Emergency Care Journal

Official Journal of the Academy of Emergency Medicine and Care (AcEMC)

eISSN 2282-2054

https://www.pagepressjournals.org/index.php/ecj/index

Emerg Care J 2026 [Online ahead of print]

To cite this Article:

Shinde V, Kanani P, Bhimani K. **Potential and limitations of large language models in acute chest pain triage. Response to *Evaluating the predictive accuracy of ChatGPT in risk stratification for chest pain in the emergency department*.** *Emerg Care J* doi: 10.4081/ecj.2026.15130

Submitted: 8 March 2026 Accepted: 2 April 2026 Early access: 15 May 2026

Potential and limitations of large language models in acute chest pain triage. Response to *Evaluating the predictive accuracy of ChatGPT in risk stratification for chest pain in the emergency department*

Varsha Shinde, Pratik Kanani, Keyur Bhimani

Department of Emergency Medicine, Dr. D. Y. Patil Medical College, Hospital and Research Centre, Dr. D. Y. Patil Vidyapeeth (Deemed to be University), Pimpri–Chinchwad, Maharashtra, India

Correspondence: Pratik Kanani, Department of Emergency Medicine, Dr. D. Y. Patil Medical College, Hospital and Research Centre, Dr. D. Y. Patil Vidyapeeth (Deemed to be University), Pimpri Colony, Pimpri–Chinchwad 411018, Maharashtra, India. Email: dr.pratik243@gmail.com

Key words: ChatGPT, emergency department, chest pain, MACE, artificial intelligence.

Dear Editor-in-Chief,

We were especially interested to find out about this recently published article by Malalan *et al.*, which discussed the evaluation of ChatGPT 4.0 as a clinical decision-support tool for predicting Major Adverse Cardiac Events (MACE) in patients presenting with chest pain to the Emergency Department (ED).[1] This novel study examines an innovative application of Large Language Models (LLMs) in a critical and time-sensitive clinical setting.

With 178 patients, the authors conducted and have succeeded in a prospective observational study, where it evaluated ChatGPT's score at three successive stages of patient evaluation — initial clinical history and ECG, first troponin test, and second troponin test. The results showed an encouraging pattern with ChatGPT's predictive accuracy (Area Under the ROC curve) improving steadily from 0.699 to 0.776 ($p=0.039$) with the addition of clinical data. The model also showed some high negative predictive value, encouraging the possibility of targeting low-risk patients who could avoid avoidable hospitalisation.

However, in spite of these positive aspects, important limitations identified by the report that currently prevent ChatGPT to be utilized as clinical routine in the ED as its application are well understood. Most significantly, ChatGPT exhibited suboptimal sensitivity and specificity trade-offs with a large proportion of MACE case mis-reported being non-urgent or urgent, as opposed to very urgent. These misclassifications threaten substantial clinical risk due to the rapid progression of acute coronary syndromes. This low positive predictive value highlights the complexity in the process of risk stratification (still a key requirement to promote effective and safe patient care through this disease paradigm). In addition, the exclusion of persons presenting with atypical features or symptoms limits the generalizability of the results.

Also, the emphasis given by the authors on ChatGPT being so simple and accessible compared to the improved prediction accuracy achieved by more advanced Machine Learning (ML) models is highly significant. Recent studies that leveraged ML algorithms like random forest classifiers and dimensionality reduction methods yielded AUROC predictions over 0.9 in comparable cohorts, suggesting better-quality discrimination when predicting MACE.[2,3] However, such ML tools require complex infrastructure, integration with electronic health records, and expertise; thus, they are not commonly adopted. In this sense, however, LLMs such as ChatGPT are seen as an appealing, user-friendly complementary that can bring closer the future of advanced predictive analytics and clinical operation. However, clinical integration will only be effective if clinicians can realize the limitations of AI tools and how to enhance them with clinical judgment.[4,5]

Continuous dynamic reassessment in the course of the ED visit is standard practice in clinical practice. The design of the study, to feed sequential data stages into ChatGPT and update these risk estimates as necessary, fits comfortably into this paradigm. It shows a hopeful path for AI tools not simply to examine static snapshots, but to learn their clinical judgment as more information accumulates. But the results also highlight how much LLMs need refining, whether that is in terms of more advanced clinical reasoning, or more specific risk estimation, for their role in providing robust support for real-time decision-making in high-stakes settings such as emergency cardiology.

We encourage future research to address the following areas: i) comparative studies against LLM performance versus clinicians' judgment and risk scores from current studies (HEART, TIMI) that are observed in real-time. Such studies are crucial for understanding the areas in which, and how AI tools could potentially help providers improve effectiveness by boosting their performance without undermining patient safety;[4] ii) extending the findings to diverse patient presentation, including atypical chest pain and other associated symptoms, in order to make the models more robust and applicable to a wider range of clinical individuals; iii) Integrate AI workflows that leverage the interpretative complexity of ML models, along with the natural language processing conversational interface and accessibility of LLMs, often via hybrid systems; iv) exploring explainability properties of LLMs for enhanced transparency and clinician trust necessary for adoption in emergency settings.

We commend Malalan *et al.* for their expertise and thought-leadership in the advancement of our understanding of AI in emergency contexts, a space that is under critical demand for change. Their work will ultimately open on the groundwork for building and responsibly deploying AI decision-support systems that would help to increase diagnostic precision, use resources more optimally, and better the outcome for patients.

Thank you for the opportunity to engage in this important discourse.

References

1. Malalan F, Zaboli A, Fiore A, et al. Evaluating the predictive accuracy of ChatGPT in risk stratification for chest pain in the emergency department. Emerg Care J 2025;21:13829

2. Zhang PI, Hsu CC, Kao Y, et al. Real-time AI prediction for major adverse cardiac events in emergency department patients with chest pain. Scand J Trauma Resusc Emerg Med 2020;28:93.

3. Liu N, Chee ML, Koh ZX, et al. Utilizing machine learning dimensionality reduction for risk stratification of chest pain patients in the emergency department. BMC Med Res Methodol 2021;21:74.

4. Backus BE, Six AJ, Kelder JC, et al. A prospective validation of the HEART score for chest pain patients at the emergency department. Int J Cardiol 2013;168:2153-8.

5. Sim JZT, Fong QW, Huang W, Tan CH. Machine learning in medicine: what clinicians should know. Singapore Med J 2023;64:91-7.

Contributions: Pratik Kanani, conceptualization, literature review, drafting, and critical revision; Varsha Shinde, literature review, critical revision, and supervision; Keyur Bhimani, conceptualization, editing, and final approval. All authors approved the final manuscript and agree to be accountable for all aspects of the work.

Conflict of interest: the authors declare that they have no conflict of interest.

Ethics approval, consent to participate, and consent for publication: not applicable.

Availability of data and materials: all data are available in the present paper.

Funding: the authors received no external funding for this work.