NEU PRESS

SELÇUK TIP DERGİSİ
SELCUK MEDICAL JOURNAL

DOI:http://dx.doi.org/10.30733/std.2025.01852
Selcuk Med J 2025;41(4): 201-204

OPEN ARAŞTIRMA MAKALESİ / RESEARCH ARTICLE

A Comparative Analysis of Large Language Models in Managing Disorders of Sex Development: Evaluation Based on Clinical Guidelines

Cinsiyet Gelişim Bozukluklarının Yönetiminde Büyük Dil Modellerinin Karşılaştırmalı Analizi: Klinik Kılavuzlara Dayalı Bir Değerlendirme

Saime Sundus Uygun[1], Fatma Ozcan Siki[2]

[1]Selcuk University, Faculty of Medicine, Department of Naonatology, Konya, Türkiye
[2]Selcuk University, Faculty of Medicine, Department of Pediatric Surgery, Konya, Türkiye

ÖZET

Amaç: Bu çalışma, günümüzde yaygın olarak kullanılan iki yapay zekâ tabanlı sohbet sistemi olan ChatGPT ve Bing AI'nin, Türk Neonatoloji Derneği tarafından yayımlanan Cinsiyet Gelişim Bozuklukları kılavuzunda yer alan klinik önerilerle uyum düzeylerini karşılaştırmayı amaçlamaktadır.
Gereç ve Yöntemler: Türk Neonatoloji Derneği Cinsiyet Gelişim Bozuklukları kılavuzuna dayalı olarak hazırlanmış 40 sorudan oluşan standart bir değerlendirme seti kullanılmıştır. Sorular, klinik karar verme süreçlerini yansıtan altı ana kategori altında gruplandırılmış ve tamamı hem ChatGPT hem de Bing AI'ya yöneltilmiştir. Tüm sorular yazılı metin formatında iletilmiştir. Yanıtlar, kılavuzla uyum açısından biri neonatoloji, diğeri çocuk cerrahisi uzmanı olmak üzere iki bağımsız uzman tarafından 5 puanlık Likert ölçeği ile değerlendirilmiştir. Her kategori için ChatGPT ve Bing AI'nin ortalama puanları hesaplanmış, bu puanlar arasındaki fark Wilcoxon işaretli sıra testi ile istatistiksel olarak karşılaştırılmıştır.
Bulgular: ChatGPT, altı kategorinin tamamında Türk Neonatoloji Derneği'nin Cinsiyet Gelişim Bozuklukları kılavuzu ile yüksek düzeyde uyum göstermiştir (ortalama puan: 4,88). Buna karşın, Bing AI bazı kategorilerde daha düşük uyum sergilemiştir (ortalama puan: 3,25). İki sistem arasındaki ortalama puan farkları tüm kategorilerde istatistiksel olarak anlamlı bulunmuştur (p<0,05). Özellikle tanı süreci/laboratuvar testleri ve tedavide multidisipliner yaklaşım kategorilerinde Bing AI'nın performansı belirgin şekilde düşüktür.
Sonuç: ChatGPT, Cinsiyet Gelişim Bozuklukları konusunda kılavuz temelli klinik destek sağlama açısından Bing AI'ya göre daha yüksek doğruluk ve tutarlılık göstermiştir. Güncel klinik kılavuzlarla uyumlu yapay zekâ destekli sistemlerin kullanımı, karmaşık ve multidisipliner karar verme süreçlerinde hekimleri destekleme potansiyeline sahiptir. Bu nedenle, klinik uygulamalarda kullanılacak yapay zekâ araçlarının seçimi, bu tür sistematik değerlendirmelere dayanmalıdır. Güvenilir yapay zekâ tabanlı sistemler, hasta yönetiminde klinisyenlere önemli katkılar sağlayabilir.

Anahtar Kelimeler: Büyük Dil Modelleri, Cinsiyet Gelişim Bozuklukları, Yapay Zekâ, Neonatoloji, Karar Destek Sistemleri

ABSTRACT

Objective: This study aims to compare the guideline compliance of two widely used AI-based chatbot systems, ChatGPT and Bing AI, with the clinical recommendations outlined in the Disorders of Sex Development (DSD) guideline published by the Turkish Neonatal Society.
Materials and Methods: A standardized evaluation set comprising 40 questions based on the DSD guideline was utilized. The questions were grouped under six main categories reflecting clinical decision-making processes and were presented to both ChatGPT and Bing AI. Responses were scored on a 5-point Likert scale by two independent experts, assessing their alignment with the guideline. Mean scores were calculated for each category, and statistical comparisons were made using the Wilcoxon signed-rank test.
Results: ChatGPT demonstrated high consistency with the guideline across all categories (mean score: 4.88), while Bing AI showed lower compliance in several areas (mean score: 3.25). The differences in scores between the two systems were statistically significant across all categories ($p < 0.05$), with Bing AI performing particularly poorly in the areas of diagnosis/laboratory testing and multidisciplinary approach.
Conclusion: ChatGPT demonstrated higher accuracy and consistency than Bing AI in providing guideline-based clinical support regarding DSD. The use of AI-supported systems aligned with current guidelines holds significant potential in supporting complex, multidisciplinary decision-making processes. Therefore, the selection of AI tools in clinical settings should be informed by such systematic evaluations.

Keywords: Large Language Models, Disorders of Sex Development, Artificial Intelligence, Neonatology, Decision Support Systems

Geliş Tarihi/Received: 10 May/Mayıs 2025 **Kabul Tarihi/Accepted:** 19 October/Ekim 2025 **Yayın Tarihi/Published Online:** 12 December/Aralık 2025

Sorumlu Yazar/Corresponding Author: Saime Sundus Uygun, Selcuk University, Faculty of Medicine, Department of Naonatology, Konya, Türkiye
e-mail: uygunsaime@hotmail.com

Atıf yapmak için/ Cite this article as: Uygun SS, Ozcan Siki F. A Comparative Analysis of Large Language Models in Managing Disorders of Sex Development: Evaluation Based on Clinical Guidelines. Selcuk Med J 2025;41(4): 201-204

Disclosure: *Author has not a financial interest in any of the products, devices, or drugs mentioned in this article. The research was not sponsored by an outside organization. Author has agreed to allow full access to the primary data and to allow the journal to review the data if requested.*

INTRODUCTION

Disorders of Sex Development (DSD) represent a group of rare but clinically complex conditions in which chromosomal, gonadal, and anatomical sex do not align. The diagnostic and management process typically begins immediately after birth and requires a multidisciplinary approach that incorporates not only physical findings but also genetic, hormonal, radiological, and ethical evaluations. These cases often create significant psychosocial pressure for both families and healthcare providers, particularly in terms of diagnosis and gender assignment. For this reason, the diagnosis and management of DSD must be conducted in strict accordance with up-to-date clinical guidelines (1–4).

In recent years, the use of artificial intelligence (AI) applications in the healthcare field has rapidly increased, emerging as valuable tools in clinical decision support. Among these, large language model (LLM)-based chatbots have gained attention for their potential to provide rapid and easily accessible medical information. However, the degree to which these systems offer recommendations and answers that align with evidence-based clinical guidelines remains a matter of concern (5).

Several studies have investigated the guideline compliance of popular chatbots such as ChatGPT and Bing AI across different medical disciplines (6,7). This study aims to systematically evaluate the responses of these two AI systems—both built on LLMs—to questions derived from the Turkish Neonatal Society's clinical guideline for the management of DSD. In doing so, the study not only tests the accuracy of these technological tools but also provides insight into their suitability as future clinical decision support systems.

MATERIALS AND METHODS

In this study, a standardized evaluation set consisting of 40 questions was developed based on the clinical guideline titled "Clinical Approach to Disorders of Sex Development" published by the Turkish Neonatal Society. The questions were categorized into six thematic domains that reflect the diagnostic and management steps outlined in the guideline:

1. Definition and Classification (Questions 1–5)
2. Postnatal Evaluation (Questions 6–12)
3. Diagnostic Process and Laboratory Testing (Questions 13–22)
4. Decision-Making and Multidisciplinary Approach (Questions 23–30)
5. Family Counseling and Social Considerations (Questions 31–35)
6. Outcomes and Recommendations (Questions 36–40)

Both AI systems—ChatGPT and Bing AI—were tested separately using this set of questions. All items were presented in identical format and sequence to each system via written input. The responses were independently evaluated by two experts (a pediatric surgeon and a neonatologist) for their compliance with the guideline. Each response was scored using a 5-point Likert scale as follows:

Score	***Description***
5	Complete alignment with the guideline
4	Largely consistent – minor omissions
3	Moderate alignment – key omissions present
2	Low consistency – significant contradictions
1	No alignment – incorrect or irrelevant answer

Following evaluation, mean scores were calculated for both AI systems across all questions. Additionally, subgroup analyses were conducted to assess the average score within each thematic category.

Each response was independently scored by two clinicians (a pediatric surgeon and a neonatologist). The final score was calculated as the average of the two ratings. To assess inter-rater reliability, the intraclass correlation coefficient (ICC) was computed, revealing excellent agreement (ICC = 0.91).

Statistical Analysis

Since the distribution of scores did not meet the assumption of normality, the Wilcoxon signed-rank test was used to compare the paired scores of ChatGPT and Bing AI across the questions. Mean scores were also analyzed separately for each category, and a significance threshold of $p < 0.05$ was applied.

The Wilcoxon signed-rank test was chosen because each question was answered by both AI systems under identical conditions, and their paired scores were evaluated by the same raters. As the data are dependent, a paired non-parametric test was appropriate.

To further illustrate the comparative performance of the two systems, a summary table and a bar chart were generated to display average scores by category. Data analysis was conducted using Python, and visualizations were created with the matplotlib library.

RESULTS

The comparison of both AI systems revealed that ChatGPT exhibited high consistency with the national DSD guideline across all categories, with an overall mean score of 4.88 out of 5. In contrast, Bing AI demonstrated lower levels of compliance, with an overall mean score of 3.25. The group-based analysis showed that ChatGPT consistently outperformed Bing AI in every thematic category.

The Wilcoxon signed-rank test revealed statistically significant differences in mean scores between ChatGPT and Bing AI for all six categories ($p < 0.05$). The most prominent discrepancies were observed in the "Diagnostic Process and Laboratory Testing" and "Decision-Making and Multidisciplinary Approach" categories, where Bing AI performed notably poorly. A total score of 195 was calculated for ChatGPT across all 40 questions, whereas Bing AI received a cumulative score of 130. This substantial difference supports the conclusion that ChatGPT provides more guideline-consistent responses and may be better suited as a clinical decision support tool in the context of DSD.

Details of the category-based performance and statistical

Table 1. Statistical Comparison of AI Responses According to the National DSD Guideline

Category	chatGPT Mean Score	Bing AI Mean Score	Wilcoxon p-value
Definition and Clasification	5±0.00	3,4±0.18	0,0431
Postnatal Evaluation	5±0.00	3,29±0.22	0,0277
Diagnosis and Laboratory	4,8±0.12	2,9±0.15	0,0078
Desicion and Multidisciplinary	5±0.00	3,13±0.16	0,0123
Family and Social Approach	4,8±0.11	3,2±0.18	0,0417
Outcomes and Recommendations	5±0.00	3,2±0.14	0,0346

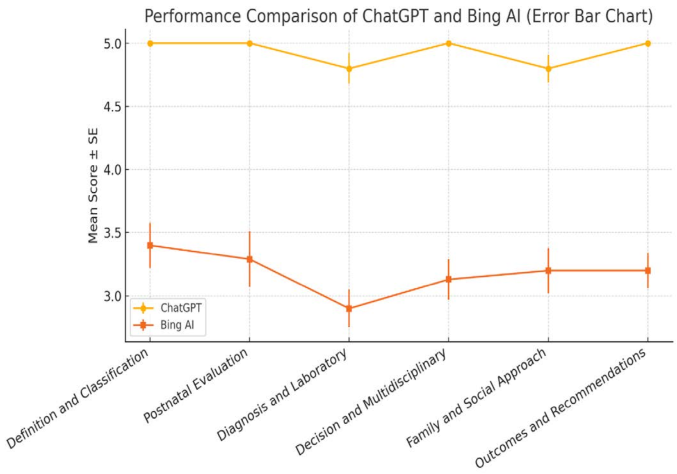

Figure 1. Comparative Analysis of ChatGPT and Bing AI Performance by Category

analysis are presented in Table 1, and a visual comparison of mean scores is provided in Figure 1.

DISCUSSION

Technological advancements in recent years have led to the widespread use of online resources for accessing medical information. The integration of artificial intelligence (AI) into these platforms has further accelerated this trend, making information retrieval more accessible for both physicians and patients. However, uncertainties remain regarding the clinical effectiveness and reliability of AI-based chatbots in healthcare settings. In this study, we evaluated the responses of two popular AI chatbot systems—ChatGPT and Bing AI—on the complex and multidisciplinary topic of Disorders of Sex Development (DSD) in newborns, by assessing their alignment with the guideline published by the Turkish Neonatal Society. ChatGPT demonstrated superior performance, particularly in categories involving diagnostic processes, laboratory testing, and multidisciplinary planning. This suggests that ChatGPT may have been trained with more comprehensive medical datasets.

Previous studies have reported similar findings. For example, ChatGPT has been shown to outperform Bing AI in the accuracy and comprehensiveness of information provided on topics such as prostate cancer (6) and microbiology (8). In the pediatric domain, studies have shown that ChatGPT demonstrated higher compliance with guidelines for conditions such as vesicoureteral reflux (VUR), while Bing AI often omitted critical elements (9). In another comparison of four AI systems, all models demonstrated guideline adherence for VUR; however, consistency varied across platforms (10). Additionally, ChatGPT was found to make more frequent and accurate references to scientific sources in pediatric urology evaluations (7).

In thyroid nodule management, ChatGPT provided consistent answers across different time points, suggesting temporal stability in its clinical guidance (11), which is a valuable feature for reliability. On the other hand, Bing AI's performance has been reported as more variable and limited, particularly in areas such as urolithiasis, where it failed to provide adequate references or comprehensive guideline-based answers (12,13). These findings align with the outcomes of our study. A similar trend has been observed in other medical disciplines. For instance, in obesity surgery, ChatGPT was more successful in identifying appropriate surgical options (14), and in gastroenterology, it was found to have a high informative capacity, though some limitations were also noted (15). Large language models have also been shown to provide higher-quality recommendations in managing complex, multidisciplinary cases (16). Given that DSD management inherently requires multidisciplinary collaboration, ChatGPT's consistent and evidence-based responses support its potential role in assisting clinical decision-making.

Nonetheless, several studies have also highlighted limitations of AI systems, such as delays in incorporating the latest guideline updates, lack of patient-specific personalization, and ethical decision-making challenges (15,17). Some of these limitations were also partially observed in our study, where both systems occasionally offered vague or superficial responses. A key strength of our study is the use of a double-blind evaluation design, in which both AI systems were assessed under identical conditions by two independent experts. However, one limitation is that only text-based responses were evaluated; visual or interactive capabilities of the systems were not included in the assessment.

CONCLUSION

In today's healthcare landscape, rapid access to medical information is essential; however, it is equally important that such information is accurate, evidence-based, and personalized. Within this context, ChatGPT demonstrated notable superiority over Bing AI in terms of guideline adherence and consistency, particularly in domains requiring multidisciplinary coordination. These findings highlight ChatGPT's potential utility in clinical education, family counseling, and decision-support systems.

Nevertheless, it remains critical that AI systems maintain alignment with the most recent clinical guidelines, exhibit ethical sensitivity, and operate under expert supervision. As artificial intelligence continues to integrate into healthcare practice, the evaluation and selection of these systems should rely on structured, guideline-based analyses. Future studies involving diverse clinical guidelines and larger datasets will be instrumental in enhancing the reliability, standardization, and clinical integration of AI-supported tools. In this regard, guideline-oriented AI evaluations may represent a new paradigm for medical education and patient management.

Conflict of interest: *The authors declared no conflicts of interest with respect to the authorship and/or publication of this article.*

Financial conflict of interest: *Author declares that he did not receive any financial support in this study.*

Address correspondence to: *Saime Sundus Uygun, Selcuk University, Faculty of Medicine, Department of Naonatology, Konya, Türkiye*
e-mail: *uygunsaime@hotmail.com*

REFERENCES

1. Özalp Kızılay D, Özen S. Current diagnostic approaches in the genetic diagnosis of disorders of sex development. J Clin Res Pediatr Endocrinol. 2024;16(4):401-10. doi: 10.4274/jcrpe.galenos.2024.2024-3-3.
2. Yavas AZ, Guran T. Diagnosis and management of non-CAH 46,XX disorders/differences in sex development. Front Endocrinol (Lausanne). 2024;15:1354759. doi: 10.3389/fendo.2024.1354759.
3. Ndoye NA, Diallo BA, Sylla M, et al. Ovotesticular disorders of sexual development: Diagnostic, therapeutic, and evolutionary aspects. J Pediatr Surg. 2025;60(4):162187. doi: 10.1016/j.jpedsurg.2025.162187
4. Khorashad BS, Goodarzi H, Greenfield D, et al. Recommendations for 46,XY disorders/differences of sex development across two decades: Insights from North American pediatric endocrinologists and urologists. Arch Sex Behav. 2024;53(8):2939-56. doi.org/10.1007/s10508-024-02942-1
5. Altıntaş E, Kılıç R, Gürlek C, et al. Comparative analysis of artificial intelligence chatbot recommendations for urolithiasis management: A study of EAU guideline compliance. Fr J Urol. 2024;34(7-8):102666. doi: 10.1016/j.fjurol.2024.102666
6. Alasker A, Alsalamah S, Alshathri N, et al. Performance of large language models (LLMs) in providing prostate cancer information. BMC Urol 24, 177 (2024). doi:10.1186/s12894-024-01570-0.
7. Caglar U, Tunç F, Özen B, et al. Evaluating the performance of ChatGPT in answering questions related to pediatric urology. J Pediatr Urol. 2024;20(1):26.e1-5. doi: 10.1016/j.jpurol.2023.08.003.
8. Ranjan J, Gupta A, Mehta S, et al. Assessment of artificial intelligence platforms with regard to medical microbiology knowledge: An analysis of ChatGPT and Gemini. Cureus. 2024;16(5):e60675. doi: 10.7759/cureus.60675.
9. Akyol Onder EN, Ensari E, Ertan P. ChatGPT-4o's performance on pediatric vesicoureteral reflux. J Pediatr Urol. 2025 Apr;21(2):504-09. doi: 10.1016/j.jpurol.2024.12.002.
10. 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 Mar 30;14(7):2378. doi: 10.3390/jcm14072378.
11. Deniz MS, Guler BY. Assessment of ChatGPT's adherence to ETA-thyroid nodule management guideline over two different time intervals 14 days apart: In binary and multiple-choice queries. Endocrine. 2024;85(2):794-02. doi: 10.1007/s12020-024-03750-2.
12. Cakir H, Kaya Z, Tunçer M, et al. Evaluating the performance of ChatGPT in answering questions related to urolithiasis. Int Urol Nephrol. 2024;56(1):17-21. doi: 10.1007/s11255-023-03773-0
13. Cil G, Dogan K. The efficacy of artificial intelligence in urology: A detailed analysis of kidney stone-related queries. World J Urol. 2024;42(1):158. doi: 10.1007/s00345-024-04847-z.
14. Lee Y, Tessier L, Brar K, et al. Performance of artificial intelligence in bariatric surgery: Comparative analysis of ChatGPT-4, Bing, and Bard in the American Society for Metabolic and Bariatric Surgery textbook of bariatric surgery questions. Surg Obes Relat Dis. 2024;20(7):609-13. doi: 10.1016/j.soard.2024.04.014.
15. Klang E, Sourosh A, Nadkarni GN, et al. Evaluating the role of ChatGPT in gastroenterology: A comprehensive systematic review of applications, benefits, and limitations. Therap Adv Gastroenterol. 2023;16:17562848231218618. doi: 10.1177/17562848231218618.
16. Ríos-Hoyo A, Shan NL, Li A, et al. Evaluation of large language models as a diagnostic aid for complex medical cases. Front Med (Lausanne). 2024 Jun 20;11:1380148. doi: 10.3389/fmed.2024.1380148.
17. Lopez-Gonzalez R, Sanchez-Cordero S, Pujol-Gebellí J, et al. Evaluation of the Impact of ChatGPT on the Selection of Surgical Technique in Bariatric Surgery. Obes Surg. 2025 Jan;35(1):19-24. doi: 10.1007/s11695-024-07279-1.