openheart

# Risk stratification of chest pain in the emergency department using artificial intelligence applied to electrocardiograms

Julian S. Haimovich [1,2] Márton Kolossváry,[3,4,5] Ridwan Alam,[1,6] Raimon Padrós-Valls,[1,7] Michael T. Lu,[3] Aaron D Aguirre [1,2,8,9]

► Additional supplemental material is published online only. To view, please visit the journal online (https://doi.org/10.1136/openhrt-2025-003343).

**To cite:** Haimovich JS, Kolossváry M, Alam R, *et al*. Risk stratification of chest pain in the emergency department using artificial intelligence applied to electrocardiograms. *Open Heart* 2025;**12**:e003343. doi:10.1136/openhrt-2025-003343

Received 26 March 2025
Accepted 4 August 2025



For numbered affiliations see end of article.

**Correspondence to**
Dr. Aaron D Aguirre; aaguirre1@mgh.harvard.edu

## ABSTRACT

**Background** Despite standardised approaches, subjective assessment and inconsistent diagnostic testing for chest pain in the emergency department (ED) drive costs, disparities and adverse outcomes. Artificial intelligence offers potential to automate and improve risk stratification.

**Methods and results** Using a retrospective cohort of 15 048 patients presenting to the ED of a tertiary care hospital, we trained a neural network classifier ('Chest Pain-AI' or 'CP-AI') to predict a 7-day composite endpoint of major cardiovascular diagnoses including myocardial infarction, pulmonary embolism, aortic dissection and all-cause mortality. Inputs to CP-AI included age, sex, cardiac biomarkers (D-dimer or troponin I or T positivity) and numerical representations of presenting 12-lead ECGs. ECG representations were derived using a publicly available deep learning model known as patient contrastive learning of representations. In an external validation set of 14 476 patients, we evaluated CP-AI against comparator models, including a 'Biomarker Model' incorporating clinical data (age, sex, biomarker positivity), based on both the area under the receiver operating characteristic curve (AUROC) and area under the precision-recall curve (AUPRC). CP-AI outperformed the Biomarker Model in prediction of the 7-day composite endpoint with an AUROC of 0.82 (95% CI 0.81 to 0.83) vs 0.79 (95% CI 0.78 to 0.81) and an AUPRC of 0.46 (95% CI 0.44 to 0.49) vs 0.35 (95% CI 0.33 to 0.37) ($p<0.05$ for both comparisons).

**Conclusions** CP-AI, a fully automated neural network classifier, demonstrated superior performance in the prediction of 7-day major cardiovascular diagnoses for patients presenting with acute chest pain compared with conventional models trained on demographics and cardiac biomarkers. CP-AI may standardise and expedite risk stratification of patients presenting to the ED with chest pain.

### WHAT IS ALREADY KNOWN ON THIS TOPIC

⇒ Despite the use of standardised protocols, chest pain evaluation in the emergency department remains highly variable due to subjective clinical judgement and inconsistent diagnostic testing, contributing to increased costs, healthcare disparities and adverse patient outcomes.

### WHAT THIS STUDY ADDS

⇒ This study introduces a fully automated neural network classifier (Chest Pain-AI) that integrates ECG waveforms, demographics and cardiac biomarkers to outperform conventional models in predicting a 7-day composite endpoint of major cardiovascular diagnoses comprising myocardial infarction, pulmonary embolism, aortic dissection and all-cause mortality.

### HOW THIS STUDY MIGHT AFFECT RESEARCH, PRACTICE OR POLICY

⇒ Chest Pain-AI has the potential to standardise and improve risk stratification of chest pain in the emergency department, reducing unnecessary diagnostic testing and improving care efficiency.

## INTRODUCTION

Chest pain is a leading cause of presentation to the emergency department (ED), accounting for 6.5 million visits in the USA annually.[1] Causes of chest pain can vary widely, from benign conditions to life-threatening emergencies such as myocardial infarction, pulmonary embolism and aortic dissection.[2] Rapid identification of high-risk patients is therefore crucial to prevent morbidity and mortality.[1 3 4] Despite the widespread use of standardised chest pain evaluation protocols, variability in diagnostic testing has resulted in preventable morbidity and contributed to significant healthcare spending.[5–8] There remains a critical need to improve risk stratification of patients presenting with acute chest pain.

Deep learning models leveraging the 12-lead ECG have demonstrated predictive capabilities in the diagnosis of acute cardiovascular conditions including pulmonary embolism, myocardial infarction and heart failure.[9–11] However, the use of ECG-based deep learning models for risk stratification of acute chest pain remains underexplored.

Using a retrospective cohort of 33 287 ED patients from two tertiary care centres, we

developed and externally validated a fully automated neural network model that uses 12-lead ECG waveforms to predict a 7-day composite endpoint of major cardiovascular diagnoses composed of myocardial infarction, pulmonary embolism, aortic dissection and all-cause mortality in patients presenting with chest pain. To assess potential improvement in risk stratification, we compared the deep learning model to conventional models trained on demographic information and cardiac biomarkers.

## METHODS

### Study sample

We leveraged a previously defined retrospective electronic health record cohort of patients presenting to the EDs of Mass General Brigham Hospitals, including Brigham and Women's Hospital (BWH) and Massachusetts General Hospital (MGH), with chest pain.[12] The dataset was assembled using the Research Patient Data Registry (RPDR) to identify patient encounters between 1 January 2006 and 31 December 2015 and link these encounters with corresponding demographics, laboratory testing, narrative notes, as well as radiology and cardiology diagnostic testing. RPDR mortality data are linked to the National Death Index, which introduces a multiyear delay in achieving dataset completeness.[13] Using RPDR, we identified adults presenting to the ED with chest pain using the following criteria: (1) age greater than 18 years, (2) biomarker testing including troponin I or T, D-dimer, or both, (3) and additional diagnostic testing during or immediately after the ED visit with cardiovascular studies (cardiac MRI, invasive coronary angiography, stress testing with or without perfusion imaging) or thoracic imaging evaluation (CT angiography, pulmonary perfusion imaging). We additionally excluded patients who did not have at least one 12-lead ECG performed while in the ED (n=17 680). When multiple ED visits were available, we selected the earliest encounter. We identified laboratory testing using the Logical Observation Identifiers Names and Codes (online supplemental table S1). We identified cardiovascular and pulmonary diagnostic testing using Current Procedural Terminology codes (online supplemental table S2).

Demographic data including age, sex and self-reported race, as well as laboratory results, were obtained from RPDR. Troponin I or T and D-dimer positivity were determined using the assay-specific age and sex thresholds for identifying abnormal values. We defined serum biomarker positivity as having at least one abnormal troponin I or T, or D-dimer test within 1 day of ED presentation.

### Study outcomes

Outcomes were defined using *International Classification of Diseases* codes 9 and 10 (online supplemental table S3). Mortality data were obtained through linkage to the National Death Index . Our primary outcome was a composite endpoint of major cardiovascular diagnoses within 7 days of ED presentation, including myocardial infarction, pulmonary embolism, aortic dissection and all-cause mortality. Secondary outcomes included each of the individual endpoints at 7 days. In a supplementary analysis, we also assessed both primary and secondary endpoints at 30 days.

### ECG sampling

Study sample ECGs were obtained during routine clinical care. ECG waveforms were accessed using the institutional MUSE database (GE Healthcare, 'MUSE' software V.8.0 and V.9.0). The MUSE database includes automated and physician-entered diagnostic statements; tabular ECG measures including axes, intervals and lead-specific voltage amplitudes; and raw ECG waveforms. When multiple ECGs were available, the first ECG available for the clinical encounter was selected. No exclusion criteria were applied to ECGs. The ECGs were converted to .xml file format and analysed at native resolution, as previously described.[14]

### Model training and testing

We randomly divided patients presenting to the BWH ED into training and internal test sets with an 80/20 split. Patients presenting to the MGH ED were reserved as an external validation set. Since the earliest ED presentation for each patient was selected for analysis, there was no overlap between the datasets.

For the neural network classifier, we first transformed each ECG into a 320-dimensional numerical vector representation using a previously described, publicly available deep learning approach known as patient contrastive learning of representations (PCLR).[14 15] PCLR is a convolutional neural network trained on over 3 million inpatient and outpatient ECGs from greater than 400 000 patients that creates numerical representations of ECG waveforms which emphasise differences between ECGs from different individuals and similarities among ECGs from the same individual (online supplemental methods). This approach enables self-supervised learning across the entirety of the PCLR ECG training database (3 million ECGs) to capture rich embeddings of ECGs without specific labels. These embeddings can subsequently be fine-tuned for specific prediction tasks using supervised learning models on smaller datasets. Using this approach, PCLR has been shown to outperform task-specific models when training data are limited.[14 16]

We input the 320-dimensional numerical PCLR ECG representations along with age, sex and biomarker results into a neural network classifier ('Chest Pain-AI' or 'CP-AI') to predict the primary and secondary outcomes using a multilabel approach (figure 1). All analyses were performed in Python using the scikit-learn package.[17] Details on the rationale for feature selection, model training and optimisation are available in the online supplemental methods.

We also trained three comparator neural network models including: (1) age and sex ('Age and Sex'), (2) age, sex and cardiac biomarker positivity ('Biomarker

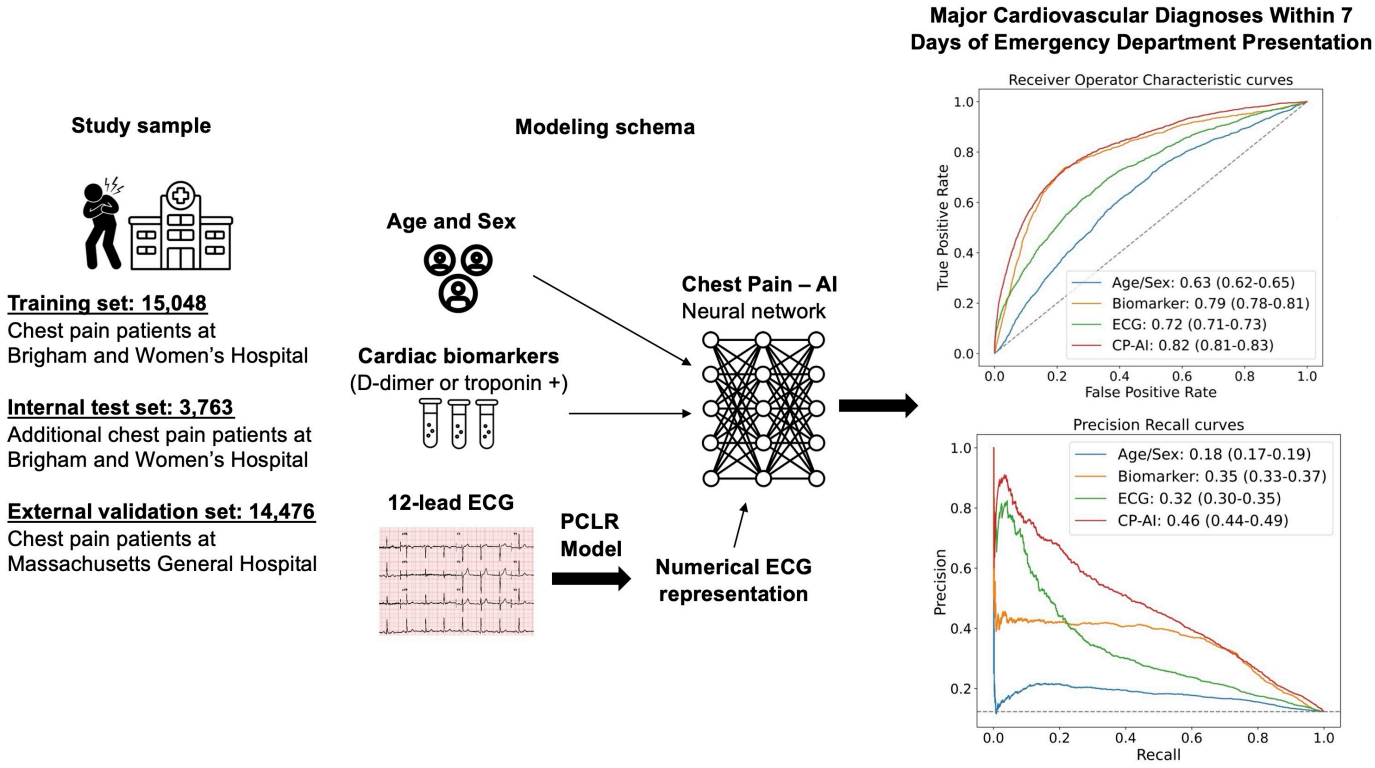


**Figure 1** Graphical abstract. The training set comprised 15 048 patients with chest pain presenting to BWH. We trained a neural network classifier ('Chest Pain-AI') to predict a 7-day composite endpoint of major cardiovascular diagnoses including myocardial infarction, pulmonary embolism, aortic dissection and all-cause mortality. Model inputs included demographics (age and sex), cardiac biomarkers (D-dimer or troponin I or T positivity) and 12-lead ECG waveforms processed by an algorithm using PCLR into 320-dimensional numerical representations. Model performance was assessed in both an internal test set of an additional 3736 patients presenting to BWH and an external validation dataset of 14 476 patients presenting to the Massachusetts General Hospital using the area under the receiver operating characteristic curve and area under the precision-recall curve. CP-AI demonstrated superior performance to comparison models trained on (1) demographics alone ('Age/Sex'), (2) demographics and cardiac biomarkers ('Biomarker') and (3) demographics and numerical ECG representations ('ECG'). BWH, Brigham and Women's Hospital; CP-AI, Chest Pain-AI; PCLR, patient contrastive learning of representations.

Model') and (3) age, sex and numerical ECG representations ('ECG Model'). Model training and optimisation of the comparator models followed the same approach as that of the CP-AI model.

After model training, we assessed model performance in both the internal test set and the external validation set using the area under the receiver operating characteristic curve (AUROC) and area under the precision-recall curve (AUPRC). Precision-recall curves and the AUPRC evaluate model performance specifically for positive predictions by plotting precision (positive predictive value) against recall (sensitivity or true positive rate). These metrics are particularly useful for imbalanced datasets with low event rates, as they focus on the model's ability to identify positive cases without being influenced by the large number of true negatives.[18]

We used bootstrapping to calculate CIs for AUROC and AUPRC and to compare these metrics between pairs of models. For the bootstrapping procedure, we used sampling with replacement over 1000 iterations to generate an empirical distribution of the AUROC (or AUPRC) and used that distribution to calculate the mean and 95% CI of the AUROC (or AUPRC). We applied a similar approach to statistically compare pairs of models: for each bootstrapped sample, we calculated the AUROC (or AUPRC) for both models and computed the difference in AUROCs (or AUPRCs). We then estimated the p value for the pairwise comparison by calculating the proportion of bootstrapped differences that were less extreme than zero (two-tailed test).[19] We also conducted a sensitivity analysis where we evaluated the AUROC of the models among sex, age and race-specific strata in the external validation set.

Finally, to evaluate how model performance compared among individuals at low risk of the composite endpoint, we compared predictions between CP-AI and the Biomarker Model in non-cases (true negatives) in the external validation set. To facilitate direct model comparison, we defined a probability threshold for high and low risks by identifying the model-specific threshold corresponding to 98% sensitivity for predicting 7-day major cardiovascular diagnoses. A 98% sensitivity level was selected based on precedent in the emergency medicine literature.[20] We then evaluated individuals for whom model predictions agreed and disagreed using a 'low risk' designation as the ground truth. Finally, we measured

corresponding test characteristics including specificity, positive predictive value and negative predictive value at 98% model sensitivity across the external validation set (including both cases and non-cases).

## RESULTS

### Study sample

The BWH dataset comprised 18811 patients who were divided 80/20 into training and internal testing sets. The training set included 15048 individuals (51% female, mean age 57±17 years) and the internal test set included 3763 individuals (50% female, mean age 57±17 years) (table 1). The MGH external validation set included 14476 individuals (45% female, mean age 58±16 years). Rates of troponin testing (training set: 94%, internal test set: 94%, validation set: 95%) and D-dimer testing (training set: 33%, internal test set: 33%, validation set: 35%) were similar. The primary composite endpoint of 7-day major cardiovascular diagnoses occurred in 2443 (16%) patients in the training set, 638 (17%) patients in the internal test set and 1784 (12%) patients in the external validation set. Rates of the 30-day composite endpoint were also similar among the modelling sets (table 1).

**Table 1** Study baseline characteristics

| | Training set (n=15 048) | Internal test set (n=3763) | External validation set (n=14 476) |
|---|---|---|---|
| Age (years)* | 57±17 | 57±17 | 58±16 |
| Female | 7727 (51) | 1896 (50) | 6498 (45) |
| Ethnicity | | | |
| White | 9164 (61) | 2313 (62) | 10743 (74) |
| Black | 2755 (18) | 678 (18) | 1213 (8) |
| Hispanic | 1289 (9) | 303 (8) | 737 (5) |
| Asian | 388 (3) | 106 (3) | 529 (4) |
| Biomarkers | | | |
| Troponin or D-dimer | 15048 (100) | 3763 (100) | 14476 (100) |
| Troponin or D-dimer positive | 5335 (36) | 1341 (36) | 4642 (32) |
| Troponin | 14161 (94) | 3534 (94) | 13779 (95) |
| Troponin positive | 2888 (19) | 701 (19) | 1813 (13) |
| D-dimer | 4954 (33) | 1230 (33) | 5009 (35) |
| D-dimer positive | 2800 (19) | 719 (19) | 3128 (22) |
| 7-day outcomes | | | |
| Major cardiovascular diagnoses | 2443 (16) | 638 (17) | 1784 (12) |
| Myocardial infarction | 1410 (9) | 367 (10) | 1037 (7) |
| Pulmonary embolism | 983 (7) | 258 (7) | 653 (5) |
| Aortic dissection | 113 (1) | 30 (1) | 146 (1) |
| Mortality | 307 (2) | 90 (2) | 294 (2) |
| 30-day outcomes | | | |
| Major cardiovascular diagnoses | 2540 (17) | 653 (17) | 1853 (13) |
| Myocardial infarction | 1448 (10) | 374 (10) | 1071 (7) |
| Pulmonary embolism | 1051 (7) | 268 (7) | 692 (5) |
| Aortic dissection | 115 (1) | 30 (1) | 149 (1) |
| Mortality | 738 (5) | 195 (5) | 589 (4) |

Data presented as n (%) unless otherwise noted.
*Data displayed as mean±SD.

### Model performance

CP-AI outperformed the comparison models for prediction of the 7-day composite endpoint with an AUROC of 0.80 (95% CI 0.78 to 0.82) in the internal test set and AUROC of 0.82 (95% CI 0.81 to 0.83) in the external validation set (figure 2). The AUROCs of the comparator models for the prediction of the 7-day composite endpoint in the external validation set were: age and sex 0.63 (95% CI 0.60 to 0.65), Biomarker Model 0.78 (95% CI 0.76 to 0.8) and ECG Model 0.70 (95% CI 0.68 to 0.72) ($p<0.05$ for all comparisons with CP-AI) (figure 2). CP-AI also yielded the highest AUPRC of 0.50 (95% CI 0.46 to 0.54) and 0.46 (95% CI 0.44 to 0.49) for prediction of 7-day MACE versus the Biomarker Model 0.35 (95% CI 0.33 to 0.37) and 0.42 (95% CI 0.39 to 0.46) in the internal test set and external validation set, respectively ($p<0.05$ for both comparisons) (figure 2). CP-AI similarly outperformed the comparison models for prediction of the 30-day composite endpoint in the internal test set and external validation set (online supplemental figure S1).

In addition, CP-AI outperformed the comparison models for prediction of the secondary 7-day outcomes in the external validation set, including: myocardial infarction (AUROC 0.91 (95% CI 0.90 to 0.92)), pulmonary embolism (AUROC 0.69 (95% CI 0.67 to 0.71)), aortic dissection (AUROC 0.71 (95% CI 0.67 to 0.75)) and mortality (AUROC 0.82 (95% CI 0.80 to 0.84)) (online supplemental figure S2). Similar findings were observed for predicting the secondary 30-day outcomes (online supplemental figure S3). The precision-recall curves for prediction of the secondary 7 and 30-day outcomes are presented in online supplemental figures S4 and S5. Performance was maintained across strata of age, sex and race (online supplemental tables S4 and S5).

We then compared test characteristics and prediction of 7-day major cardiovascular diagnoses by the

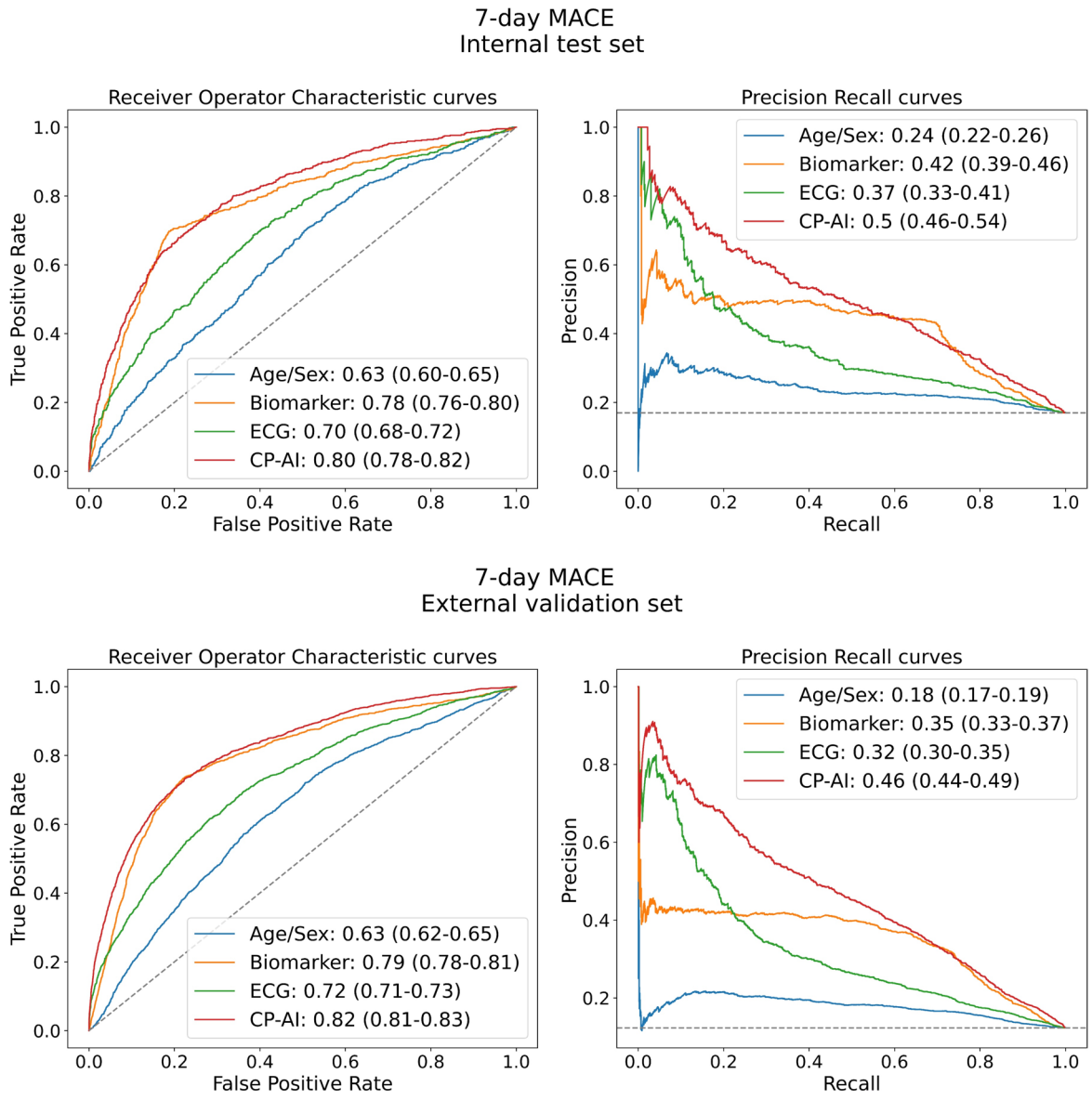


**Figure 2** Receiver operating characteristic and precision-recall curves for prediction of 7-day major cardiovascular diagnoses. Top: In the internal test set (n=3746), CP-AI outperformed comparison models trained on (1) demographics alone ('Age/Sex'), (2) demographics and cardiac biomarkers ('Biomarker') and (3) demographics and PCLR-derived ECG representations ('ECG') for prediction of the 7-day composite endpoint, as shown in both the ROC curves (left) and PR curves (right). The ROC curve legend shows the area under the curve and corresponding 95% CI of CP-AI and the comparator models. The dashed line represents the expected curve of a 'no skill' model equal to an area of 0.5. The PR curve legend shows the area under the PR curve and corresponding 95% CI of CP-AI and the comparator models. The dashed line represents the average precision of a 'no skill' model equal to the event rate in the validation sample. Bottom: ROC curves (left) and PR curves (right) for CP-AI and the comparator models for prediction of the 7-day composite endpoint in the external validation set. CP-AI, Chest Pain-AI; PCLR, patient contrastive learning of representations; PR, precision-recall; ROC, receiver operating characteristic.

CP-AI and the Biomarker Model among non-cases in the external validation set at a predetermined 98% sensitivity level (table 2). Of 12 692 non-cases in the external validation set, both models agreed in the correct categorisation of 254 (2%) patients as low risk and the incorrect categorisation of 10 233 (81%) patients as high risk. Of the remaining 2205 (17%) of patients for which the models disagreed, 1702 (78%) were correctly categorised as low risk by CP-AI and incorrectly categorised as high risk by the Biomarker Model. Conversely, 503 (22%) patients were incorrectly categorised as high risk by CP-AI and correctly categorised as low risk by the Biomarker Model. Taken together, a net of 1199 (9%) of non-cases in the external validation set were correctly reclassified from high risk to low risk by CP-AI compared with the Biomarker Model.

Finally, we compared the test characteristics of CP-AI and the Biomarker Model at 98% sensitivity. The corresponding specificities for CP-AI and the Biomarker Model were 15.4% and 6.0%, positive

**Table 2** Model comparison for risk stratification of 7-day major cardiovascular diagnoses

| **Non-cases in external validation set (true negatives), n=12 692** | | |
|---|---|---|
| Biomarker Model | Chest Pain-AI | |
| | Low risk | High risk |
| Low risk | 254 | **503** |
| High risk | **1702** | 10 233 |
| **Cases in external validation set (true positives), n=1784** | | |
| Biomarker Model | Chest Pain-AI | |
| | Low risk | High risk |
| Low risk | 4 | **32** |
| High risk | **32** | 1716 |

Comparison of Chest Pain-AI and Biomarker Model risk predictions for non-cases (true negatives; top) and cases (true positives; bottom) in the external validation set. Model-specific risk thresholds delineating high and low risks were determined by the model-specific threshold corresponding to a sensitivity of 98%. In total, 1199 of non-cases (9%) were correctly reclassified to low risk by Chest Pain-AI. Since model sensitivities were predefined, no difference in case (true positive) reclassification is expected.

predictive values were 15.4% and 13% and negative predictive values were 98% and 96%, respectively.

## DISCUSSION

We present the findings of CP-AI, a fully automated neural network classifier, which integrates demographic and cardiac biomarker data with ECG waveforms to risk stratify patients with acute chest pain. In an external validation set, CP-AI achieved an AUROC of 0.82 and an AUPRC of 0.50 for predicting the primary outcome of 7-day major cardiovascular diagnoses composite MACE outcome encompassing myocardial infarction, pulmonary embolism, aortic dissection and mortality. When compared with conventional models that do not incorporate ECG waveform data, CP-AI showed superior overall performance. CP-AI further demonstrated significant improvements in risk stratification among low-risk patients, correctly reclassifying 9% of low-risk patients when compared with a conventional model trained on demographics and cardiac biomarkers. CP-AI may have utility as a rapid, fully automated risk stratification tool for patients presenting with acute chest pain.

Previous studies have highlighted the effectiveness of deep learning models leveraging ECG data to predict specific causes of chest pain (eg, myocardial infarction and pulmonary embolism) as well as mortality. Gustafsson *et al* used the SWEDEHEART registry to develop a deep learning model that integrated ECG waveforms with age and sex to distinguish between ST-elevation myocardial infarction (STEMI) and non-ST-elevation myocardial infarction (NSTEMI), achieving AUROCs of 0.99 and 0.83 for STEMI and NSTEMI classifications, respectively.[9] Somani *et al* trained a deep learning model combining clinical data (demographics, comorbidities, vital signs and labs) with ECG waveforms to predict pulmonary embolism, achieving an AUROC of 0.81. Notably, though, when only ECG data were used, the AUROC dropped to 0.59, underscoring the importance of clinical context to their model.[11] In another study, Sun *et al* developed a series of deep learning models based on demographic and ECG data to predict mortality, reporting AUROCs of 0.84, 0.81 and 0.90 for 30-day, 1-year and 5-year mortality predictions, respectively.[21] While there are notable variations in study populations, input features and methodologies, CP-AI exhibited comparable performance to these state-of-the-art models, highlighting the potential for a single model to simultaneously predict multiple outcomes without loss in performance.

Despite extensive efforts to standardise the modern evaluation of acute chest pain, variability in provider experience and inherent biases have led to inconsistent application and outcomes. A study of patients undergoing rapid chest pain assessment in an ED setting revealed significant discrepancies in bedside risk scores between ED physicians and cardiologists in more than half of patients.[22] The study found that ED physicians tended to overestimate risk compared with cardiologists in at least 25% of the study population, suggesting that ED providers may bias towards more risk-averse diagnostic algorithms. Implicit biases have further contributed to pronounced differences in race and sex-specific outcomes.[8 23] For instance, an analysis of a national survey of ambulatory patients in the USA showed that women presenting with chest pain waited longer, had lower rates of ECG testing and were less likely to be admitted to the hospital.[24] Similarly, in a large meta-analysis comprising 705 000 patients, Shah *et al* found significant differences in time to evaluation, door-to-balloon time and mortality in females presenting with STEMI.[23] Artificial intelligence, though prone to its own biases, has the potential to automate and standardise both data interpretation and the application of diagnostic testing algorithms, and may therefore help to offset provider biases and improve chest pain risk stratification among diverse patient populations.[25 26]

Improving patient selection and utilisation of diagnostic testing for patients across the risk spectrum is essential to both preventing morbidity and reducing healthcare spending. In studies of low-risk patient cohorts, the yield of obstructive coronary artery disease was a mere 1–3%, highlighting that overuse of diagnostic testing in low-risk patients is a potential source of excessive spending.[27 28] When comparing risk stratification among low-risk individuals, we found that CP-AI correctly reclassified 9% of low-risk patients compared with a model trained on conventional data. This improvement has the potential to reduce unnecessary diagnostic testing in nearly 1 in 10 patients presenting with chest pain. Surprisingly, prior studies have suggested that diagnostic testing is overused in low-risk patients and underused in high-risk patients resulting in potentially avoidable morbidity.[5 29]

While only a relative minority of patients presenting with chest pain will have adverse outcomes, it is essential to correctly identify these diagnoses. In situations like this, where there is a class imbalance between cases and non-cases, the precision-recall curve provides critical metrics of evaluation beyond the receiver operating characteristic curve. The CP-AI model notably outperforms all other models in the AUPRC as well as AUROC. Taken together, these results suggest that artificial intelligence models like CP-AI have the potential to enhance risk stratification and assist healthcare providers in guiding the appropriate utilisation of diagnostic testing across the risk spectrum.

Our findings must be interpreted in the context of the study's limitations. The use of billing codes to ascertain clinical outcomes may result in misclassification. Notably, though, consistent improvements in CP-AI prediction were also observed in the secondary mortality outcome, which is not prone to misclassification bias. The study sample was also limited to patients with available ECG and biomarker data, which may under-represent low-risk patients. Although stratified analysis was performed, the study population was largely composed of older white individuals, which may limit generalisability. However, in stratified analyses based on age, sex and ethnicity, CP-AI showed consistent performance (online supplemental table S4). Additionally, our dataset was limited to initial biomarker values, whereas chest pain algorithms commonly use serial measures. One might expect the performance of models, like CP-AI, to improve if given access to serial biomarker data. Lastly, while the model was developed and validated within two distinct hospitals, further work is required to show that our results are generalisable across many healthcare settings in diverse geographical locations, and to prospectively validate our findings.

In conclusion, we demonstrate that a fully automated neural network classifier, CP-AI, which integrates patient demographics, cardiac biomarkers and ECG waveform data, effectively stratifies risk among ED patients presenting with acute chest pain. CP-AI outperformed comparator models trained on conventional risk stratification data and accurately reclassified patients across the risk spectrum. CP-AI offers a promising tool for enhancing clinical decision-making and utilisation of healthcare resources through more accurate and individualised risk assessment.

**Author affiliations**
[1]Cardiology Division, Massachusetts General Hospital, Harvard Medical School, Boston, Massachusetts, USA
[2]Heart and Vascular Institute, Cardiovascular Research Center, Mass General Brigham, Boston, Massachusetts, USA
[3]Cardiovascular Imaging Research Center, Massachusetts General Hospital, Harvard Medical School, Boston, Massachusetts, USA
[4]Gottsegen National Cardiovascular Center, Budapest, Hungary
[5]Physiological Controls Research Center, University Research and Innovation Center, Óbuda University, Budapest, Hungary
[6]Research Laboratory of Electronics, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA
[7]Bioinformatics and Systems Biology, University of California San Diego, La Jolla, California, USA
[8]Center for Systems Biology, Massachusetts General Hospital, Harvard Medical School, Boston, Massachusetts, USA
[9]Wellman Center for Photomedicine, Massachusetts General Hospital, Harvard Medical School, Boston, Massachusetts, USA

**Social media** Julian S. Haimovich, X @jhaimovichmd; Aaron D Aguirre, X @aaguirremd

**Contributors** Study design: JSH, MK, RA, RP-VMS, MTL, ADA. Methodology: JSH, MK, RA, RP-VMS, MTL, ADA. Data curation: JSH, MK. Analysis: JSH, MK, RA, RP-VMS. Manuscript drafting: JSH, MK, RA, RP-VMS, MTL, ADA. ADA (corresponding author) is the guarantor.

**Funding** This work was supported by the MGH Primary Care Innovation Fund and by the Wellman Center for Photomedicine. ADA was also supported by the National Institutes of Health (R01HL173544). JSH is supported by the T32 Training Grant (T32HL007208). MK was supported by the János Bolyai Research Scholarship of the Hungarian Academy of Sciences and by the 2024-2.1.1 University Research Scholarship Program of the Ministry of Culture and Innovation from the National Research, Development and Innovation Fund of Hungary. Project number PD 147269 and Excellence 151118 has been implemented with the support provided by the Ministry of Culture and Innovation of Hungary from the National Research, Development and Innovation Fund, financed under the PD_23 and Excellence_24 funding scheme.

**Disclaimer** The author is a staff member of the World Health Organization. The author alone is responsible for the views expressed in this publication and they do not necessarily represent the views, decisions or policies of the World Health Organization.

**Competing interests** MTL reported receiving grants to his institution from the American Heart Association, AstraZeneca, Ionis, Johnson & Johnson Innovation, Kowa Pharmaceuticals America, MedImmune, the National Academy of Medicine, the National Heart, Lung and Blood Institute, and the Risk Management Foundation of the Harvard Medical Institutions, outside the submitted work. ADA has received research funding from Amgen and Philips Healthcare for work unrelated to this study.

**Patient and public involvement statement** Patients and/or the public were involved in the design, or conduct, or reporting, or dissemination plans of this research. Refer to the Methods section for further details.

**Patient consent for publication** Not applicable.

**Ethics approval** This retrospective study was performed under approval of the Institutional Review Board (IRB) at Mass General Brigham with a waiver of informed consent from the IRB.

**Provenance and peer review** Part of a Topic Collection; Not commissioned; externally peer-reviewed.

**Data availability statement** No data are available.



**ORCID iDs**
Julian S. Haimovich http://orcid.org/0000-0003-4346-6241
Aaron D Aguirre http://orcid.org/0000-0002-5509-1646

## REFERENCES

1 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. *Circulation* 2021;144:e368–454.
2 Amsterdam EA, Kirk JD, Bluemke DA, *et al*. Testing of low-risk patients presenting to the emergency department with chest pain: a scientific statement from the American Heart Association. *Circulation* 2010;122:1756–76.
3 Menees DS, Peterson ED, Wang Y, *et al*. Door-to-balloon time and mortality among patients undergoing primary PCI. *N Engl J Med* 2013;369:901–9.
4 Isselbacher EM, Preventza O, Hamilton Black III J, *et al*. 2022 ACC/AHA guideline for the diagnosis and management of aortic disease. *J Am Coll Cardiol* 2022;80:e223–393.
5 Mullainathan S, Obermeyer Z. Diagnosing physician error: a machine learning approach to low-value health care. *Q J Econ* 2022;137:679–727.
6 Goodacre S, Cross E, Arnold J, *et al*. The health care burden of acute chest pain. *Heart* 2005;91:229–30.
7 Mnatzaganian G, Hiller JE, Braitberg G, *et al*. Sex disparities in the assessment and outcomes of chest pain presentations in emergency departments. *Heart* 2020;106:111–8.
8 Popp LM, Ashburn NP, Snavely AC, *et al*. Race differences in cardiac testing rates for patients with chest pain in a multisite cohort. *Acad Emerg Med* 2023;30:1020–8.
9 Gustafsson S, Gedon D, Lampa E, *et al*. Development and validation of deep learning ECG-based prediction of myocardial infarction in emergency department patients. *Sci Rep* 2022;12.
10 Yao X, Rushlow DR, Inselman JW, *et al*. Artificial intelligence-enabled electrocardiograms for identification of patients with low ejection fraction: a pragmatic, randomized clinical trial. *Nat Med* 2021;27:815–9.
11 Somani SS, Honarvar H, Narula S, *et al*. Development of a machine learning model using electrocardiogram signals to improve acute pulmonary embolism screening. *Eur Heart J Digit Health* 2022;3:56–66.
12 Kolossváry M, Raghu VK, Nagurney JT, *et al*. Deep learning analysis of chest radiographs to triage patients with acute chest pain syndrome. *Radiology* 2023;306.
13 Identify subjects / request data. Available: https://rc.partners.org/research-apps-services/identify-subjects-request-data [Accessed 30 Dec 2024].
14 Diamant N, Reinertsen E, Song S, *et al*. Patient contrastive learning: a performant, expressive, and practical approach to electrocardiogram modeling. *PLoS Comput Biol* 2022;18.
15 Model_zoo/pclr at master · broadinstitute/ml4h. Github. n.d. Available: https://github.com/broadinstitute/ml4h/tree/master/model_zoo/PCLR
16 Haimovich JS, Diamant N, Khurshid S, *et al*. Artificial intelligence-enabled classification of hypertrophic heart diseases using electrocardiograms. *Cardiovasc Digit Health J* 2023;4:48–59.
17 Pedregosa F, Varoquaux G, Gramfort A, *et al*. Scikit-learn: machine learning in python, Available: http://arxiv.org/abs/1201.0490
18 Davis J, Goadrich M. The relationship between Precision-Recall and ROC curves. Proceedings of the 23rd international conference on Machine learning - ICML '06. New York, New York, USA: ACM Press, 2006
19 Efron B, Tibshirani RJ. *An Introduction to the Bootstrap*. Chapman and Hall/CRC, 1994.
20 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.
21 Sun W, Kalmady SV, Sepehrvand N, *et al*. Towards artificial intelligence-based learning health system for population-level mortality prediction using electrocardiograms. *NPJ Digit Med* 2023;6.
22 Wu WK, Yiadom M, Collins SP, *et al*. Documentation of HEART score discordance between emergency physician and cardiologist evaluations of ED patients with chest pain. *Am J Emerg Med* 2017;35:132–5.
23 Shah T, Haimi I, Yang Y, *et al*. Meta-analysis of gender disparities in in-hospital care and outcomes in patients with ST-segment elevation myocardial infarction. *Am J Cardiol* 2021;147:23–32.
24 Banco D, Chang J, Talmor N, *et al*. Sex and race differences in the evaluation and treatment of young adults presenting to the emergency department with chest pain. *J Am Heart Assoc* 2022;11.
25 van Assen M, Beecy A, Gershon G, *et al*. Implications of bias in artificial intelligence: considerations for cardiovascular imaging. *Curr Atheroscler Rep* 2024;26:91–102.
26 Armoundas AA, Narayan SM, Arnett DK, *et al*. Use of artificial intelligence in improving outcomes in heart disease: a scientific statement from the american heart association. *Circulation* 2024;149:e1028–50.
27 Hermann LK, Newman DH, Pleasant WA, *et al*. Yield of routine provocative cardiac testing among patients in an emergency department-based chest pain unit. *JAMA Intern Med* 2013;173:1128–33.
28 Winchester DE, Brandt J, Schmidt C, *et al*. Diagnostic yield of routine noninvasive cardiovascular testing in low-risk acute chest pain patients. *Am J Cardiol* 2015;116:204–7.
29 Abaluck J, Agha L, Kabrhel C, *et al*. The determinants of productivity in medical testing: intensity and allocation of care. *Am Econ Rev* 2016;106:3730–64.