.e1-373.e9.
14. Jeblick K, Schachtner B, Dexl J, et al. ChatGPT makes medicine easy to swallow: an exploratory case study on simplified radiology reports. Eur Radiol. 2024;34(5):2817-25.