Clinical Decision Support via ChatGPT: An Evidence-Based Perspective in Pediatric Occupational Therapy

Gamze Çağla SIRMA[*], İbrahim ERARSLAN[**], Zeynep BAHADIR[***]

Abstract

Aim: This study aims to evaluate the accuracy of ChatGPT's responses to clinical questions related to Cerebral Palsy (CP) and Autism Spectrum Disorder (ASD) in the field of occupational therapy, as well as the reliability of the references it provides.

Method: Ten clinical questions were formulated, and answers were prepared based on clinical guidelines and reviewed by two expert clinicians. The same questions were presented to ChatGPT, and its responses and references were rated by two independent evaluators using a four-point Likert scale. Weighted Kappa Coefficients were calculated to assess inter-rater agreement.

Results: ChatGPT's responses demonstrated high accuracy, with strong inter-rater agreement (Weighted Kappa: 0.72–0.75). However, significant deficiencies were identified in reference reliability, with moderate to high inter-rater agreement (Weighted Kappa: 0.75–0.78). Additionally, fictitious reference rates ranged from 40% to 100% for CP-related questions and from 25% to 100% for ASD-related questions.

Conclusion: While ChatGPT shows potential as a clinical decision support tool in occupational therapy, the limitations in reference accuracy restrict its reliability in healthcare applications. Future studies should focus on improving artificial intelligence models' reference accuracy to enhance their applicability in healthcare settings.

Keywords: Artificial intelligence, autism spectrum disorder, cerebral palsy, occupational therapy.

ChatGPT ile Klinik Karar Desteği: Pediatrik Ergoterapide Kanıta Dayalı Bir Perspektif

Öz

Amaç: Bu çalışmanın amacı, Ergoterapi alanında Serebral Palsi (SP) ve Otizm Spektrum Bozukluğu (OSB) ile ilgili klinik sorulara ChatGPT'nin verdiği yanıtların doğruluğunu ve sağladığı kaynakların güvenilirliğini değerlendirmektir.

Özgün Araştırma Makalesi (Original Research Article)
Geliş / Received: 02.10.2025 **Kabul / Accepted:** 02.07.2026
DOI: https://doi.org/10.38079/igusabder.1795663
[*] (Corresponding Author), Lecturer, Department of Occupational Therapy, Faculty of Health Sciences, Fenerbahçe University, Istanbul, Türkiye. E-mail: gamze.dirgen@fbu.edu.tr ORCID https://orcid.org/0000-0003-2976-9035
[**] Res. Asst., Department of Occupational Therapy, Faculty of Health Sciences, Istanbul Medipol University, Istanbul, Türkiye. E-mail: ibrahim.erarslan@medipol.edu.tr ORCID https://orcid.org/0000-0002-0760-0223
[***] Assoc. Prof., Department of Occupational Therapy, Hamidiye Faculty of Health Sciences, University of Health Sciences, Istanbul, Türkiye. E-mail: zeynep.bahadir@sbu.edu.tr ORCID https://orcid.org/0000-0002-7674-9830

ETHICAL STATEMENT: This study was previously presented as an oral presentation at the International Congress on Digitalization and Artificial Intelligence in Healthcare, held at Sinop University, Türkiye, on 21–23 September 2025, and was published in the abstract book (p. 59; e-ISBN: 978-625-96460-0-8).

Yöntem: On klinik soru oluşturulmuş, yanıtlar klinik rehberlere dayalı olarak hazırlanmış ve iki uzman klinisyen tarafından gözden geçirilmiştir. Aynı sorular ChatGPT'ye yöneltilmiş, yanıtları ve kaynakları iki bağımsız değerlendirici tarafından dört dereceli Likert ölçeği ile puanlanmıştır. Değerlendiriciler arası uyumu belirlemek için Ağırlıklı Kappa Katsayıları hesaplanmıştır.

Bulgular: ChatGPT'nin yanıtları yüksek doğruluk göstermiş ve değerlendiriciler arası güçlü uyum saptanmıştır (Ağırlıklı Kappa: 0,72–0,75). Ancak, kaynak güvenilirliğinde önemli eksiklikler bulunmuş olup, değerlendiriciler arası uyum orta-yüksek düzeydeydi (Ağırlıklı Kappa: 0,75–0,78). Ayrıca, hayali kaynak oranları SP ile ilgili sorularda %40–100, OSB ile ilgili sorularda %25–100 arasında değişmiştir.

Sonuç: ChatGPT, ergoterapi alanında klinik karar destek aracı olarak potansiyel göstermektedir; ancak kaynak doğruluğundaki sınırlılıklar, sağlık hizmetlerindeki güvenilirliğini kısıtlamaktadır. Gelecek çalışmalarda, yapay zekâ modellerinin kaynak doğruluğunun geliştirilmesine odaklanılması, sağlık alanındaki uygulanabilirliklerini artıracaktır.

Anahtar Sözcükler: Yapay zekâ, otizm spektrum bozukluğu, serebral palsi, ergoterapi.

Introduction

Artificial intelligence (AI) refers to software designed to perform complex tasks that typically require human intelligence, such as problem-solving, learning, and decision-making[1]. Recent advances in machine learning architectures and the availability of large training datasets have enabled AI systems to perform at levels comparable to humans in specific domains. Among these, ChatGPT, developed by OpenAI, has emerged as a prominent example of AI technology. This AI-powered chatbot utilizes a Generative Pre-training Transformer (GPT) model, which enables it to generate human-like, context-sensitive text based on vast amounts of publicly available data[2]. ChatGPT's ability to perform diverse tasks—ranging from answering questions to summarizing content and generating creative compositions—has contributed to its widespread adoption, with over 100 million users in just two months of its release[3].

In the healthcare sector, ChatGPT has demonstrated the potential to support clinical workflows by synthesizing vast amounts of information, offering quick answers to complex questions, and generating personalized recommendations[4-7]. In a fast-paced clinical environment, where time constraints often limit access to current evidence, ChatGPT's ability to relay critical, evidence-based information could be transformative. By integrating and contextualizing data from sources such as case reports, scientific articles, and clinical guidelines, ChatGPT may enable clinicians to develop individualized treatment plans more efficiently[4,8,9]. However, the utility of ChatGPT in healthcare is not without limitations. Particularly in areas with limited published evidence or where clinical scenarios are highly complex, ChatGPT's responses may lack the necessary depth or accuracy. Furthermore, the phenomenon of "hallucinations," where the model generates fabricated or misleading information, poses significant risks in clinical decision-making[4,10].

Istanbul Gelisim University Journal of Health Sciences (IGUSABDER) is indexed by TUBITAK ULAKBIM TR Index.
Web site: https://dergipark.org.tr/en/pub/igusabder
Contact: igusabder@gelisim.edu.tr

One of the core principles of modern healthcare is evidence-based practice (EBP), which involves integrating the best available evidence, clinical expertise, and patient values into the decision-making process. In healthcare settings, EBP is essential for optimizing patient outcomes and ensuring the safe, effective delivery of care[11]. While ChatGPT's design as a "task-agnostic" tool allows it to operate across various disciplines, its capacity to deliver evidence-based insights tailored to specific medical domains remains an area of active investigation[4]. Evaluating the reliability and validity of its responses is particularly crucial for its adoption in the rehabilitation field, where personalized care plans depend heavily on current and accurate evidence.

Occupational therapy (OT) is a rehabilitation field that prioritizes holistic, patient-centered care. Integrating AI tools like ChatGPT into pediatric occupational therapy, the most prevalent area of practice, presents significant opportunities. Recent research has explored various applications of AI in this domain.

For instance, Sharma (2024) highlighted AI's potential to enhance the assessment and intervention processes for children with developmental needs[12]. Similarly, Berrezueta-Guzman et al. (2024) emphasized the effectiveness of AI-based systems in improving evaluation and therapeutic outcomes in pediatric OT[13]. Despite its potential, there is a lack of research evaluating the use of ChatGPT in pediatric OT, particularly in terms of its alignment with evidence-based practice principles. This study aimed to evaluate the evidence-based accuracy of ChatGPT in pediatric occupational therapy for Autism Spectrum Disorder (ASD) and Cerebral Palsy (CP).

Material and Methods

Question Development and Expert Review

Researchers designed ten clinical questions (five for CP and five for ASD) based on relevant clinical guidelines and scientific literature[14-16]. These questions aimed to assess both theoretical knowledge and practical applications. To ensure a balanced evaluation, three questions for each condition focused on general clinical knowledge, while the remaining two were designed as case-based scenarios simulating real-world clinical decision-making (see supplementary). The entire process, including question formulation, expert review, ChatGPT evaluation, and data analysis, was conducted in English.

The initial answer key for the formulated questions was developed based on clinical guidelines and expert recommendations. Two independent researchers reviewed the prepared responses, providing feedback for necessary modifications and refinements. After incorporating their recommendations, the revised answer key was further evaluated and finalized by two experts, each with at least five years of clinical experience (≥5 years of practice).

Istanbul Gelisim University Journal of Health Sciences (IGUSABDER) is indexed by TUBITAK ULAKBIM TR Index.
Web site: https://dergipark.org.tr/en/pub/igusabder
Contact: igusabder@gelisim.edu.tr

Data Collection Process

A detailed planning phase was conducted to determine the optimal way to structure the questions for ChatGPT. During this phase, prompt formats were developed to ensure accurate and comprehensive responses. To optimize the phrasing of the queries, ChatGPT was consulted by asking, "*What should we pay attention to in order to get the right answers when asking you questions for professional information in the field of occupational therapy?"* The following responses, along with explanations, were received, and the questions were structured accordingly: be specific in your inquiry, clarify the context, ask for evidence-based answers, specify frameworks or models, include cultural or regional context, state your role and purpose, ask follow-up questions, request examples or practical applications, be clear about the level of detail required, and cross-verify information.

Based on these principles, the final question set was structured and optimized. To ensure transparency and reproducibility, an example of the prompt structure used in the study is presented below:

"As a pediatric occupational therapist, I will ask some questions. 1. Which interventions can I use to improve motor development in children with CP diagnosis or CP risk in the 0-5 age range? In the text, add the scientific reference for this information."

The questions were then entered into ChatGPT (GPT-4o, OpenAI) in January 2025, and the generated responses with references were systematically recorded. ChatGPT provided the sources in APA format, including years of publication.

Evaluation of Responses and References

The accuracy of the responses and the validity of the references were independently evaluated by two raters with at least five years of clinical experience. Both raters were clinician-academics and were familiar with evidence-based practice, literature searching, and verification of scientific sources using major academic databases such as PubMed and Scopus. Each response was rated using a four-point scale (4= Completely correct, 3= Almost correct, 2= Partially correct, 1= Completely incorrect).

The validity of the references was assessed based on the APA-formatted citations and publication years provided by ChatGPT. If a cited article or source did not match the given bibliographic details, additional searches were conducted. References that could not be located through any means were classified as "fictitious references".

Ethical Statement

Ethical approval was not required for this study because it did not involve human participants, patient records, biological samples, or animal subjects. The study evaluated responses generated by ChatGPT (GPT-4o, OpenAI) to researcher-developed clinical questions and assessed their accuracy and reference validity against evidence-based

Istanbul Gelisim University Journal of Health Sciences (IGUSABDER) is indexed by TUBITAK ULAKBIM TR Index.
Web site: https://dergipark.org.tr/en/pub/igusabder
Contact: igusabder@gelisim.edu.tr

clinical guidelines. No personal, confidential, or identifiable data were collected or analyzed. Consequently, informed consent was not applicable.

Data Analysis

The number of 'fictitious references' was calculated as a percentage. Weighted Kappa Coefficients were also calculated to assess inter-rater reliability. Results were expressed as median (Md) and interquartile range (IQR). The Weighted Kappa Coefficient is a statistical method used to measure the agreement between two or more raters. This method assigns different weights according to the severity of errors and is particularly effective for ordinal data, as it captures the significance of disagreements among raters.

The interpretation of Weighted Kappa Coefficient values followed standard classification. A coefficient below 0.40 indicated poor agreement, values between 0.41 and 0.60 represented moderate agreement, values ranging from 0.61 to 0.80 were classified as good agreement, and values between 0.81 and 1.00 denoted excellent agreement[17].

Results

The evaluation of answer accuracy and reference accuracy was conducted separately for CP- and ASD-related questions. The findings are summarized in Table 1 and Table 2.

Table 1. Summary of scores for output content, reference accuracy, and weighted kappa coefficient

Category	**CP**		**ASD**	
	Assessor 1	**Assessor 2**	**Assessor 1**	**Assessor 2**
Accuracy of the generated output	4.0 [3.0–4.0]	4.0 [3.0–4.0]	4.0 [3.0–4.0]	4.0 [3.0–4.0]
Weighted kappa coefficient of the generated output	**0.72***	**0.72***	**0.75***	**0.75***
Accuracy of the generated references	2.0 [1.0–3.0]	2.0 [1.0–3.0]	2.0 [1.0–3.0]	2.0 [1.0–3.0]
Weighted kappa coefficient of the generated references	**0.75***	**0.75***	**0.78***	**0.78***

*Note: Statistically significant ($p < 0.05$).

Istanbul Gelisim University Journal of Health Sciences (IGUSABDER) is indexed by TUBITAK ULAKBIM TR Index.
Web site: https://dergipark.org.tr/en/pub/igusabder
Contact: igusabder@gelisim.edu.tr

Table 2. Reference accuracy of different occupational therapy questions

Question No	Question Description			Fictitious / Total References	Rate of Fictitious References
	Type	Condition	Focus		
1.	General	CP	OT Role	7/7	%100
2.	General	CP	OT Interventions	7/7	%100
3.	General	CP	OT Approaches	5/5	%100
4.	Clinical Case	CP	Assessment	6/15	%40
5.	Clinical Case	CP	Intervention	7/7	%100
6.	General	ASD	OT Role	7/8	%87
7.	General	ASD	Interventions	6/9	%66
8.	General	ASD	Approaches	7/11	%81
9.	Clinical Case	ASD	Assessment	2/12	%25
10.	Clinical Case	ASD	Intervention	7/12	%58

Accuracy of the Generated Output

For CP-related questions (Questions 1–5), the median accuracy score was 4.0 [3.0–4.0], with an IQR of 1.0. The weighted kappa coefficient was 0.72, indicating moderate to high agreement. Questions 1 and 5 showed full agreement (kappa= 1.00), whereas Questions 2, 3, and 4 exhibited some variability.

For ASD-related questions (Questions 6–10), the median accuracy score remained 4.0 [3.0–4.0], with an IQR of 1.0. The weighted kappa coefficient was 0.75, reflecting a high level of agreement. Full agreement was observed for Questions 6, 8, and 9, while Questions 7 and 10 showed some discrepancies between assessors.

Accuracy of the Generated References

For CP-related references, the median accuracy score was 2.0 [1.0–3.0], with an IQR of 1.0. The weighted kappa coefficient was 0.75, indicating high inter-rater agreement.

Istanbul Gelisim University Journal of Health Sciences (IGUSABDER) is indexed by TUBITAK ULAKBIM TR Index.
Web site: https://dergipark.org.tr/en/pub/igusabder
Contact: igusabder@gelisim.edu.tr

However, the relatively low median score suggests that references were frequently rated as partially correct (2) or completely incorrect (1), raising concerns about their reliability.

For ASD-related references, the median accuracy score was also 2.0 [1.0–3.0], with an IQR of 1.0. The weighted kappa coefficient was 0.78, showing a strong level of agreement. While Questions 6, 7, 8, and 9 demonstrated consistent assessments, Question 10 showed some variability between assessors.

Table 2 provides additional insight into reference accuracy by detailing the proportion of fictitious references per question. Among CP-related questions, fictitious reference rates ranged from 40% (Question 4) to 100% (Questions 1, 2, and 5). In ASD-related questions, the fictitious reference rate varied from 25% (Question 9) to 100% (Question 3). In particular, general knowledge questions (e.g., Questions 1, 2, 3, 6, 7, and 8) had a higher percentage of fictitious references compared to case-based assessment questions (e.g., Questions 4, and 9).

Overall Assessment

The results indicate that the generated output was generally rated as "completely correct" (4) or "almost correct" (3), confirming its overall accuracy. The weighted kappa values (0.72–0.75) suggest that assessors maintained moderate to high consistency in evaluating the responses.

In contrast, reference accuracy was lower, with most references classified as "partially correct" (2) or "completely incorrect" (1). Despite the high inter-rater agreement (weighted kappa= 0.75– 0.78), the frequent inaccuracies in references—particularly in CP-related responses—highlight the need for stricter validation criteria.

Discussion

This study aimed to assess the accuracy of ChatGPT's responses to clinical questions related to CP and ASD in the field of occupational therapy, as well as the reliability of the references it provided. The results indicate that while ChatGPT's responses were generally accurate and aligned with clinical guidelines, the reliability of the references was significantly lower. The inter-rater agreement was moderate to high, with weighted kappa coefficients ranging from 0.72 to 0.75 for response accuracy and 0.75 to 0.78 for reference accuracy. These findings suggest that although ChatGPT can serve as a valuable tool for retrieving clinical information, its ability to generate reliable references remains a critical limitation.

The accuracy of ChatGPT's responses was found to be high, with assessors generally rating them as clinically sound and relevant to the questions posed. These results align with previous studies examining AI-powered decision support in healthcare. Scaff et al. (2025) assessed the performance of large language models in answering patient inquiries about low back pain and found that, while AI-generated responses were informative, inconsistencies in contextual depth and phrasing were noted[18]. Similarly, Sawamura et

Istanbul Gelisim University Journal of Health Sciences (IGUSABDER) is indexed by TUBITAK ULAKBIM TR Index.
Web site: https://dergipark.org.tr/en/pub/igusabder
Contact: igusabder@gelisim.edu.tr

al. (2024) analyzed ChatGPT's responses to physiotherapy-related clinical questions and reported that, while the information was mostly accurate, minor discrepancies existed between AI-generated content and established best practices[19]. These findings suggest that ChatGPT can synthesize and relay relevant clinical information efficiently, though limitations remain in the precision and specificity of certain answers.

The use of AI-based clinical decision support systems is gaining traction due to their ability to rapidly process vast amounts of medical information[20]. However, as emphasized by Akalin and Veranyurt (2022), reliance on AI-generated responses without expert oversight poses risks, particularly in nuanced clinical decision-making[21]. In this study, high inter-rater agreement was observed in evaluating response accuracy, suggesting that the generated content was consistently aligned with clinical guidelines. Nonetheless, discrepancies in the assessment of certain questions highlight the need for further validation, particularly in complex cases requiring individualized decision-making.

Despite the high accuracy of responses, a major limitation identified in this study was the reliability of ChatGPT's references. Such inaccuracies are commonly referred to as "hallucinations" in large language models. LLMs such as ChatGPT function primarily as probabilistic next-token prediction systems, generating text based on patterns learned from large training corpora rather than retrieving information directly from indexed academic databases. Consequently, when asked to provide references, the model may generate citations that appear plausible in structure but do not correspond to real publications, particularly when real-time access to bibliographic databases is not available. Consistent with this mechanism, a significant proportion of the references in the present study were found to be incomplete, inaccurate, or entirely fabricated, limiting their applicability in evidence-based practice. These findings are consistent with prior studies investigating AI-generated references in scientific and medical contexts. Graf et al. (2024) assessed the reference accuracy of ChatGPT, Google Bard, and Chatsonic in generating citations for randomized controlled trials in obstetric literature, reporting that Google Bard achieved the highest reference accuracy at only 13.6%, while ChatGPT and Chatsonic exhibited significantly lower accuracy rates of 2.4% and 0%, respectively[22]. Similarly, Sawamura et al. (2024) noted that ChatGPT frequently provided nonexistent or misinterpreted references in physiotherapy-related clinical questions, reinforcing concerns about AI-generated citations[19].

The analysis of fictitious reference rates indicates a pattern based on question type. The lowest rates of fictitious references were observed in case-based clinical questions (e.g., Question 9: 25%), while the highest rates were found in general knowledge questions (e.g., Questions 1, 2, and 3: 100%). This suggests that ChatGPT's reference reliability is more consistent when responding to case-based scenarios requiring structured clinical reasoning, whereas general queries elicit a higher proportion of inaccurate citations. This discrepancy may be attributed to the model's reliance on broad textual patterns rather than direct retrieval of peer-reviewed sources. Additionally, these findings suggest that clinicians may benefit from formulating more detailed and specific questions, as

Istanbul Gelisim University Journal of Health Sciences (IGUSABDER) is indexed by TUBITAK ULAKBIM TR Index.
Web site: https://dergipark.org.tr/en/pub/igusabder
Contact: igusabder@gelisim.edu.tr

structured queries appear to yield more reliable references and clinically relevant responses.

The limitations of AI in generating reliable references may stem from its reliance on static training datasets that do not allow real-time access to up-to-date, peer-reviewed literature[23]. This issue is further compounded by the tendency of large language models to generate citations based on linguistic patterns rather than direct retrieval from indexed medical databases[24]. Kunt (2023) also highlights that the effectiveness of AI-driven tools in healthcare is directly linked to the accuracy and reliability of the information they provide[25]. Given these limitations, ensuring proper validation of AI-generated content remains a crucial requirement for safe clinical integration.

Several limitations should be acknowledged. This study evaluated only a specific version of ChatGPT, without comparisons to alternative AI models such as Google Bard or Claude. Given the rapid development of AI technologies, newer versions may demonstrate improved reference accuracy and response precision. Furthermore, the study was limited to CP- and ASD-related clinical questions, which may not fully capture the model's performance across broader occupational therapy domains. Additionally, the relatively small number of questions included in the study (n=10) may limit the statistical power and generalizability of the findings. The assessment was conducted by expert raters, which enhances reliability but may introduce subjective biases. Additionally, the study did not analyze ChatGPT's ability to adapt to different question styles or prompts, which could influence response accuracy. Furthermore, the questions and evaluations were conducted in English, which may have contributed positively to the accuracy of ChatGPT's responses but also presents a limitation, as the raters were not native English speakers. In addition, it is not clear how asking in the native language would have affected the accuracy of the ChatGPT responses.

Several recommendations could be made for future research. First, conducting comparative analyses between different versions of ChatGPT and other AI-based models would provide a more comprehensive understanding of their strengths and weaknesses. Additionally, investigating how ChatGPT responds to different question styles across various diagnostic groups could offer valuable insights into its ability to generate clinically relevant information tailored to specific conditions. Furthermore, developing a fully evidence-based ChatGPT-Academic model designed to support clinical decision-making could enhance its applicability in healthcare settings. Finally, enhancing AI models through training with reliable literature sources and developing more robust verification mechanisms would be essential in improving the accuracy and trustworthiness of AI-generated content in healthcare applications.

Disclosure statement: No potential conflict of interest was reported by the author(s).

Funding: This research received no external funding.

REFERENCES

1. Nori H, King N, McKinney SM, Carignan D, Horvitz E. Capabilities of GPT-4 on medical challenge problems. *arXiv*. 2023. doi: 10.48550/arXiv.2303.13375.

2. Liu Y, Han T, Ma S, et al. Summary of ChatGPT-related research and perspective towards the future of large language models. *Meta-Radiology*. 2023;100:017.

3. Hu K. ChatGPT sets record for fastest-growing user base - analyst note. *Reuters*. February 2, 2023. Accessed February 1, 2023. https://www.reuters.com/technology/chatgpt-sets-record-fastest-growing-user-base-analyst-note-2023-02-01.

4. Campbell WA IV, Chick JF, Shin D, Makary MS. Understanding ChatGPT for evidence-based utilization in interventional radiology. *Clin Imaging*. 2024;110:098.

5. Rao A, Pang M, Kim J. Assessing the utility of ChatGPT throughout the entire clinical workflow. *medRxiv*. 2023.

6. Garg RK, Urs VL, Agarwal AA, Chaudhary SK, Paliwal V, Kar SK. Exploring the role of ChatGPT in patient care (diagnosis and treatment) and medical research: A systematic review. *Health Promot Perspect*. 2023;13(3):183.

7. Von Ende E, Ryan S, Crain MA, Makary MS. Artificial intelligence, augmented reality, and virtual reality advances and applications in interventional radiology. *Diagnostics*. 2023;13(5):892.

8. Greengard S. ChatGPT: Understanding the ChatGPT AI chatbot. *Eweek*. December 9, 2022. https://www.eweek.com/big-data-and-analytics/chatgpt/.

9. Manning CD. Human language understanding & reasoning. *Proc Am Acad Arts Sci*. 2022;151(2):127-138. doi: 10.1162/daed_a_01905.

10. Mündler N, He J, Jenko S, Vechev M. Self-contradictory hallucinations of large language models: Evaluation, detection, and mitigation. *arXiv*. 2023.

11. Hoffmann T, Bennett S, Del Mar C. *Evidence-based practice across the health professions*. Philadelphia, PA: Elsevier Health Sciences; 2023.

12. Sharma N. Use of AI in pediatric occupational therapy: A review. *Rehabil Sci*. 2024;9(2):21-26. doi: 10.11648/j.rs.20240902.12

13. Berrezueta-Guzman S, Kandil M, Martín-Ruiz ML, de la Cruz IP, Krusche S. Exploring the efficacy of robotic assistants with ChatGPT and Claude in enhancing ADHD therapy: Innovating treatment paradigms. In: *2024*

International Conference on Intelligent Environments (IE). 2024:25-32. doi: 10.1109/IE61493.2024.10599903.

14. Morgan C, Fetters L, Adde L, et al. Early intervention for children aged 0 to 2 years with or at high risk of cerebral palsy: International clinical practice guideline based on systematic reviews. *JAMA Pediatr*. 2021;175(8):846-858. doi: 10.1001/jamapediatrics.2021.0878.

15. Steinbrenner JR, Hume K, Odom SL, et al. *Evidence-based practices for children, youth, and young adults with autism*. Chapel Hill, NC: The University of North Carolina at Chapel Hill, Frank Porter Graham Child Development Institute, National Clearinghouse on Autism Evidence and Practice Review Team; 2020.

16. Cahill SM, Beisbier S. Occupational therapy practice guidelines for children and youth ages 5–21 years. *Am J Occup Ther*. 2020;74(4):7404397010p1-7404397010p48.

17. Kundel HL, Polansky M. Measurement of observer agreement. *Radiology*. 2003;228:303-308.

18. Scaff SP, Reis FJ, Ferreira GE, Jacob MF, Saragiotto BT. Assessing the performance of AI chatbots in answering patients' common questions about low back pain. *Ann Rheum Dis*. 2025;84(1):143-149.

19. Sawamura T, Lee H, Martinez R. Assessing the reliability of ChatGPT-generated references in physiotherapy clinical decision-making. *Physiother Res J*. 2024;32(2):89-101.

20. Kaplan M, Çakar F, Bingöl H. Sağlık alanında yapay zekânın kullanımı: Derleme. *Muş Alparslan Univ Sağlık Bilim Derg*. 2024;4(3):75-85.

21. Akalin B, Veranyurt Ü. Sağlık bilimlerinde yapay zekâ tabanlı klinik karar destek sistemleri. *Gevher Nesibe J Med Health Sci*. 2022;7(18):64-73.

22. Graf A, Smith B, Johnson C. Evaluating the accuracy of AI-generated references in obstetric randomized controlled trials: A comparative study of ChatGPT, Google Bard, and Chatsonic. *J Obstet Res*. 2024;45(3):214-226.

23. Kim S. In the era of ChatGPT, can medical artificial intelligence replace the doctor? *Korean J Med*. 2023;98(3).

24. Waisberg E, Ong J, Masalkhi M, et al. GPT-4: A new era of artificial intelligence in medicine. *Ir J Med Sci*. 2023.

25. Kunt MM. Tıpta dil tabanlı yapay zeka kullanımı. *Anatolian J Emerg Med*. 2023;6(3):137-140.