ISSN: 3068-8965
Journal of Pediatric Advance Research

ATHENAEUM SCIENTIFIC PUBLISHERS

*Research Article*

# A Retrospective Cross Sectional Study Comparing the Accuracy of Artificial Intelligence Tools in the Diagnosis of Common Pediatric Emergencies

**Neha R Mahindrakar[1], Srutdi Kamalam Natarajan[2*], Arundhati Negi[3], Aniket Khamar[4], Aakansha Maria Rajeev[5]**

[1]SS Institute of Medical Sciences and Research Centre, Davangere, Karnataka, India
[2]Sri Ramachandra Institute of Higher Education and Research, Chennai,Tamil Nadu, India
[3]Kasturba Medical College, Manipal, Karnataka, India
[4]Kempegowda Institute of Medical Sciences, Bengaluru, Karnataka, India
[5]SDM College of Medical Sciences and Hospital, India

***Corresponding author:** Srutdi Kamalam Natarajan, Sri Ramachandra Institute of Higher Education and Research, Chennai, Tamil Nadu, India; Email: Srutdi@gmail.com

Citation: Mahindrakar NR, et al. A Retrospective Cross Sectional Study Comparing the Accuracy of Artificial Intelligence Tools in the Diagnosis of Common Pediatric Emergencies. J Pediatric Adv Res. 2025;4(2):1-6.

http://dx.doi.org/10.46889/JPAR.2025.4210

Received Date: 21-07-2025

Accepted Date: 04-08-2025

Published Date: 11-08-2025



## Abstract

Background: Artificial Intelligence (AI) has advanced tremendously over the past few years and has demonstrated great potential in supporting existing healthcare systems. Due to their rapid data processing, AI tools are increasingly being used by people to assess health concerns and make preliminary decisions about seeking health care. This study aims to analyse the diagnostic accuracy of AI tools in determining common pediatric emergencies.

Methods: 120 pediatric case reports were collected from open access, peer reviewed journals based on the inclusion criteria. The cases were rewritten in layman language and entered into three AI tools : ChatGPT, WebMD Symptom Checker and Mayo Clinic symptom checker and the top three differential diagnoses were analyzed. A standard 3-point scoring system was used to assess the diagnostic performance of each tool by two independent reviewers. Statistical analysis of the scores were done using Friedman's test followed by post hoc pairwise comparisons with Bonferroni correction.

Results: Analysis showed that ChatGPT had the highest average mean of 0.95 ± 0.94 and was ranked the highest with a mean rank of 2.42 out of the three tools studied. Friedman's test and post hoc analysis with Bonferroni correction confirmed the statistical significance ($p < .001$). There was no statistical difference between the scores for WebMD and Mayo Clinic symptom checkers (adjusted $p = 1.000$).

Conclusion: With increasing access and a large number of people resorting to AI as a reliable tool for diagnosis, ChatGPT seems to show potential in this regard. Although the diagnostic accuracy seems significant , it is important to consider the limitations of such tools which do not take into consideration the contextual information, social determinants of health and emotional aspects of clinical presentation.

**Keywords:** Artificial Intelligence; Pediatrics; Emergency; Diagnosis; Symptom Checkers; Accuracy

## Introduction

Artificial Intelligence (AI) is a machine-learning process that mimics human cognitive abilities by drawing inferences from new and existing data. Over the years, the use of AI has exhibited great potential in the field of medical diagnostics, which is quite complex and requires extensive training and experience to be able to gauge the differentials from presenting signs and symptoms. AI has shown promise in the context of medical diagnosis, appropriate testing and treatment options, even the possibility of identifying management plans for rare conditions. This can be attributed to the large datasets that AI has access to

and its ability to recognize patterns which might be beyond human capacity [1]. While AI tools offer an easy-to-use, rapid and accessible technology to its users, their diagnostic accuracy, especially when it comes to the pediatric age group, is a matter of concern. Inaccurate assessments and incorrect diagnoses given by them could lead to delay in patients getting critical care or unnecessary visits to the Emergency Department on the other hand. Limited studies exist in current literature that study the reliability and diagnostic accuracy of these tools in assessing pediatric emergency cases. In order to address this gap, we conducted a retrospective, cross-sectional study to evaluate and compare the diagnostic accuracy of three publicly available AI tools most used by caregivers in the United States for pediatric emergency health concerns- ChatGPT-4$^{o}$ (OpenAI), WebMD Symptom Checker and Mayo Clinic Symptom Checker.

**Methodology**

This retrospective, cross-sectional study was conducted to evaluate the diagnostic accuracy of three publicly available Artificial Intelligence (AI) symptom checkers which are most frequently used by caregivers for pediatric health concerns in the United States. The AI tools selected were ChatGPT-4$^{o}$ (OpenAI), WebMD Symptom Checker and Mayo Clinic Symptom Checker.

A total of 120 pediatric case reports were sourced from open access, peer-reviewed journals including SAGE Journals, Cureus, American Journal of Case Reports, BMC and Wiley Online Library. Cases were included based on the following criteria: (1) patients were under 18 years of age, (2) the presentation occurred in an emergency or acute care setting, (3) the clinical condition was relevant and commonly encountered. No rare or ambiguous presentations were included, (4) the report was published in or translated into English and (5) the case was published within the last 10 years.

Each case was rewritten in a standardized, caregiver friendly format. The language and tone were adapted to simulate a layperson speech, avoiding medical terminology to simulate realistic queries. The case descriptions were carefully constructed to include the presenting complaints, the duration and progression of symptoms, the impact on the child's daily activities (e.g., being too tired to play or having a poor appetite) and pertinent negatives.

Each standardized vignette was individually entered into the three AI symptom checkers. To ensure consistency, all reviewers used the same version of each tool throughout the study. For every interaction, the top three diagnoses suggested were recorded. Diagnostic performance was then assessed using a structured 3-point scoring system: a score of 2 was assigned if the correct diagnosis was listed first, 1 if it was included among the top three suggestions but not ranked first and 0 if it was not included in the top three at all. For each tool, two independent reviewers assessed the AI-generated outputs and assigned scores accordingly. Any disagreements in scoring were resolved through discussion to ensure inter-rater reliability.

*Data Analysis*

Null hypothesis - There would be no significant difference in diagnostic accuracy among the three AI tools.
Alternate hypothesis - There would be significant difference in the diagnostic accuracy among the three AI tools.
Descriptive statistics were used to calculate the mean diagnostic score for each tool - bar charts for average scores and stacked bar charts for the distribution of individual scores (0, 1, 2). To compare diagnostic performance, the Friedman test was used, followed by post hoc pairwise comparisons with Bonferroni correction.

**Results**

A total of 120 publicly available case reports were evaluated by ChatGPT, Mayo Clinic and WebMD to analyse the reliability of the tools. Scores obtained by each tool with respect to the standard scoring system is represented graphically (Fig. 1). Descriptive analysis showed that ChatGPT received the highest average score (Mean = 0.95, SD = 0.942), followed by WebMD (Mean = 0.24, SD = 0.622) and Mayo Clinic (Mean = 0.17, SD = 0.508) (Fig. 2).

Given the non-normal distribution or ordinal nature of the data, Friedman's Two-Way ANOVA by ranks was applied to compare the three sources. The test revealed a statistically significant difference in scores among the sources ($\chi^2 = 80.687$, df = 2, $p < .001$). The results are statistically significant at the 0.001 level, leading to rejection of the null hypothesis (Table 1). This indicates that there are significant differences in the distributions of scores across the three sources. The mean ranks further supported this finding, with ChatGPT ranked highest (Mean Rank = 2.42), followed by WebMD (1.82) and Mayo Clinic (1.75) (Table 2).

Post hoc pairwise comparisons with Bonferroni correction demonstrated that ChatGPT was rated significantly higher than both Mayo Clinic (adjusted $p < 0.001$) and WebMD (adjusted $p < 0.001$). However, no significant difference was found between the scores for Mayo Clinic and WebMD (adjusted $p = 1.000$) (Table 3, Fig. 3).

| | Null Hypothesis | Test | Sig.[a,b] | Decision |
|---|---|---|---|---|
| 1 | The distributions of Chat GPT Score, Mayo Score and WebMD Score are the same. | Related-Samples Friedman's Two-Way Analysis of Variance by Ranks | <.001 | Reject the null hypothesis. |
| a. The significance level is .050.<br>b. Asymptotic significance is displayed. | | | | |

**Table 1:** Hypothesis test summary.

| Mean Rank | |
|---|---|
| **Chat GPT Score** | 2.42 |
| **Mayo Score** | 1.75 |
| **WebMD Score** | 1.82 |

**Table 2:** Mean ranks of the AI tools.

| Sample 1-Sample 2 | Test Statistic | Std. Error | Std. Test Statistic | Sig. | Adj. Sig.[a] |
|---|---|---|---|---|---|
| Mayo Score-WebMD Score | -.067 | .129 | -.516 | .606 | 1.000 |
| Mayo Score-Chat GPT Score | .671 | .129 | 5.196 | <.001 | .000 |
| WebMD Score-Chat GPT Score | .604 | .129 | 4.680 | <.001 | .000 |
| Each row tests the null hypothesis that the Sample 1 and Sample 2 distributions are the same. Asymptotic significance (2-sided tests) are displayed. The significance level is .050.<br>a. Significance values have been adjusted by the Bonferroni correction for multiple tests. | | | | | |

**Table 3:** Pairwise comparisons.

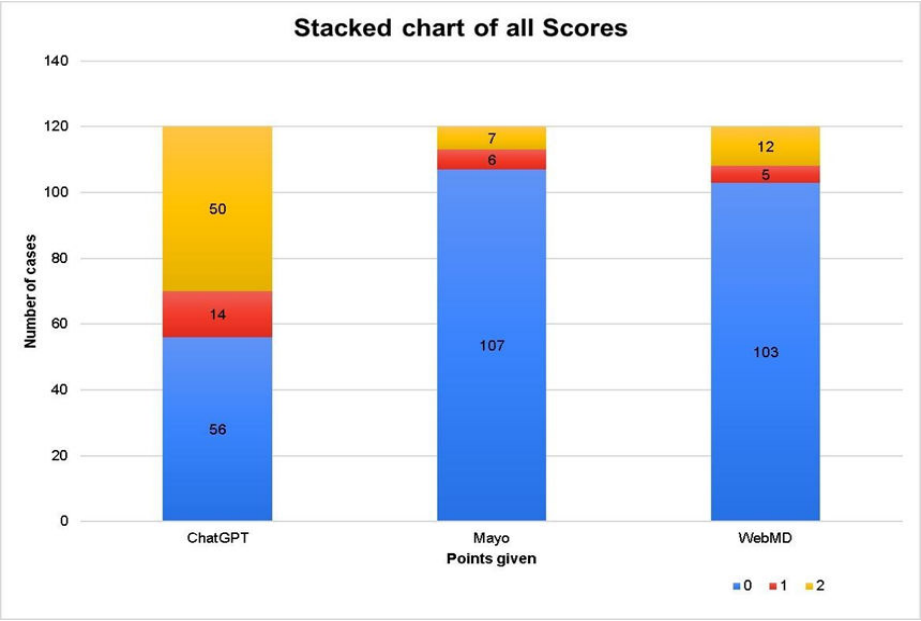


**Figure 1:** Stacked bar graph showing scores obtained by each individual AI tool.

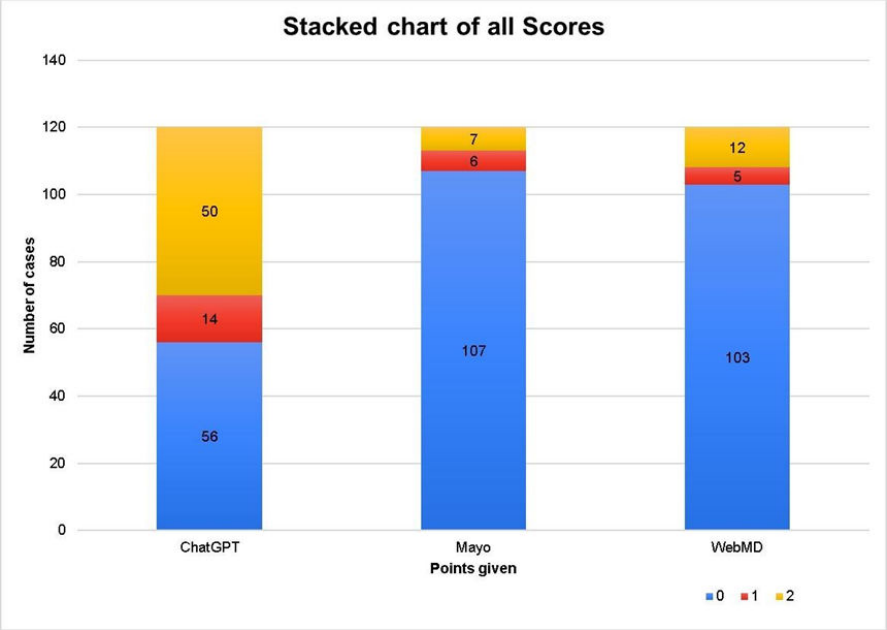


**Figure 2:** Mean scores of each individual AI tool.

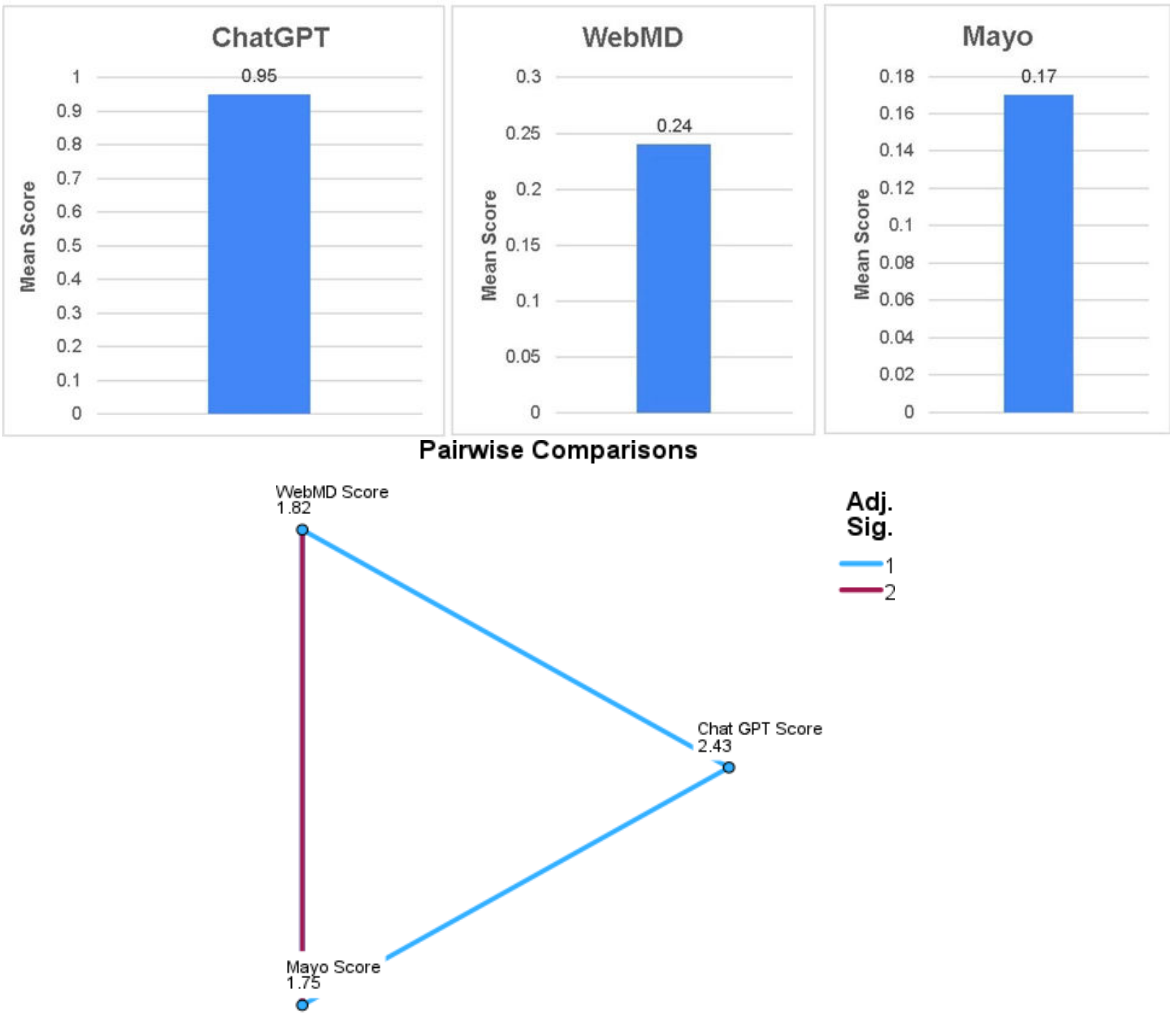


**Figure 3:** Pairwise comparisons.

## Discussion

This study evaluated the reliability of three online diagnostic tools - ChatGPT, WebMD and Mayo Clinic using 120 peer reviewed publically available pediatric case reports. Our findings indicate that ChatGPT significantly outperformed both WebMD and Mayo Clinic with a mean score of 0.95 ± 0.94 and mean rank of 2.42. The statistical significant differences were confirmed through Friedman's test and post hoc Bonferroni comparisons ($p < .001$).

Our results align with recent studies which demonstrate the high diagnostic accuracy of AI systems, particularly language models like ChatGPT. A study by Gasmi I, et al., found that ChatGPT achieved 93.94% diagnostic accuracy across 132 emergency scenarios and correctly identified 54.5% of high-urgency emergencies, though with inconsistencies in critical care advice [2]. Similarly, Huang SW, et al., found that a deep learning model trained on over 1.3 million pediatric visits demonstrated diagnostic accuracy that was on par with experienced pediatricians, showing the potential of AI in processing complex data [3].

The relevance of AI in specific pediatric scenarios like appendicitis further supports ChatGPT's utility. Various studies have described how AI models, using a combination of biomarkers, outperformed traditional indicators like CRP or leucocyte count [4,5]. The diagnostic power of AI is further supported by studies done in pediatric radiology. An AI algorithm achieved 95.6% sensitivity and performance close to pediatric radiologist in fracture detection [6].

ChatGPT is also shown to play a role in clinical decision making and triage patients with high accuracy, achieving Cohen's Kappa of 0.899 and classification accuracies between 85% and 90% for imaging and triage tasks [7,8]. Additionally, AI offers an opportunity to reduce over testing and misdiagnosis through standardized evidence based algorithms in areas where existing diagnostic tools show low sensitivity (0-17.8%), such as pediatric chest pain [9].

Beyond diagnostic accuracy, studies have demonstrated a favourable public perception and growing user trust in AI tools. A large majority of parents were comfortable with AI being used to interpret radiographs, decide on antibiotics or determine hospitalization needs for their children in studies done by Liang H and Sarno DL [10,11]. However, comfort levels varied - particularly among minority and younger caregiver populations.

In our study, WebMD and Mayo Clinic scored lower and showed no statistically significant difference between each other. This is consistent with the broader literature, where traditional digital health platforms often lack context sensitivity and adaptive reasoning that newer AI tools like ChatGPT provide [2]. Despite its promising capabilities, several studies demonstrate the limitations to the widespread clinical adoption of ChatGPT [4,12]. The collective findings of our study aligns with the existing literature and concludes that even though effective, AI tools can be ambiguous in real time. There is a need for model validation of AI tools and physician collaboration to ensure effective integration into diagnostic settings.

## Conflict of Interests

The authors declare that they have no conflicts of interest.

## Funding

This research did not receive any specific grant from funding agencies in the public, commercial or non-profit sectors.

## Acknowledgements

The authors would like to acknowledge Dr Gaurav Joshi, Associate Professor in Marketing, Lal Bahadur Shastri Institute of Management for their expert assistance with the statistical analysis conducted in this study.

## References

1. Alowais SA, Alghamdi SS, Alsuhebany N. Revolutionizing healthcare: the role of artificial intelligence in clinical practice. BMC Med Educ. 2023;23:689.
2. Gasmi I, Calinghen A, Parienti JJ. Comparison of diagnostic performance of a deep learning algorithm, emergency physicians, junior radiologists and senior radiologists in the detection of appendicular fractures in children. Pediatr Radiol. 2023;53:1675-84.

3. Huang S-W, Liu YK. Pediatric chest pain: a review of diagnostic tools in the pediatric emergency department. Diagnostics. 2024;14(5):526.
4. Berikol GB, Kanbakan A, Ilhan B, Doğanay F. Mapping artificial intelligence models in emergency medicine: A scoping review on artificial intelligence performance in emergency care and education. Turk J Emerg Med. 2025;25(2):67-91.
5. Bushuven S, Bentele M, Bentele S. ChatGPT, can you help me save my child's life?- Diagnostic accuracy and supportive capabilities to lay rescuers by ChatGPT in prehospital basic life support and paediatric advanced life support cases - an in-silico analysis. J Med Syst. 2023;47:123.
6. Chekmeyan M, Liu SH. Artificial intelligence for the diagnosis of pediatric appendicitis: A systematic review. Am J Emerg Med. 2025;92:18-31.
7. Thompson G, deForest E, Eccles R. Ensuring diagnostic accuracy in pediatric emergency medicine. Clin Pediatr Emerg Med. 2011;12(2):121-32.
8. Ramgopal S, Kapes J, Alpern ER. Perceptions of artificial intelligence-assisted care for children with a respiratory complaint. Hosp Pediatr. 2023;13(9):802-10.
9. Reismann J, Romualdi A, Kiss N. Diagnosis and classification of pediatric acute appendicitis by artificial intelligence methods: An investigator-independent approach. PLoS One. 2019;14(9):e0222030.
10. Liang H, Tsui BY, Ni H. Evaluation and accurate diagnoses of pediatric diseases using artificial intelligence. Nat Med. 2019;25:433-8.
11. Di Sarno L, Caroselli A, Tonin G. Artificial intelligence in pediatric emergency medicine: applications, challenges and future perspectives. Biomedicines. 2024;12(6):1220.
12. Ziegner M, Pape J, Lacher M. Real-life benefit of artificial intelligence-based fracture detection in a pediatric emergency department. Eur Radiol. 2025.