Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Multimodal machine learning bridges the cardiology–ICU gap predicts in- hospital mortality and outperforms severity scores in critically ill CAD patients

Bing Cui
Peking University Third Hospital, Peking University, Peking University
Xinfang Zhang
Peking University Third Hospital, Peking University, Peking University
Zhaokui Kou
Shandong Second Medical University
Tianlan Wang
Hong Kong Generative AI Research & Development Center
Ge Chen
Hong Kong Generative AI Research & Development Center
Yulong Liu
Hong Kong Generative AI Research & Development Center
Hua Xu
The Hong Kong University of Science and Technology
Yaohua Zhang
yaohuaz@hotmail.com

Peking University Third Hospital, Peking University, Peking University

Research Article

Keywords: coronary artery disease, acute coronary syndrome, intensive care, mortality prediction, machine learning, external validation

Posted Date: November 23rd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11331717/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Title page

Multimodal machine learning bridges the cardiology–ICU gap predicts in-hospital mortality and outperforms severity scores in critically ill CAD patients

Bing Cui[1 †], Xinfang Zhang[1†], Zhaokui Kou[2], Tianlan Wang[3], Ge Chen[3], Yulong Liu[3], Hua Xu[4], Yaohua Zhang[1*]

† These authors contributed equally to this work.

*Corresponding author: Yaohua Zhang; e-mail：yaohuaz@hotmail.com;

Affiliations

1. Department of Cardiology and Institute of Vascular Medicine, Peking University Third Hospital; State Key Laboratory of Vascular Homeostasis and Remodeling, Peking University; NHC Key Laboratory of Cardiovascular Molecular Biology and Regulatory Peptides, Peking University; Beijing Key Laboratory of Cardiovascular Receptors Research. Beijing 100191, China.

2. Shandong Second Medical University. Shang dong 261053, China.

3. Hong Kong Generative AI Research & Development Center, Hong Kong, China.

4. Division of Emerging Interdisciplinary Areas, The Hong Kong University of Science and Technology, Hong Kong, China.

Abstract

Background Critically ill patients with coronary artery disease (CAD) occupy an intermediate position between two risk assessment frameworks: cartographical specific scores established outside the intensive care unit (ICU), and the general ICU severity scores that do not account for cardiac-specific characteristics. To address this gap, we developed and externally validated machine-learning models predicting in-hospital mortality from the first 24 h of ICU data.

Methods Adults diagnosed with coronary artery disease (CAD) or acute coronary syndrome, who experienced an ICU stay of 24 hours or more, were identified within the eICU Collaborative Research Database, encompassing data from 208 hospitals across the United States during the years 2014 to 2015. Predictors were hourly vital-sign and laboratory trajectories (24 h × 41 variables), static and comorbidity data, first-day treatments and diagnosis text. Logistic regression, LightGBM, XGBoost and CHD-MMNet were developed in 146 hospitals with hospital-grouped cross-validation and validated in 42 unseen hospitals, in 2015 discharges after training on 2014, and externally in MIMIC-IV v3.1 (2008-2022), where a frozen portable model was applied without updating. Comparators were APACHE IV, day-1 SOFA and OASIS. Analyses followed the pre-specified plan and TRIPOD+AI.

Results Of 200,859 ICU stays, 17586 were included (188 hospitals, mean age 67.6 years, 65.0% male) and 1315 (7.5%) died. We first tested whether the model improved on existing severity scores in hold-out hospitals (n = 4653, 308 deaths). The AUROC of LightGBM was 0.897 (95% CI 0.879-0.914) versus 0.864 (0.842-0.887) for APACHE IV (ΔAUROC 0.032, $P = 1.2\times10^{-4}$), with better precision–recall, calibration and net benefit, demonstrating a consistent advantage over the clinical comparator; XGBoost was equivalent whereas CHD-MMNet inferior. To assess

robustness over time, we performed temporal validation, which yielded an AUROC of 0.886 versus 0.854 for APACHE IV, indicating that the advantage persisted in a subsequent calendar year. We then applied the frozen model externally to an independent health system (24,129 MIMIC-IV CAD stays, 2785 deaths). It retained an AUROC of 0.871 (0.864-0.877), showed no decay across admission-year groups, and exceeded day-1 SOFA (0.658) and OASIS (0.688). Ablation showed that only removal of the hourly time series reduced discrimination (ΔAUROC −0.036). The highest risk tier contained 6.5% of patients and 47.4% of deaths. Together, these findings indicate that the model improved on existing severity scores, remained stable over time, transported across health systems, relied mainly on hourly physiological data, and enabled clinically meaningful risk stratification.

Conclusions Machine learning using first-day ICU data predicts in-hospital death in critically ill CAD patients more accurately than traditional scores over 14 years. Gradient boosting on physiological trajectories performed as well as or better than deep sequence networks, favoring simpler models for implementation.

Keywords: coronary artery disease; acute coronary syndrome; intensive care; mortality prediction; machine learning; external validation

Introduction

Ischaemic heart disease continues to be the leading cause of death worldwide, causing nearly 9 million deaths annually, and its total impact is increasing with the ageing of the population(1, 2) . A considerable number of patients with coronary artery disease (CAD) requires intensive care after a myocardial infarction complicated by shock or arrest, after coronary surgery, or when CAD coexists with sepsis, respiratory failure or other critical illness. In contemporary cardiac ICU registries overall mortality is 8-9%, rising above 30% with cardiogenic shock or cardiac arrest (3, 4). Early personalized risk assessment in this population directs monitoring intensity, timing of invasive strategies, escalation to mechanical support, and care goals discussions, with current guidelines recommending formal risk stratification (5, 6).

The available instruments are inadequately aligned with the needs of the critically ill CAD patients. Cardiology scores such as GRACE and TIMI originated from unselected acute coronary syndrome (ACS) populations from a handful of admission variables, achieving C-statistics of from 0.78 to 0.84 for predicting in-hospital mortality, and they neither capture the physiological trajectory of the first ICU Day nor perform uniformly across sexes and subgroups(7-10). General ICU scores such as APACHE IV and SAPS 3 were constructed for case-mix adjustment and benchmarking rather than individual prognosis (11-13). In cardiac ICUs, its discrimination dropped to 0.82–0.86 and the calibration degree drifts with time and case-mix (14-16). Dedicated cardiac-ICU scores (M-CARS, the CCCTN laboratory score, IABP-SHOCK II) improve specificity but still rely on single admission snapshots and have been validated in few centers (17-19). Therefore, recent

literature reviews call for dynamic, data-driven risk assessment frameworks specifically designed for cardiac critical care(20-22).

Applying machine learning to electronic health records enables the integration of irregularly sampled vital-sign and laboratory trajectories, treatments and free text. In the ICU populations, machine learning has demonstrated superior performance compared to traditional severity scoring(23-26). Multimodal fusion of structured and unstructured data has further improved mortality prediction (27-30). However, there are still significant gaps remain in the evidence base. Only about 6% of ICU machine-learning studies were validated on independent data (31, 32). When the model is migrated to a new institution, it usually loses 5 to 14 AUROC points (33-35), and systematic reviews of low-risk-of-bias comparisons find little or no discrimination gain of flexible algorithms over logistic regression (36). In ACS specifically, registry-scale gradient boosting improved the C-statistic by only 0.01 over regression(37) , and a MIMIC-derived model for myocardial infarction with cardiogenic shock fell from an AUROC of 0.84 to 0.76 on eICU external validation (38). No study has combined first-day physiological trajectories, treatment exposures and diagnosis text for critically ill CAD patients across more than 100 hospitals with hold-out-hospital and temporal validation.

We therefore used the eICU Collaborative Research Database (eICU-CRD), a comprehensive database encompassing 208 hospitals across the United States. It provides offers a unique multicenter framework conducive to leave-hospital-out validation(39, 40). There were three key objectives we pursued: (i) to develop multimodal machine-learning models, specifically employing gradient boosting on engineered features and a deep late-fusion sequence network, to predict in-hospital mortality after the first 24 h of ICU care in patients with CAD; (ii) confirm its validity in unseen hospitals and in a later calendar year against APACHE IV; (iii) characterize the data

modalities and variables drive the predictions. The study has been pre-specified in a statistical analysis plan and is reported according to TRIPOD+AI (41).

Methods

Study design and data source

This study was a multicenter retrospective cohort study aimed at developing and conducting internal–external validation of prognostic models. Data came from eICU-CRD v2.0, which contains 200,859 ICU admissions of 139,367 patients treated in 208 US hospitals participating in the Philips eICU telehealth programme between 2014 and 2015, including time-stamped vital signs, laboratory results, treatments, medications, structured diagnoses, past history and APACHE IV outputs (39, 42). The database is de-identified under HIPAA Safe Harbor, its creation was approved by the institutional review boards of the Massachusetts Institute of Technology and Philips with a waiver granted for the requirement of individual consent. Access to the database was subsequently obtained through PhysioNet following the completion of credentialing procedures. A statistical analysis plan (SAP) specifying the cohort, outcome, predictors, splits, models and metrics was finalised before any outcome–model association was examined (**Supplementary Appendix S1**). The deviation situations are listed in Supplementary **Table S9**.

Participants

We included ICU stays of adults (≥18 years) with CAD or ACS, identified from the structured diagnosis table (Philips diagnosis strings under *cardiovascular | chest pain/ASHD | coronary artery disease* or *acute coronary syndrome*, and *cardiogenic shock due to myocardial infarction*, excluding "myocardial infarction ruled out") or from the APACHE admission diagnosis (acute myocardial infarction, unstable or stable angina, or coronary artery bypass grafting). Only the first

ICU stay of each hospitalisation was eligible, and the stay had to last at least 24 h so that the full prediction window was observed. Stays with unknown hospital discharge status, missing sex, or no vital-sign record in the first 24 h were excluded. To preserve separation at the patient level, any patient appearing in both the development and hold-out hospital groups (multi-hospital health systems) had all their stays removed.

Outcome and prediction time

The primary outcome was all-cause in-hospital mortality. Secondary outcomes were ICU mortality and ICU stay longer than 7 days. Predictions were made at 24 h after ICU admission, using only data time-stamped between 60 min before and 1440 min after ICU admission.

Predictors

Four modalities were extracted (**Supplementary Table S3**). (1) *Hourly physiological time series*: 8 vital signs (heart rate, respiratory rate, SpO_2, temperature, systolic, diastolic and mean arterial pressure—invasive values preferred, non-invasive otherwise—and central venous pressure) and 33 laboratory variables (including troponin I, lactate, arterial blood gases, renal and hepatic panels, electrolytes, glucose, haematology and coagulation) were cleaned with physiological plausibility ranges, aggregated to hourly means (with hourly minimum and maximum for vitals) and stored as a 24 × 41 tensor with an observation mask; variables measured in <5% of stays were dropped; measurement coverage of every retained variable is shown in **Supplementary Figure S1**. (2) *Static data*: age, sex, ethnicity, height, weight, body-mass index, admission source, ICU type, time from hospital to ICU admission, coronary phenotype in the first 24 h (STEMI, NSTEMI, unstable angina, post-CABG surgery, cardiogenic shock, cardiac arrest, PCI), concurrent ICU diagnoses (heart failure, atrial fibrillation, sepsis, acute kidney injury, respiratory failure), 19 past-history comorbidities, first-day Glasgow Coma Scale, ventilation and intubation status, elective surgery

and ejection fraction where recorded. (3) *First-day treatments*: 33 binary indicators (plus three record-availability counts) derived from the treatment, medication and infusion tables (vasopressors, inotropes, mechanical circulatory support, mechanical and non-invasive ventilation, coronary angiography/PCI, CABG, thrombolysis, antiplatelet, anticoagulant, statin, β-blocker, ACE-inhibitor/ARB, nitrate, diuretic, antiarrhythmic, insulin infusion, sedation, renal replacement, transfusion, pacing and others). (4) *Diagnosis text*: the admission diagnosis, all priority-labelled ICU diagnosis strings entered within 24 h and past-history entries concatenated into one document (median 415 characters).

For the tabular models the time series were summarised per variable by first, last, minimum, maximum, mean, standard deviation, slope (last − first) and count, and three derived indices were added (maximum shock index, BUN/creatinine ratio, minimum PaO_2/FiO_2), giving 521 features after one-hot encoding and text embedding; text was embedded with term-frequency–inverse-document-frequency (TF-IDF) followed by truncated singular value decomposition (20 components), fitted on development data only. Hospital identifiers and hospital characteristics (region, size, teaching status) were never used as predictors. APACHE IV outputs were used only as the comparator.

Sample size

With an outcome prevalence of 7.5% and up to 521 candidate parameters, the Riley criteria for a Cox–Snell R^2 of 0.25 and expected shrinkage ≥0.9 require approximately 1600 events for unpenalised development (43, 44); the development set contained 1,007 events and the full cohort 1,315, and regularisation, hospital-grouped cross-validation and hold-out validation were therefore used to control optimism. The hold-out and temporal validation sets each exceeded the ≥200 events recommended for external validation (45).

Missing data

Missingness was treated as informative and handled without multiple imputation, mirroring deployment conditions. Gradient-boosting models used native missing-value handling; logistic regression and the deep network used median imputation (training data only) with missing-indicator variables; in the hourly tensor, values were forward-filled within the 24 h, hours before the first observation were set to the training median, and observation masks and time-since-last-observation channels were supplied to the network(46). Features missing in >95% of development stays were removed (Supplementary Table S3).

Model development

Data splits. Hospitals were randomly assigned, stratified by US region, to a development group (146 hospitals, ~80% of stays) and a hold-out group (42 hospitals). Within the development group hospitals were assigned to five folds (GroupKFold by hospital) that were used for all hyper-parameter tuning, early stopping and out-of-fold (OOF) prediction. For temporal validation, models were retrained with the same hyper-parameters on development stays discharged in 2014 and tested on all stays discharged in 2015(47, 48).

Tabular models. We fitted (a) L2-regularised logistic regression (regularisation strength selected by OOF log-loss); (b) LightGBM(49)and (c) XGBoost(50), each tuned by Bayesian optimisation with Optuna(51)(100 and 60 trials respectively, objective OOF log-loss, per-fold early stopping; selected hyper-parameters in Supplementary Table S4). Hold-out and temporal predictions were the mean of the fivefold models. The gradient-boosting model with the lower development OOF log-loss (LightGBM) was pre-specified as the primary model before hold-out results were examined.

Deep multimodal network (CHD-MMNet). The network architecture employed a late fusion strategy involving three distinct branches: a two-layer bidirectional gated recurrent unit (GRU; hidden size 128)(52) over the hourly tensor (value, mask and time-since-observation channels, 123 inputs per hour) with concatenated final-state and mean pooling; a two-layer multilayer perceptron for static and treatment features; and a two-layer perceptron for a 64-dimensional TF-IDF–SVD text embedding. Branch outputs (3 × 128) were concatenated and passed to a fusion layer and sigmoid output (666,497 parameters). Training used focal loss ($\gamma = 2$, $\alpha = 0.75$), AdamW (learning rate 10^{-3}, weight decay 10^{-4}, cosine schedule), batch size 256, dropout 0.3, gradient clipping and early stopping on fold-validation area under the precision–recall curve (patience 10). A Transformer encoder alternative and 24 hyper-parameter configurations were compared on development folds only. Because focal loss yields deliberately under-confident probabilities, a logistic (Platt) recalibration fitted on development OOF predictions was applied to hold-out and temporal predictions. Ablated variants (no text, no time series, no treatment, static only, time series only) followed the same protocol. Training was performed in PyTorch 2.14 (53) on an NVIDIA GB10 GPU (about 4.5 s per fold model).

Deployable reduced model and risk tiers

Given that the primary model employs 521 features, we have derived a reduced model for settings where features must be assembled explicitly. Candidate variables were ranked by mean absolute SHAP value in the primary model, excluding the diagnosis-text components (which cannot be entered at the bedside and do not transport across databases), and models with 5, 10, 15, 20, 25 and 30 variables were fitted with the same hyper-parameters and folds; we retained the smallest set whose development out-of-fold AUROC was within an absolute difference of 0.01 of the full model (**Supplementary Table S14**). Predicted risks were also grouped into four pre-defined

clinical tiers (<2.5%, 2.5-10%, 10-25% and ≥25%) based on established clinical thresholds, and the observed mortality, share of patients and share of deaths in each tier were tabulated in all three cohorts.

Exploratory analysis: a language-model paradigm

Due to the bag-of-words encoder in the pre-specified text branch and recent work reports that encoding structured intensive-care variables as natural language and fine-tuning a pretrained transformer can outperform tabular baselines(54), we added a post-hoc, non-pre-specified experiment (**Supplementary Table S9**). The eICU-CRD and MIMIC-IV modules we accessed do not include narrative clinical notes, as eICU records note entries using a controlled vocabulary (comprising 469 distinct values with a median length of 9 characters). Consequently, a free-text analysis was not possible. Instead, each stay was serialised into a standardised natural-language description (demographics, coronary presentation, first-24-hour conditions, past history, Glasgow Coma Scale, vital-sign ranges with verbal qualifiers derived from development-set deciles, laboratory peaks and last values, variables not measured, and treatments; median 1,400 characters, 453 tokens). Bio_ClinicalBERT was fine-tuned using low-rank adaptation ($r = 8$ on query and value projections; 0.49 of 108.8 million parameters trainable) and an adaptive softmax weighting over all 13 encoder layers, using the same hospital-grouped folds, focal-free class-weighted loss, and early stopping on validation AUPRC. Fine-tuning ran on a local GPU; no patient data left the analysis environment, consistent with the PhysioNet data-use agreement. We evaluated the language model alone, gradient boosting on its 768-dimensional fused embedding, an "LM-boost" model combining the embedding with 521 tabular features, and a logit-average probability fusion approach.

External validation in MIMIC-IV

To account for the dated nature of the eICU-CRD dataset (2014–2015) and to evaluate its cross-institutional applicability, we pre-specified an addendum to the analysis plan (**Supplementary Appendix S1b**) before examining any external outcome. Due to the unavailability or non-comparability of 80 out of the 521 features outside eICU (diagnosis-text components, APACHE-derived variables, troponin I and BNP — MIMIC-IV measures troponin T and NT-proBNP — eICU record-structure counts and fine-grained admission source), *we defined a portable model, LightGBM-P, using the 441 features that map across databases plus five coarse-grained substitutes (admission source in four categories; cardiac ICU)*. LightGBM-P was trained on the eICU development set with the same hyper-parameters and hospital-grouped folds, achieving the same performance as the full model (hold-out AUROC 0.897; temporal 0.885), and was then frozen (specification in **Supplementary Table S11**; the external and eICU cohorts are compared in **Supplementary Table S12**).

The external cohort was drawn from MIMIC-IV v3.1 (Beth Israel Deaconess Medical Center, 2008–2022) (55) using the same eligibility rules: adults, first ICU stay of the hospitalisation, ICU length of stay ≥24 h, coronary artery disease or acute coronary syndrome identified from ICD-9 410–414 or ICD-10 I20–I25 in any diagnosis position or from CABG/PCI procedure codes, and at least one vital sign in the first 24 h. Vital-sign and laboratory variables were mapped by *itemid* (every mapping verified against *d_items*/*d_labitems* at run time), harmonised to eICU units and processed with the identical plausibility ranges, hourly aggregation and summary statistics; treatments were derived from *inputevents*, *procedureevents*, *prescriptions* and procedure codes within the prediction window; phenotypes and comorbidities were derived from ICD codes. The outcome was *hospital_expire_flag*.

As MIMIC-IV contains no APACHE IV output, we computed two standard intensive-care severity scores on the same patients and in the same window as clinical comparators: the day-1 Sequential Organ Failure Assessment (SOFA) and the Oxford Acute Severity of Illness Score (OASIS), the latter converted to a probability by its published logistic calibration. Following the convention of the MIMIC concept code, components without a measurement were scored zero; the PaO_2/FiO_2 ratio was derivable in 55.7% and bilirubin in 42.4% of stays, whereas urine output, platelets and Glasgow Coma Scale components were available in more than 97%.

The analyses adhered to the pre-specified hierarchy: (i) the frozen model applied without any updating; (ii) intercept-only recalibration; (iii) intercept-and-slope (logistic) recalibration; and (iv) a LightGBM retrained within MIMIC-IV by 5-fold cross-validation, giving the locally attainable upper bound. Recalibration and local retraining were cross-fitted over five folds to avoid optimism. Because MIMIC-IV records an admission-year group for each stay, performance was additionally reported across five three-year strata spanning 2008–2022, directly assessing calendar-time transportability.

Comparator

APACHE IV predicted hospital mortality, as provided in eICU-CRD, was evaluated on the 88.8% of stays for which it was available. All machine-learning models were additionally evaluated on this same subset for like-for-like comparison, and APACHE IV re-calibrated by logistic (with intercepts and slopes fitted on the development (intercept and slope fitted on development data) was also reported.

Statistical analysis

Baseline characteristics are summarised as mean ± SD, median [IQR] or n (%), with standardised mean differences. Discrimination was quantified by the area under the receiver-operating-

characteristic curve (AUROC) and the area under the precision–recall curve (AUPRC), considering the low event rate(56). Comparisons of AUROCs were inducted using the DeLong test (57) with Bonferroni correction within each validation set. Calibration was assessed by calibration plots (deciles with 95% CI and flexible curves), calibration intercept and slope, the integrated calibration index (ICI) and the Brier score(58-60). Clinical utility was evaluated by decision-curve analysis across threshold probabilities of 1–60%(61, 62). Sensitivity, specificity and predictive values were reported at the Youden-optimal threshold and at 90% sensitivity. Uncertainty was expressed as 95% confidence intervals from 1000 patient-level bootstrap resamples. Pre-specified subgroup analyses (coronary presentation, age, sex, mechanical ventilation, cardiogenic shock, ICU type, ethnicity) and fairness metrics (AUROC and calibration-in-the-large by sex and ethnicity) were performed in hold-out hospitals, and between-hospital heterogeneity was summarised by per-hospital AUROC (hospitals with ≥10 deaths) with a random-effects pooled estimate and I^2. Model explanations used TreeSHAP for LightGBM (63, 64) and Integrated Gradients for CHD-MMNet(65), reported as descriptive summaries of model behaviour rather than causal inference (66). Analyses were performed in Python 3.12 with duckdb, scikit-learn(67), LightGBM, XGBoost, Optuna, SHAP, Captum and PyTorch; the random seed was 2026 throughout. Reporting followed TRIPOD+AI (**Supplementary Table S10**)(41, 68).

Results

Cohort

Among 200,859 ICU hospitalization records, 27,736 carried a CAD/ACS diagnosis. After restricting to adults (27,731), first ICU stays (24,250), stays of ≥24 h (17,974), known outcome (17,861), known sex (17,859), presence of first-day vital signs (17,717) and patient-level split

integrity (17,586), the final cohort comprised 17,586 stays of 16,792 patients from 188 hospitals (**Figure 1A**). The in-hospital mortality rate was 7.5% (1,315 deaths) and ICU mortality 4.5%. The development set comprised 12,933 stays (1,007 deaths, 7.8%) from 146 hospitals and the hold-out set 4,653 stays (308 deaths, 6.6%) from 42 hospitals; 8,716 stays (628 deaths) were discharged in 2014 and 8,870 (687 deaths) in 2015. Baseline characteristics of development and hold-out cohorts are compared in **Supplementary Table S2**.

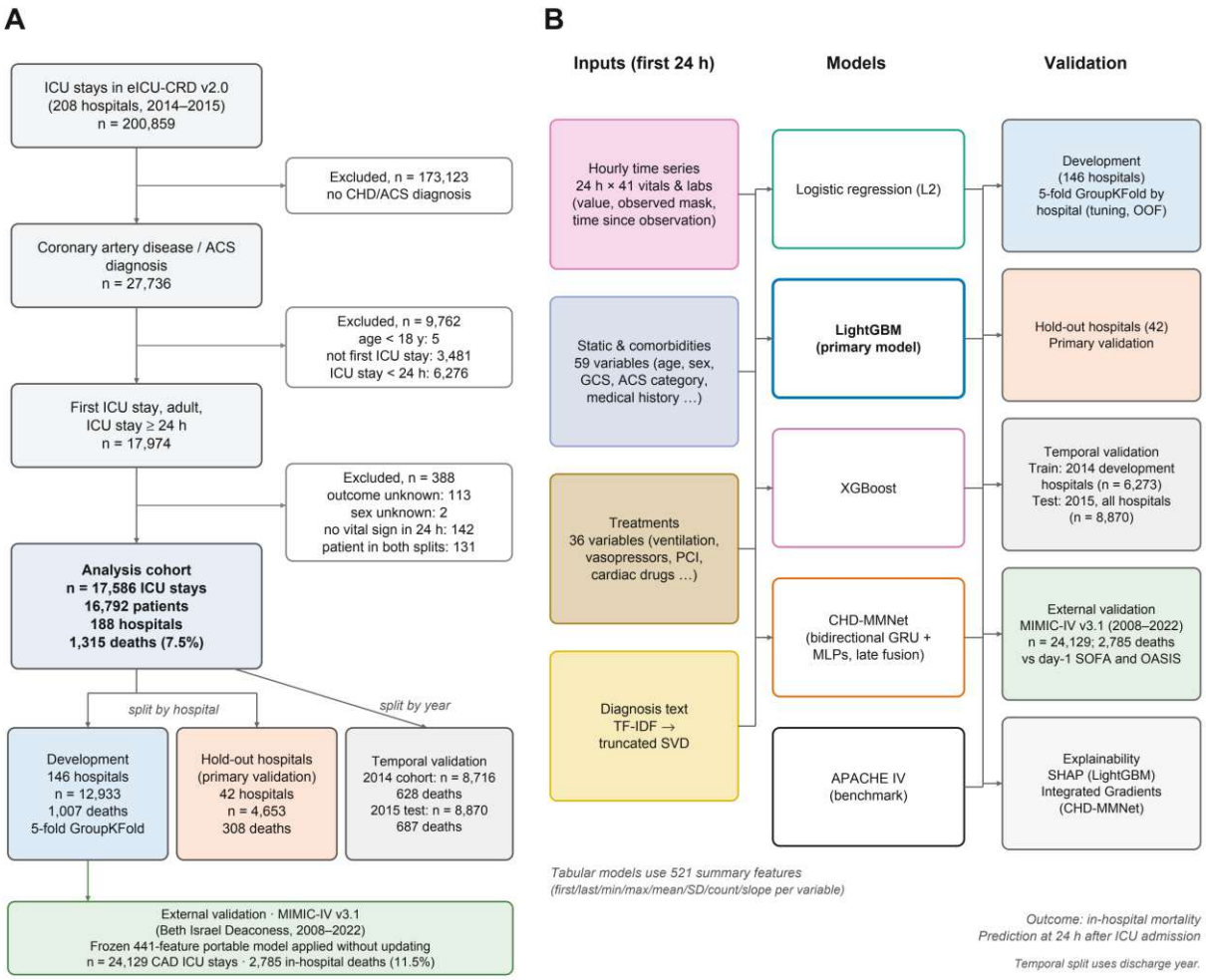

Figure 1. Study design. (**A**) Selection of the analysis cohort from the eICU Collaborative Research Database v2.0 (208 hospitals, 2014–2015). Of 200,859 ICU stays, 27,736 carried a coronary artery disease / acute coronary syndrome diagnosis; after restriction to adult first ICU stays with ICU length of stay ≥24 h (n = 17,974) and

exclusion of stays with unknown outcome (113), unknown sex (2), no vital sign in the first 24 h (142) or patients present in both hospital splits (131), 17,586 stays (16,792 patients, 188 hospitals; 1,315 in-hospital deaths, 7.5%) remained. Hospitals were split into a development set (146 hospitals, n = 12,933, 1,007 deaths; 5-fold GroupKFold by hospital for tuning and out-of-fold estimates) and hold-out hospitals (42 hospitals, n = 4,653, 308 deaths; primary validation). For temporal validation, models were re-trained on 2014 admissions in the development hospitals (n = 6,273) and tested on all 2015 admissions (n = 8,870, 687 deaths; 2014 cohort n = 8,716, 628 deaths). (**B**) Data modalities from the first 24 h of the ICU stay (hourly 24 × 41 time-series tensor of vital signs and laboratory values with value, observed-mask and time-since-observation channels; 59 static/comorbidity variables; 36 treatment variables; free-text diagnosis strings encoded by TF-IDF and truncated SVD), the models compared (L2 logistic regression, LightGBM (primary model], XGBoost, CHD-MMNet — a late-fusion network with a bidirectional GRU time-series encoder — and the APACHE IV benchmark), the validation scheme and the explainability analyses. Tabular models use 521 summary features (first/last/min/max/mean/SD/count/slope per variable). Outcome: in-hospital mortality, predicted 24 h after ICU admission.

Table 1 shows baseline characteristics (for the complete 125 variable versions, see **Supplementary Table S1**). Mean age was 67.6 ± 12.6 years and 65.0% were men. Coronary presentation was NSTEMI in 3,835 (21.8%), STEMI in 2,604 (14.8%), unstable angina in 1,251 (7.1%), post-CABG surgery in 4,322 (24.6%) and CAD without ACS admitted with other critical illness ("stable CAD/other") in 5,574 (31.7%); in-hospital mortality by presentation was 8.5%, 6.7%, 1.9%, 1.8% and 12.8% respectively. Cardiogenic shock was documented within 24 h in 2.6% and cardiac arrest in 4.2%. Non-survivors were older (73.2 vs 67.1 years), had lower first-day Glasgow Coma Scale (median 13 vs 15), lower minimum systolic pressure (78 vs 86 mmHg) and SpO_2 (85% vs 90%), and higher maximum lactate (3.0 vs 1.9 mmol/L), blood urea nitrogen (34 vs 19 mg/dL) and creatinine (1.7 vs 1.1 mg/dL); 31.5% of the cohort received invasive mechanical ventilation and 29.7% vasopressors in the first 24 h. Maximum troponin I concentrations did not differ between survivors and non-survivors, although troponin was measured more often in non-survivors (56% vs 41%). APACHE IV predictions were available for 88.8% of stays.

Of the 24,250 eligible first-time ICU hospitalizations, 6276 (25.9%) lasted less than 24 h and were therefore outside the scope of a 24-hour prediction model. These patients were slightly younger

(mean 65.7 vs 67.6 years) and had a lower in-hospital mortality than those retained (6.2% vs 7.5%), although 285 of them died in the ICU within the first day (median 11.2 h after admission); 22.3% of all in-hospital deaths among eligible first stays occurred in this excluded group. The model therefore applies to patients still in the ICU at 24 h and does not address deaths occurring before that point.

Table 1. Baseline characteristics of the analysis cohort by in-hospital outcome.

Characteristic	Overall (n = 17,586)	Survivors (n = 16,271)	Non-survivors (n = 1,315)	SMD	P value	Missing, %
Age, years	67.6 ± 12.6	67.1 ± 12.6	73.2 ± 12.1	0.49	<0.001	0.0
Male sex	11438 (65.0)	10674 (65.6)	764 (58.1)	0.15	<0.001	
Body-mass index, kg/m²	28.5 (24.8–33.2)	28.6 (24.9–33.2)	27.3 (23.4–32.3)	0.12	<0.001	3.7
Coronary presentation					<0.001	
NSTEMI	3835 (21.8)	3508 (21.6)	327 (24.9)	0.08		
Post-CABG surgery	4322 (24.6)	4246 (26.1)	76 (5.8)	0.58		
STEMI	2604 (14.8)	2430 (14.9)	174 (13.2)	0.05		
Stable CAD / other	5574 (31.7)	4860 (29.9)	714 (54.3)	0.51		
Unstable angina	1251 (7.1)	1227 (7.5)	24 (1.8)	0.27		
ICU type					<0.001	
Cardiac ICU (CCU/CTICU/CSICU)	8432 (47.9)	7936 (48.8)	496 (37.7)	0.22		
Medical/Surgical/Neuro ICU	9154 (52.1)	8335 (51.2)	819 (62.3)	0.22		
Glasgow Coma Scale, day 1	15.0 (13.0-15.0)	15.0 (14.0-15.0)	13.0 (6.5-15.0)	0.65	<0.001	4.1
APACHE IV score	52.0 (39.0-68.0)	51.0 (38.0-65.0)	81.0 (62.5-106.0)	1.20	<0.001	10.3
Heart rate, mean 24 h, bpm	79.8 (70.1-90.0)	79.3 (69.8-89.1)	87.1 (76.0-99.8)	0.48	<0.001	0.0
Systolic BP, minimum 24 h, mmHg	86.0 (75.0-99.0)	86.0 (75.0-100.0)	78.0 (66.0-88.0)	0.53	<0.001	0.2
Mean arterial pressure, minimum 24 h, mmHg	58.0 (50.0-67.0)	58.0 (51.0-68.0)	52.0 (43.0-60.0)	0.54	<0.001	0.2
SpO_2, minimum 24 h, %	90.0 (85.0-93.0)	90.0 (86.0-93.0)	85.0 (77.0-91.0)	0.55	<0.001	0.8
Shock index, maximum 24 h	1.2 (1.0-1.5)	1.2 (0.9-1.4)	1.4 (1.2-1.9)	0.62	<0.001	0.2
Lactate, maximum 24 h,mmol/L	2.0 (1.3-3.6)	1.9 (1.3-3.1)	3.0 (1.7-6.2)	0.57	<0.001	80.0
Troponin I, maximum 24 h, ng/mL	3.7 (0.5-18.9)	3.7 (0.5-18.9)	3.7 (0.5-18.9)	0.03	0.883	57.8
Creatinine, maximum 24 h, mg/dL	1.1 (0.8-1.6)	1.1 (0.8-1.5)	1.7 (1.1-2.8)	0.46	<0.001	8.0
Blood urea nitrogen, maximum 24 h, mg/dL	20.0 (14.0-31.0)	19.0 (14.0-29.0)	34.0 (23.0-52.0)	0.70	<0.001	9.2
Glucose, maximum 24 h, mg/dL	164.0 (128.0-211.0)	163.0 (127.0-208.0)	191.0 (140.0-261.0)	0.34	<0.001	4.1
Bicarbonate, minimum 24 h, mmol/L	22.9 (20.0-25.0)	23.0 (20.1-25.0)	20.0 (16.0-24.0)	0.53	<0.001	8.5
White-cell count, maximum 24 h, $\times 10^9$/L	12.0 (8.9-16.1)	11.8 (8.8-15.8)	14.6 (10.7-20.5)	0.37	<0.001	10.0
Haemoglobin, minimum 24 h, g/dL	10.9 (9.2-12.7)	10.9 (9.2-12.7)	10.5 (8.8-12.2)	0.16	<0.001	9.0
Albumin, minimum 24 h, g/dL	3.1 (2.7-3.5)	3.2 (2.7-3.5)	2.7 (2.2-3.1)	0.72	<0.001	59.0
Cardiogenic shock (24 h)	459 (2.6)	341 (2.1)	118 (9.0)	0.30	<0.001	

Cardiac arrest	737 (4.2)	499 (3.1)	238 (18.1)	0.50	<0.001	
Heart failure (24 h)	2108 (12.0)	1836 (11.3)	272 (20.7)	0.26	<0.001	
Sepsis (24 h)	1598 (9.1)	1259 (7.7)	339 (25.8)	0.50	<0.001	
Acute kidney injury (24 h)	1764 (10.0)	1417 (8.7)	347 (26.4)	0.48	<0.001	
Respiratory failure (24 h)	3092 (17.6)	2523 (15.5)	569 (43.3)	0.64	<0.001	
Hypertension	11049 (62.8)	10229 (62.9)	820 (62.4)	0.01	0.736	
Diabetes mellitus	6419 (36.5)	5915 (36.4)	504 (38.3)	0.04	0.161	
Chronic heart failure	3272 (18.6)	2908 (17.9)	364 (27.7)	0.24	<0.001	
Chronic kidney disease	2568 (14.6)	2268 (13.9)	300 (22.8)	0.23	<0.001	
COPD	2577 (14.7)	2307 (14.2)	270 (20.5)	0.17	<0.001	
Cancer	1946 (11.1)	1730 (10.6)	216 (16.4)	0.17	<0.001	
Invasive mechanical ventilation	5547 (31.5)	4941 (30.4)	606 (46.1)	0.33	<0.001	
Non-invasive ventilation	1560 (8.9)	1325 (8.1)	235 (17.9)	0.29	<0.001	
Vasopressor	5216 (29.7)	4535 (27.9)	681 (51.8)	0.50	<0.001	
Inotrope	2146 (12.2)	1939 (11.9)	207 (15.7)	0.11	<0.001	
Mechanical circulatory support (IABP/ECMO/Impella)	697 (4.0)	598 (3.7)	99 (7.5)	0.17	<0.001	
Coronary angiography/PCI	3386 (19.3)	3227 (19.8)	159 (12.1)	0.21	<0.001	
CABG surgery	2362 (13.4)	2316 (14.2)	46 (3.5)	0.38	<0.001	
Renal replacement therapy	459 (2.6)	388 (2.4)	71 (5.4)	0.16	<0.001	
ICU length of stay, days	2.2 (1.5-3.9)	2.1 (1.5-3.7)	3.8 (2.1-7.1)	0.51	<0.001	0.0
Hospital length of stay, days	5.9 (3.4-9.7)	5.9 (3.4-9.5)	6.1 (3.2-11.4)	0.12	0.032	0.0
ICU mortality	785 (4.5)	14 (0.1)	771 (58.6)	1.68	<0.001	

Values are mean ± SD, median (IQR) or n (%). SMD, standardised mean difference; P from t test / Mann–Whitney U / χ^2 as appropriate. 24-h variables refer to the first 24 h after ICU admission; APACHE IV score available for 89.7% of stays. The full list of variables is given in Supplementary Table S1.

Model performance in the development set

In hospital-grouped cross-validation, OOF AUROCs were 0.884 (95% CI 0.875-0.893) for LightGBM, 0.883 (0.874-0.892) for XGBoost, 0.871 (0.861-0.882) for logistic regression, 0.866 (0.856-0.878) for CHD-MMNet and 0.847 (0.834-0.858) for APACHE IV (**Table 2**). Tuning histories and deep-model training curves are shown in **Supplementary Figure S2**, and development-set calibration for all models in **Supplementary Figure S3**. LightGBM had the lowest OOF log-loss and is therefore selected as the primary model for further analysis.

Table 2. Discrimination and calibration of the prediction models on the three validation sets.

Validation set	Model	n / deaths	AUROC (95% CI)	AUPRC (95% CI)	Brier score (95% CI)	Calibration slope (95% CI)	Calibration intercept (95% CI)	DeLong P vs APACHE IV (ΔAUROC)	DeLong P vs CHD-MMNet (ΔAUROC)

Development, out-of-fold (n = 12,933; 1,007 deaths)	Logistic regression	12,933 / 1,007	0.871 (0.861-0.882)	0.404 (0.372-0.437)	0.057 (0.054-0.060)	1.07 (1.02-1.13)	0.01 (−0.06 to 0.08)	<0.001 (Δ = +0.025)	
Development, out-of-fold (n = 12,933; 1,007 deaths)	LightGBM (primary model)	12,933 / 1,007	0.884 (0.875-0.893)	0.431 (0.400-0.463)	0.056 (0.053-0.059)	1.01 (0.97-1.06)	0.13 (0.06-0.20)	<0.001 (Δ = +0.037)	<0.001 (Δ = +0.017)
Development, out-of-fold (n = 12,933; 1,007 deaths)	XGBoost	12,933 / 1,007	0.883 (0.874-0.892)	0.435 (0.404-0.467)	0.055 (0.052-0.059)	1.03 (0.99-1.08)	0.07 (0.00-0.15)	<0.001 (Δ = +0.036)	<0.001 (Δ = +0.017)
Development, out-of-fold (n = 12,933; 1,007 deaths)	CHD-MMNet (deep multimodal)†	12,933 / 1,007	0.866 (0.856-0.878)	0.419 (0.388-0.453)	0.056 (0.053-0.059)	1.00 (0.94-1.07)	−0.00 (−0.06 to 0.07)	<0.001 (Δ = +0.020)	reference
Development, out-of-fold (n = 12,933; 1,007 deaths)	APACHE IV‡	11,641 / 917	0.847 (0.834-0.858)	0.380 (0.350-0.414)	0.065 (0.062-0.068)	0.82 (0.78-0.87)	−0.65 (−0.74 to −0.58)	reference	
Development, out-of-fold (n = 12,933; 1,007 deaths)	APACHE IV, recalibrated‡	11,641 / 917	0.847 (0.834-0.858)	0.380 (0.350-0.414)	0.060 (0.056-0.063)	1.00 (0.95-1.06)	0.00 (−0.08 to 0.07)	reference	
Hold-out hospitals (n = 4,653; 308 deaths)	Logistic regression	4,653 / 308	0.886 (0.864–0.904)	0.448 (0.392-0.510)	0.047 (0.042-0.052)	1.17 (1.08-1.29)	−0.06 (−0.20 to 0.07)	0.021 (Δ = +0.020)	
Hold-out hospitals (n = 4,653; 308 deaths)	LightGBM (primary model)	4,653 / 308	0.897 (0.879-0.914)	0.461 (0.405-0.519)	0.046 (0.041-0.051)	1.10 (1.01-1.20)	0.07 (−0.07 to 0.20)	<0.001 (Δ = +0.032)	0.014 (Δ = +0.013)
Hold-out hospitals (n = 4,653; 308 deaths)	XGBoost	4,653 / 308	0.899 (0.880-0.915)	0.464 (0.408-0.523)	0.046 (0.041-0.051)	1.12 (1.03-1.23)	0.01 (−0.13 to 0.14)	<0.001 (Δ = +0.034)	0.003 (Δ = +0.015)
Hold-out hospitals (n = 4,653; 308 deaths)	CHD-MMNet (deep multimodal)†	4,653 / 308	0.884 (0.864-0.904)	0.448 (0.396-0.509)	0.047 (0.043-0.052)	1.14 (1.04-1.28)	−0.01 (−0.13 to 0.10)	0.041 (Δ = +0.016)	reference
Hold-out hospitals (n = 4,653; 308 deaths)	APACHE IV‡	3,974 / 253	0.864 (0.842–0.887)	0.346 (0.291-0.415)	0.055 (0.050-0.060)	0.82 (0.74-0.91)	−0.74 (−0.92 to −0.58)	reference	
Hold-out hospitals (n = 4,653; 308 deaths)	APACHE IV, recalibrated‡	3,974 / 253	0.864 (0.842-0.887)	0.346 (0.291-0.415)	0.050 (0.044-0.055)	0.99 (0.90-1.10)	−0.10 (−0.27 to 0.03)	reference	
Temporal validation, 2015 (n = 8,870; 687 deaths)	Logistic regression	8,870 / 687	0.875 (0.863-0.887)	0.411 (0.375-0.450)	0.057 (0.053-0.060)	1.20 (1.13-1.28)	0.03 (−0.06 to 0.11)	<0.001 (Δ = +0.021)	
Temporal validation, 2015 (n = 8,870; 687 deaths)	LightGBM (primary model)	8,870 / 687	0.886 (0.874-0.897)	0.439 (0.402-0.480)	0.055 (0.051-0.059)	1.08 (1.02-1.15)	0.14 (0.05-0.23)	<0.001 (Δ = +0.033)	0.005 (Δ = +0.011)
Temporal validation, 2015 (n = 8,870; 687 deaths)	XGBoost	8,870 / 687	0.887 (0.876-0.898)	0.439 (0.401-0.479)	0.055 (0.051-0.059)	1.13 (1.07-1.21)	0.11 (0.02-0.20)	<0.001 (Δ = +0.034)	0.001 (Δ = +0.012)

Temporal validation, 2015 (n = 8,870; 687 deaths)	CHD-MMNet (deep multimodal)†	8,870 / 687	0.875 (0.864-0.887)	0.411 (0.375-0.451)	0.057 (0.053-0.061)	1.23 (1.15-1.31)	0.07 (0.00-0.16)	<0.001 (Δ = +0.023)	reference
Temporal validation, 2015 (n = 8,870; 687 deaths)	APACHE IV‡	7,957 / 626	0.854 (0.840-0.868)	0.383 (0.345-0.425)	0.064 (0.060-0.068)	0.83 (0.78-0.89)	−0.61 (−0.71 to −0.52)	reference	
Temporal validation, 2015 (n = 8,870; 687 deaths)	APACHE IV, recalibrated‡	7,957 / 626	0.854 (0.840-0.868)	0.383 (0.345-0.425)	0.059 (0.055-0.063)	1.01 (0.95-1.08)	0.03 (−0.06 to 0.12)	reference	

AUROC, area under the receiver-operating-characteristic curve; AUPRC, area under the precision–recall curve; 95% CIs from 1,000 patient-level bootstrap resamples. Calibration intercept = calibration-in-the-large (logistic intercept with slope fixed at 1); slope = logistic recalibration slope.

†CHD-MMNet: AUROC/AUPRC from the raw network output; Brier score, slope and intercept after Platt recalibration fitted on development out-of-fold predictions (in-sample for the development row; out-of-sample for hold-out and temporal sets).

‡APACHE IV evaluated on stays with an available APACHE IV prediction (development 11,641/917; hold-out 3,974/253; 2015 7,957/626); recalibrated APACHE IV = logistic recalibration fitted on development (intercept −0.862, slope 0.824). DeLong tests vs APACHE IV are computed on the APACHE-available subset; DeLong tests vs CHD-MMNet on all rows. Development-set estimates for CHD-MMNet, LightGBM and XGBoost are out-of-fold (5-fold GroupKFold by hospital).

340

Hold-out-hospital validation

Among the 42 unseen hospitals, LightGBM achieved an AUROC of 0.897 (0.879-0.914) and AUPRC of 0.461 (0.405-0.519), compared with 0.864 (0.842-0.887) and 0.346 (0.291-0.415) for APACHE IV (**Figure 2A-B, Table 2**). On the APACHE-available subset (n = 3974, 253 deaths) the AUROC difference was 0.032 (DeLong p = 1.2×10^{-4}; Bonferroni-adjusted $P = 7\times10^{-4}$). XGBoost was indistinguishable from LightGBM (AUROC 0.899, 0.880-0.915; $P = 0.34$), logistic regression reached 0.886 (0.864-0.904), and CHD-MMNet 0.884 (0.864-0.904)—superior to APACHE IV on the common subset (Δ 0.016, $P = 0.041$) but inferior to LightGBM (Δ 0.013, $P = 0.014$).

LightGBM was well calibrated and does not require recalibration (intercept 0.07, 95% CI −0.07 to 0.20; slope 1.10, 1.01-1.20; ICI 0.008; Brier 0.046, 0.041-0.051), whereas APACHE IV systematically over-predicted risk (intercept −0.74, −0.92 to −0.58; slope 0.82, 0.74-0.91; observed-to-expected ratio 0.62; ICI 0.040; Brier 0.055) (**Figure 3A**). After logistic recalibration of APACHE IV, its Brier score improved to 0.050, but did not change discrimination. Platt-recalibrated CHD-MMNet probabilities transported to unseen hospitals with a slope of 1.14 and intercept of −0.01 (Brier 0.047). In decision-curve analysis LightGBM provided higher net benefit than APACHE IV and than treat-all/treat-none strategies across threshold probabilities from 2% to 60% (net benefit at 5%, 10%, 20% and 30% thresholds: 0.046, 0.037, 0.024 and 0.016 vs 0.038, 0.027, 0.014 and 0.007 for APACHE IV; **Figure 3B**). At the Youden-optimal threshold, the sensitivity of LightGBM was 0.857 (0.816-0.895), the specificity was 0.812 (0.801-0.823), the positive predictive value was 0.244 and negative predictive value 0.988. At 90% sensitivity, its specificity was 0.694 versus 0.636 for APACHE IV (**Supplementary Table S7**).

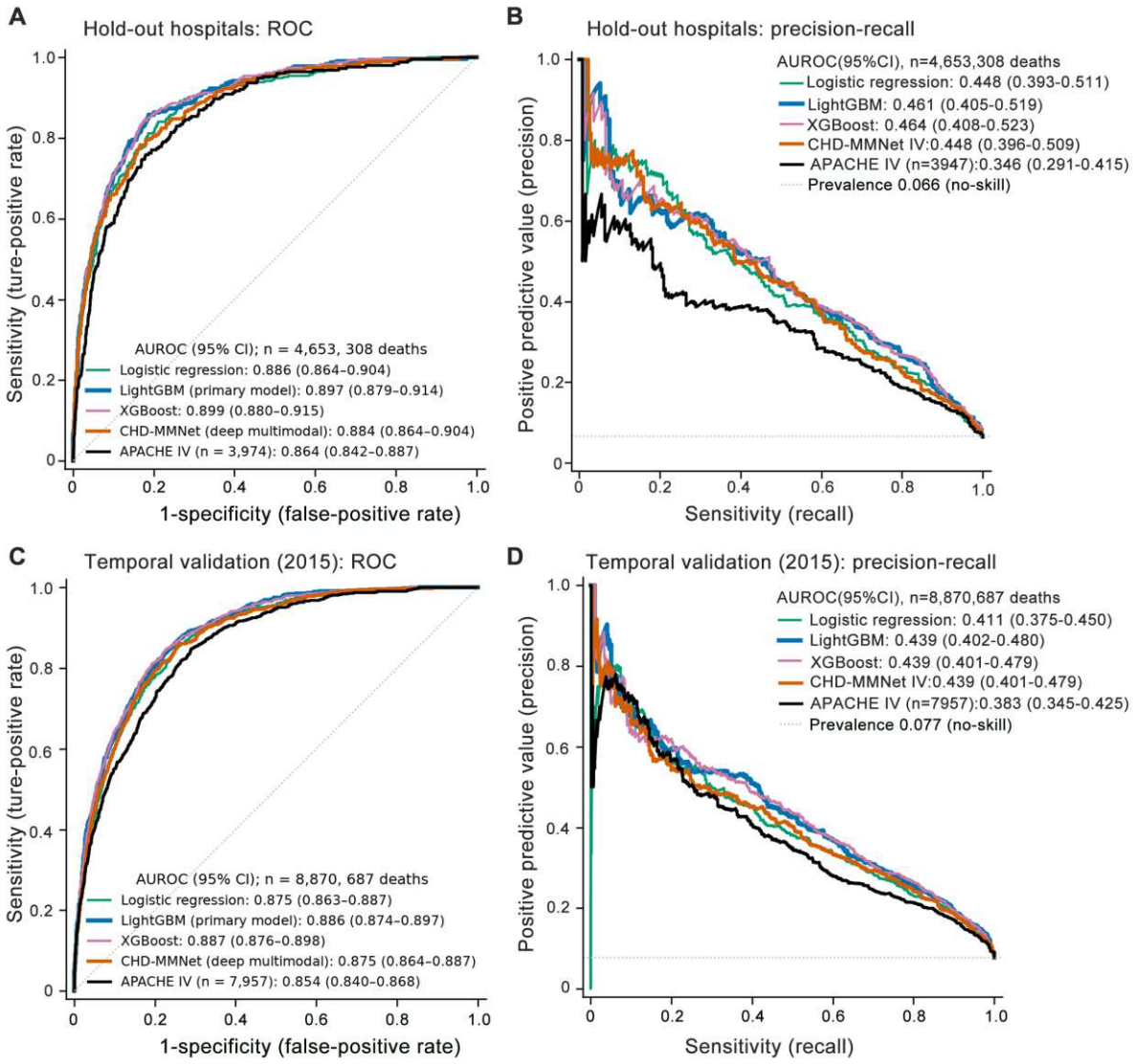

Figure 2. LightGBM and XGBoost outperform APACHE IV in discrimination in both hold-out hospitals and temporal validation. Receiver-operating-characteristic (ROC) curves (**A**, **C**) and precision–recall curves (**B**, **D**) for the five models in the 42 hold-out hospitals (**A**, **B**; n = 4,653, 308 deaths) and in the 2015 temporal test set (**C**, **D**; n = 8,870, 687 deaths). Curves are computed from patient-level predictions; legends give the area under the curve with 1,000-resample patient-level bootstrap 95% CI. APACHE IV is evaluated on stays with an available APACHE IV prediction (hold-out n = 3,974, 253 deaths; 2015 n = 7,957, 626 deaths); all other models are evaluated on all stays. Dotted lines indicate chance (ROC) and the outcome prevalence (precision–recall; hold-out 0.066, 2015 0.077). CHD-MMNet curves use the raw network output (ranking metrics are unaffected by recalibration).

Temporal validation

Models retrained on 2014 discharges and tested on the 8,870 stays discharged in 2015 (687 deaths) retained their ranking: AUROC 0.886 (0.874-0.897) for LightGBM, 0.887 (0.876-0.898) for

XGBoost, 0.875 (0.863-0.887) for logistic regression, 0.875 (0.864-0.887) for CHD-MMNet and 0.854 (0.840-0.868) for APACHE IV (**Figure 2C-D; Table 2**). LightGBM calibration remained acceptable (slope 1.08, intercept 0.14; Brier 0.055), while APACHE IV showed overprediction again (intercept −0.61, slope 0.84; Brier 0.064) (**Figure 3C-D**).

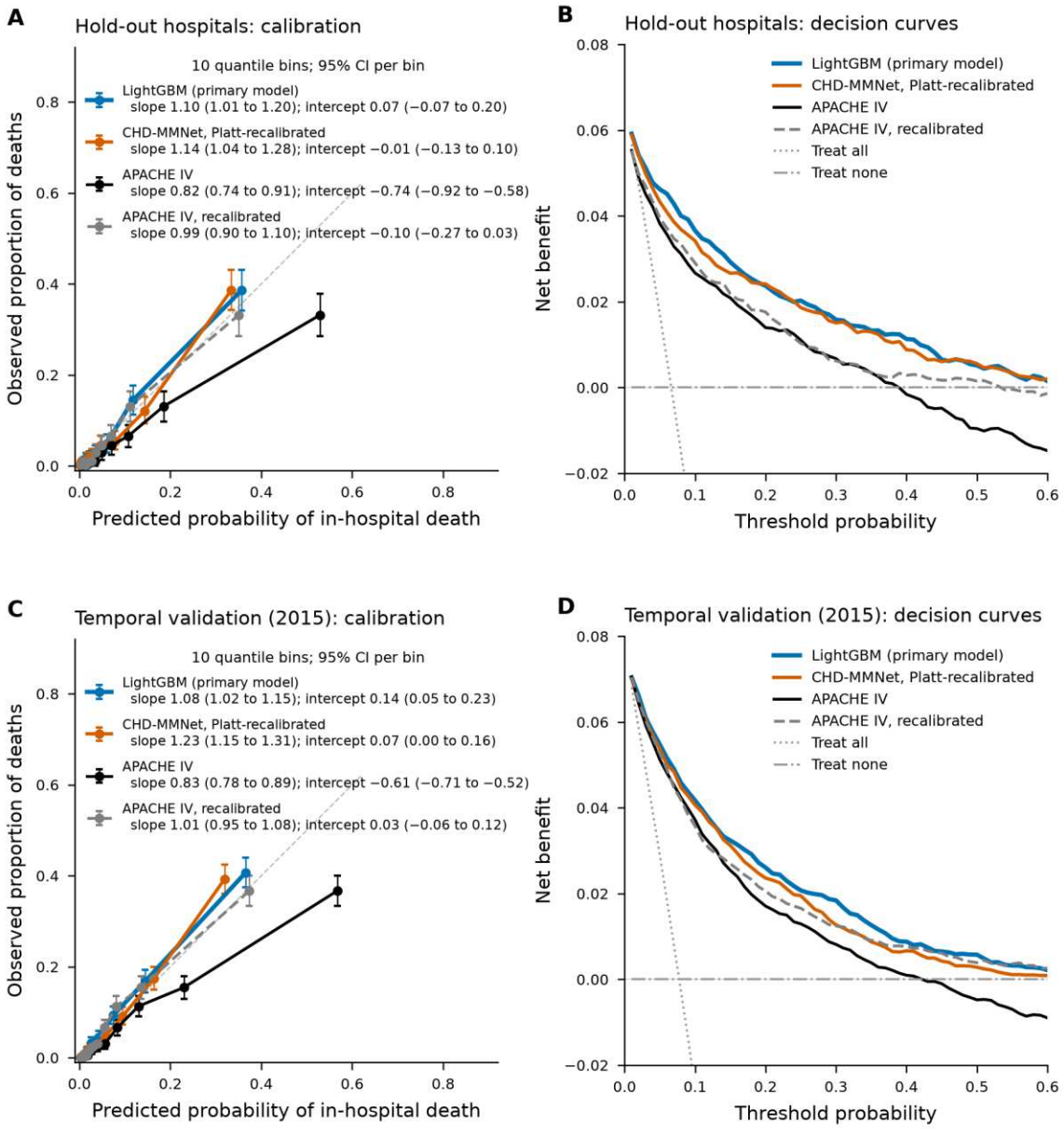

Figure 3. LightGBM achieves better calibration and higher net benefit than APACHE IV and CHD-MMNet in hold-out and temporal validation. (**A**, **C**) Calibration curves with 10 quantile bins of predicted risk (points = mean predicted vs observed proportion of deaths; error bars = 95% CI per bin) for LightGBM, Platt-recalibrated CHD-MMNet, APACHE IV and recalibrated APACHE IV in the hold-out hospitals (**A**) and

the 2015 temporal test set (**C**). Legend entries give the logistic calibration slope and the calibration-in-the-large intercept (logistic intercept with slope fixed at 1) with bootstrap 95% CI; the dashed grey line is perfect calibration. (**B**, **D**) Decision curves: net benefit across threshold probabilities 0.01–0.60 for the same models compared with treating all or no patients as high risk (hold-out **B**, 2015 **D**). CHD-MMNet was recalibrated by Platt scaling fitted on development out-of-fold predictions (hold-out) or train-2014 out-of-fold predictions (2015); APACHE IV was recalibrated by logistic recalibration fitted on development stays (intercept −0.862, slope 0.824). APACHE IV curves use APACHE-available stays; treat-all/treat-none reference lines use all stays.

Explainability and modality contributions

Among the 20 most influential LightGBM features, 15 are physiological or laboratory-aggregated indicators (**Figure 4A–B**). Age had the largest mean |SHAP| (0.274 log-odds), followed by a diagnosis-text component loading on respiratory failure, renal failure, shock and hypotension diagnoses (0.218). Other prominent features included the first-day Glasgow Coma Scale (0.111), last BUN (0.092), a second text component separating post-CABG from acute infarction (0.090), last heart rate (0.090), SpO_2 variability (0.089) and minimum white-cell count (0.085) (**Figure 4C**). By feature group, laboratory summaries contributed 43% of total |SHAP|, vital signs 24%, static variables 17%, text 13% and treatments 3%. For CHD-MMNet, Integrated Gradients assigned 49% of attribution to the time-series branch, 24% to static variables, 14% to treatments and 13% to text. Attribution showed a U-shaped over time, concentrating on the admission hour and on hours 22–24, and the highest-weighted variables were BUN, heart rate, respiratory rate, SpO_2, PaO_2, lactate, FiO_2 and white-cell count (**Figure 4D–E**). For blood gases and lactate, a large share of attribution fell on the time-since-measurement channel, indicating that the decision to measure carried prognostic information.

The ablation experiment confirmed the primacy of physiology (**Figure 5C; Supplementary Table S5**). Removing the hourly time series reduced CHD-MMNet AUROC from 0.884 to 0.848 in hold-out hospitals (Δ 0.036, $p = 5\times10^{-10}$) and from 0.875 to 0.845 in 2015; the static-only network reached 0.844, below APACHE IV. Removing the text branch (0.884, $P = 0.82$) or the treatment

branch (0.884, $P = 0.91$) did not change discrimination. Likewise, LightGBM without treatment features achieved the same AUROC as the full model (0.897), and LightGBM on static features alone 0.845 (0.824–0.866). A post-hoc average of LightGBM and CHD-MMNet predictions improved AUROC by only 0.001 (0.898), and the two models' predictions were highly correlated ($r = 0.90$).

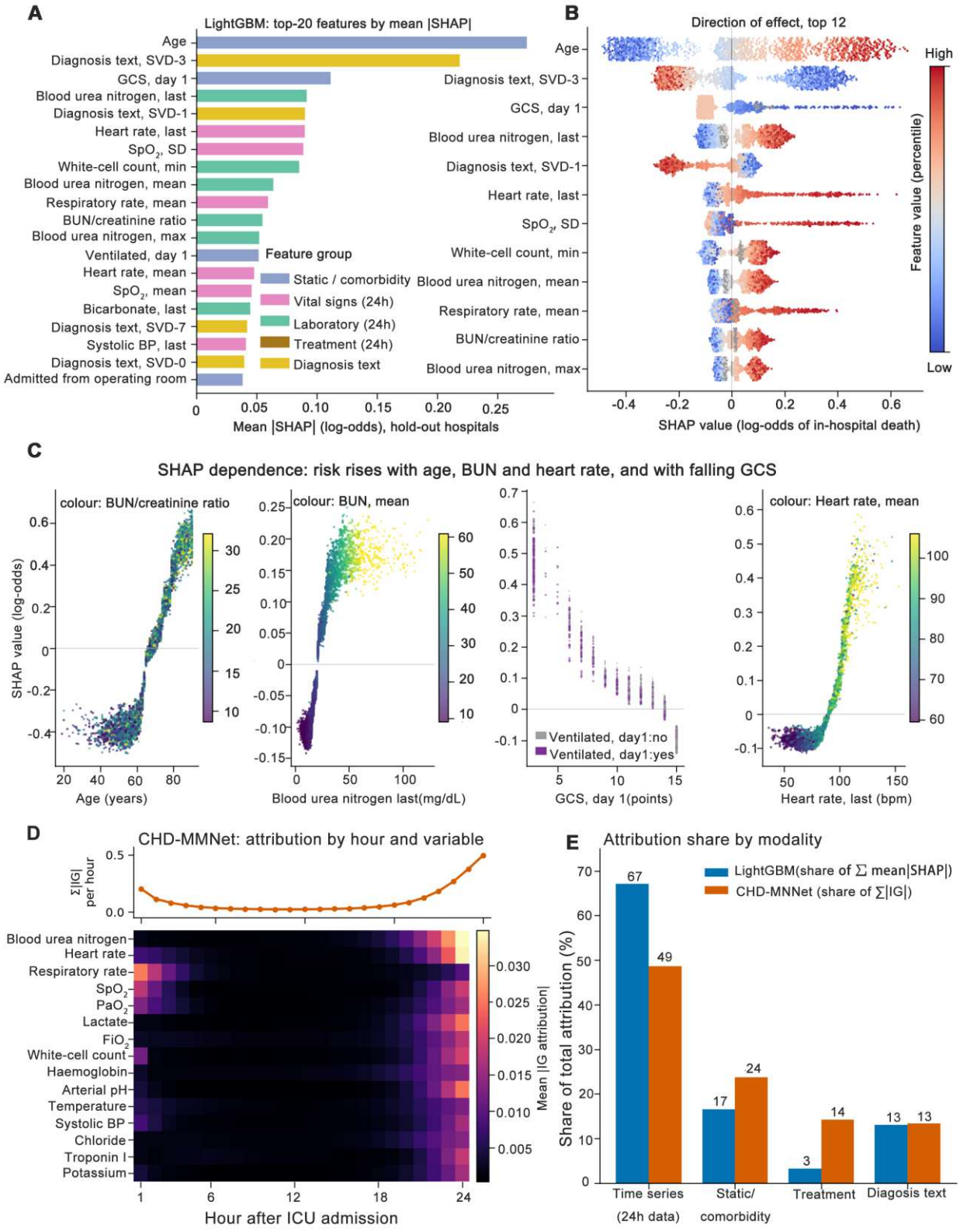

Figure 4. Explainability analysis identifies physiological and laboratory time-series data as the primary driver of predictions in both LightGBM and CHD-MMNet. (**A**) The 20 features with the largest mean absolute SHAP value (log-odds

units; TreeSHAP averaged over the five fold models), coloured by feature group. Text SVD components are truncated-SVD dimensions of the TF-IDF-encoded diagnosis text (SVD-3 loads negatively on respiratory/renal failure, shock and hypotension terms and positively on unstable angina; SVD-1 separates CABG from acute myocardial infarction). (**B**) SHAP values of the top-12 features for every hold-out stay (up to 1,500 randomly sampled points per feature), coloured by the within-feature percentile of the feature value (grey = missing). (**C**) SHAP dependence plots for age, last blood urea nitrogen (BUN; values above the 99.5th percentile omitted), day-1 Glasgow Coma Scale and last heart rate (30–180 bpm shown), coloured by the interaction feature identified from the SHAP values. (**D**) Mean absolute Integrated-Gradients attribution of CHD-MMNet by hour after ICU admission and time-series variable (15 variables with the largest total attribution; sum of value, mask and time-since-observation channels; 308 hold-out deaths and 308 randomly sampled survivors, ensemble of five fold models, 50 Riemann steps, all-zero baseline), with the total attribution per hour above. (**E**) Share of total attribution by input modality: LightGBM (share of the summed mean |SHAP| over features; 24-h vital-sign and laboratory summaries combined as "time series") vs CHD-MMNet (share of summed |IG| attribution). GCS, Glasgow Coma Scale; SpO_2, peripheral oxygen saturation.

Subgroups, fairness and between-hospital heterogeneity

LightGBM discrimination in hold-out hospitals was consistent across coronary presentations—STEMI 0.950 (0.922-0.973), NSTEMI 0.901 (0.868-0.931), post-CABG surgery 0.886 (0.819-0.945), CAD with other critical illness 0.844 (0.812-0.875)—and exceeded APACHE IV in each, with the largest gains in NSTEMI (+0.057), post-CABG (+0.096) and cardiogenic shock (0.868 vs 0.781) (**Table 3; Figure 5B**). There were only 9 deaths due to unstable angina pectoris and the estimated values were inaccurate (0.723, 0.508–0.912). The performance in females (0.899) was similar to that in males (0.895). The performance in patients <65 years old (0.928) was higher than that in patients ≥65 years (0.872), and higher in mechanically ventilated patients (0.932 vs 0.870). AUROCs were 0.896 in Caucasian, 0.894 in African-American, 0.942 in Hispanic and 0.827 in Asian patients (12 deaths). Except for the smaller Asian subgroup, the calibration-maximum values of all gender and ethnicity groups were close to zero (**Supplementary Table S8**). In the 10 hold-out hospitals with ≥10 deaths, LightGBM AUROC ranged from 0.842 to 0.966 (random-effects pooled 0.917, 0.888-0.946; I^2 = 67%) versus 0.874 (0.839–0.908) for APACHE IV，indicating superior pooled discrimination for LightGBM despite moderate between-hospital heterogeneity (**Figure 5A; Supplementary Table S6**).

448 **Table 3. Subgroup discrimination of LightGBM and APACHE IV in the hold-out hospitals.**

Subgroup	Level	n (deaths), all rows	LightGBM AUROC (95% CI)	n (deaths), APACHE-available	APACHE IV AUROC (95% CI)	ΔAUROC (LightGBM − APACHE IV)§	DeLong P§
Coronary presentation	STEMI	619 (33)	0.950 (0.922–0.973)	552 (28)	0.935 (0.899–0.966)	+0.016	0.072
Coronary presentation	NSTEMI	984 (90)	0.901 (0.868–0.931)	820 (67)	0.844 (0.794–0.887)	+0.055	0.005
Coronary presentation	Unstable angina	291 (9)	0.723 (0.508–0.912)	246 (7)	0.817 (0.621–0.960)	−0.014	0.752
Coronary presentation	Post-CABG surgery	1,306 (22)	0.886 (0.819–0.945)	1,071 (18)	0.790 (0.686–0.883)	+0.085	0.083
Coronary presentation	Stable CAD / other	1,453 (154)	0.844 (0.812–0.875)	1,285 (133)	0.824 (0.790–0.857)	+0.016	0.318
Age	<65 years	1,802 (63)	0.928 (0.888–0.959)	1,560 (49)	0.885 (0.824–0.933)	+0.050	0.027
Age	≥65 years	2,851 (245)	0.872 (0.851–0.894)	2,414 (204)	0.838 (0.808–0.862)	+0.032	0.003
Sex	Female	1,640 (130)	0.899 (0.874–0.923)	1,419 (109)	0.841 (0.798–0.880)	+0.048	0.001
Sex	Male	3,013 (178)	0.895 (0.872–0.920)	2,555 (144)	0.877 (0.851–0.900)	+0.023	0.025
Mechanical ventilation (24 h)	Yes	1,242 (125)	0.932 (0.910–0.953)	964 (101)	0.904 (0.871–0.929)	+0.025	0.009
Mechanical ventilation (24 h)	No	3,411 (183)	0.870 (0.844–0.896)	3,010 (152)	0.833 (0.802–0.864)	+0.040	0.002
Cardiogenic shock (24 h)	Yes	120 (40)	0.868 (0.801–0.927)	104 (35)	0.781 (0.692–0.865)	+0.083	0.025
Cardiogenic shock (24 h)	No	4,533 (268)	0.890 (0.869–0.908)	3,870 (218)	0.856 (0.830–0.880)	+0.032	<0.001
ICU type	Cardiac ICU (CCU/CTICU/CSICU)	2,302 (120)	0.894 (0.861–0.922)	1,872 (93)	0.847 (0.805–0.883)	+0.040	0.007
ICU type	Medical/Surgical /Neuro ICU	2,351 (188)	0.894 (0.872–0.913)	2,102 (160)	0.871 (0.845–0.895)	+0.028	0.013
Ethnicity	Caucasian	3,435 (223)	0.896 (0.875–0.915)	2,926 (181)	0.858 (0.831–0.885)	+0.039	<0.001
Ethnicity	African American	386 (26)	0.894 (0.829–0.949)	349 (24)	0.838 (0.756–0.909)	+0.060	0.009
Ethnicity	Hispanic	237 (25)	0.942 (0.895–0.975)	196 (19)	0.917 (0.864–0.967)	+0.021	0.541
Ethnicity	Asian	126 (12)	0.827 (0.667–0.948)	72 (9)	0.834 (0.728–0.924)	−0.042	0.481

Ethnicity	Other/Unknown	453 (20)	0.889 (0.809–0.952)	417 (18)	0.893 (0.806–0.959)	−0.003	0.846

AUROC with 1,000-resample bootstrap 95% CI. §ΔAUROC and DeLong P computed on the APACHE-available rows of each subgroup (paired comparison). Unstable angina (9 deaths) and Asian ethnicity (12 deaths) have very few events; estimates are imprecise.

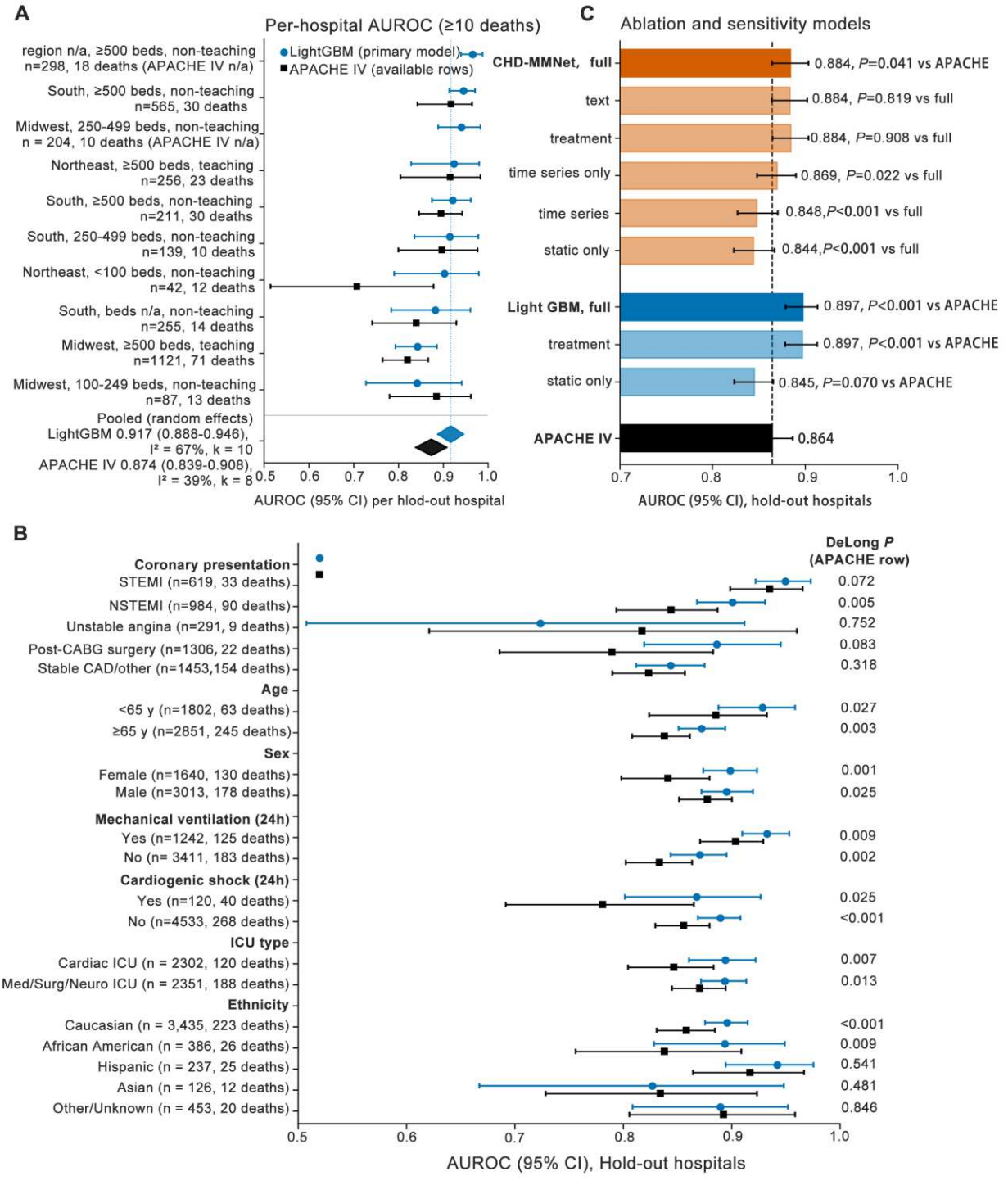

Figure 5. The primary model transports across hospitals and clinical subgroups, outperforms APACHE IV, and derives its predictive signal mainly from time-series data. (**A**) Hospital-level AUROC (95% CI) of LightGBM

(circles) and APACHE IV (squares) in the 10 hold-out hospitals with ≥10 in-hospital deaths, labelled by US region, bed-size category and teaching status (n, deaths). Two hospitals had <10 deaths among APACHE-available stays (APACHE IV n/a). Diamonds show the DerSimonian–Laird random-effects pooled AUROC: LightGBM 0.917 (0.888–0.946), I^2 = 67%, k = 10; APACHE IV 0.874 (0.839–0.908), I^2 = 39%, k = 8. Dotted vertical line = pooled LightGBM estimate. (**B**) Subgroup AUROC (bootstrap 95% CI) of LightGBM (all stays) and APACHE IV (APACHE-available stays) by coronary presentation, age, sex, mechanical ventilation and cardiogenic shock in the first 24 h, ICU type and ethnicity; the right-hand column gives the DeLong P value for the paired comparison on APACHE-available stays. (**C**) AUROC (95% CI) of CHD-MMNet input ablations (no text, no treatment, time series only, no time series, static only) and of LightGBM sensitivity models (retrained without the 36 treatment features, or on the 99 static features only) compared with the full models and with APACHE IV (dashed line, 0.864). Annotations give point estimates and DeLong P values versus the full CHD-MMNet (ablations, all stays) or versus APACHE IV (full models and LightGBM sensitivity models, APACHE-available stays).

External validation in MIMIC-IV (2008–2022)

To assess whether the reduced model generalized beyond eICU, we performed external validation in MIMIC-IV, an independent ICU database that differs in documentation practices, patient population and case-mix, thereby providing a stringent test of model transportability. The MIMIC-IV cohort comprised 24,129 ICU stays (median age 72 years; 66.1% male) of whom 2,785 (11.5%) died in hospital — a considerably sicker case-mix than eICU, with more heart failure (46.0% vs 12.8%), acute kidney injury (36.5% vs 10.2%), cardiogenic shock (7.7% vs 2.6%), mechanical ventilation (44.0% vs 33.3%) and a lower proportion of classic ACS presentations (STEMI 5.6%, NSTEMI 11.3%, unstable angina 0.6%; **Table 4**).

Table 4. External validation in MIMIC-IV v3.1.

(A) Cohort characteristics compared with the eICU development and hold-out-hospital sets. Continuous variables are median (IQR); binary variables n (%).

Characteristic	eICU development (146 hospitals)	eICU hold-out (42 hospitals)	MIMIC-IV (external)
No. of ICU stays	12,933	4,653	24,129
Age, years	68.0 (59.0–77.0)	68.0 (59.0–77.0)	72.0 (64.0–80.0)
Body-mass index, kg/m²	28.6 (24.8–33.3)	28.4 (24.7–32.9)	28.1 (24.8–32.1)
Glasgow Coma Scale, day 1	15.0 (13.0–15.0)	15.0 (15.0–15.0)	10.0 (3.0–15.0)
ICU length of stay, days	2.1 (1.5–3.9)	2.2 (1.6–3.9)	2.3 (1.5–4.2)
Heart rate, mean 24 h	79.9 (70.1–90.1)	79.5 (70.2–89.6)	81.1 (72.8–90.7)

Systolic BP, minimum 24 h	85.0 (74.0–99.0)	86.0 (75.2–100.0)	89.0 (80.0–97.0)
SpO_2, minimum 24 h	90.0 (85.0–93.0)	91.0 (86.0–94.0)	92.0 (90.0–94.0)
Lactate, maximum 24 h	2.1 (1.3–3.5)	2.0 (1.3–3.6)	2.2 (1.6–3.2)
Blood urea nitrogen, maximum 24 h	19.0 (14.0–31.0)	20.0 (15.0–31.0)	23.0 (16.0–39.0)
Creatinine, maximum 24 h	1.1 (0.8–1.6)	1.1 (0.9–1.6)	1.1 (0.9–1.8)
White-cell count, maximum 24 h	11.9 (8.9–16.0)	12.2 (9.0–16.2)	12.7 (9.1–17.3)
Haemoglobin, minimum 24 h	10.9 (9.2–12.7)	10.9 (9.1–12.7)	9.5 (8.0–11.1]
Male sex	8425 (65.1)	3013 (64.8)	15939 (66.1)
STEMI	1985 (15.3)	619 (13.3)	1363 (5.6)
NSTEMI	2851 (22.0)	984 (21.1)	2730 (11.3)
Unstable angina	960 (7.4)	291 (6.3)	140 (0.6)
Post-CABG surgery	3016 (23.3)	1306 (28.1)	6133 (25.4)
Stable CAD / other	4121 (31.9)	1453 (31.2)	13763 (57.0)
Cardiogenic shock (24 h)	339 (2.6)	120 (2.6)	1858 (7.7)
Cardiac arrest	552 (4.3)	185 (4.0)	938 (3.9)
Heart failure	1661 (12.8)	447 (9.6)	11089 (46.0)
Sepsis	1221 (9.4)	377 (8.1)	4040 (16.7)
Acute kidney injury	1325 (10.2)	439 (9.4)	8816 (36.5)
Respiratory failure	2274 (17.6)	818 (17.6)	6425 (26.6)
Cardiac ICU	6130 (47.4)	2302 (49.5)	13135 (54.4)
Invasive mechanical ventilation	4305 (33.3)	1242 (26.7)	10614 (44.0)
Vasopressor	4156 (32.1)	1060 (22.8)	9684 (40.1)
Coronary angiography/PCI	2613 (20.2)	773 (16.6)	4562 (18.9)
Renal replacement therapy	380 (2.9)	79 (1.7)	622 (2.6)
Mechanical circulatory support	511 (4.0)	186 (4.0)	978 (4.1)
In-hospital death	1007 (7.8)	308 (6.6)	2785 (11.5)
ICU death	629 (4.9)	156 (3.4)	1853 (7.7)

(B) Performance of the frozen portable model and of the pre-specified updating hierarchy in MIMIC-IV (n = 24,129, 2,785 deaths). Values are estimate (95% bootstrap CI); ICI is a point estimate.

Model	AUROC	AUPRC	Brier	Calibration slope	Calibration intercept	ICI	O:E ratio
LightGBM-P, frozen (no updating)	0.871 (0.864–0.877)	0.507 (0.487–0.527)	0.076 (0.074–0.078)	1.202 (1.163–1.243)	-0.097 (-0.142–-0.056)	0.014	0.935 (0.905–0.962)
LightGBM-P + intercept recalibration	0.871 (0.865–0.877)	0.507 (0.489–0.528)	0.076 (0.074–0.078)	1.202 (1.167–1.245)	-0.000 (-0.044–0.040)	0.009	1.000 (0.970–1.028)
LightGBM-P + intercept & slope recalibration	0.871 (0.865–0.877)	0.507 (0.488–0.526)	0.076 (0.074–0.078)	0.999 (0.968–1.034)	-0.000 (-0.049–0.045)	0.013	1.000 (0.969–1.029)
LightGBM retrained in MIMIC-IV (5-fold CV)	0.911 (0.906–0.916)	0.591 (0.571–0.608)	0.068 (0.066–0.070)	1.145 (1.110–1.184)	0.093 (0.047–0.137)	0.016	1.060 (1.030–1.088)
Logistic regression-P, frozen	0.843 (0.836–0.850)	0.442 (0.422–0.462)	0.091 (0.089–0.093)	0.701 (0.676–0.728)	-0.695 (-0.750–-0.641)	0.057	0.682 (0.661–0.704)

(C) Comparison with standard intensive-care severity scores computed on the same MIMIC-IV patients and window. Negative deltas indicate worse discrimination than the frozen model.

Model / score	AUROC (95% CI)	AUPRC	ΔAUROC vs frozen model (DeLong p)
LightGBM-P, frozen (no updating)	0.871 (0.864–0.878)	0.507	reference
LightGBM-P, intercept+slope recalibrated	0.871 (0.865–0.877)	0.507	-0.000 (p = 0.0198)
LightGBM retrained in MIMIC-IV (5-fold CV)	0.911 (0.906–0.916)	0.591	+0.040 (p = 1.88e-84)
OASIS (calibrated probability)	0.688 (0.677–0.699)	0.262	-0.183 (p = 8.95e-298)
Day-1 SOFA (score)	0.658 (0.648–0.669)	0.244	-0.213 (p = 0)
OASIS (score)	0.688 (0.677–0.698)	0.262	-0.183 (p = 8.95e-298)

Without any updates, the frozen LightGBM-P achieved an AUROC of 0.871 (95% CI 0.864-0.877) and AUPRC of 0.507 (0.487-0.527) in MIMIC-IV (**Figure 6B**, **Table 4**) — a loss of only 0.026 AUROC relative to the eICU hold-out hospitals, at the favourable end of the 0.02–0.06 loss anticipated in the addendum and far smaller than the 0.05-0.14 typically reported for cross-institution transfer. Calibration transported remarkably well despite the 1.5-fold higher event rate: calibration intercept −0.10 (−0.14 to −0.06), slope 1.20 (1.16–1.24), ICI 0.014, observed-to-expected ratio 0.94 (0.91-0.96) (**Figure 6C**). Intercept-only recalibration removed the residual offset (intercept 0.00, ICI 0.009) and intercept-and-slope recalibration additionally corrected the slope (0.999, 0.968-1.034) without changing discrimination. A LightGBM retrained inside MIMIC-IV reached an AUROC of 0.911 (0.906-0.916), thus the frozen model recovered 90% of the locally attainable discrimination above chance. The frozen portable logistic regression performed distinctly worse (0.843, 0.836-0.850) and the calibration was severely inaccurate (intercept −0.70, slope 0.70; ICI 0.057). The net benefit of the frozen model exceeded treat-all and treat-none strategies across thresholds of 1–60% and lay close to the locally retrained model (**Figure 6E**).

When compared with the clinical control indicators calculated for the same batch of patients, the frozen model discriminated far better than either standard severity score: day-1 SOFA reached an AUROC of 0.658 (0.648-0.669) and OASIS 0.688 (0.677-0.699), versus 0.871 for the model (ΔAUROC +0.213 and +0.183 respectively, both $p < 10^{-15}$; Table 4C, Figure 7C). The discriminatory ability showed no decay across calendar time — AUROC 0.857 (0.845-0.869) in 2008–2010, 0.873 in 2011-2013, 0.873 in 2014-2016, 0.881 in 2017–2019 and 0.886 (0.870-0.903) in 2020–2022 (**Figure 6A**) — and after intercept-and-slope recalibration the slope was 0.96–1.04 in every stratum. The pre-specified criterion for contemporary usability (2020–2022 AUROC

≥0.85 with recalibrated slope 0.8-1.2) was therefore met (0.886; slope 1.04). The performance was highest in patients admitted to a cardiac ICU (0.918, 0.910-0.925) and lower in other ICUs (0.812, 0.802-0.822), highest in STEMI (0.914), unstable angina (0.956) and post-CABG patients (0.944) and lowest in cardiogenic shock (0.763, 0.738-0.786), where the frozen model also under-predicted risk (calibration intercept +0.27, observed-to-expected ratio 1.16) (**Figure 6D**). AUROC ranged from 0.849 to 0.893 across ethnic groups and was 0.855 in women versus 0.879 in men. Overall, the frozen model outperformed both SOFA and OASIS, remained stable over 15 years, met the pre-specified usability criterion, and was weakest in cardiogenic shock, with the lowest discrimination and under-predicted risk.

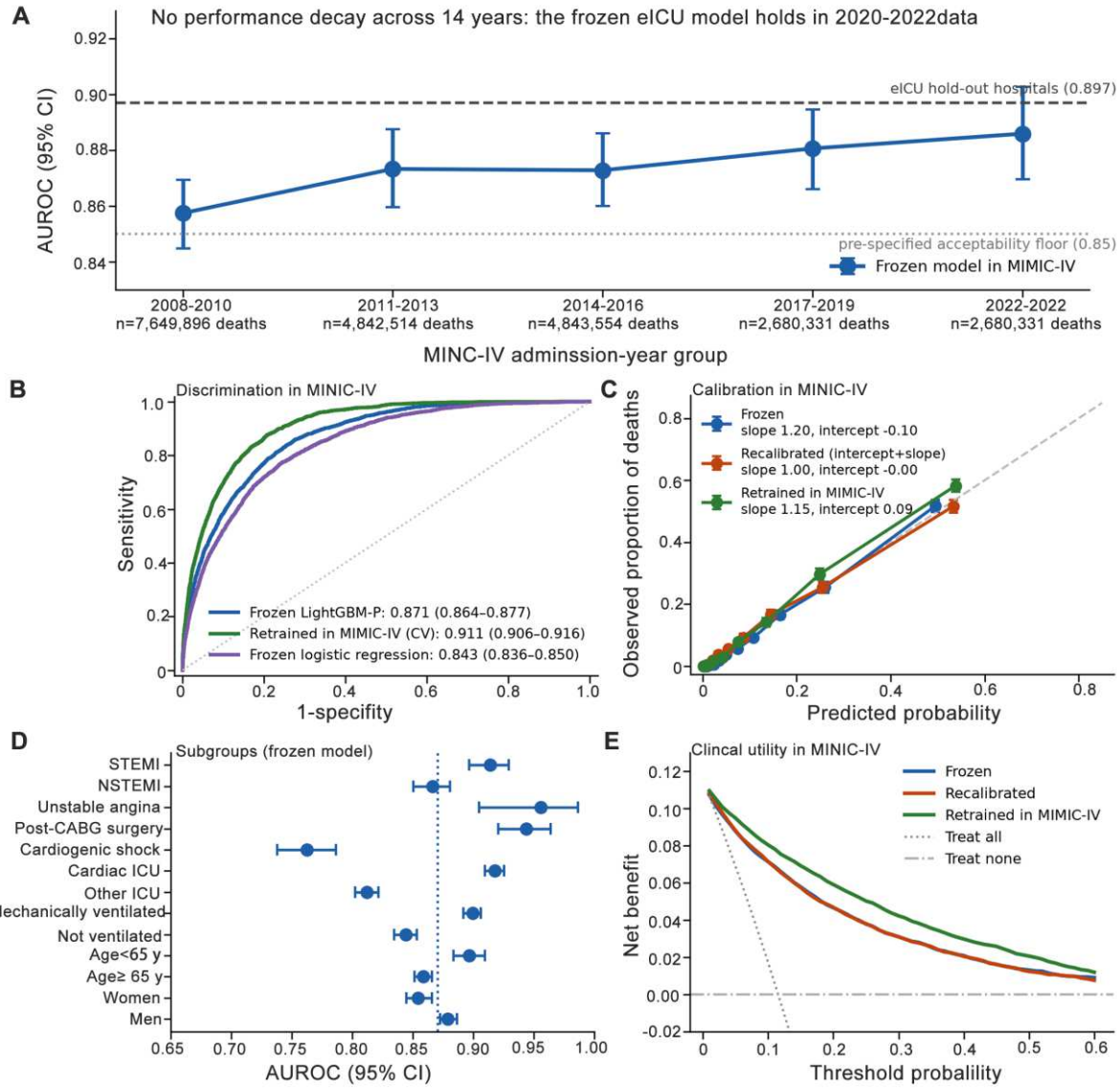

Figure 6. External validation confirms that the frozen portable model transports to MIMIC-IV with stable performance over 14 years and across subgroups (24,129 CAD ICU stays, 2,785 in-hospital deaths, 2008–2022). (**A**) AUROC with 95% bootstrap CI by admission-year group for LightGBM-P applied without updating; the dashed line is the eICU hold-out-hospital estimate (0.897) and the dotted line the pre-specified acceptability floor (0.85). (B) ROC curves for the frozen portable model, the same architecture retrained inside MIMIC-IV (5-fold cross-validation, locally attainable upper bound) and the frozen portable logistic regression. (**C**) Calibration in ten quantile bins (95% CI per bin) for the frozen, recalibrated (intercept and slope) and locally retrained models, with calibration slope and calibration-in-the-large intercept. (**D**) Subgroup AUROC of the frozen model; the dotted line marks the overall estimate (0.871). (**E**) Decision curves relative to treat-all and treat-none strategies. Recalibration and local retraining were cross-fitted across five folds.

Reduced model and risk tiers

A 25-variable model met the pre-specified criterion (development out-of-fold AUROC 0.875, 0.009 below the full model) and comprised age, first-day Glasgow Coma Scale, blood urea nitrogen (last, mean, maximum, minimum, and the BUN/creatinine ratio), heart rate (last, mean, maximum), oxygen-saturation variability and mean, respiratory rate, white-cell count, lactate, bicarbonate, systolic blood pressure, red-cell distribution width, partial-thromboplastin-time variability, glucose, day-1 ventilation and admission from the operating room (**Supplementary Table S15; Figure 7D**). In hold-out hospitals, it achieved an AUROC of 0.882 versus 0.897 for the full model, with a larger relative loss in AUPRC (0.406 vs 0.461) and a Brier score of 0.048. Therefore, this reduction is limited in terms of discrimination, but its accuracy in high-risk areas cannot be ignored. To assess the clinical implications of this trade-off, we next examined tier-level risk stratification.

The prediction results were divided into four clinical tiers (**Table 5, Figure 7A-B**). The model placed 58.1% of hold-out patients in a low-risk tier with an observed mortality of 0.7%. While the 6.5% classified as very-high risk had an observed mortality of 48.5% and accounted for 47.4% of all deaths. Tier-level observed risks tracked predicted risks closely in the external cohort after recalibration (0.3%, 3.7%, 15.8% and 43.0% observed versus 1.4%, 5.4%, 16.1% and 42.0% predicted), among which the very-high-risk tier comprised 15.7% of patients and contained 58.6% of deaths. Overall, the reduced model achieved effective risk stratification, concentrating nearly half of all deaths into a small very-high-risk tier while identifying a large low-risk population with an extremely low mortality rate.

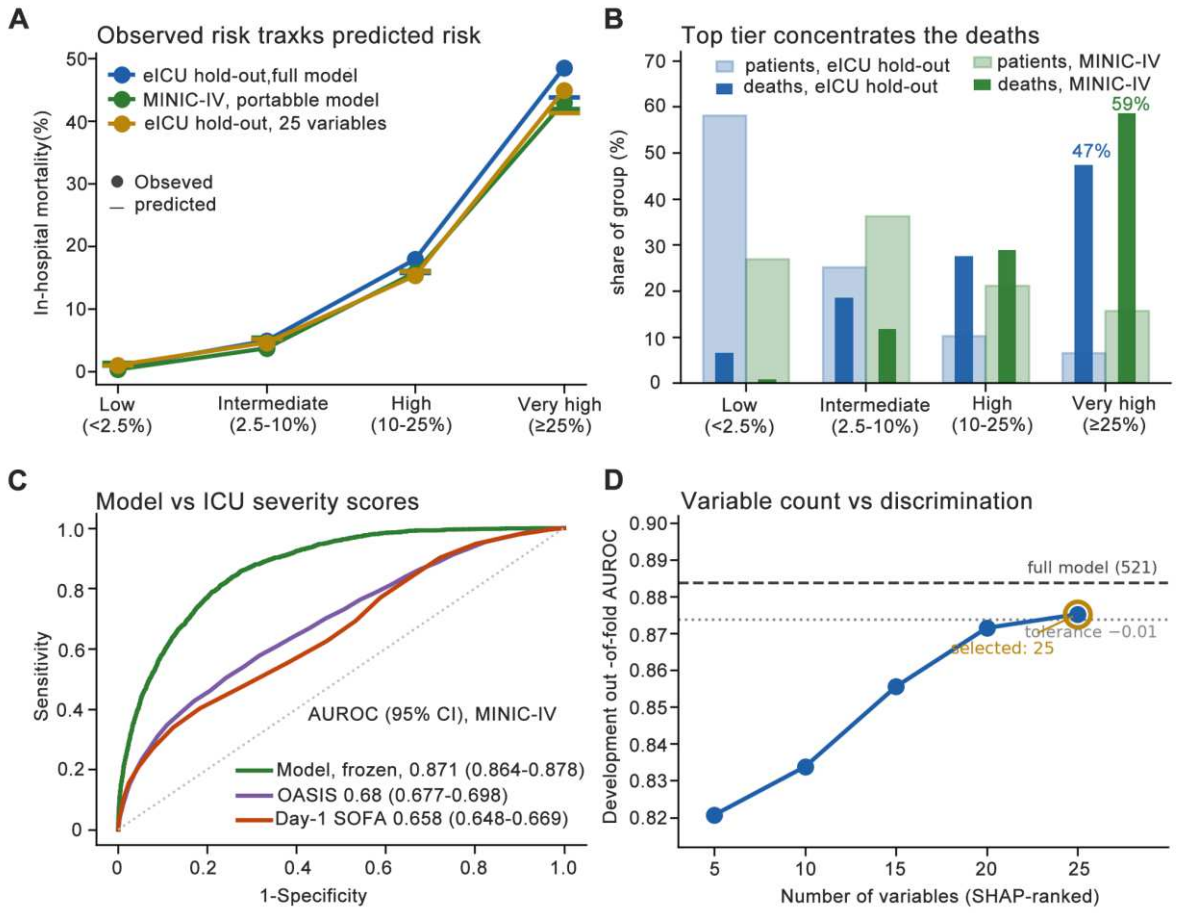

Figure 7. The 25-variable reduced model preserves clinical utility and risk stratification, with discrimination superior to day-1 SOFA and OASIS. (**A**) Observed in-hospital mortality (filled circles) and mean predicted risk (horizontal dashes) within four pre-defined risk tiers, for the full model in eICU hold-out hospitals, the frozen portable model in MIMIC-IV after recalibration, and the 25-variable reduced model in hold-out hospitals. (**B**) Share of patients (pale bars) and share of all deaths (solid bars) falling in each tier, for eICU hold-out hospitals (blue) and MIMIC-IV (green); the value above the rightmost bars is the proportion of deaths in the very-high-risk tier. (**C**) Receiver-operating-characteristic curves in MIMIC-IV for the frozen model and for day-1 SOFA and OASIS computed on the same patients and time window. (**D**) Development out-of-fold AUROC of reduced models against the number of SHAP-ranked variables retained; dashed line, full 521-feature model; dotted line, the pre-specified 0.01 tolerance; circled point, the selected 25-variable model.

Table 5. Clinical risk tiers. Observed mortality, share of the cohort and share of deaths within pre-defined tiers of predicted risk, in the eICU development (out-of-fold) and hold-out-hospital cohorts and in the MIMIC-IV external cohort after recalibration.

cohort	model	tier	n	deaths	observed_pct	mean_predicted_pct	share_of_cohort_pct	share_of_deaths_pct
eICU development (out-of-fold)	LightGBM (521 features)	Low (<2.5%)	7054	50	0.7	0.9	54.5	5.0
eICU development (out-of-fold)	LightGBM (521 features)	Intermediate (2.5–10%)	3283	214	6.5	5.2	25.4	21.3
eICU development (out-of-fold)	LightGBM (521 features)	High (10–25%)	1580	297	18.8	16.0	12.2	29.5
eICU development (out-of-fold)	LightGBM (521 features)	Very high (≥25%)	1016	446	43.9	42.5	7.9	44.3
eICU hold-out hospitals	LightGBM (521 features)	Low (<2.5%)	2705	20	0.7	1.0	58.1	6.5
eICU hold-out hospitals	LightGBM (521 features)	Intermediate (2.5–10%)	1171	57	4.9	5.1	25.2	18.5
eICU hold-out hospitals	LightGBM (521 features)	High (10–25%)	476	85	17.9	15.8	10.2	27.6
eICU hold-out hospitals	LightGBM (521 features)	Very high (≥25%)	301	146	48.5	43.8	6.5	47.4
eICU hold-out hospitals	Parsimonious (25 features)	Low (<2.5%)	2563	25	1.0	1.0	55.1	8.1
eICU hold-out hospitals	Parsimonious (25 features)	Intermediate (2.5–10%)	1238	57	4.6	5.2	26.6	18.5
eICU hold-out hospitals	Parsimonious (25 features)	High (10–25%)	529	81	15.3	16.0	11.4	26.3
eICU hold-out hospitals	Parsimonious (25 features)	Very high (≥25%)	323	145	44.9	41.4	6.9	47.1
MIMIC-IV external (recalibrated)	LightGBM-P (441 features)	Low (<2.5%)	6506	21	0.3	1.4	27.0	0.8
MIMIC-IV external (recalibrated)	LightGBM-P (441 features)	Intermediate (2.5–10%)	8723	326	3.7	5.4	36.2	11.7
MIMIC-IV external (recalibrated)	LightGBM-P (441 features)	High (10–25%)	5108	806	15.8	16.1	21.2	28.9
MIMIC-IV external (recalibrated)	LightGBM-P (441 features)	Very high (≥25%)	3792	1632	43.0	42.0	15.7	58.6

Exploratory language-model paradigm

Fine-tuning a clinical language model on serialized patient descriptions did not improve on gradient boosting (**Supplementary Table S13; Supplementary Figure S6**). In hold-out hospitals, the language model alone reached an AUROC of 0.860 (0.839-0.881) and AUPRC of 0.365 (0.313-0.421) versus 0.897 and 0.461 for LightGBM (ΔAUROC −0.038, $P = 9\times10^{-8}$). Moreover, the calibration of its predicted probability is poor (with a slope of 0.70 and an intercept of -1.88). Gradient boosting applied to the 768-dimensional fused embedding alone performed similarly (0.856, 0.833-0.876), indicating that the embedding is a lossy re-encoding of information already present in the tabular features rather than a source of new signal. Adding the embedding to the tabular features (LM-boost) did not change discrimination (0.893, 0.874–0.910; ΔAUROC −0.005,

$P = 0.11$; AUPRC 0.455) and probability-level fusion was slightly worse (0.887; $P = 0.005$). The same ordering held in development out-of-fold estimates and in temporal validation (LM alone 0.838, LM-boost 0.882, LightGBM 0.886). The learned layer weights were nearly uniform (0.076-0.079 across the 13 layers), consistent with no particular depth of linguistic representation being informative for this task.

Discussion

In 17,586 critically ill patients with coronary artery disease from 188 US hospitals, multimodal machine-learning models trained on the first 24 h of routinely collected ICU data predicted the first 24 hours predicted in-hospital mortality. The AUROC was approximately 0.90 in 42 hospitals that never seen during development and of about 0.89 in a subsequent calendar year. It was 0.03 AUROC higher than APACHE IV, and had significantly better calibration and higher net benefits within the range of clinically relevant decision thresholds. When the model was frozen and applied to 24,129 CAD ICU stays from an unrelated academic health system spanning 2008–2022 (MIMIC-IV), the model retained an AUROC of 0.871 with near-preserved calibration and showed no performance decay in the most recent years. Performance held across STEMI, NSTEMI, post-surgical and mixed critical-illness presentations, across sexes and ethnic groups, and in every hold-out hospital with sufficient events. Gradient boosting on engineered physiological summaries matched or exceeded a purpose-built deep sequence network, and ablation experiments identified the hourly physiological trajectory-not diagnosis text or treatment exposures-as the modality that carried the incremental information.

The discrimination observed for APACHE IV (0.85-0.86) is consistent with its reported performance in contemporary cardiac ICUs (0.82–0.86)(14, 16) and lower than in its derivation

cohort (0.88)(11); its systematic over-prediction (observed-to-expected ratio 0.62) is expected for a score derived in 2002-2003 and illustrates the calibration drift that limits legacy severity scores (13, 69). Cardiology risk scores achieve c-statistics of 0.78–0.84 for in-hospital mortality in ACS registries (7-9) and cardiac-ICU-specific tools 0.79-0.86 (17-19). Our models exceeded these values while addressing a broader and sicker population that includes CAD patients admitted for non-cardiac critical illness, in whom mortality (12.8%) was highest and for whom neither GRACE nor TIMI is applicable. Prior machine-learning investigations into mortality from infarction have reported AUROCs of 0.87–0.91, mostly in single institutions or single databases (37, 70-72), and the one CAD model validated across databases lost 0.09 AUROC in eICU(38). In contrast, our hold-out-hospital AUROC was marginally higher than the cross-validated development estimate, cross-database transfer to MIMIC-IV cost only 0.026 AUROC (versus 0.09 for the comparable prior model (38). Additionally, per-hospital AUROCs remained above 0.84, in line with the observation that a hospital-grouped development strategy and physiologically anchored features reduce dependence on site-specific practice patterns (47, 73, 74).

Contrary to our expectation that a recurrent network operating on the raw hourly tensor would extract more from the trajectories, CHD-MMNet was outperformed by LightGBM by 0.013 AUROC in hold-out hospitals and by 0.011 in temporal validation, with no added benefit from combining the two. Several factors likely contributed: with about 1000 development events, a network of 0.7 million parameters is data-limited and converged within 3-7 epochs; hourly aggregation and eight summary statistics per variable already capture most of the prognostically relevant trajectory information over a 24-h horizon; and gradient boosting handles heterogeneous, heavily missing tabular data natively. Our finding agrees with systematic reviews showing no consistent advantage of complex algorithms over well-specified simpler models in low-risk-of-bias

comparisons(36), with registry-scale ACS analyses in which boosting improved on regression by 0.01(37), and with benchmark studies in which the gain of sequence models over gradient boosting on 24-h windows is small (75, 76). It does not exclude a benefit of deep sequence models for longer or continuously updated horizons(26, 27, 77), but for a fixed 24-h prediction the simpler model is preferable for transparency, calibration and ease of deployment (78).

Encoding structured variables as text and fine-tuning a clinical transformer - a paradigm reported to beat tabular baselines for intensive-care mortality(54) - did not help here, either alone or fused with gradient boosting, and the embedding-only model performed no better than the language model itself. Two features of our setting probably explain the difference from published reports: the comparator is a Bayesian-optimised gradient-boosting model over 521 engineered trajectory features rather than a lightly tuned baseline, and our serialization necessarily contains no information beyond those features, given that neither database provided narrative notes. The experiment therefore strengthens rather than weakens the study's main message: with roughly 1,000 development events and no free text, additional model capacity — recurrent, transformer or language-model based - does not extract more prognostic signal from the first ICU day. Whether a language model adds value where genuine narrative documentation exists, or where transfer from large-scale pretraining on event sequences can substitute for scarce events, remains an open question. This gap is best addressed with note-containing datasets and self-supervised pretraining rather than with larger supervised models.

Both explanation methods converged on a coherent clinical picture: age, depressed consciousness, renal dysfunction (BUN, creatinine, BUN/creatinine ratio), tachycardia and haemodynamic instability at the end of the first day, hypoxaemia and its variability, leucocytosis and metabolic acidosis. Detailed patient-specific analyses for two cases are shown in **Supplementary Figure S4**

and attribution maps for deaths versus survivors in **Supplementary Figure S5**. The highest time-sensitive attributions were noted during the admission hour and the final hours before prediction, indicating that the model weighs both initial severity and the response to 24 h of treatment. The diagnosis-text components ranked highly in SHAP yet removing text from the deep network cost nothing, and treatment indicators contributed little in either model once physiology was included. This suggests redundancy rather than irrelevance, as the text encodes shock, respiratory and renal failure diagnoses that are already expressed in the vital signs and laboratory values, and vasopressor or ventilator use is a downstream marker of the same physiology. Importantly, troponin concentration—the cornerstone of ACS diagnosis—carried almost no prognostic information in this ICU population, whereas the fact that blood gases or lactate were measured did, reflecting informative missingness that models must handle explicitly(46, 79).

In the external cohort, where APACHE IV is unavailable, the model out-discriminated day-1 SOFA and OASIS by 0.21 and 0.18 AUROC. Both scores performed below their usual range in general intensive-care populations, which is expected in a cardiac cohort restricted to patients alive at 24 h and reflects their design for case-mix adjustment rather than individual prognosis; the comparison nevertheless shows that the information the model adds is not available from the severity scores a clinician would otherwise consult.

For deployment, a simplified model with 25 variables achieved an AUROC of 0.882 in unseen hospitals, losing 0.015 AUROC and 0.055 AUPRC relative to the full model — an acceptable trade when features must be assembled explicitly, but an argument for the full model wherever the pipeline can run automatically. The risk tiers make the operating characteristics concrete: a low-risk tier containing more than half of patients with under 1% mortality, and a very-high-risk tier of about one in fifteen patients carrying nearly half of all deaths.

A model that runs automatically on data already in the electronic record at 24 h could stratify critically ill CAD patients into risk tiers, aiding in monitoring intensity, early cardiology–intensive care co-management, timing of invasive or mechanical-support decisions and family communication. At the Youden threshold, the primary model identified 86% of eventual deaths while flagging 19% of survivors, and its negative predictive value was 0.99. Decision-curve analysis indicated a clinically meaningful net benefit over both APACHE IV and default strategies across thresholds of 2–60%, the range in which such stratification decisions are plausibly made(62). Since calibration was preserved in unseen hospitals, the risk estimates can be interpreted as probabilities, which is a prerequisite for shared decision-making. Nevertheless, local calibration monitoring is advisable before deployment (35, 59). Crucially, whether acting on these predictions improves outcomes requires prospective, ideally randomized, evaluation.

Key strengths of this study include a pre-specified analysis plan, the largest multicentre CAD-specific ICU cohort analyzed to date, hospital-grouped cross-validation with genuine leave-hospital-out and temporal validation, benchmarking against the standard severity score on the same patients, comprehensive calibration and decision-analytic assessment, subgroup and fairness reporting, and open, documented code. However, there are notable limitations. Firstly, the development data (eICU-CRD, 2014–2015) predate current practice; we addressed this directly by validating the frozen model in MIMIC-IV up to 2022, where discrimination was stable and, if anything, slightly higher in 2020–2022, but MIMIC-IV is a single academic centre and generalizability to non-US health systems and to the contemporary era of routine mechanical circulatory support remains untested (33, 80). Observed between-site heterogeneity (I^2=67%) also requires ongoing local performance monitoring after clinical deployment. Sparse recording of key physiological variables and the absence of structured echocardiographic, angiographic and ECG

687 data across both databases also limit the full feature completeness of the model. Second, our CAD
688 cohort was defined via structured diagnostic codes rather than adjudicated angiography, and spans
689 a broad spectrum of acuity from elective post-CABG cases to patients with septic shock. Though
690 prespecified subgroup analyses were performed, the model underperformed in cardiogenic shock
691 (external AUC 0.763, 95%CI 0.738–0.786) with systematic risk under-prediction, indicating
692 dedicated shock-specific stratification is needed. Restriction to ICU stays ≥24 h excluded 6,276
693 admissions accounting for 22.3% of eligible in-hospital deaths, so the model only applies to
694 patients remaining in care at 24 h. Third, several high-impact clinical variables were sparsely
695 recorded, structured imaging and ECG data were unavailable across databases, and MIMIC-IV
696 phenotypes rely on whole-admission billing codes that may include post-prediction information.
697 Post-hoc testing found no performance gain from a pretrained clinical language model on
698 structured diagnosis strings, but free-text clinical notes were not accessible, leaving the value of
699 narrative data untested. All post-hoc model explanations reflect associative patterns only, and do
700 not represent causal clinical effects (66).

701 ## Conclusions

702 The multimodal machine learning model using the routine record data of the first 24 hours in the
703 ICU is more accurate and better calibrated in predicting in-hospital deaths of critically ill patients
704 with coronary heart disease than APACHE IV. Its performance can be transferred to unseen
705 hospitals, older years, and across 14 calendar years to different medical systems without any
706 updates. The key advantage lies in the hourly physiological and laboratory data trends, with
707 gradient boosting on these engineered features performing at least as effectively as a deep
708 multimodal network. Most notably, a gradient boosting model built on rigorously engineered
709 trajectory features delivers equivalent performance to a far more resource-intensive deep

multimodal network. These models warrant prospective evaluation as decision-support tools in both cardiac and general intensive care.

Ethics approval and consent to participate

This study was conducted in accordance with the ethical principles of the Declaration of Helsinki and the applicable PhysioNet data use agreements. The study involved a retrospective secondary analysis of de-identified data from the eICU Collaborative Research Database and MIMIC-IV. And the requirement for individual patient consent was waived. The authors completed the required training and signed the PhysioNet data-use agreement.

Consent for publication

Not applicable.

Availability of data and materials

eICU-CRD v2.0 is available to credentialed researchers at https://physionet.org/content/eicu-crd/2.0/. Derived cohort definitions, feature-extraction code, model code, trained model weights and result tables are provided in the Supplementary Appendix and will be deposited in a public repository upon acceptance. All code (cohort extraction, feature engineering, model training, evaluation and figures) is available with the Supplementary Appendix. Large-language-model-based tools were used to assist with code generation and manuscript drafting under author supervision; all analyses, results and interpretations were verified by the authors.

Competing interests

The authors declare no competing interests.

Funding

This work was supported by the National Natural Science Foundation of China (Grant No. 82303234).

Acknowledgements

Not applicable.

Author contributions

Conceptualization, B.C., X.F.Z. and Y.H.Z.; methodology, T.L.W. and G.C.; writing—original draft preparation, B.C. and Z.H.K.; writing—review and editing, B.C. and X.F.Z.; data curation, T.L.W. and G.C.; formal analysis, H.X.; interpretation of the results, H.X. and Y.H.Z.; supervision, Y.H.Z. All authors have read and approved the final version of the manuscript.

References

1. Roth GA, Mensah GA, Johnson CO, Addolorato G, Ammirati E, Baddour LM, et al. Global Burden of Cardiovascular Diseases and Risk Factors, 1990–2019. Journal of the American College of Cardiology. 2020;76(25):2982-3021.
2. Roth GA, Mensah GA, Fuster V. The Global Burden of Cardiovascular Diseases and Risks. Journal of the American College of Cardiology. 2020;76(25):2980-1.
3. Bohula EA, Katz JN, van Diepen S, Alviar CL, Baird-Zars VM, Park J-G, et al. Demographics, Care Patterns, and Outcomes of Patients Admitted to Cardiac Intensive Care Units. JAMA Cardiology. 2019;4(9).
4. Jentzer JC, van Diepen S, Barsness GW, Henry TD, Menon V, Rihal CS, et al. Cardiogenic Shock Classification to Predict Mortality in the Cardiac Intensive Care Unit. Journal of the American College of Cardiology. 2019;74(17):2117-28.
5. Byrne RA, Rossello X, Coughlan JJ, Barbato E, Berry C, Chieffo A, et al. 2023 ESC Guidelines for the management of acute coronary syndromes. European Heart Journal. 2023;44(38):3720-826.
6. Rao SV, O'Donoghue ML, Ruel M, Rab T, Tamis-Holland JE, Alexander JH, et al. 2025 ACC/AHA/ACEP/NAEMSP/SCAI Guideline for the Management of Patients With Acute Coronary Syndromes: A Report of the American College of Cardiology/American Heart Association Joint Committee on Clinical Practice Guidelines. Circulation. 2025;151(13).
7. Granger CB. Predictors of Hospital Mortality in the Global Registry of Acute Coronary Events. Archives of Internal Medicine. 2003;163(19):2345.
8. Morrow DA, Antman EM, Charlesworth A, Cairns R, Murphy SA, de Lemos JA, et al. TIMI Risk Score for ST-Elevation Myocardial Infarction: A Convenient, Bedside, Clinical Score for Risk Assessment at Presentation. Circulation. 2000;102(17):2031-7.
9. D'Ascenzo F, Biondi-Zoccai G, Moretti C, Bollati M, Omedè P, Sciuto F, et al. TIMI, GRACE and alternative risk scores in Acute Coronary Syndromes: A meta-analysis of 40 derivation studies on 216,552 patients and of 42 validation studies on 31,625 patients. Contemporary Clinical Trials. 2012;33(3):507-14.
10. Wenzl FA, Kraler S, Ambler G, Weston C, Herzog SA, Räber L, et al. Sex-specific evaluation and redevelopment of the GRACE score in non-ST-segment elevation acute coronary syndromes in populations from the UK and Switzerland: a multinational analysis with external cohort validation. The Lancet. 2022;400(10354):744-56.
11. Zimmerman JE, Kramer AA, McNair DS, Malila FM. Acute Physiology and Chronic Health Evaluation (APACHE) IV: Hospital mortality assessment for today's critically ill patients*. Critical Care Medicine. 2006;34(5):1297-310.
12. Moreno RP, Metnitz PGH, Almeida E, Jordan B, Bauer P, Campos RA, et al. SAPS 3—From evaluation of the patient to evaluation of the intensive care unit. Part 2: Development of a prognostic model for hospital mortality at ICU admission. Intensive Care Medicine. 2005;31(10):1345-55.
13. Salluh JIF, Soares M. ICU severity of illness scores. Current Opinion in Critical Care. 2014;20(5):557-65.
14. Bennett CE, Wright RS, Jentzer J, Gajic O, Murphree DH, Murphy JG, et al. Severity of illness assessment with application of the APACHE IV predicted mortality and outcome trends analysis in an academic cardiac intensive care unit. Journal of Critical Care. 2019;50:242-6.

15. Jentzer JC, Bennett C, Wiley BM, Murphree DH, Keegan MT, Gajic O, et al. Predictive Value of the Sequential Organ Failure Assessment Score for Mortality in a Contemporary Cardiac Intensive Care Unit Population. Journal of the American Heart Association. 2018;7(6).
16. Jentzer JC, van Diepen S, Murphree DH, Ismail AS, Keegan MT, Morrow DA, et al. Admission diagnosis and mortality risk prediction in a contemporary cardiac intensive care unit population. American Heart Journal. 2020;224:57-64.
17. Jentzer JC, Anavekar NS, Bennett C, Murphree DH, Keegan MT, Wiley B, et al. Derivation and Validation of a Novel Cardiac Intensive Care Unit Admission Risk Score for Mortality. Journal of the American Heart Association. 2019;8(17).
18. Patel SM, Jentzer JC, Alviar CL, Baird-Zars VM, Barsness GW, Berg DD, et al. A pragmatic lab-based tool for risk assessment in cardiac critical care: data from the Critical Care Cardiology Trials Network (CCCTN) Registry. European Heart Journal Acute Cardiovascular Care. 2022;11(3):252-7.
19. Pöss J, Köster J, Fuernau G, Eitel I, de Waha S, Ouarrak T, et al. Risk Stratification for Patients in Cardiogenic Shock After Acute Myocardial Infarction. Journal of the American College of Cardiology. 2017;69(15):1913-20.
20. Jentzer JC, Rossello X. Past, present, and future of mortality risk scores in the contemporary cardiac intensive care unit. European Heart Journal Acute Cardiovascular Care. 2021;10(8):940-6.
21. Rafie N, Jentzer JC, Noseworthy PA, Kashou AH. Mortality Prediction in Cardiac Intensive Care Unit Patients: A Systematic Review of Existing and Artificial Intelligence Augmented Approaches. Frontiers in Artificial Intelligence. 2022;5.
22. Jentzer JC, Kashou AH, Murphree DH. Clinical applications of artificial intelligence and machine learning in the modern cardiac intensive care unit. Intelligence-Based Medicine. 2023;7:100089.
23. Pirracchio R, Petersen ML, Carone M, Rigon MR, Chevret S, van der Laan MJ. Mortality prediction in intensive care units with the Super ICU Learner Algorithm (SICULA): a population-based study. The Lancet Respiratory Medicine. 2015;3(1):42-52.
24. Meiring C, Dixit A, Harris S, MacCallum NS, Brealey DA, Watkinson PJ, et al. Optimal intensive care outcome prediction over time using machine learning. PLOS ONE. 2018;13(11):e0206862.
25. Rajkomar A, Oren E, Chen K, Dai AM, Hajaj N, Hardt M, et al. Scalable and accurate deep learning with electronic health records. npj Digital Medicine. 2018;1(1).
26. Thorsen-Meyer H-C, Nielsen AB, Nielsen AP, Kaas-Hansen BS, Toft P, Schierbeck J, et al. Dynamic and explainable machine learning prediction of mortality in patients in the intensive care unit: a retrospective study of high-frequency data in electronic patient records. The Lancet Digital Health. 2020;2(4):e179-e91.
27. Zhang D, Yin C, Zeng J, Yuan X, Zhang P. Combining structured and unstructured data for predictive models: a deep learning approach. BMC Medical Informatics and Decision Making. 2020;20(1).
28. Soenksen LR, Ma Y, Zeng C, Boussioux L, Villalobos Carballo K, Na L, et al. Integrated multimodal artificial intelligence framework for healthcare applications. npj Digital Medicine. 2022;5(1).
29. Kline A, Wang H, Li Y, Dennis S, Hutch M, Xu Z, et al. Multimodal machine learning in precision health: A scoping review. npj Digital Medicine. 2022;5(1).

30. Mamandipoor B, Hsu C-N, Krause M, Schmidt UH, Gabriel RA. Development and External Validation of a Multimodal Artificial Intelligence Mortality Prediction Model of Critically Ill Patients Using Multicenter Data. Anesthesiology. 2026.
31. Shillan D, Sterne JAC, Champneys A, Gibbison B. Use of machine learning to analyse routinely collected intensive care unit data: a systematic review. Critical Care. 2019;23(1).
32. van de Sande D, van Genderen ME, Huiskens J, Gommers D, van Bommel J. Moving from bytes to bedside: a systematic review on the use of artificial intelligence in the intensive care unit. Intensive Care Medicine. 2021;47(7):750-60.
33. Allyn J, Oliver M, Cerveau R, Allou N, Barennes T, Ferdynus C. Multi-source data integration through pooling and transfer learning improves generalizability and specialization of deep learning models for ICU mortality and length of stay prediction: a four-database external validation study. Critical Care. 2026;30(1).
34. Wong A, Otles E, Donnelly JP, Krumm A, McCullough J, DeTroyer-Cooley O, et al. External Validation of a Widely Implemented Proprietary Sepsis Prediction Model in Hospitalized Patients. JAMA Internal Medicine. 2021;181(8):1065.
35. Finlayson SG, Subbaswamy A, Singh K, Bowers J, Kupke A, Zittrain J, et al. The Clinician and Dataset Shift in Artificial Intelligence. New England Journal of Medicine. 2021;385(3):283-6.
36. Christodoulou E, Ma J, Collins GS, Steyerberg EW, Verbakel JY, Van Calster B. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. Journal of Clinical Epidemiology. 2019;110:12-22.
37. Khera R, Haimovich J, Hurley NC, McNamara R, Spertus JA, Desai N, et al. Use of Machine Learning Models to Predict Death After Acute Myocardial Infarction. JAMA Cardiology. 2021;6(6):633.
38. Zhang Q, Xu L, Xie Z, He W, Huang X. Machine learning-based prediction of mortality in acute myocardial infarction with cardiogenic shock. Frontiers in Cardiovascular Medicine. 2024;11.
39. Pollard TJ, Johnson AEW, Raffa JD, Celi LA, Mark RG, Badawi O. The eICU Collaborative Research Database, a freely available multi-center database for critical care research. Scientific Data. 2018;5(1).
40. Sauer CM, Dam TA, Celi LA, Faltys M, de la Hoz MAA, Adhikari L, et al. Systematic Review and Comparison of Publicly Available ICU Data Sets—A Decision Guide for Clinicians and Data Scientists. Critical Care Medicine. 2022;50(6):e581-e8.
41. Gary S Collins KGMM, Paula Dhiman, Richard D Riley, Andrew L Beam, Ben Van Calster, Marzyeh Ghassemi, Xiaoxuan Liu, Johannes B Reitsma, Maarten van Smeden, Anne-Laure Boulesteix, Jennifer Catherine Camaradou, Leo Anthony Celi, Spiros Denaxas, Alastair K Denniston, Ben Glocker, Robert M Golub, Hugh Harvey, Georg Heinze, Michael M Hoffman, André Pascal Kengne, Emily Lam, Naomi Lee, Elizabeth W Loder, Lena Maier-Hein, Bilal A Mateen, Melissa D McCradden, Lauren Oakden-Rayner, Johan Ordish, Richard Parnell, Sherri Rose,38 Karandeep Singh, Laure Wynants, Patricia Logullo. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ. 2024:q902.
42. Goldberger AL, Amaral LAN, Glass L, Hausdorff JM, Ivanov PC, Mark RG, et al. PhysioBank, PhysioToolkit, and PhysioNet. Circulation. 2000;101(23).

43. Riley RD. Correction to: Minimum sample size for developing a multivariable prediction model: Part II - binary and time - to - event outcomes by Riley RD, Snell KI, Ensor J, et al. Statistics in Medicine. 2019;38(30):5672-.
44. Riley RD, Ensor J, Snell KIE, Harrell FE, Martin GP, Reitsma JB, et al. Calculating the sample size required for developing a clinical prediction model. BMJ. 2020:m441.
45. Riley RD, Ensor J, Snell KIE, Debray TPA, Altman DG, Moons KGM, et al. External validation of clinical prediction models using big datasets from e-health records or IPD meta-analysis: opportunities and challenges. BMJ. 2016:i3140.
46. Che Z, Purushotham S, Cho K, Sontag D, Liu Y. Recurrent Neural Networks for Multivariate Time Series with Missing Values. Scientific Reports. 2018;8(1).
47. Steyerberg EW, Harrell FE. Prediction models need appropriate internal, internal–external, and external validation. Journal of Clinical Epidemiology. 2016;69:245-7.
48. Debray TPA, Vergouwe Y, Koffijberg H, Nieboer D, Steyerberg EW, Moons KGM. A new framework to enhance the interpretation of external validation studies of clinical prediction models. Journal of Clinical Epidemiology. 2015;68(3):279-89.
49. Shi Y, Ke G, Soukhavong D, Lamb J, Meng Q, Finley T, et al. lightgbm: Light Gradient Boosting Machine. CRAN: Contributed Packages: The R Foundation; 2020.
50. Chen T, Guestrin C. XGBoost. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; 2016/08/13: ACM; 2016. p. 785-94.
51. Akiba T, Sano S, Yanase T, Ohta T, Koyama M. Optuna. Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining; 2019/07/25: ACM; 2019. p. 2623-31.
52. Cho K, van Merrienboer B, Gulcehre C, Bahdanau D, Bougares F, Schwenk H, et al. Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP): Association for Computational Linguistics; 2014. p. 1724-34.
53. Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, et al. PyTorch: An Imperative Style, High-Performance Deep Learning Library. Adv Neur In. 2019;32.
54. Wang H, Lao G, He R, Liu T, Luo H, Qin C, et al. Predicting ICU in-hospital mortality from text-encoded structured EHR data using adaptive transformer layer fusion. iScience. 2026;29(6):116135.
55. Johnson AEW, Bulgarelli L, Shen L, Gayles A, Shammout A, Horng S, et al. MIMIC-IV, a freely accessible electronic health record dataset. Scientific Data. 2023;10(1).
56. Saito T, Rehmsmeier M. The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets. PLOS ONE. 2015;10(3):e0118432.
57. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the Areas under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach. Biometrics. 1988;44(3):837.
58. Van Calster B, Nieboer D, Vergouwe Y, De Cock B, Pencina MJ, Steyerberg EW. A calibration hierarchy for risk models was defined: from utopia to empirical data. Journal of Clinical Epidemiology. 2016;74:167-76.
59. Van Calster B, McLernon DJ, van Smeden M, Wynants L, Steyerberg EW. Calibration: the Achilles heel of predictive analytics. BMC Medicine. 2019;17(1).

60. Austin PC, Steyerberg EW. The Integrated Calibration Index (ICI) and related metrics for quantifying the calibration of logistic regression models. Statistics in Medicine. 2019;38(21):4051-65.
61. Vickers AJ, Elkin EB. Decision Curve Analysis: A Novel Method for Evaluating Prediction Models. Medical Decision Making. 2006;26(6):565-74.
62. Vickers AJ, Van Calster B, Steyerberg EW. Net benefit approaches to the evaluation of prediction models, molecular markers, and diagnostic tests. BMJ. 2016:i6.
63. Lundberg SM, Lee SI. A Unified Approach to Interpreting Model Predictions. Advances in Neural Information Processing Systems 30 (Nips 2017). 2017;30.
64. Lundberg SM, Erion G, Chen H, DeGrave A, Prutkin JM, Nair B, et al. From local explanations to global understanding with explainable AI for trees. Nature Machine Intelligence. 2020;2(1):56-67.
65. Sundararajan M, Taly A, Yan QQ. Axiomatic Attribution for Deep Networks. Pr Mach Learn Res. 2017;70.
66. Ghassemi M, Oakden-Rayner L, Beam AL. The false hope of current approaches to explainable artificial intelligence in health care. The Lancet Digital Health. 2021;3(11):e745-e50.
67. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. J Mach Learn Res. 2011;12:2825-30.
68. Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. BMJ. 2015;350(jan07 4):g7594-g.
69. Vincent J-L, Moreno R. Clinical review: Scoring systems in the critically ill. Critical Care. 2010;14(2):207.
70. Kwon J-m, Jeon K-H, Kim HM, Kim MJ, Lim S, Kim K-H, et al. Deep-learning-based risk stratification for mortality of patients with acute myocardial infarction. PLOS ONE. 2019;14(10):e0224502.
71. Hernesniemi JA, Mahdiani S, Tynkkynen JA, Lyytikäinen L-P, Mishra PP, Lehtimäki T, et al. Extensive phenotype data and machine learning in prediction of mortality in acute coronary syndrome – the MADDEC study. Annals of Medicine. 2019;51(2):156-63.
72. Lin X, Pan X, Yang Y, Yang W, Wang X, Zou K, et al. Machine learning models to predict 30-day mortality for critical patients with myocardial infarction: a retrospective analysis from MIMIC-IV database. Frontiers in Cardiovascular Medicine. 2024;11.
73. Futoma J, Simons M, Panch T, Doshi-Velez F, Celi LA. The myth of generalisability in clinical research and machine learning in health care. The Lancet Digital Health. 2020;2(9):e489-e92.
74. Subbaswamy A, Saria S. From development to deployment: dataset shift, causality, and shift-stable models in health AI. Biostatistics. 2019.
75. Harutyunyan H, Khachatrian H, Kale DC, Ver Steeg G, Galstyan A. Multitask learning and benchmarking with clinical time series data. Scientific Data. 2019;6(1).
76. Sheikhalishahi S, Balaraman V, Osmani V. Benchmarking machine learning models on multi-centre eICU critical care dataset. PLOS ONE. 2020;15(7):e0235424.
77. Hyland SL, Faltys M, Hüser M, Lyu X, Gumbsch T, Esteban C, et al. Early prediction of circulatory failure in the intensive care unit using machine learning. Nature Medicine. 2020;26(3):364-73.
78. Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence. 2019;1(5):206-15.

79. Xiao C, Choi E, Sun J. Opportunities and challenges in developing deep learning models using electronic health records data: a systematic review. Journal of the American Medical Informatics Association. 2018;25(10):1419-28.
80. Kelly CJ, Karthikesalingam A, Suleyman M, Corrado G, King D. Key challenges for delivering clinical impact with artificial intelligence. BMC Medicine. 2019;17(1).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryAppendix1010.docx