http://medicaleducation-bulletin.ir
Liter to Editor (Pages: 1417-1420)

Artificial Intelligence in Pediatric Emergency Education: Opportunities, Limitations, and Validation Requirements

*Hanieh Soltani[1]

[1]Department of Pediatrics, Clinical Research Development Center of Children's Hospital, Hormozgan University of Medical Sciences, Bandar Abbas, Iran.

Abstract

Artificial intelligence, particularly large language models, can support scenario generation, virtual patients, personalized learning, and feedback in pediatric emergency education. However, the time-sensitive and age-dependent nature of pediatric emergency care requires rigorous evaluation of these tools for accuracy, safety, and educational effectiveness. Key limitations include plausible misinformation, omission of critical actions, inconsistent responses, algorithmic bias, and risks to patient confidentiality. Strong performance on examinations does not establish practical competence or improved patient outcomes. This letter reviews evidence from medical education and pediatric resuscitation and proposes a staged validation framework encompassing the intended use and target learners, expert content review, response consistency, generalizability, assessment validity, learning retention and transfer, and data governance. Responsible integration requires faculty oversight and continuous monitoring, positioning artificial intelligence as a complement to clinical teaching and simulation rather than a replacement for professional judgment and hands-on practice.

Key Words: Artificial intelligence, Clinical simulation, Medical education, Pediatric emergency medicine, Validation.

*Please cite this article as: Soltani H. Artificial Intelligence in Pediatric Emergency Education: Opportunities, Limitations, and Validation Requirements. Med Edu Bull 2026; 7(2): 1417-20. DOI: **10.22034/meb.2026.612417.1154**

Corresponding Author:

Hanieh Soltani, MD, Faculty of Medicine, Hormozgan University of Medical Sciences, Bandar Abbas, Iran.

Email: Hani.sti2424@gmail.com

Received date: Jul. 25, 2026; Accepted date: Nov.12, 2026

Dear Editor,

The growing use of artificial intelligence (AI) in medical education offers opportunities to redesign educational content, support clinical reasoning, and provide feedback. The scoping review by Gordon et al. and the systematic review by Hallquist et al. identified diverse applications, including question generation, clinical scenario development, virtual patients, and learner assessment. However, heterogeneity in the tools studied, research methods, and outcome measures limits general conclusions regarding their effectiveness (1, 2). In pediatric education, large language models (LLMs) have similarly been proposed as resources for educators and learners, although their effective use requires an understanding of their limitations (3).

Educational opportunities: Interactive practice and targeted feedback

LLMs can assist with drafting educational scenarios, portraying patients or caregivers, supporting history-taking practice, and providing feedback. Adjusting the complexity of questions and explanations to learners' training levels may also support personalized learning. These applications have been discussed in medical and pediatric education, although their educational value depends on instructional design and implementation (2, 3).

In pediatric emergency education, these capabilities could support exercises involving information gathering from parents, prioritization of clinical problems, and explanation of clinical decision-making. We propose that AI-generated scenarios undergo review for alignment with learning objectives, learner level, and scientific accuracy before use. Within this approach, AI can expand opportunities for practice, while faculty retain responsibility for approving educational content.

Importantly, a model's ability to answer questions should be distinguished from its ability to improve learning. For example, Gilson et al. evaluated ChatGPT's performance on questions related to the United States Medical Licensing Examination. Such research examines a model's performance on knowledge-based tasks but does not, by itself, establish improvement in learners' clinical skills (4). Model examination performance should therefore not be treated as equivalent to educational effectiveness.

Limitations: The gap between correct diagnosis and safe action

Studies of pediatric emergencies demonstrate that diagnostic accuracy does not necessarily translate into adequate practical guidance. Bushuven et al. evaluated 22 basic life support and pediatric advanced life support scenarios and reported a diagnostic accuracy of 94%. Nevertheless, reliable recommendations to contact emergency medical services were provided in only 12 of the 22 scenarios, while inappropriate advanced life support interventions were recommended to parents in three scenarios (5). This was not an educational intervention involving students or residents; consequently, it does not directly measure educational effectiveness. It nevertheless illustrates the distinction between identifying a diagnosis and recommending safe action.

Kular and Kumar compared responses from two versions of ChatGPT to 13 questions on pediatric basic life support, yielding 52 outputs, against the 2020 American Heart Association guidelines. Only 9.61% of outputs were classified as fully comprehensive, and 5.77% fully conformed to the study's criteria (6). These findings are specific to the evaluated model versions, questions, reference guideline, and scoring approach and should not be generalized to all AI systems. Furthermore, conformity with the 2020 guidelines should not be interpreted as conformity with the most recent available recommendations.

In adult resuscitation, Beck et al. similarly demonstrated that conformity of individual statements with guidelines does not necessarily indicate comprehensive coverage of essential recommendations; the evaluated models omitted numerous key guideline messages (7). Although this study did not address pediatric resuscitation, its methodological implication is relevant: the correctness of the information provided and the completeness of the information required should be evaluated separately.

From an educational perspective, incomplete or incorrect outputs delivered confidently may encourage uncritical acceptance. Evaluation should therefore extend beyond fluency and overall

accuracy to include omitted actions, advice inappropriate for the intended audience, and potentially harmful errors (5–7).

Validation requirements: From content accuracy to educational outcomes

Based on the available evidence, we propose six domains for validating AI tools in pediatric emergency education. These domains represent an author-proposed framework rather than an established, independently validated standard.

1. Define the intended use, audience, and boundaries

Before evaluation, educators should specify whether the tool will generate content, portray a virtual patient, provide feedback, or assign scores. Learner level, language, educational setting, and the degree of faculty supervision should also be defined. Validation for a low-risk application should not be interpreted as authorization for high-stakes assessment or independent clinical guidance.

2. Evaluate content accuracy, completeness, and safety

Outputs should be reviewed by relevant experts against authoritative, current sources. We propose reporting critical errors separately from overall accuracy, because a favorable average score may conceal a dangerous recommendation. Pediatric resuscitation studies demonstrate the importance of independently evaluating diagnostic accuracy, adequacy of advice, and guideline conformity (5, 6).

3. Assess consistency and generalizability

We propose testing performance across repeated questions, alternative wording, incomplete information, and varied scenarios. The model name and version, evaluation date, prompts, and reference sources should be documented. For non-English applications, terminology, comprehensibility, and alignment with local resources should be evaluated independently. Linguistic and contextual limitations identified in existing research caution against automatically transferring findings between settings (2, 5).

4. Establish assessment validity and reliability

When AI assigns scores to learner performance, educators should establish what those scores represent. Johnsson et al. used Kane's validity framework to examine AI-based assessment through the inferences of scoring, generalization, extrapolation, and implications. Their study, conducted in a simulated procedural setting outside pediatric emergency medicine, identified limitations involving construct representation, explainability, and transfer to other datasets (8). Numerical agreement with human raters alone is therefore insufficient to establish the validity of an assessment tool.

5. Measure learning, retention, and transfer

Satisfaction and confidence should not be the only research outcomes. We propose that comparative studies measure immediate learning, knowledge retention, performance in unfamiliar scenarios, and the ability to complete tasks without AI assistance. Control groups and objective assessments are particularly important because reliance on self-assessment and the absence of comparison groups have limited parts of the medical education literature (2). Claims of improved patient outcomes require direct clinical evidence rather than improvements in examination scores alone.

6. Implement data governance and post-deployment monitoring

The World Health Organization's guidance on large multimodal models and Meskó and Topol's analysis of regulatory oversight emphasize data protection, accountability, and ongoing monitoring (9, 10). Accordingly, we propose using synthetic scenarios or appropriately de-identified data whenever possible. Entry of identifiable pediatric information into public systems should be governed by applicable institutional authorization and data-protection arrangements. Responsibility for content review, error-reporting pathways, and criteria for suspending use should be established in advance. Changes to model versions or reference sources should trigger reassessment.

Responsible integration into the curriculum

We propose beginning with lower-risk applications, such as generating draft materials subject to expert review. Tools should subsequently be evaluated in supervised simulation settings, with wider implementation contingent on acceptable evidence of safety and educational effectiveness. Learners should also receive instruction in source verification, recognition of incorrect responses, and independent explanation of their clinical reasoning. This approach aligns with the medical education literature's emphasis on understanding technological limitations and with the requirements of responsible governance (1, 2, 9, 10).

CONCLUSION

AI may expand opportunities for interactive practice, content development, and feedback in pediatric emergency education. However, fluent responses, diagnostic accuracy, and examination success do not independently establish educational validity or safe performance. Adoption should depend on a clearly defined intended use, expert review, evaluation of critical errors, evidence supporting assessment validity, measurement of retention and transfer, and protection of pediatric data. Within this framework, AI can complement faculty instruction, simulation, and clinical education rather than replace professional judgment and hands-on practice.

CONFLICT OF INTEREST: None.

REFERENCES

1. Gordon M, Daniel M, Ajiboye A, Uraiby H, Xu NY, Bartlett R, et al. A scoping review of artificial intelligence in medical education: BEME Guide No. 84. Med Teach. 2024 Apr;46(4):446-470. doi: 10.1080/0142159X.2024.2314198.

2. Hallquist E, Gupta I, Montalbano M, Loukas M. Applications of Artificial Intelligence in Medical Education: A Systematic Review. Cureus. 2025 Mar 1;17(3):e79878. doi: 10.7759/cureus.79878.

3. Suresh S, Misra SM. Large Language Models in Pediatric Education: Current Uses and Future Potential. Pediatrics. 2024 Sep 1;154(3):e2023064683. doi: 10.1542/peds.2023-064683. PMID: 39108227.

4. Gilson A, Safranek CW, Huang T, Socrates V, Chi L, Taylor RA, Chartash D. How Does ChatGPT Perform on the United States Medical Licensing Examination (USMLE)? The Implications of Large Language Models for Medical Education and Knowledge Assessment. JMIR Med Educ. 2023 Feb 8;9:e45312. doi: 10.2196/45312. Erratum in: JMIR Med Educ. 2024 Feb 27;10:e57594. doi: 10.2196/57594.

5. Bushuven S, Bentele M, Bentele S, Gerber B, Bansbach J, Ganter J, et al. "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 Nov 21;47(1):123. doi: 10.1007/s10916-023-02019-x.

6. Kular S, Kumar V. Artificial Intelligence Chatbots in Pediatric Emergencies: A Reliable Lifeline or a Risk? Cureus. 2025 Aug 1;17(8):e89234. doi: 10.7759/cureus.89234. PMID: 40901221; PMCID: PMC12401188.

7. Beck S, Kuhner M, Haar M, Daubmann A, Semmann M, Kluge S. Evaluating the accuracy and reliability of AI chatbots in disseminating the content of current resuscitation guidelines: a comparative analysis between the ERC 2021 guidelines and both ChatGPTs 3.5 and 4. Scand J Trauma Resusc Emerg Med. 2024 Sep 26;32(1):95. doi: 10.1186/s13049-024-01266-2.

8. Johnsson, V., Søndergaard, M. B., Kulasegaram, K., Sundberg, K., Tiblad, E., Herling, L., Petersen, O. B., & Tolsgaard, M. G. Validity evidence supporting clinical skills assessment by artificial intelligence compared with trained clinician raters. Medical Education, 2024;*58*(1): 105-117. https://doi.org/10.1111/medu.15190.

9. Malpani, R., Reis, A. A., Pujari, S., Reeder, J., Vayena, E., Labrique, A., et al. Ethics and governance of artificial intelligence for health: Guidance on large multi-modal models (LMMs). World Health Organization; 2024.

10. Meskó B, Topol EJ. The imperative for regulatory oversight of large language models (or generative AI) in healthcare. NPJ Digit Med. 2023 Jul 6;6(1):120. doi: 10.1038/s41746-023-00873-0.