Evaluating the predictive accuracy of ChatGPT in risk stratification for chest pain in the emergency department

Fabio Malalan,[1] Arian Zaboli,[2] Armando Fiore,[1] Nicola Fanton,[1] Laura Sponga,[1] Barbara Zaia,[1] Giovanni Mantelli,[1] Caterina Barberi,[1] Paola Pisani,[1] Laura Giorio,[1] Gianni Turcato[3]

[1]Emergency Department, S. Maria del Carmine Hospital, Azienda per i Servizi Sanitari di Trento, Rovereto; [2]Innovation, Research and Teaching Service, Teaching Hospital of the Paracelsus Medical Private University, Bolzano; [3]Department of Internal Medicine, Intermediate Care Unit, Hospital Alto Vicentino, Santorso, Italy

Abstract

Chest pain is a frequent cause of emergency department (ED) visits, yet accurately assessing the risk of major adverse cardiac events (MACE) remains challenging. This study evaluated the potential of ChatGPT 4.0 as a clinical decision-support tool for predicting MACE in patients with chest pain. We conducted a prospective observational study at the Rovereto Hospital ED from March to August 2024, analyzing 178 patients. ChatGPT received patient data in three sequential phases: initial clinical history and ECG, first troponin test, and second troponin test. Its predictive performance improved with additional data, with the area under the receiver operating characteristic curve (AUROC) increasing from 0.699 in the first phase to 0.776 after the second troponin test (p=0.039). However, ChatGPT misclassified several MACE cases as non-urgent, raising concerns about sensitivity and risk stratification. While ChatGPT demonstrated potential in identifying MACE cases, its current performance does not support routine ED implementation. Further refinements in AI-based models are needed to improve real-time risk assessment and ensure safer, more reliable clinical integration.

Correspondence: Arian Zaboli, Innovation, Research and Teaching Service, Teaching Hospital of the Paracelsus Medical Private University, via A. Volta 13, Bolzano, Italy.
E-mail: zaboliarian@gmail.com

Key words: ChatGPT, emergency department, chest pain, MACE, artificial intelligence, decision support, risk assessment.

Contributions: FM, AF, conceptualization, methodology, investigation, data curation; AZ, formal analysis, writing-original draft preparation, writing - review & editing; NF, LS, BZ, GM, CB, PP, LG, investigation; GT, conceptualization, methodology, writing-original draft preparation.

Ethics approval and consent to participate: the study was approved by the local ethics committee (Comitato Etico Territoriale della Provincia Autonoma di Trento per le sperimentazioni clinche, Trento, Italia, approval number 02/2024) and was conducted according to the Declaration of Helsinki, adhering to the ethical principles for medical research involving human subjects. All participants provided written consent to participate in the study.

Conflict of interest: the authors have no conflict of interest to declare.

Availability of data and materials: data generated or analyzed during the study are included in this published article.

Received: 17 March 2025.
Accepted: 29 April 2025.
Early view: 6 June 2025.

Emergency Care Journal 2025; 21:13829
doi:10.4081/ecj.2025.13829

Introduction

Chest pain is one of the primary reasons for Emergency Department (ED) visits, accounting for over 5% of daily presentations. Although the percentage of patients who actually experience a Major Adverse Cardiovascular Event (MACE) is less than 10%, accurately and safely ruling out MACE in patients presenting with chest pain remains a significant challenge.[1-3] Following a comprehensive clinical, electrocardiographic, and laboratory evaluation, a substantial proportion of patients can already be stratified as low or no risk and safely discharged.[1-3] However, there remains a non-negligible group of patients for whom risk management remains uncertain, as current clinical tools do not always provide definitive information regarding the risk of MACE, particularly in cases with atypical or ambiguous presentations.[1-3] In recent years, various risk stratification systems, such as the HEART score and the Emergency Department Assessment of Chest Pain Score (EDACS), have been introduced to assist clinicians in identifying low-risk patients who can be safely discharged from the hospital.[4,5] However, despite the advancements offered by these tools, a significant percentage of patients are still admitted at the conclusion of the diagnostic process due to the inability to fully exclude a potential MACE.[6,7] This uncertainty can lead to an increase in hospital admissions or the performance of invasive procedures that ultimately yield negative results.

Artificial Intelligence (AI) is making significant inroads into various areas of medicine. Among the many proposed applications, AI's support in complex decision-making processes could help clarify patient risk levels, particularly in cases of chest pain.[8] Recently, a systematic review examined the effectiveness of Machine Learning (ML) in the assessment of chest pain, demonstrating that it is an excellent strategy for risk stratification and for identifying patients at risk of acute myocardial infarction or

MACE.[9] By providing advanced decision support, AI could help to more accurately identify low-risk patients, thereby improving the operational efficiency of ED and reducing costs associated with unnecessary hospital admissions or negative invasive procedures.[9,10] Nevertheless, despite extensive research, AI has not yet been standardized for use in ED settings. This may be due to challenges in the utilization and implementation of certain AI systems, such as machine learning. For this reason, the study explored the capability of a user-friendly Large Language Model (LLM), such as ChatGPT, to support clinical decision-making during the complex assessment of patients with chest pain in the ED.

Materials and Methods

Study design and setting

This is a prospective observational study conducted between March 1st, 2024, and August 31st, 2024, at the ED of Rovereto Hospital (Italy), involving patients presenting with chest pain. All patients visiting the ED undergo triage procedures. Excluding those patients who, according to specific protocols, are directly referred for specialized evaluation *via* a fast-track process, patients are assessed by the ED medical team. There is no specific management protocol in place for chest pain; thus, after triage assessments, patients with this symptom are evaluated in the ED's medical examination room.

Participants

All patients presenting to the ED with chest pain were considered for inclusion in the study. Patients who, in addition to chest pain, reported other symptoms indicative of possible cardio-respiratory conditions, such as dyspnea, syncope, palpitations, or abdominal pain (unless considered potential secondary or associated symptoms of chest pain), were excluded. Additional exclusion criteria were as follows: i) age under 18 years; ii) direct admission to the Shock Room due to emergency services alert (112) or immediately critical/unstable clinical conditions upon ED arrival, according to triage protocols; iii) failure to provide informed consent, or inability to do so independently; iv) known or suspected pregnancy; v) life expectancy of less than six months due to chronic or oncological disease; vi) patients who, at the time of ED evaluation, were deemed unable to undergo potential coronary angiography within the next three months due to comorbidities or frailty; vii) presence of ST-Elevation Myocardial Infarction (STEMI).

Patients with cardiovascular symptoms or abdominal pain complaints were excluded to focus the study on cases of pure chest pain presentation, reducing heterogeneity and ensuring a more focused evaluation of ChatGPT's predictive capabilities. Each patient was required to provide informed consent to participate in the study.

Study protocol and data collection

The study was conducted during daytime shifts on weekdays. Two dedicated nurses assisted ED physicians with data collection and interaction with ChatGPT. The nurses were provided with a hospital computer equipped with 4.0 and a standardized data collection form. This form specified the required information and outlined the procedural steps to follow. When a patient with chest pain arrived in the ED, they were flagged by triage nurses for enrollment in the study. The study was structured into three phases.

Initial evaluation

During the initial medical assessment, the physician completed a form that included the following information: i) onset of pain: time elapsed, in hours, from the onset of chest pain symptoms to arrival at the ED; ii) presence of chest pain in the ED: whether chest pain was currently experienced by the patient in the ED; iii) duration of pain: duration of the pain either currently felt by the patient in the ED; iv) pain quality: nature of the pain, categorized as oppressive/compressive, heavy/tight, or penetrating/stabbing/pleuritic; v) primary pain location: initial site of the pain, specified as retrosternal, precordial, hemithorax/ neck/jugular/mandible, or apical; vi) pain radiation: presence and location of any pain radiation, specified as bilateral arms, left arm or shoulder, dorsal, neck, or jaw; vii) associated symptoms: any additional symptoms present, including dyspnea, palpitations, nausea, vomiting, or sweating; viii) comorbidities: patient's comorbid conditions, with a focus on cardiovascular risk factors as outlined by major international guidelines; ix) vital signs: key vital signs recorded at the time of evaluation; x) ECG interpretation: textual description and binary assessment indicating the presence of ST abnormalities or T-wave changes.

This data was then input into ChatGPT 4.0, along with the patient's sex and age, maintaining anonymity throughout. Three questions were subsequently posed to ChatGPT: i) Is the patient currently experiencing a MACE? (binary response: yes/no); ii) How would you classify the severity of the patient's condition? (three levels: non-urgent, urgent, very urgent); iii) What percentage risk would you assign to the patient for developing a MACE? (0 to 100 percent). The first question, framed as a binary query, was designed to simulate ChatGPT's ability to support MACE risk stratification, rather than to diagnose an event in real time. The second question aimed to assess the model's ability to infer clinical urgency, with categories inspired by triage logic. These were not derived from a formal triage system, but served as a proxy for care prioritization based on available clinical data.

Intermediate evaluation

After obtaining the results of the initial troponin test, this information was incorporated into ChatGPT. The same three questions were then posed again to ChatGPT: i) Is the patient currently experiencing a MACE? (binary response: yes/no); ii) How would you classify the severity of the patient's condition? (three levels: non-urgent, urgent, very urgent); iii) What percentage risk would you assign to the patient for developing a MACE? (0 to 100 percent). These questions allowed ChatGPT to update its assessment of the patient's condition based on the additional troponin data. The cardiac biomarker used in this study was the high-sensitivity cardiac troponin I (hs-cTnI) assay developed by Abbott (ARCHITECT STAT High Sensitive Troponin-I). This assay detects troponin I levels with a 99th percentile upper reference limit of 34 ng/L for males and 16 ng/L for females, and a limit of detection of 1.2 ng/L. All measurements were conducted using the standardized laboratory protocol of our institution.

Final evaluation

If a second troponin test was required for the patient, the result was once again entered into ChatGPT, and the three questions were repeated. ChatGPT processed data up to phase B for all patients, while additional details were incorporated for those requiring a second troponin test (hs-cTnI). The decision to perform a second troponin test was based on the 2023 ESC Guidelines on the management of acute coronary syndromes.[1] Specifically, a second

test was performed in cases where the first hs-cTnI result was below or close to the upper reference limit, but clinical suspicion of myocardial infarction remained. This approach allowed for the detection of dynamic changes in troponin levels, enabling rule-in or rule-out decisions in line with current diagnostic algorithms. The nurses, in addition to interfacing with ChatGPT, recorded all information in a database to link the AI assessments with the clinical outcomes of the patients. This allowed for a comprehensive dataset that aligned ChatGPT's evaluations with actual patient results, enabling a deeper analysis of AI's effectiveness in supporting clinical decision-making in chest pain cases.

AI Structuring

ChatGPT 4.0 was specifically configured to support this study, employing a highly customized approach for the analysis of clinical data. The model was supplemented with open-access guidelines from the American Heart Association and the European Society of Cardiology to ensure that its recommendations were based on the most recent scientific evidence. In addition to specific guidelines for chest pain and myocardial infarction, the system was enhanced by including comprehensive cardiological guidelines covering chronic disease management, arrhythmias, heart failure, and cardiovascular prevention. This extensive information base improved the model's predictive capabilities, ensuring that the decision support provided was applicable not only to acute emergencies but also to more complex and nuanced clinical scenarios. To ensure accurate and consistent data interpretation, an information integration process was implemented, allowing ChatGPT to process multiple, context-specific clinical inputs such as medical history, vital signs, ECG findings, and biomarkers. The system updated its assessments as new data became available, such as troponin results. The AI was designed with stringent data confidentiality protocols, with a particular focus on patient anonymization, and was programmed to provide standardized responses, minimizing the risk of interpretive variability.

Additionally, ChatGPT 4.0 was trained to recognize recurrent patterns associated with MACE, grounded in established guidelines. This enabled the system to deliver a personalized risk assessment for each patient, offering both dichotomous indications regarding the presence of a MACE and a quantitative estimate of the likelihood of developing one.

Outcome

The primary outcome of the study was the occurrence of a MACE within three months following the patient's evaluation in the ED, as outlined in previous research.[5] The MACE outcomes were defined as follows: i) acute myocardial infarction, according to the universal definition established in the dedicated ESC consensus; ii) percutaneous coronary intervention, referring to non-surgical revascularization procedures using PTCA techniques and/or stent placement; iii) coronary artery bypass grafting, referring to surgical revascularization procedures; iv) coronary angiography indicating stenoses that could not be corrected by PCI but were managed with conservative medical therapy; v) death from any cause.

Two secondary outcomes were also evaluated: i) the ability of ChatGPT to assign an appropriate urgency level to each patient (non-urgent, urgent, very urgent), based on the clinical information available at each phase; ii) the model's yes/no classification of whether a MACE was likely, evaluated against actual outcomes.

Statistical analysis

Continuous variables in the study were reported as median and Interquartile Range (IQR) or as mean and Standard Deviation (SD), depending on their distribution. Categorical variables were presented as counts and percentages. Univariate comparisons were conducted using the chi-square test, Mann-Whitney U test, or Student's *t*-test, as appropriate.

ChatGPT's performance was evaluated using 2x2 contingency tables, comparing the actual MACE diagnoses with those patients identified by ChatGPT as having MACE across all three study phases. Sensitivity, specificity, Negative Predictive Value (NPV), and Positive Predictive Value (PPV) were reported with their 95% confidence intervals (95% CI).

The severity assigned by ChatGPT to each patient was analyzed using contingency tables. The percentage risk assigned by ChatGPT across the three study phases was used to assess its predictive ability against actual MACE outcomes using Receiver Operating Characteristic (ROC) curves, which were reported with their 95% CI. Finally, the clinical utility of ChatGPT at each study phase was analyzed using Decision Curve Analysis (DCA). DCA is used to assess the practical utility of a tool in clinical settings by comparing it to two opposing strategies and evaluating net clinical benefit across varying threshold probabilities.

Results with $p<0.05$ were considered statistically significant. All analyses were conducted using STATA 16.1 statistical software (StataCorp. College Station, Texas: StataCorp LLC).

Ethical consideration

The study was approved by the Local Ethics Committee (Comitato etico territoriale della provincia autonoma di Trento per le sperimentazioni clinche, Trento, Italia, approval number 02/2024) and was conducted according to the Declaration of Helsinki, adhering to the ethical principles for medical research involving human subjects.

Results

During the study period, 178 patients with chest pain were enrolled. Of these, 12.9% (23/178) experienced a MACE, while the remaining 87.1% (155/178) did not. The clinical and historical characteristics of the patients are detailed in Table 1.

As shown in Table 1, there were no significant imbalances in clinical or historical characteristics between the two groups. However, an ECG identified as abnormal by the ED physician and a higher HEART score were more frequently observed in patients who experienced a MACE compared to those who did not. Table 2 presents the 2x2 contingency tables where ChatGPT was tasked with identifying patients at risk for MACE.

As reported in Table 2, during the initial evaluation, where only clinical history and an already-interpreted ECG were available, ChatGPT demonstrated high specificity (70.9%; 95% CI: 64.3-77.6) and a high Negative Predictive Value (NPV) (91.7%; 95% CI: 87.6-95.7). However, this was at the expense of lower sensitivity (56.5%; 95% CI: 49.2-63.8) and Positive Predictive Value (PPV) (22.4%; 95% CI: 16.3-28.5). Following the second and third evaluations, which incorporated the results of the first and second troponin tests, ChatGPT demonstrated improved sensitivity (82.6%; 95% CI: 77.0-88.2) and NPV (95.6%; 95% CI: 92.6-98.6), though with decreased specificity (56.1%; 95% CI: 48.8-63.4) and PPV (21.8%; 95% CI: 15.8-27.9). ChatGPT's ability to distinguish between more and less severe cases is detailed in Table 3. Table 3 shows that in the first

phase, ChatGPT classified 43.5% (10/23) of patients who eventually experienced a MACE as non-urgent. However, after the first troponin test, only 17.4% (4/23) of patients with a MACE were classified as non-urgent. Even after all available information was considered, 17.4% (4/23) of patients with a MACE continued to be categorized as non-urgent. Conversely, among patients who did not experience a MACE, changes in urgency classification were also observed across the three study phases. In the initial phase, only 10.9% (17/155) of patients without MACE were classified as very urgent. After the second evaluation, this figure rose to 16.1% (25/155), and by the third evaluation, 18.7% (29/155) were categorized as very urgent.

Additionally, ChatGPT was asked to estimate the probability of a MACE on a scale from 0 to 100 during each of the three study phases. The predictive ability of ChatGPT was assessed using ROC curves, as presented in Figure 1.

ChatGPT's predictive ability improved significantly over the three phases, with its accuracy progressively increasing at each step. In the initial evaluation, where only clinical history and ECG data were available, ChatGPT demonstrated an area under the receiver operating characteristic curve (AUROC) of 0.699 (95% CI: 0.602-0.797). With the addition of the first troponin test during the second evaluation, the AUROC rose to 0.757 (95% CI: 0.662-0.852). Finally, upon incorporating the second troponin test, the AUROC reached 0.776 (95% CI: 0.681-0.871). This improvement in predictive ability across the phases was statistically significant (p=0.039). Furthermore, ChatGPT's performance at each of the three evaluation points in identifying MACE was assessed using DCA. As shown in Figure 2, the DCA illustrates an increase in the clinical utility of ChatGPT as more information became available, with the best performance observed during the final evaluation. Figure 2 demonstrates that the improvement in predictive ability is evident at lower threshold probabilities and proves useful only during the intermediate and final evaluations. This indicates that

Table 1. Clinical and anamnestic characteristics of study patients, divided between those with and without a MACE.

Variables	No MACE	MACE	p
Patients, n (%)	155 (87.1)	23 (12.9)	
Sex, n (%)			0.245
Male	88 (56.8)	16 (69.6)	
Female	67 (43.2)	7 (30.4)	
Age in years, median (IQR)	64 (48-78)	70 (51-81)	0.274
Duration of chest pain in minutes, median (IQR)	8 (2-48)	12 (1-72)	0.992
Location of the chest pain, n (%)			
Retrosternal	81 (52.3)	17 (73.9)	0.051
Precordial	11 (7.1)	2 (8.7)	0.783
Neck/jugulat/mandible	49 (31.6)	4 (17.4)	0.164
Apical	20 (12.9)	0 (0.0)	0.067
Quality of the chest pain, n (%)			
Oppressive/compressive	36 (23.2)	8 (34.8)	0.231
Heavy/tight	86 (55.5)	14 (60.9)	0.627
Penetrating/stabbing/pleuritic	43 (27.7)	2 (8.7)	0.050
Pain radiation, n (%)	46 (29.7)	6 (26.1)	0.724
Associated symptoms, n (%)			
Palpitations	14 (9.0)	5 (21.7)	0.066
Nausea and/or vomiting	18 (11.6)	1 (4.3)	0.292
Dyspnea	26 (16.8)	7 (30.4)	0.116
Sweating	5 (3.2)	0 (0.0)	0.382
Previous pathology, n (%)			
Hypertension	61 (39.3)	11 (47.8)	0.440
Ischemic heart disease	25 (16.1)	7 (30.4)	0.095
Chronic heart failure	4 (2.6)	1 (4.3)	0.632
Dementia	1 (0.6)	0 (0.0)	0.699
Chronic kidney failure	1 (0.6)	0 (0.0)	0.699
Diabetes	18 (11.6)	6 (26.1)	0.058
Tumor	14 (9.0)	4 (17.4)	0.215
ECG, n (%)			
Abnormal	42 (27.1)	13 (56.5)	0.004
ST abnormalities	19 (12.3)	6 (26.1)	0.075
T-wave changes	6 (3.9)	3 (13.0)	0.061
Vital parameters, median (IQR)			
Systolic BP (mmHg)	144 (130-160)	155 (135-190)	0.069
Diastolic BP (mmHg)	80 (75-90)	85 (70-97)	0.368
Heart rate (bpm)	80 (67-90)	78 (70-95)	0.986
Peripheral oxygen saturation (SpO_2)	97 (96-98)	97 (96-99)	0.652
Respiratory rate (breath per minute)	18 (17-18)	18 (16-19)	0.595
HEART score, mean (SD)	3.5 (1.9)	5.8 (1.8)	<0.001

MACE, major adverse cardiac events; IQR, interquartile range; BP, blood pressure; SD, standard deviation.

ChatGPT's predictive utility becomes more clinically relevant as the model incorporates additional data, such as sequential troponin results.

Discussion

This study was the first to evaluate the use of ChatGPT in the clinical practice of the ED for assessing patients with chest pain. The findings demonstrate that ChatGPT's predictive capabilities for MACE improve as more information becomes available. However, while the predictive ability is promising, it is not yet sufficiently high to support the safe implementation and practical use of this tool in clinical settings, particularly regarding the reliable exclusion of risk.

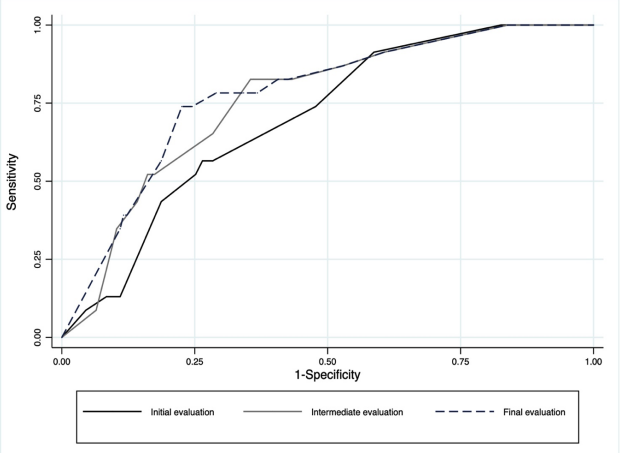

Figure 1. Predictive ability of ChatGPT across study phases, assessed using ROC curves for identifying patients with MACE.

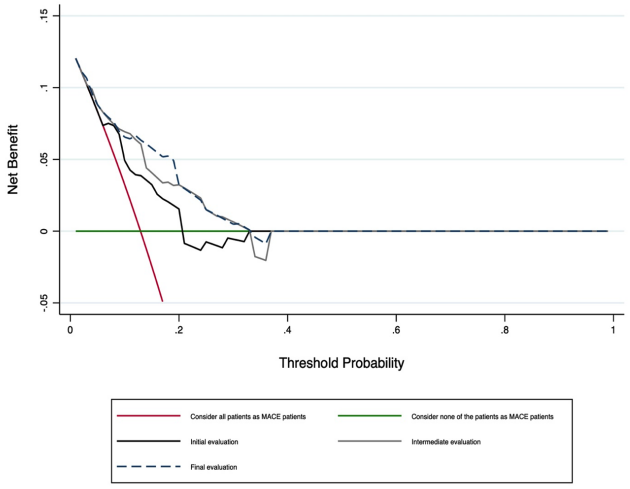

Figure 2. Decision curve analysis for determining the clinical benefit of severity predictions (0 to 100) by ChatGPT across study phases. In this analysis, the initial evaluation is represented by a solid black line, the intermediate evaluation by a gray line, and the final evaluation by a dark blue dashed line. The analysis focuses on patients presenting to the ED with chest pain and subsequently diagnosed with MACE. The x-axis displays the threshold probability for MACE, while the y-axis represents the net clinical benefit. ChatGPT's predictive ability is compared against two opposing strategies: the red line, which assumes that all patients presenting to the ED have a MACE, and the green line, which assumes that none of the patients are considered at risk for MACE.

Table 2. 2x2 contingency tables comparing ChatGPT predictions to MACE diagnoses across different evaluation phases.

	ChatGPT Initial Evaluation	
	No MACE	MACE
No MACE according to ChatGPT	110	10
MACE according to ChatGPT	45	13
Sensitivity: 56.5% (95%CI: 49.2-63.8)		
Specificity: 70.9% (95%CI: 64.3-77.6)		
PPV: 22.4% (95%CI: 16.3-28.5)		
NPV: 91.7% (87.6-95.7)		
	ChatGPT Intermediate Evaluation	
	No MACE	**MACE**
No MACE according to ChatGPT	87	4
MACE according to ChatGPT	68	19
Sensitivity: 82.6% (95%CI: 77.0-88.2)		
Specificity: 56.1% (95%CI: 48.8-63.4)		
PPV: 21.8% (95%CI: 15.8-27.9)		
NPV: 95.6% (92.6-98.6)		
	ChatGPT Final Evaluation	
	No MACE	**MACE**
No MACE according to ChatGPT	87	4
MACE according to ChatGPT	68	19
Sensitivity: 82.6% (95%CI: 77.0-88.2)		
Specificity: 56.1% (95%CI: 48.8-63.4)		
PPV: 21.8% (95%CI: 15.8-27.9)		
NPV: 95.6% (92.6-98.6)		

MACE, major adverse cardiac events; PPV, positive predictive value; NPV, negative predictive value.

This study introduces significant insights for clinical practice. Firstly, it demonstrates that a specially trained ChatGPT model can achieve good performance in identifying MACE, with predictive ability improving as more information is provided. Nonetheless, the level of performance attained is not yet comparable to that of other artificial intelligence models used in chest pain assessment.[11,12] For instance, a recent study by Zhang *et al.* utilizing a machine learning model based on a random forest algorithm reported a high system capability, achieving an ROC of 0.915 for predicting the risk of acute myocardial infarction within one month of ED admission for chest pain.[13] Another study conducted by Liu *et al.*, which evaluated chest pain in the ED using machine learning, demonstrated that by employing eight dimensionality reduction methods, a ROC of 0.901 was achieved for predicting MACE in patients with chest pain in the ED.[14] This highlights the strong predictive capability of machine learning models in risk assessment. The results of previous studies indicate that ML is capable of accurately and safely stratifying patients with chest pain who present to the ED.[15,16] However, as highlighted by a recent systematic review, these excellent results face the challenge that ML models are rarely implemented in a standardized manner within electronic health records. This may be due to the inherent complexity involved in the use and implementation of advanced tools like ML, which can pose significant practical and logistical hurdles for widespread clinical adoption.[15,16] In contrast, an LLM like ChatGPT, being readily accessible and easy to use, could represent a more convenient alternative to the more complex ML models.[17]

Secondly, the study demonstrated that LLMs like ChatGPT are not reliably able to accurately determine the severity of a patient's condition. In the initial evaluation, which included only ECG and clinical history, ChatGPT classified only 3 MACE patients as very urgent, while the remaining 20 were categorized as either urgent or non-urgent. At the second evaluation, with the addition of the first troponin test, ChatGPT's performance improved, classifying 12 MACE patients as very urgent; however, 11 MACE patients were still identified as urgent or non-urgent. Finally, in the third evaluation, after the second troponin test, ChatGPT identified 13 MACE patients as very urgent, leaving the remaining 10 classified as either urgent or non-urgent. This indicates a difficulty for ChatGPT in accurately determining the urgency of patients, a weakness that has been noted in other studies and contexts. For instance, a simulated study on ChatGPT's ability to triage patients showed a tendency to overestimate patient severity compared to human operators, suggesting that ChatGPT may not yet be a fully realistic alternative to human judgment.[18] Additionally, a recent study by Heston *et al.*, which examined simulated clinical cases of chest pain, found that when ChatGPT was provided with the same information used in TIMI and HEART scores, it did not perform comparably.[19] In fact, correlations between ChatGPT's assessments and these scoring systems were not fully aligned, highlighting its limitations relative to established clinical tools. This demonstrates ChatGPT's current inability to safely quantify patient risk and to stratify patients in accordance with clinical needs. Despite showing promise as a supplementary tool, ChatGPT lacks the precision required for reliable risk assessment and urgency classification, which are critical for effective decision-making in clinical settings.

Third, the integration of a decision-support tool like ChatGPT into the clinical decision-making process for managing complex chest pain patients could provide clinicians, both nurses and physicians, with an intelligent, analytical resource for comparing their own decisions. Such a tool may help to clarify clinical uncertainties often encountered when assessing complex and complicated patients. While this study does not establish the effectiveness of such a tool, it opens the door to future research on applying dynamic decision support in the evaluation of ED patients with chest pain. This could allow AI tools to be further enhanced with richer information to better support clinical decision-making. Currently, available studies have created AI tools focused on patient stratification at a single point in time, without accounting for the dynamic nature of patient evaluation or the information gathered at different stages, as done in this study.[20,21] The strategy applied here could inform future development of ED support tools that are not only effective but also adaptive, capable of incorporating data continuously throughout the patient evaluation process. This dynamic approach may ultimately provide more realistic, real-time support for ED clinicians, potentially transforming AI's role in emergency care.

Table 3. Urgency assigned by ChatGPT at each study phase, with comparison between patients who experienced a MACE and those who did not.

ChatGPT initial evaluation		
	No MACE	MACE
Non-urgent	110 (70.9)	10 (43.5)
Urgent	28 (18.1)	10 (43.5)
Very-urgent	17 (10.9)	3 (13.0)
ChatGPT intermediate evaluation		
	No MACE	MACE
Non-urgent	87 (56.1)	4 (17.4)
Urgent	43 (27.7)	7 (30.4)
Very-urgent	25 (16.1)	12 (52.2)
ChatGPT final evaluation		
	No MACE	MACE
Non-urgent	87 (56.1)	4 (17.4)
Urgent	39 (25.2)	6 (26.1)
Very-urgent	29 (18.7)	13 (56.5)

MACE, major adverse cardiac events.

Limitations

The study has important limitations. First, its single-center design subjects it to all the inherent limitations associated with this type of study, including reduced generalizability. Second, the study evaluated only the performance of ChatGPT without directly comparing it to that of ED physicians, leaving it unclear whether ChatGPT could outperform human judgment. Nevertheless, the study's objective was to assess ChatGPT's potential as a supportive tool, rather than to directly compare it with clinician performance. We acknowledge that this represents a major limitation. A direct comparison with physician decision-making would have offered a more comprehensive understanding of the tool's actual utility and clinical relevance. Future studies should be designed to explore this comparison, ideally within prospective randomized or observational frameworks, to evaluate not only predictive accuracy but also real-world usability, complementarity with human judgment, and impact on clinical outcomes. Third, the study excluded patients presenting with cardiovascular symptoms or abdominal pain, even though it may be a manifestation of acute coronary syndromes. This choice was made to ensure a more homogeneous sample of patients with pure chest pain, allowing for a clearer assessment of ChatGPT's performance in this specific context. However, this may limit the generalizability of our findings to patients with atypical or non-classical presentations of acute coronary syndromes.

Conclusions

This study represents the first use of ChatGPT in a clinical setting, providing it with information throughout the evaluation of patients presenting with chest pain. Although ChatGPT demonstrated a reasonable ability to identify patients with MACE, its performance was not sufficient to justify its integration into emergency department workflows at this time. The study suggests that potential advancements and improvements in ChatGPT may, in the future, bring it to a level where its integration into clinical practice could become viable.

References

1. Byrne RA, Rossello X, Coughlan JJ, et al.; ESC Scientific Document Group. 2023 ESC Guidelines for the management of acute coronary syndromes. Eur Heart J 2023;44:3720-826.
2. Virani SS, Newby LK, Arnold SV, et al.; Peer Review Committee Members. 2023 AHA/ACC/ACCP/ASPC/NLA/PCNA Guideline for the Management of Patients With Chronic Coronary Disease: A Report of the American Heart Association/American College of Cardiology Joint Committee on Clinical Practice Guidelines. Circulation 2023;148:e9-e119.
3. Writing Committee Members; Gulati M, Levy PD, Mukherjee D, et al. 2021 AHA/ACC/ASE/CHEST/SAEM/SCCT/SCMR Guideline for the Evaluation and Diagnosis of Chest Pain: A Report of the American College of Cardiology/American Heart Association Joint Committee on Clinical Practice Guidelines. J Am Coll Cardiol 2021;78:e187-e285.
4. Zaboli A, Ausserhofer D, Sibilio S, et al. Effect of the emergency department assessment of chest pain score on the triage performance in patients with chest pain. Am J Cardiol 2021;161:2-18.
5. 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.
6. Laureano-Phillips J, Robinson RD, Aryal S, et al.. HEART score risk stratification of low-risk chest pain patients in the emergency department: a systematic review and meta-analysis. Ann Emerg Med 2019;74:187-203.
7. Nilsson T, Johannesson E, Lundager Forberg J, et al. Diagnostic accuracy of the HEART Pathway and EDACS-ADP when combined with a 0-hour/1-hour hs-cTnT protocol for assessment of acute chest pain patients. Emerg Med J 2021;38:808-13.
8. Thangaraj PM, Khera R. Accelerating chest pain evaluation with machine learning. Eur Heart J Acute Cardiovasc Care 2023;12:753-4.
9. Stewart J, Lu J, Goudie A, et al. Applications of machine learning to undifferentiated chest pain in the emergency department: A systematic review. PLoS One 2021;16:e0252612.
10. Günay S, Öztürk A, Özerol H, et al. Comparison of emergency medicine specialist, cardiologist, and chat-GPT in electrocardiography assessment. Am J Emerg Med 2024;80:51-60.
11. Hinson JS, Taylor RA, Venkatesh A, et al. Accelerated chest pain treatment with artificial intelligence-informed, risk-driven triage. JAMA Intern Med 2024;184:1125-7.
12. Dawson LP, Smith K, Cullen L, et al. Care models for acute chest pain that improve outcomes and efficiency: JACC state-of-the-art review. J Am Coll Cardiol 2022;79:2333-48.
13. 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.
14. 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.
15. Ting Sim JZ, Fong QW, Huang W, Tan CH. Machine learning in medicine: what clinicians should know. Singapore Med J 2023;64:91-7.
16. Sidey-Gibbons JAM, Sidey-Gibbons CJ. Machine learning in medicine: a practical introduction. BMC Med Res Methodol 2019;19:64.
17. Tan S, Xin X, Wu D. ChatGPT in medicine: prospects and challenges: a review article. Int J Surg 2024;110:3701-6.
18. Zaboli A, Brigo F, Sibilio S, et al. Human intelligence versus Chat-GPT: who performs better in correctly classifying patients in triage? Am J Emerg Med 2024;79:44-7.
19. Heston TF, Lewis LM. ChatGPT provides inconsistent risk-stratification of patients with atraumatic chest pain. PLoS One 2024;19:e0301854.
20. Wu CC, Hsu WD, Islam MM, et al. An artificial intelligence approach to early predict non-ST-elevation myocardial infarction patients with chest pain. Comput Methods Programs Biomed. 2019;173:109-17.
21. Tsai DJ, Tsai SH, Chiang HH, et al. Development and validation of an artificial intelligence electrocardiogram recommendation system in the emergency department. J Pers Med 2022;12:700.