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.

Pulmonary hypertension records and repeated pulmonary artery pressure elevation identify distinct prognostic populations in intensive care

Shenpeng Zhou
Guangzhou Medical University
Lixian Xu
Guangzhou Medical University
Peilian Wu
Guangzhou Medical University
Junyan Han
Guangzhou Medical University
Meiqi Li
Guangzhou Medical University
Xiaoying Zhang
Guangzhou Medical University
Zhenxiang Li
Guangzhou Medical University
Shiwen Li
Guangzhou Medical University
Fei Peng
Guangzhou Medical University
Mingmei Xiong
qnwmh@126.com

Guangzhou Medical University

Research Article

Keywords: pulmonary hypertension, pulmonary artery pressure, intensive care, hospital mortality, case ascertainment, prediction model

Posted Date: November 25th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11629288/v1

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

Additional Declarations: No competing interests reported.

Abstract

Background

Pulmonary hypertension (PH) diagnoses and pulmonary artery pressure monitoring identify patients through different clinical pathways. We examined how these sources define intensive care populations and whether pressure information improves hospital mortality prediction beyond routine clinical data and recording availability.

Methods

We developed penalised logistic regression and XGBoost models in MIMIC-IV and evaluated them in the multicentre eICU database. Adults alive and remaining in intensive care at 24 hours qualified through a PH diagnostic algorithm or repeated mean pulmonary artery pressure above 20 mmHg. We evaluated ascertainment groups and compared clinical, availability-only and pressure-augmented specifications. Frozen and refitted models were compared in identical external patients. Confidence intervals used hospital-cluster resampling. A separate analysis assessed patients remaining at 72 hours.

Results

MIMIC-IV included 5403 patients with 499 deaths; eICU included 4830 patients with 310 deaths across 128 hospitals. Mortality in haemodynamics-only versus diagnosis-only patients was 5.2% versus 18.3% in MIMIC-IV and 4.7% versus 20.2% in eICU. At 24 hours, pressure-augmented XGBoost achieved an external area under the receiver operating characteristic curve (AUROC) of 0.849 (95% CI 0.823–0.871) and area under the precision–recall curve of 0.376 against an event proportion of 0.064. The calibration slope was 1.057 and calibration-in-the-large was 0.231, indicating overall underprediction. Subgroup AUROCs were 0.698 in diagnosis-only and 0.837 in haemodynamics-only patients. The full pressure specification added 0.0133 (0.0055–0.0216) to AUROC beyond recording availability. The paired refitted-minus-frozen difference was 0.0076 (− 0.0041 to 0.0224). The separate 72-hour model achieved an AUROC of 0.812.

Conclusions

PH records and repeated pressure elevation identify clinically distinct populations with different mortality and prediction profiles. Pressure augmentation adds predictive information beyond recording availability at 24 hours. Case definitions, monitoring context and subgroup evaluation are central to interpreting prognostic models in respiratory critical care.

Background

Pulmonary hypertension (PH) complicates cardiovascular and respiratory disease and contributes to the clinical complexity of the intensive care unit (ICU), where right ventricular dysfunction interacts with acute circulatory and respiratory failure [1, 2]. The haemodynamic definition is a resting mean pulmonary artery pressure (mPAP) above 20 mmHg measured by right heart catheterisation; wedge pressure, pulmonary vascular resistance and clinical context guide further classification [1, 3]. Prognostic assessment in ICU populations therefore depends on how patients are identified and which measurements inform prediction.

ICU cohorts of patients with PH have documented high mortality, with outcomes varying by admission context and acute organ dysfunction [4, 5, 6]. REVEAL Lite 2 and COMPERA 2.0 stratify mortality risk in pulmonary arterial hypertension (PAH) using disease-specific clinical assessments [7, 8]. Hospital mortality during an acute ICU admission requires assessment of the patient's current critical illness and a shorter prediction horizon.

Diagnostic documentation and pulmonary artery monitoring identify patients through different clinical pathways. Codes and structured diagnoses capture recognised disease, while monitoring captures pressures in patients selected for measurement. An expanded population defined by either source can therefore include different clinical presentations and baseline risks. Whether a model performs well in this population depends on its ability to rank risk within these groups and distinguish between them. Comparisons in identical external patients help separate changes in model development from changes in the population assessed [9, 10].

Previous database studies established a basis for ICU prognostic research in PH. Deng and colleagues reported higher hospital mortality in patients with a PH diagnosis in MIMIC-IV [11]. Zhang and colleagues developed a mortality nomogram for patients labelled as having primary PH using MIMIC and eICU [12]. Extending such models to patients identified through pressure monitoring raises a specific question: how do diagnostic records, repeated pressure elevation and the availability of measurements contribute to the resulting prognostic estimates?

We examined how PH diagnostic records and repeated pulmonary artery pressure elevation identify prognostically different ICU populations, and whether pressure information improves prediction beyond routine clinical data and recording availability. We compared clinical and pressure-augmented models in a common external population and evaluated performance within each ascertainment group. Paired comparisons of frozen diagnosis-cohort models and models refitted in expanded cohorts quantified the refitting contrast while holding the evaluation population constant.

Methods

Study design and data sources

This retrospective cohort study used MIMIC-IV version 3.1 from Beth Israel Deaconess Medical Center (2008–2022) for development and the multicentre eICU Collaborative Research Database version 2.0

(2014–2015) for external evaluation [13–16]. We extracted data from their PhysioNet BigQuery datasets. Models from an earlier, unpublished diagnostic-cohort analysis were retained with their fitted parameters and preprocessing for comparison with the expanded-cohort models. The cohort extension and supplementary comparisons were exploratory and used an eICU database previously examined in this project [17, 18].

Eligibility and prediction landmarks

We identified adults aged at least 18 years and selected the first ICU stay in MIMIC-IV. In eICU, we selected the first available patient record ordered by hospital discharge year, hospital admission identifier, unit visit number and ICU stay identifier. Age was capped at 90 years. Eligibility required survival and continued admission to both ICU and hospital beyond the landmark, with a known hospital discharge outcome. The primary landmark was 24 hours; a separate cohort was constructed using the same eligibility rules at 72 hours. Patient selection is shown in Fig. 1 for the primary analysis and Fig. S12 for the 72-hour analysis.

Diagnostic and haemodynamic ascertainment

The diagnostic algorithm used ICD-9 code 416.0 and ICD-10 codes I27.0, I27.2, I27.20–I27.24 and I27.29. eICU also included explicit structured pulmonary hypertension diagnoses. The broader ICD-9 code 416.8 was audited separately. Full matching rules appear in Additional file 1. MIMIC-IV current-admission discharge codes have no diagnosis timestamp. eICU diagnoses are ICU clinical documentation with an entry offset, which denotes recording time rather than disease onset [32]. The primary retrospective ascertainment used all qualifying current-admission diagnoses; supplementary analyses restricted eICU diagnostic eligibility to documentation before each landmark.

Pulmonary artery systolic, diastolic and mean pressures were obtained from dedicated structured monitoring fields. Measurements were restricted to the observation window; MIMIC-IV also required storage before the landmark. We excluded values outside the prespecified plausibility ranges of 5–200, 0–100 and 3–150 mmHg, respectively. After calculating hourly medians, we removed hourly pressure sets with a diastolic value exceeding the mean or systolic value, or a mean exceeding the systolic value. Quality checks operated on the aggregated structured measurements.

The primary pressure rule required at least three valid hourly mPAP observations, a median of these hourly medians above 20 mmHg, and at least two elevated hours whose earliest and latest indices differed by at least two. This operational phenotype represents repeated pressure elevation during ICU monitoring. Chronic PH, PAH and haemodynamic subtypes require clinical adjudication beyond these monitoring fields. The expanded cohort met either the diagnostic algorithm or the pressure rule. Mutually exclusive groups comprised diagnosis only, haemodynamics only, and both. Diagnosis-only patients included those with absent or insufficient monitoring and those below the pressure threshold.

Outcome and candidate predictors

The outcome was all-cause death before discharge from the index hospital admission among patients eligible at the landmark. Follow-up extended from the landmark to hospital discharge. Candidate predictors covered demographics, vital signs and laboratory measurements. Dynamic summaries captured early circulatory and respiratory instability. Pressure augmentation comprised systolic, diastolic and mean pulmonary artery pressure, together with an indicator of valid pressure recording. Complete candidate lists, model specifications, units, aggregation rules and missingness are supplied in Additional files 1 and 2.

We summarised laboratory measurements by their within-window medians after unit harmonisation and plausibility checks, and vital signs from hourly medians. Availability timestamps determined eligibility where provided. Each fitting partition supplied its imputation medians. Least absolute shrinkage and selection operator (LASSO) regression used median imputation and training-partition standardisation. XGBoost also included indicators for variables missing in its fitting partition. Framework comparisons therefore evaluated both model form and input representation; comparisons within each framework assessed the addition of the specified pressure-related inputs.

Model development and matched external comparisons

We evaluated LASSO logistic regression using routine clinical predictors and XGBoost using clinical predictors plus dynamic summaries. Each framework included a clinical and a pressure-augmented specification, yielding four models. LASSO applied shrinkage and feature selection within the regression pipeline [19]. XGBoost used the complete specified predictor sets to capture nonlinear relationships and interactions [20].

Development used five outer and three inner stratified cross-validation folds, with imputation, scaling and penalty or hyperparameter selection repeated within training folds [21]. LASSO used a one-standard-error rule favouring stronger regularisation. XGBoost used a compact grid of tree depths and L2 penalties with fixed tree count and learning rate. We used the original class distribution without synthetic oversampling or class weighting. Final models were fitted to all eligible MIMIC-IV development patients after inner tuning. Development out-of-fold log loss guided tree-model selection; these estimates formed part of model selection. Final fits were locked before external prediction.

Frozen diagnosis-cohort models retained their coefficients, preprocessing, hyperparameters and calibration. We applied frozen and expanded-cohort models to identical eICU patients to estimate paired performance differences while holding the evaluation population constant. Separate analyses evaluated performance within each ascertainment group.

Performance and statistical analysis

We assessed discrimination using area under the receiver operating characteristic curve (AUROC) and area under the precision–recall curve (AUPRC, calculated as average precision) [22], and overall prediction error using the Brier score [23]. Calibration measures comprised a logistic intercept and slope, calibration-in-the-large with slope fixed at one, and plots using eight predicted-risk groups [24]. Ideal intercept and calibration-in-the-large values are zero; the ideal slope is one. Exploratory decision curves quantified net benefit across risk thresholds from 5% to 40%, comparing model-guided classification with classifying all or no patients as high risk [25]. Thresholds represented hypothetical preferences for risk-directed actions.

We resampled entire eICU hospitals 1000 times to estimate 95% confidence intervals and paired differences in AUROC, AUPRC and Brier score. Positive AUROC or AUPRC differences and negative Brier differences favour the first model. The principal pressure model and its frozen comparator also received hospital-cluster intervals for calibration. The sample comprised all eligible source records, without an a priori sample-size calculation. Exploratory confidence intervals were not adjusted for multiplicity. Baseline comparisons were descriptive; candidate eligibility followed the specified predictor lists. Clinical profiles were summarised by ascertainment group, hospital outcome and database using observed-value medians and interquartile ranges or documented positive/observed counts, accompanied by variable-specific missingness.

Explanations and sensitivity analyses

Shapley additive explanations (SHAP) were calculated using TreeSHAP in reproducible random samples of 300 patients from each 24-hour expanded cohort [26]. Contributions were expressed on the log-odds scale, included recording indicators, and were checked for numerical additivity to the model prediction. Regression coefficients and graphical nomograms represented the final penalised models. We ranked features by mean absolute SHAP contribution and plotted eight observed continuous predictors with quartile-median summaries in the same external sample. For LASSO, we reported all coefficients, development imputation values and selection counts across five outer folds; odds ratios were calculated as exp(coefficient × specified increment).

Sensitivity analyses used a repeated mPAP threshold of at least 25 mmHg and a broader definition requiring at least one valid hourly mPAP median above 20 mmHg. The primary models were applied without refitting to these alternative populations. We audited ICU type and supplementary catheter-related fields to characterise measurement context, followed by a post hoc analysis excluding eICU units labelled CTICU, CCU-CTICU or CSICU. The 72-hour models used observations from admission to 72 hours in patients eligible at that landmark.

Additional ascertainment and benchmark analyses

Supplementary ascertainment analyses applied the fixed models to populations defined by diagnostic timing. In eICU, timely diagnostic eligibility required a qualifying record with 0 ≤ diagnosisOffset < 1440 or 4320 minutes; a separate variant included negative offsets. The pressure eligibility branch remained fixed. Prior-admission PH in MIMIC-IV was described separately because current-admission coding lacks timestamps.

Pressure-information analyses compared the original dynamic and full pulmonary artery pressure (PAP) XGBoost models with availability-only and pressure-value specifications. The value specification omitted explicit and automatically generated PAP-recording indicators and retained automatic indicators for other inputs. The two new specifications followed the original nested MIMIC-only development procedure at each landmark and were locked before supplementary external prediction. All specifications were also evaluated in identical external subsets with complete PAP measurements.

We compared recorded Acute Physiology and Chronic Health Evaluation (APACHE) IV and IVa hospital-mortality predictions separately with the four original models in matched patients [31, 33], auditing invalid codes, duplicates and coverage. At 72 hours, APACHE retained its first-day prediction while the study models used updated inputs. Hospital-cluster resampling supplied paired intervals. We decomposed pooled AUROC into the weighted concordance of within-group and between-group event–non-event pairs, counting ties as one half. Calibration intervals were added for the principal pressure model and its frozen comparator. Full specifications and results appear in Additional file 6.

Missing data sensitivity analyses

Missing-data sensitivity analyses used the same cohorts and predictor sets at both landmarks. For each LASSO specification, 20 posterior Bayesian ridge chained-equation imputations generated 20 fitted models; their predicted probabilities were averaged. Imputation, scaling and penalty selection were repeated within the original nested training folds, with outcomes excluded from imputation. Native-NaN XGBoost specifications evaluated training-derived automatic missingness indicators while retaining the explicit PAP-recording variable in PAP models. Supplementary fits were locked before prediction in the previously examined eICU cohorts. Paired comparisons, calibration, imputation diagnostics and selection stability appear in Additional file 5.

Exploratory predictor expansion and model simplification

Predictor expansion evaluated body mass index (BMI), 11 categories of documented medical history, central venous pressure (CVP), and five recorded treatments: mechanical ventilation, non-invasive ventilation, renal replacement therapy, vasopressors and selected inotropes. Historical diagnoses came from completed prior admissions or timestamped history available by the landmark; current-admission discharge codes were excluded as predictors. Predictor blocks and single-block increments used the same nested development procedure; medication orders were audited separately. Compact penalised logistic models used 14 routine candidates with size limits of four, six or eight nonzero predictors. A development-only one-standard-error rule favoured fewer predictors, and five outer folds assessed the

complete selection procedure. Fits were locked before external prediction and evaluated without external recalibration. Comparisons estimated performance differences without a prespecified non-inferiority margin. Methods and results appear in Additional file 3, with code and aggregate outputs in Additional file 2.

Software and reproducibility

Supplementary analyses used Python 3.12.1, pandas 3.0.6, scikit-learn 1.9.1 and XGBoost 3.4.1. Additional file 2 supplies dependency versions, extraction rules, model specifications and aggregate outputs. Integrity checks covered cohort overlap, feature agreement for the original diagnostic subset, pressure-summary consistency, frozen-prediction reproduction, coefficient-based regression predictions and SHAP additivity. Supplementary checks verified development locks, matched score joins, bootstrap calculations and agreement between aggregate outputs and reported results.

Results

Cohort ascertainment and mortality

Among 65366 adult first ICU records in MIMIC-IV, 51552 met the 24-hour landmark eligibility criteria. The corresponding eICU counts were 138852 first available adult records and 95443 eligible patients. mPAP was observed in 5831 and 6729 eligible patients, respectively. Diagnostic ascertainment identified 1921 MIMIC-IV patients and 597 eICU patients. Adding the repeated haemodynamic definition expanded these cohorts to 5403 patients with 499 deaths and 4830 patients with 310 deaths, respectively; the eICU cohort covered 128 hospitals (Table 1; Figs. 1 and 2).

Table 1

a. Clinical profiles by ascertainment at 24 hours in MIMIC-IV.

Characteristic	Diagnosis only	Haemodynamics only	Both
Patients	1559	3482	362
Hospital death	285/1559 (18.3%)	180/3482 (5.2%)	34/362 (9.4%)
Age, years	74 [64–83]	69 [61–77]	68 [60–75]
Male sex	750/1559 (48.1%)	2321/3482 (66.7%)	207/362 (57.2%)
Body mass index, kg/m²	27.48 [23.75–33.15]	29.01 [25.47–33.23]	28.95 [25.35–33.43]
Heart rate, /min	82 [72–95]	82 [77.5–90]	80 [75–88]
Respiratory rate, /min	20 [17–22.75]	18 [16–20]	18 [16–20.94]
Mean arterial pressure, mmHg	76 [70–85]	73 [69–77]	72.5 [68–77.5]
Oxygen saturation, %	96 [95–98]	98 [96.25–99]	98 [96–99]
Blood urea nitrogen, mg/dL	25 [17–43]	17.5 [13.5–23]	20 [15–31.88]
Creatinine, mg/dL	1.1 [0.8–1.8]	0.95 [0.75–1.25]	1.1 [0.85–1.6]
Albumin, g/dL	3.2 [2.9–3.6]	3.1 [2.7–3.6]	3.25 [2.8–3.6]
Lactate, mmol/L	1.7 [1.2–2.3]	2.2 [1.65–3]	2.08 [1.5–2.8]
Systolic PAP, mmHg	28 [25.44–35]	36 [32–41]	43 [36.5–52]
Diastolic PAP, mmHg	13 [11.62–18.75]	19 [17–22]	22 [18–25.5]
Mean PAP, mmHg	18.5 [17–20]	25.25 [23–29]	29.5 [25–35]
Central venous pressure, mmHg	8.75 [6–13]	12 [9.75–14]	11 [9–14]
History of heart failure	402/1146 (35.1%)	576/2716 (21.2%)	98/226 (43.4%)
History of chronic lung disease	234/1146 (20.4%)	284/2716 (10.5%)	26/226 (11.5%)
History of chronic kidney disease	283/1146 (24.7%)	306/2716 (11.3%)	52/226 (23.0%)
Recorded mechanical ventilation	154/1559 (9.9%)	2005/3482 (57.6%)	141/362 (39.0%)
Recorded vasopressor	407/1559 (26.1%)	2771/3482 (79.6%)	223/362 (61.6%)

Continuous variables are median [interquartile range] among observed values; categorical variables are positive/observed (%). Recorded history and treatments describe documentation within available sources. Missingness is reported in Additional file 4; percentages use observed denominators.

Table 1
b. Clinical profiles by ascertainment at 24 hours in eICU.

Characteristic	Diagnosis only	Haemodynamics only	Both
Patients	510	4233	87
Hospital death	103/510 (20.2%)	197/4233 (4.7%)	10/87 (11.5%)
Age, years	69 [59.25–79]	67 [59–75]	68 [56.5–77]
Male sex	214/510 (42.0%)	2766/4233 (65.3%)	49/87 (56.3%)
Body mass index, kg/m²	28.72 [23.99–35.78]	29.94 [26.12–34.39]	28.78 [24.08–35.53]
Heart rate, /min	83.75 [74–95.25]	82 [75.75–90.75]	80.5 [70.5–90]
Respiratory rate, /min	20 [17.25–24]	17.75 [15–20]	18 [15.88–21.12]
Mean arterial pressure, mmHg	76 [69–85]	74 [69.19–79.5]	73 [68.25–78.75]
Oxygen saturation, %	96.5 [95–98.5]	97.75 [96–99]	98 [96.56–99]
Blood urea nitrogen, mg/dL	28 [18–44.5]	18 [14–24]	25.5 [16–36]
Creatinine, mg/dL	1.26 [0.85–2.13]	1.02 [0.81–1.32]	1.3 [0.83–1.73]
Albumin, g/dL	2.8 [2.4–3.2]	3.2 [2.9–3.6]	3.12 [2.7–3.5]
Lactate, mmol/L	1.7 [1–2.35]	2.21 [1.45–4.06]	2.3 [1.17–3.18]
Systolic PAP, mmHg	23.5 [20.75–31.25]	35 [31.75–40]	47.25 [39.12–60]
Diastolic PAP, mmHg	15 [10.5–16.75]	18 [16–21]	22.5 [20–27.75]
Mean PAP, mmHg	18.75 [14.69–19.94]	25 [22.5–28.25]	32.75 [28–39.38]
Central venous pressure, mmHg	12 [7.25–17]	11.5 [9.5–14]	11 [8.12–14]
History of heart failure	208/485 (42.9%)	931/4188 (22.2%)	45/86 (52.3%)
History of chronic lung disease	168/485 (34.6%)	407/4188 (9.7%)	18/86 (20.9%)
History of chronic kidney disease	128/485 (26.4%)	479/4188 (11.4%)	23/86 (26.7%)
Recorded mechanical ventilation	178/381 (46.7%)	3066/4077 (75.2%)	36/77 (46.8%)

Continuous variables are median [interquartile range] among observed values; categorical variables are positive/observed (%). Recorded history and treatments describe documentation within available sources. Missingness is reported in Additional file 4; percentages use observed denominators.

Characteristic	Diagnosis only	Haemodynamics only	Both
Recorded vasopressor	88/197 (44.7%)	2195/2796 (78.5%)	32/40 (80.0%)

Continuous variables are median [interquartile range] among observed values; categorical variables are positive/observed (%). Recorded history and treatments describe documentation within available sources. Missingness is reported in Additional file 4; percentages use observed denominators.

The three ascertainment groups had distinct mortality profiles. Haemodynamics-only patients accounted for 3482 MIMIC-IV and 4233 eICU patients, with mortality of 5.2% and 4.7%. Diagnosis-only patients had mortality of 18.3% and 20.2%, while patients meeting both definitions had mortality of 9.4% and 11.5%. Haemodynamics-only patients were younger and more often male than diagnosis-only patients (Table 1). The separate coding audit identified the broader ICD-9 code 416.8 in 363 haemodynamics-only MIMIC-IV patients.

Haemodynamics-only ascertainment was concentrated in cardiovascular and cardiothoracic care. In MIMIC-IV, 3045 of 3482 patients (87.4%) were in cardiovascular intensive care; in eICU, 2793 of 4233 (66.0%) were in CTICU, CCU-CTICU or CSICU units. Cardiac output records were available for 3142 and 2220 patients, respectively, and wedge pressure for 238 and 740. These monitoring records characterised the clinical setting in which pressure elevation was captured (Additional file 1).

Diagnosis-only patients were older and had higher respiratory rates and renal laboratory values in both databases (Table 1). Median age was 74 versus 69 years in MIMIC-IV and 69 versus 67 years in eICU when comparing diagnosis-only with haemodynamics-only patients. Corresponding blood urea nitrogen values were 25 versus 17.5 mg/dL and 28 versus 18 mg/dL; creatinine was 1.10 versus 0.95 mg/dL and 1.26 versus 1.02 mg/dL. Table 1 also presents measured pulmonary pressures, BMI, CVP, documented history and recorded treatments.

Patients who died had higher median blood urea nitrogen than survivors in MIMIC-IV (30 versus 18 mg/dL) and eICU (28 versus 18 mg/dL), with higher respiratory rates and lower observed albumin. Clinical profiles by outcome and database, with variable-specific missingness, are provided in Additional file 4. Lactate availability varied by ascertainment group: missingness in the diagnosis-only and haemodynamics-only groups was 31.7% and 3.2% in MIMIC-IV, compared with 69.2% and 82.6% in eICU. Observed lactate medians therefore describe different measurement subsets.

Predictor selection and final model composition

The clinical and pressure-augmented LASSO models retained 21 and 23 predictors, respectively. Age, respiratory rate, oxygen saturation, blood urea nitrogen, albumin, bilirubin, lactate, glucose, partial thromboplastin time and red cell distribution width were selected in all five outer folds of both models. The pressure-augmented model also retained diastolic and mean pulmonary artery pressure and recording availability. Table 4 presents all candidate predictors, penalised odds ratios and selection frequencies. Compact model specifications are reported in Additional file 3.

Table 4
Primary LASSO predictor retention and penalised odds ratios at 24 hours.

Predictor and unit	Increment	Clinical OR (folds)	Pressure OR (folds)
Age, years	10	1.354 (5/5)	1.268 (5/5)
Male sex	1	0.970 (2/5)	Zero coefficient (1/5)
Heart rate, /min	10	1.042 (5/5)	1.034 (4/5)
Respiratory rate, /min	5	1.734 (5/5)	1.501 (5/5)
Mean arterial pressure, mmHg	10	1.089 (1/5)	1.008 (1/5)
Oxygen saturation, %	5	0.700 (5/5)	0.835 (5/5)
Creatinine, mg/dL	1	Zero coefficient (0/5)	Zero coefficient (0/5)
Blood urea nitrogen, mg/dL	10	1.212 (5/5)	1.153 (5/5)
Sodium, mmol/L	5	Zero coefficient (1/5)	Zero coefficient (1/5)
Potassium, mmol/L	1	0.937 (2/5)	0.983 (1/5)
Bicarbonate, mmol/L	5	0.944 (3/5)	0.908 (5/5)
White blood cells, 10^9/L	5	1.025 (3/5)	1.042 (5/5)
Platelets, 10^9/L	50	1.073 (5/5)	1.029 (3/5)
Haemoglobin, g/dL	1	1.006 (0/5)	1.007 (0/5)
Albumin, g/dL	1	0.661 (5/5)	0.673 (5/5)
Bilirubin, mg/dL	1	1.094 (5/5)	1.079 (5/5)
Lactate, mmol/L	1	1.226 (5/5)	1.292 (5/5)
Glucose, mg/dL	10	1.064 (5/5)	1.061 (5/5)
INR	1	1.095 (5/5)	1.036 (4/5)
Partial thromboplastin time, s	10	1.093 (5/5)	1.090 (5/5)
Red cell distribution width, %	1	1.104 (5/5)	1.058 (5/5)
Systolic blood pressure, mmHg	10	0.929 (3/5)	0.920 (5/5)
Diastolic blood pressure, mmHg	10	Zero coefficient (0/5)	Zero coefficient (0/5)
Temperature, °C	1	0.878 (4/5)	0.914 (4/5)

OR = exp(coefficient × increment), conditional on other model inputs. Parentheses give selection counts across the five outer folds. A zero coefficient denotes omission from the final penalised fit; selection counts describe the separate outer-fold fits. Binary changes are 0 to 1. Raw coefficients, intercepts and imputation values appear in Table S16.

Predictor and unit	Increment	Clinical OR (folds)	Pressure OR (folds)
PAP recorded	1	Not a candidate	0.274 (5/5)
Systolic PAP, mmHg	5	Not a candidate	Zero coefficient (0/5)
Diastolic PAP, mmHg	5	Not a candidate	1.360 (5/5)
Mean PAP, mmHg	5	Not a candidate	1.074 (4/5)

OR = exp(coefficient × increment), conditional on other model inputs. Parentheses give selection counts across the five outer folds. A zero coefficient denotes omission from the final penalised fit; selection counts describe the separate outer-fold fits. Binary changes are 0 to 1. Raw coefficients, intercepts and imputation values appear in Table S16.

Model performance at 24 hours

Development out-of-fold AUROCs were 0.834 for clinical LASSO, 0.853 for pressure-augmented LASSO, 0.857 for dynamic XGBoost and 0.864 for pressure-augmented XGBoost. The pressure-augmented XGBoost model had the lower development log loss of the two tree specifications.

In the expanded eICU cohort, pressure-augmented XGBoost achieved an AUROC of 0.849 (95% CI 0.823–0.871), an AUPRC of 0.376 and a Brier score of 0.0491 (Table 2). The AUPRC reference event proportion was 0.064. Its calibration slope was 1.057 (95% CI 0.973–1.149) and calibration-in-the-large was 0.231 (0.091–0.358), indicating overall underprediction. Clinical LASSO achieved an AUROC of 0.779, Brier score of 0.0516 and calibration-in-the-large of − 0.003. Figure 3 presents ROC, precision–recall, grouped calibration and exploratory decision curves; Table S34 gives threshold-specific net benefit.

Table 2
External performance in the same 4830 patients at 24 hours

Model	AUROC (95% CI)	AUPRC	Brier	Slope	CITL
Clinical LASSO Frozen	0.778 (0.741–0.808)	0.327	0.0539	1.356	-0.708
Clinical LASSO Refitted	0.779 (0.744–0.809)	0.317	0.0516	0.994	-0.003
Pressure LASSO Frozen	0.824 (0.792–0.850)	0.345	0.0504	1.233	-0.221
Pressure LASSO Refitted	0.830 (0.800–0.853)	0.344	0.0507	0.971	0.198
Dynamic XGBoost Frozen	0.827 (0.799–0.851)	0.349	0.0512	1.331	-0.556
Dynamic XGBoost Refitted	0.814 (0.784–0.839)	0.342	0.0503	1.129	-0.008
Pressure XGBoost Frozen	0.842 (0.812–0.864)	0.372	0.0490	1.291	-0.217
Pressure XGBoost Refitted	0.849 (0.823–0.871)	0.376	0.0491	1.057	0.231

Frozen models were fitted in the previous diagnostic cohort. Refitted models were developed in the expanded cohort. There were 310 deaths (6.4%). Confidence intervals use 1000 hospital-cluster bootstrap resamples. AUPRC is average precision. Slope is the logistic calibration slope; CITL is calibration-in-the-large with slope fixed at one. Lower Brier scores are better.

In the same expanded external patients, the frozen diagnosis-cohort pressure-augmented XGBoost model achieved an AUROC of 0.842 and Brier score of 0.0490. Refitting changed AUROC by 0.0076 (95% CI − 0.0041 to 0.0224) and Brier score by 0.00010 (− 0.00095 to 0.00122); both intervals included zero (Table 3; Fig. 4). In the original 597-patient diagnostic cohort, corresponding AUROCs were 0.726 and 0.718.

Table 3
Paired model differences in the expanded eICU cohort at 24 hours

Comparison	AUROC difference (95% CI)	Brier difference (95% CI)
Clinical LASSO Refitted minus frozen	0.0010 (-0.0068–0.0108)	-0.00229 (-0.00343–-0.00109)
Pressure LASSO Refitted minus frozen	0.0051 (-0.0005–0.0125)	0.00031 (-0.00030–0.00096)
Dynamic XGBoost Refitted minus frozen	-0.0121 (-0.0254–0.0011)	-0.00097 (-0.00220–0.00032)
Pressure XGBoost Refitted minus frozen	0.0076 (-0.0041–0.0224)	0.00010 (-0.00095–0.00122)
Pressure minus clinical LASSO	0.0510 (0.0362–0.0679)	-0.00088 (-0.00170–-0.00005)
Pressure minus dynamic XGBoost	0.0350 (0.0253–0.0472)	-0.00118 (-0.00168–-0.00073)

All comparisons use the same patients. Positive AUROC differences and negative Brier differences favour the first model. Pressure augmentation includes three pressure values and an availability indicator. Intervals are exploratory and not adjusted for multiplicity.

Discrimination and calibration by ascertainment group

The refitted pressure-augmented XGBoost model achieved AUROCs of 0.698 in diagnosis-only, 0.837 in haemodynamics-only and 0.791 in overlapping patients, compared with 0.849 overall (Table 5). Corresponding frozen-model estimates were 0.703, 0.830 and 0.812, compared with 0.842 overall. In diagnosis-only patients, the refitted calibration slope was 0.782 (95% CI 0.595–0.997) and calibration-in-the-large was 0.355 (0.118–0.569). The overlapping group contained 10 deaths and had wider intervals. Within-group pairs comprised 59.8% of all event–non-event pairs with weighted concordance 0.830; between-group pairs comprised 40.2% with concordance 0.878. Their weighted sum reproduced overall AUROC (Table S26).

Table 5
Ascertainment-group performance of pressure-augmented XGBoost at 24 hours.

Population	Patients/ deaths	Model	AUROC (95% CI)	Slope (95% CI)	CITL (95% CI)
All eligible	4830/310	Frozen	0.842 (0.812 to 0.864)	1.291 (1.170 to 1.412)	−0.217 (− 0.352 to − 0.092)
All eligible	4830/310	Refitted	0.849 (0.823 to 0.871)	1.057 (0.973 to 1.149)	0.231 (0.091 to 0.358)
Diagnosis only	510/103	Frozen	0.703 (0.657 to 0.744)	0.887 (0.656 to 1.161)	0.262 (0.036 to 0.464)
Diagnosis only	510/103	Refitted	0.698 (0.651 to 0.736)	0.782 (0.595 to 0.997)	0.355 (0.118 to 0.569)
Haemodynamics only	4233/197	Frozen	0.830 (0.788 to 0.862)	1.297 (1.153 to 1.449)	−0.409 (− 0.573 to − 0.239)
Haemodynamics only	4233/197	Refitted	0.837 (0.797 to 0.870)	1.090 (0.981 to 1.197)	0.158 (− 0.030 to 0.337)
Both	87/10	Frozen	0.812 (0.608 to 0.967)	1.742 (0.613 to 3.982)	0.337 (− 0.262 to 0.776)
Both	87/10	Refitted	0.791 (0.606 to 0.951)	1.033 (0.364 to 2.468)	0.726 (− 0.029 to 1.216)

Frozen and refitted predictions refer to the same external patients. AUROC intervals come from the original hospital-cluster evaluation; calibration intervals use 1000 whole-hospital resamples of fixed predictions. Ideal slope is one and ideal CITL is zero. The overlapping group contained 10 deaths and had wider intervals. Full 72-hour calibration results appear in Table S25.

Pressure information and model explanations

Adding pressure measurements and their availability indicator increased external AUROC by 0.0510 (95% CI 0.0362–0.0679) for LASSO and 0.0350 (0.0253–0.0472) for XGBoost within the expanded cohorts.

Corresponding Brier differences were − 0.00088 (− 0.00170 to − 0.00005) and − 0.00118 (− 0.00168 to − 0.00073).

At 24 hours, the clinical–dynamic, availability-only, pressure-value and full pressure specifications achieved AUROCs of 0.814, 0.836, 0.834 and 0.849. Adding availability increased AUROC by 0.0217 (95% CI 0.0121–0.0332); the full specification improved on availability alone by 0.0133 (0.0055–0.0216). At 72 hours, the four AUROCs were 0.781, 0.806, 0.804 and 0.812. The full-minus-availability AUROC difference was 0.0057 (− 0.0022 to 0.0132); corresponding AUPRC and Brier differences were 0.0157 (0.0021–0.0303) and − 0.00159 (− 0.00260 to − 0.00054). Among patients with all three PAP components observed, full-minus-base AUROC differences were 0.0278 (0.0180–0.0392) at 24 hours and 0.0219 (0.0102–0.0322) at 72 hours (Tables S27–S28).

Blood urea nitrogen had the largest mean absolute SHAP contribution in both 300-patient samples. In eICU, the next highest-ranked inputs were bilirubin missingness, pressure-recording availability, glucose, diastolic PAP, age, sodium and respiratory rate (Fig. 5; Additional file 4). Among observed measurements, quartile-median SHAP contributions increased with blood urea nitrogen, glucose, diastolic PAP, age and respiratory rate. High lactate values also had positive contributions, while bicarbonate and sodium showed nonuniform relationships across their observed ranges (Figs. S10–S11).

In the pressure-augmented LASSO model, fitted odds ratios were 1.27 per 10-year increase in age, 1.50 per 5/min increase in respiratory rate, 1.15 per 10 mg/dL increase in blood urea nitrogen, 1.29 per 1 mmol/L increase in lactate and 0.67 per 1 g/dL increase in albumin. Diastolic PAP had an odds ratio of 1.36 per 5 mmHg increase. These penalised associations condition on the other model inputs. Additional file 4 supplies the complete equations and development imputation values.

Threshold and landmark sensitivity analyses

The repeated 25-mmHg threshold identified 3886 MIMIC-IV patients with 458 deaths and 2800 eICU patients with 268 deaths. The broader single-hour definition identified 7002 patients with 532 deaths and 6640 with 347 deaths, respectively (Additional file 1).

Excluding the three cardiothoracic-labelled eICU unit categories left 1881 patients with 168 deaths. Pressure-augmented XGBoost achieved AUROCs of 0.848 for the expanded-cohort model and 0.838 for the frozen model, with a paired difference of 0.0097 (95% CI − 0.0053 to 0.0260).

The 72-hour cohorts comprised 2844 MIMIC-IV patients with 395 deaths and 2459 eICU patients with 241 deaths across 114 hospitals. Pressure-augmented XGBoost achieved an external AUROC of 0.812 (95% CI 0.787–0.837) and Brier score of 0.0755. Its refitted-minus-frozen AUROC difference was 0.0142 (− 0.0008 to 0.0287). Corresponding differences for clinical LASSO, pressure LASSO and dynamic XGBoost were 0.0209 (0.0035–0.0384), 0.0196 (0.0036–0.0343) and 0.0194 (0.0032–0.0353), respectively (Table S4).

Diagnostic timing and general ICU benchmark

Restricting eICU diagnostic eligibility to records entered before the landmark retained 4663 patients with 264 deaths at 24 hours and 2390 with 223 deaths at 72 hours. The fixed expanded-cohort pressure model achieved AUROCs of 0.844 (0.816 to 0.870) and 0.813 (0.791 to 0.835), respectively. At 24 hours, diagnosis-only patients with timely documentation numbered 343 with 57 deaths (16.6%); 165 patients had only nonnegative records at or after the landmark, with 46 deaths (27.9%). Another two patients had negative-offset qualifying records and no nonnegative record before 24 hours; neither died. Including negative offsets yielded a timely 24-hour union of 4665 patients with 264 deaths (Tables S29–S30).

Valid APACHE IV and IVa hospital-mortality predictions covered 4321/4830 patients (89.5%) at 24 hours and 2182/2459 (88.7%) at 72 hours. In the 24-hour matched patients, pressure-augmented XGBoost and APACHE IVa achieved AUROCs of 0.842 and 0.829, with a paired difference of 0.0131 (− 0.0186 to 0.0430). The Brier difference was − 0.00378 (− 0.00660 to − 0.00025). At 72 hours, corresponding AUROCs were 0.810 and 0.762, with a paired difference of 0.0488 (0.0146 to 0.0827). The 72-hour comparison used updated model inputs and first-day APACHE predictions. Both APACHE versions and all four model comparisons appear in Tables S31–S33.

Missing data sensitivity results

Multiple imputation yielded external clinical/PAP LASSO AUROCs of 0.767/0.814 at 24 hours (median references 0.779/0.830) and 0.718/0.757 at 72 hours (references 0.737/0.784). Native-NaN XGBoost AUROCs across indicator variants were 0.817–0.821 for dynamic clinical and 0.848–0.852 for PAP models at 24 hours; corresponding 72-hour ranges were 0.773–0.776 and 0.808–0.809. For 24-hour PAP XGBoost retaining indicators, the paired AUROC change was 0.0021 (95% CI − 0.0069 to 0.0100). Paired differences, calibration estimates and diagnostics appear in Tables S19–S24 and Figs. S13–S16.

Predictor expansion and compact models

The pressure-augmented XGBoost reference retained the highest AUROC among the models evaluated after predictor expansion at 24 hours. Adding BMI, CVP or documented history separately yielded AUROCs of 0.841, 0.840 and 0.843. The treatment-enhanced model achieved 0.813, with a paired difference from the pressure reference of − 0.0366 (95% confidence interval − 0.0549 to − 0.0205). At 72 hours, adding history increased AUROC by 0.0068 (nominal 95% confidence interval 0.0006–0.0134), while the Brier score point estimate was higher.

Both compact models retained seven predictors. The 24-hour model used age, heart rate, respiratory rate, oxygen saturation, blood urea nitrogen, white blood cell count and bicarbonate; sodium replaced oxygen saturation at 72 hours. External AUROCs were 0.691 (95% CI 0.654–0.721) and 0.647 (0.607–0.685), respectively. Paired AUROC differences from dynamic clinical XGBoost were − 0.1238 (− 0.1517 to − 0.0979) and − 0.1334 (− 0.1740 to − 0.0929). The seven-input models reduced input burden and had lower discrimination. Nomograms, exact equations, calibration and all size-constrained comparisons appear in Additional files 2 and 3.

Discussion

PH diagnostic records and repeated pulmonary artery pressure elevation identified clinically distinct ICU populations in both databases. Haemodynamics-only patients were concentrated in cardiovascular and cardiothoracic care and had lower mortality than diagnosis-only patients. The pressure-augmented model achieved an external AUROC of 0.849 overall, with different performance across ascertainment groups. At 24 hours, the full pressure specification added predictive information beyond recording availability. These findings connect mortality prediction to the patients captured by each data source and the clinical context of monitoring.

Repeated pressure monitoring identified 3482 additional patients in MIMIC-IV and 4233 in eICU. These patients had repeated pressure elevation during acute care, while the diagnostic groups reflected documented PH. Their differences in clinical setting and mortality give the expanded cohort its prognostic structure. The ASPIRE registry similarly documented variation in hospital survival across medical, surgical and obstetric critical care admissions in patients with known PH [5]. Preserving the distinction between documented disease and monitored pressure elevation clarifies the population to which each estimate applies.

Pooled discrimination combined risk ordering within ascertainment groups with comparisons between groups. Within-group pairs accounted for 59.8% of event–non-event pairs with weighted concordance 0.830; between-group pairs accounted for 40.2% with concordance 0.878. Diagnosis-only patients had lower discrimination and a calibration slope below one. Reporting subgroup performance alongside the pooled estimate therefore makes clinically relevant differences visible. Restricting diagnoses to records entered before the landmark retained similar pooled discrimination while changing cohort mortality and composition.

The full pressure specification added predictive information beyond recording availability at 24 hours. Availability increased AUROC by 0.0217, and the full pressure specification added 0.0133 (0.0055 to 0.0216) beyond availability. At 72 hours, the additional AUROC difference was 0.0057 (− 0.0022 to 0.0132), while AUPRC and Brier differences favoured the full specification. Complete-PAP subset comparisons also favoured pressure augmentation. Together, these findings link mortality prediction to measured pressures and the context in which measurements are acquired. Testing patterns can carry prognostic information alongside measured values [27].

The predictors connect mortality risk to renal, metabolic and respiratory status alongside pulmonary haemodynamics. Higher blood urea nitrogen distinguished patients who died from survivors in both databases and ranked first in the SHAP samples. Respiratory rate and lactate had positive penalised associations, whereas albumin and oxygen saturation had negative associations. This pattern accords with evaluating systemic illness alongside pulmonary pressure in critically ill patients with PH [2]. Prior PAH cohorts linked renal dysfunction to adverse haemodynamics and mortality, and lower albumin to poorer survival [28, 29]; ICU studies identified lactate and oxygenation as prognostic markers [5, 6]. These clinical domains help explain the information used by the fitted models.

Several predictors overlap with those in the nomogram reported by Zhang and colleagues, which included age, respiratory rate, red cell distribution width, glucose and SAPS II [12]. The present study evaluates how repeated pressure elevation changes the population under assessment and what additional prediction is supplied by pressure-augmented models beyond recording availability. Paired comparisons in identical external patients further distinguish the refitting contrast from differences in case mix [10].

Discrimination and calibration provide complementary measures of prognostic performance [24]. Clinical LASSO had lower AUROC and calibration-in-the-large close to zero; pressure-augmented XGBoost had stronger discrimination and overall underprediction. In the APACHE IVa-matched 24-hour patients, the Brier difference favoured pressure-augmented XGBoost at − 0.00378 (− 0.00660 to − 0.00025). The paired AUROC difference was 0.0131 (− 0.0186 to 0.0430), with an interval that included zero. This benchmark distinguishes a reduction in overall prediction error from uncertainty about improved risk ranking.

LASSO selection and SHAP attribution explain different aspects of the models. A zero LASSO coefficient marks a predictor omitted at the selected penalty; SHAP quantifies contributions to fitted tree predictions [19, 26]. Sodium had a zero coefficient in both primary LASSO models and ranked seventh in the external SHAP sample. The contributions of pressure availability and bilirubin missingness reflect information carried by monitoring and laboratory-recording patterns. Regression nomograms display the penalised equations, and the supplied coefficients and imputation values permit exact reproduction.

Pressure-augmented specifications retained higher discrimination than the corresponding clinical specifications across all evaluated missing-data strategies. Multiple imputation yielded lower LASSO AUROC point estimates at both landmarks, and native-NaN XGBoost produced no consistent gain. The persistent model ordering supports the predictive contribution of pressure-related inputs across these preprocessing approaches. Each comparison evaluates the complete pipeline, including imputation and recording patterns.

Predictor expansion and simplification quantified the relationship between input burden and external performance. At 24 hours, adding BMI, history, CVP or recorded treatments yielded lower AUROC point estimates than the pressure reference. Seven-input clinical models achieved external AUROCs of 0.691 at 24 hours and 0.647 at 72 hours. These results support retaining the complete models as the main specifications. REVEAL Lite 2 illustrates the value of a compact score whose inputs capture information for a defined PAH outcome [7]. In the ICU setting evaluated here, simplification reduced discrimination and identifies a concrete target for further model development.

The two landmarks address distinct prognostic decisions: hospital mortality assessment after the first ICU day and updated assessment among patients remaining after three days. Their eligibility criteria produce different populations and event rates. Estimating the incremental value of later measurements would require comparison of early and updated inputs in the same patients eligible at 72 hours. This

follows the landmark principle of defining prediction within the population still at risk at the assessment time [30].

The study combines parallel ascertainment across two databases with paired evaluation in identical external patients. Preprocessing was repeated within training folds, and external uncertainty was estimated by resampling entire hospitals. For respiratory critical care research, these analyses establish the population and monitoring information represented by a mortality estimate. Code, coefficient tables, aggregate outputs and supplementary protocols document case identification, pressure processing and the complete comparisons, supporting reproduction and extension of the analyses.

The estimates apply to adults alive and remaining in ICU at the specified landmark. ICU monitoring captures pressure under acute clinical conditions; standardised resting and end-expiratory measurements, complete PAWP/PVR data and clinical adjudication are needed to classify chronic PH and its subtypes. Diagnostic timestamps describe documentation availability in eICU, while current-admission MIMIC coding lacks timestamps. The observational analyses quantify predictive associations; model coefficients and SHAP contributions do not estimate effects of intervening on individual predictors. Cohort extension and supplementary comparisons were exploratory in an eICU database previously examined in this project. Evaluation in new, clinically adjudicated cohorts with a defined risk-directed action is the next step for assessing transportability and patient benefit.

Conclusions

PH diagnostic records and repeated pulmonary artery pressure elevation identify ICU populations with distinct clinical characteristics, mortality and prediction profiles. Pressure-augmented models add predictive information beyond recording availability at 24 hours. Paired external comparisons distinguish the observed refitting contrast from differences in the patients assessed. Prognostic evaluation in respiratory critical care should report case definitions, monitoring context and performance within the populations represented.

Abbreviations

APACHE
Acute Physiology and Chronic Health Evaluation
AUROC
area under the receiver operating characteristic curve
AUPRC
area under the precision–recall curve
BMI
body mass index
CITL
calibration-in-the-large

CVP
central venous pressure
ICD
International Classification of Diseases
ICU
intensive care unit
LASSO
least absolute shrinkage and selection operator
mPAP
mean pulmonary artery pressure
PAC
pulmonary artery catheter
PAH
pulmonary arterial hypertension
PAP
pulmonary artery pressure
PAWP
pulmonary artery wedge pressure
PH
pulmonary hypertension
PVR
pulmonary vascular resistance
SHAP
Shapley additive explanations
XGBoost
extreme gradient boosting.

Declarations

Ethics approval and consent to participate

This study used deidentified data from MIMIC-IV version 3.1 and eICU-CRD version 2.0. The MIMIC-IV data-sharing initiative was approved by the Institutional Review Board of Beth Israel Deaconess Medical Center with a waiver of individual informed consent. eICU-CRD was deidentified before release in accordance with the HIPAA safe-harbor standard. Access to both databases required completion of approved human-subjects research training and acceptance of the PhysioNet data-use agreements. This secondary analysis involved no participant contact or identifiable information; therefore, individual consent to participate was not required for use of the released datasets.

Consent for publication

Not applicable.

Clinical trial number

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Additional files

Additional file 1 (.docx). Supplementary methods, tables and figures. Operational definitions, cohort flow, secondary model results, missingness, regression nomograms and pressure dependence plots.

Additional file 2 (.zip). Reproducibility materials. Analysis scripts, aggregate results, coefficient tables and a research-only offline calculator. Restricted patient-level data and fitted model objects are excluded.

Additional file 3 (.docx). Exploratory predictor expansion and model simplification. Supplementary methods, Tables S9–S12 and Figures S5–S9 covering BMI, history, CVP, recorded treatments and compact logistic models.

Additional file 4 (.docx). Clinical profiles and predictor interpretation. Tables S13–S18 provide complete ascertainment, outcome and database profiles with missingness, primary LASSO specifications, SHAP rankings and observed quartile summaries. Figures S10–S11 show predictor dependence plots, and Figure S12 shows the 72-hour selection flow.

Additional file 5 (.docx). Missing data sensitivity analyses. Posterior multiple-imputation LASSO ensembles and native-NaN XGBoost comparisons at 24 and 72 hours, with paired external evaluation, calibration, imputation diagnostics and selection stability in Tables S19–S24 and Figs. S13–S16.

Additional file 6 (.docx). Ascertainment, pressure information and general ICU benchmark analyses. Tables S25–S34 and Figures S17–S21 report subgroup calibration, event–non-event pair decomposition, pressure-information comparisons, diagnosis timing, APACHE IV/IVa benchmarks and exploratory threshold-specific net benefit. Complete aggregate results and reporting-item mappings accompany Additional file 2.

Funding

This work was supported by the Research Startup Fund of Guangzhou Eighth People's Hospital, Guangzhou Medical University.

Author Contribution

Shenpeng Zhou contributed to conceptualization, methodology, software, data curation, formal analysis, visualization and preparation of the original draft. Lixian Xu, Peilian Wu, Junyan Han, Meiqi Li, Xiaoying Zhang, Zhenxiang Li, Shiwen Li and Fei Peng contributed to clinical interpretation, validation and critical revision of the manuscript. Mingmei Xiong contributed to conceptualization, supervision, project administration, interpretation and critical revision of the manuscript. Shenpeng Zhou and Mingmei Xiong verified the underlying data and analyses. All authors read and approved the final manuscript.

Acknowledgements

We thank the teams maintaining MIMIC-IV and eICU-CRD and the clinicians whose routine documentation made these research resources possible.

Data Availability

The datasets analysed during the current study are available to credentialed users through PhysioNet: MIMIC-IV version 3.1 (https://physionet.org/content/mimiciv/3.1/) and eICU-CRD version 2.0 (https://physionet.org/content/eicu-crd/2.0/). Access requires completion of the required training and acceptance of the applicable data-use agreements.

References

1. Humbert M, Kovacs G, Hoeper MM, Badagliacca R, Berger RMF, Brida M, et al. 2022 ESC/ERS Guidelines for the diagnosis and treatment of pulmonary hypertension. Eur Heart J. 2022;43:3618–731. https://doi.org/10.1093/eurheartj/ehac237.
2. Hoeper MM, Benza RL, Corris P, de Perrot M, Fadel E, Keogh AM, et al. Intensive care, right ventricular support and lung transplantation in patients with pulmonary hypertension. Eur Respir J. 2019;53:1801906. https://doi.org/10.1183/13993003.01906-2018.
3. Kovacs G, Bartolome S, Denton CP, Gatzoulis MA, Gu S, Khanna D, et al. Definition, classification and diagnosis of pulmonary hypertension. Eur Respir J. 2024;64:2401324. https://doi.org/10.1183/13993003.01324-2024.
4. Huynh TN, Weigt SS, Sugar CA, Shapiro S, Kleerup EC. Prognostic factors and outcomes of patients with pulmonary hypertension admitted to the intensive care unit. J Crit Care. 2012;27:e7397–13. https://doi.org/10.1016/j.jcrc.2012.08.006.

5. Bauchmuller K, Condliffe R, Southern J, Billings C, Charalampopoulos A, Elliot CA, et al. Critical care outcomes in patients with pre-existing pulmonary hypertension: insights from the ASPIRE registry. ERJ Open Res. 2021;7:00046–2021. https://doi.org/10.1183/23120541.00046-2021.
6. Naranjo M, Mercurio V, Hassan H, Alturaif N, Cuomo A, Attanasio U, et al. Causes and outcomes of ICU hospitalisations in patients with pulmonary arterial hypertension. ERJ Open Res. 2022;8:00002–2022. https://doi.org/10.1183/23120541.00002-2022.
7. Benza RL, Kanwar MK, Raina A, Scott JV, Zhao CL, Selej M, et al. Development and Validation of an Abridged Version of the REVEAL 2.0 Risk Score Calculator, REVEAL Lite 2, for Use in Patients With Pulmonary Arterial Hypertension. Chest. 2021;159:337–46. https://doi.org/10.1016/j.chest.2020.08.2069.
8. Hoeper MM, Pausch C, Olsson KM, Huscher D, Pittrow D, Grünig E, et al. COMPERA 2.0: a refined four-stratum risk assessment model for pulmonary arterial hypertension. Eur Respir J. 2022;60:2102311. https://doi.org/10.1183/13993003.02311-2021.
9. Riley RD, Archer L, Snell KIE, Ensor J, Dhiman P, Martin GP, et al. Evaluation of clinical prediction models (part 2): how to undertake an external validation study. BMJ. 2024;384:e074820. https://doi.org/10.1136/bmj-2023-074820.
10. Vergouwe Y, Moons KG, Steyerberg EW. External validity of risk models: Use of benchmark values to disentangle a case-mix effect from incorrect coefficients. Am J Epidemiol. 2010;172:971–80. https://doi.org/10.1093/aje/kwq223.
11. Deng H, Wang P, Liu M. The prognostic value of pulmonary hypertension in intensive care unit patients from Beth Israel Deaconess Medical Center (BIDMC). J Cardiothorac Surg. 2025;20:29. https://doi.org/10.1186/s13019-024-03301-w.
12. Zhang CY, Hu YS, Meng ZY, Lu CH, Xie YF, Yu Q, et al. Development and validation of a mortality predictive model for ICU patients with primary pulmonary hypertension. Sci Rep. 2024;14:31497. https://doi.org/10.1038/s41598-024-83139-3.
13. Johnson AEW, Bulgarelli L, Shen L, Gayles A, Shammout A, Horng S, et al. MIMIC-IV, a freely accessible electronic health record dataset. Sci Data. 2023;10:1. https://doi.org/10.1038/s41597-022-01899-x.
14. Johnson A, Bulgarelli L, Pollard T, Gow B, Moody B, Horng S et al. MIMIC-IV (version 3.1). PhysioNet. 2024. https://doi.org/10.13026/kpb9-mt58
15. 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. Sci Data. 2018;5:180178. https://doi.org/10.1038/sdata.2018.178.
16. Pollard T, Johnson A, Raffa J, Celi LA, Badawi O, Mark R. eICU Collaborative Research Database (version 2.0). PhysioNet. 2019. https://doi.org/10.13026/C2WM1R
17. Collins GS, Moons KGM, Dhiman P, Riley RD, Beam AL, Van Calster B, et al. TRIPOD + AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ. 2024;385:e078378. https://doi.org/10.1136/bmj-2023-078378.

18. Benchimol EI, Smeeth L, Guttmann A, Harron K, Moher D, Petersen I, et al. The REporting of studies Conducted using Observational Routinely-collected health Data (RECORD) Statement. PLoS Med. 2015;12:e1001885. https://doi.org/10.1371/journal.pmed.1001885.
19. Tibshirani R. Regression shrinkage and selection via the lasso. J R Stat Soc Ser B Methodol. 1996;58:267–88. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x.
20. Chen T, Guestrin C. XGBoost: a scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM; 2016. pp. 785–794. https://doi.org/10.1145/2939672.2939785
21. Varma S, Simon R. Bias in error estimation when using cross-validation for model selection. BMC Bioinformatics. 2006;7:91. https://doi.org/10.1186/1471-2105-7-91.
22. 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:e0118432. https://doi.org/10.1371/journal.pone.0118432.
23. Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gonen M, Obuchowski N, et al. Assessing the performance of prediction models: a framework for traditional and novel measures. Epidemiology. 2010;21:128–38. https://doi.org/10.1097/ede.0b013e3181c30fb2.
24. Van Calster B, McLernon DJ, van Smeden M, Wynants L, Steyerberg EW. Topic Group Evaluating Diagnostic Tests and Prediction Models of the STRATOS Initiative. Calibration: the Achilles heel of predictive analytics. BMC Med. 2019;17:230. https://doi.org/10.1186/s12916-019-1466-7.
25. Vickers AJ, Elkin EB. Decision Curve Analysis: A Novel Method for Evaluating Prediction Models. Med Decis Mak. 2006;26:565–74. https://doi.org/10.1177/0272989X06295361.
26. 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. Nat Mach Intell. 2020;2:56–67. https://doi.org/10.1038/s42256-019-0138-9.
27. Agniel D, Kohane IS, Weber GM. Biases in electronic health record data due to processes within the healthcare system: retrospective observational study. BMJ. 2018;361:k1479. https://doi.org/10.1136/bmj.k1479.
28. Shah SJ, Thenappan T, Rich S, Tian L, Archer SL, Gomberg-Maitland M. Association of serum creatinine with abnormal hemodynamics and mortality in pulmonary arterial hypertension. Circulation. 2008;117:2475–83. https://doi.org/10.1161/circulationaha.107.719500.
29. Snipelisky D, Jentzer J, Batal O, Dardari Z, Mathier M. Serum albumin concentration as an independent prognostic indicator in patients with pulmonary arterial hypertension. Clin Cardiol. 2018;41:782–7. https://doi.org/10.1002/clc.22954.
30. van Houwelingen HC. Dynamic prediction by landmarking in event history analysis. Scand J Stat. 2007;34:70–85. https://doi.org/10.1111/j.1467-9469.2006.00529.x.
31. 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. Crit Care Med. 2006;34:1297–310. https://doi.org/10.1097/01.CCM.0000215112.84523.F0.

32. MIT Laboratory for Computational Physiology. eICU Collaborative Research Database diagnosis table documentation. https://eicu.mit.edu/eicutables/diagnosis/. Accessed 24 October 2026.
33. MIT Laboratory for Computational Physiology. eICU Collaborative Research Database apachePatientResult table documentation. https://eicu.mit.edu/eicutables/apachepatientresult/. Accessed 24 October 2026.

Figures

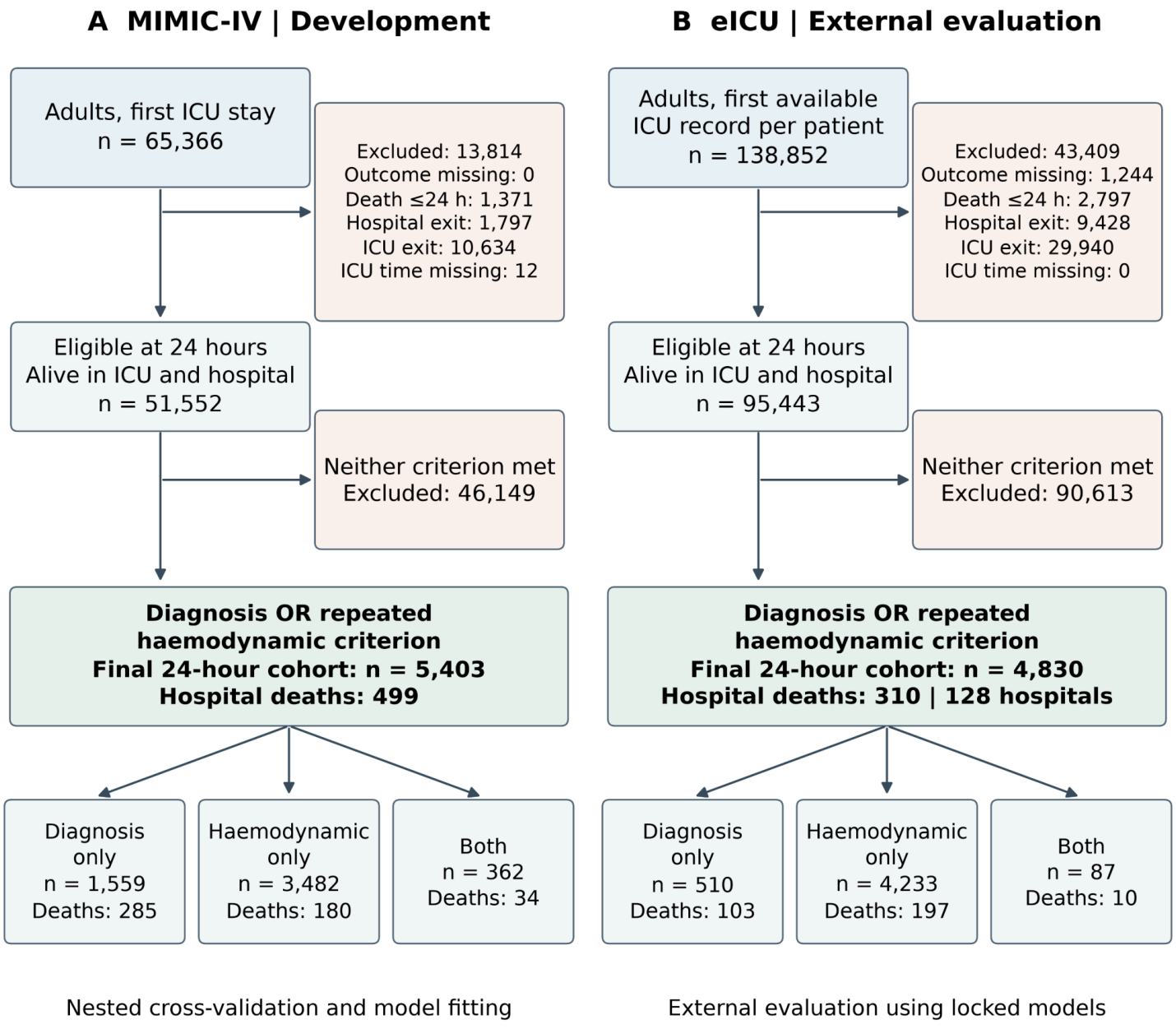

Figure 1

Patient selection for the primary 24-hour analysis. Starting populations contain adults with one ICU record per patient; MIMIC-IV uses the first ICU stay, whereas eICU uses the first available record under the specified ordering. Landmark exclusions are mutually exclusive and applied in the displayed order: unknown hospital outcome, recorded death at or before 24 hours, hospital exit at or before 24 hours, then ICU exit at or before 24 hours. Missing ICU duration is listed separately. Diagnostic ascertainment uses the specified codes or structured diagnosis entries. The repeated haemodynamic criterion requires at least three valid hourly mPAP measurements, a median above 20 mmHg, and at least two elevated hours separated by two or more hour indices. Patients meeting neither criterion include those with absent or insufficient pressure records and no qualifying diagnosis. The three final groups are mutually exclusive. Deaths are subsequent all-cause deaths during the index hospital admission. The 72-hour selection process is presented separately in Fig. S12.

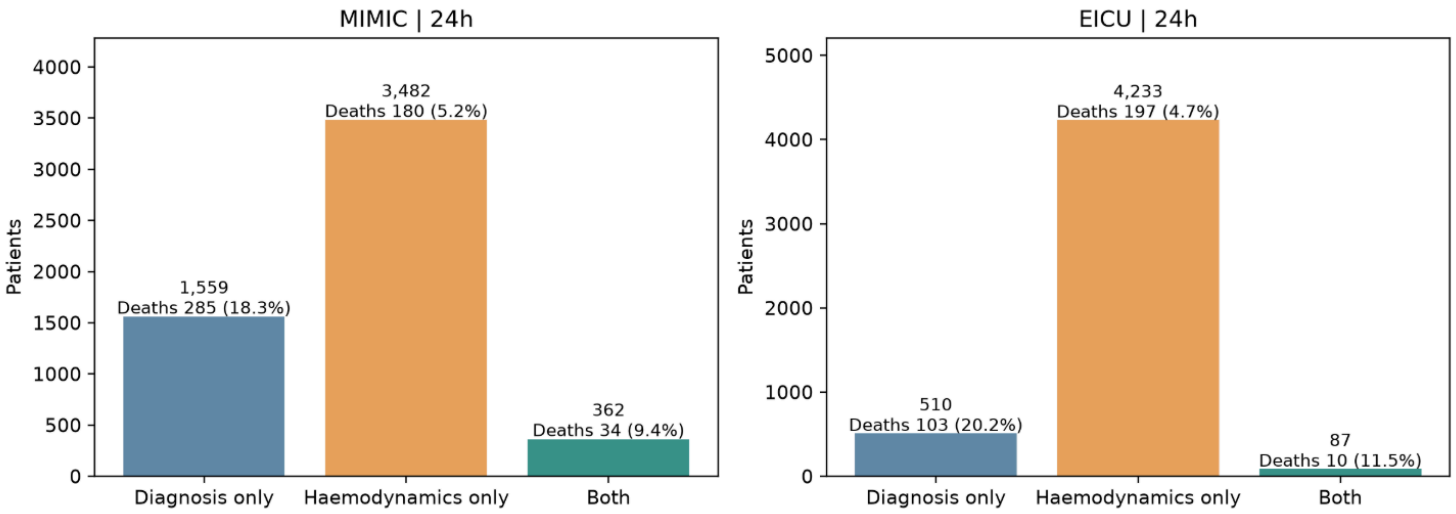

Figure 2

Ascertainment and hospital mortality at 24 hours. Mutually exclusive diagnosis-only, haemodynamics-only and overlapping groups are shown for MIMIC-IV and eICU. Labels provide patient numbers and hospital deaths. The repeated haemodynamic rule requires a median mPAP above 20 mmHg across at least three valid hours and at least two elevated hours separated by two hour indices. Diagnosis-only patients include those with absent or insufficient monitoring.

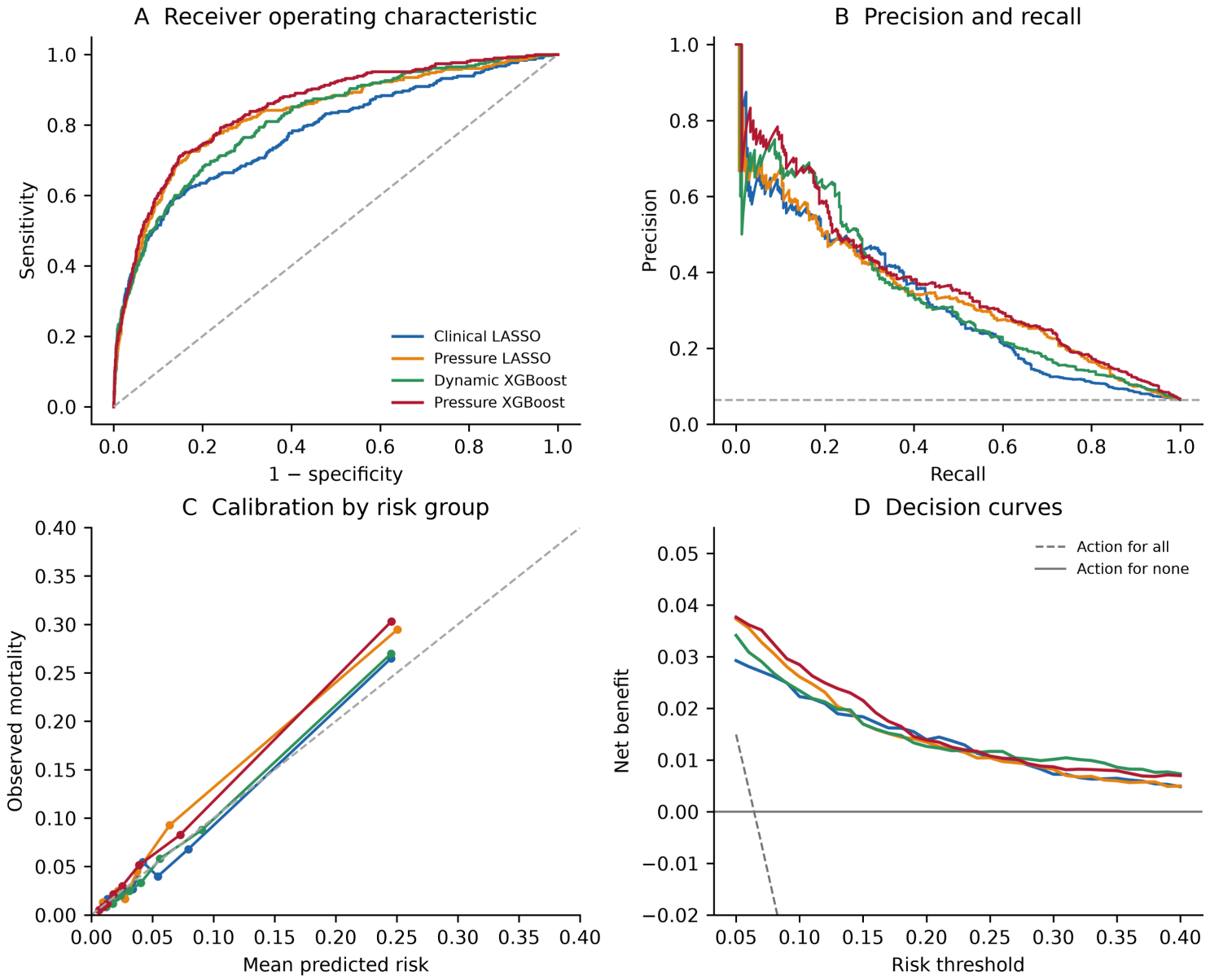

Figure 3

External evaluation of the four expanded-cohort models at 24 hours. Panels show receiver operating characteristic curves, precision–recall curves, calibration in eight predicted-risk groups and decision curves in 4830 eICU patients with 310 deaths. The grey dashed line in the precision–recall panel marks the event proportion of 0.064. The calibration diagonal indicates agreement. Decision curves compare hypothetical model-based action with action for all or no patients across exploratory thresholds. Pressure models include systolic, diastolic and mean pulmonary artery pressure and an availability indicator. Predictions retain their development calibration.

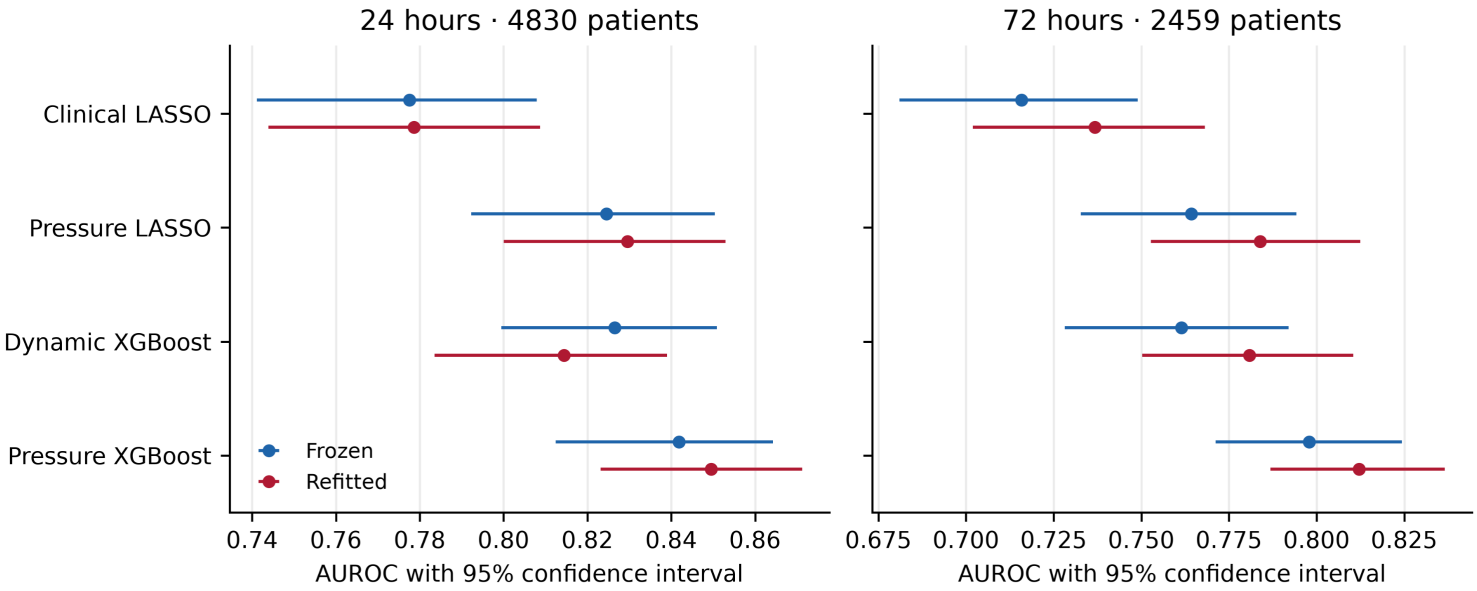

Figure 4

Frozen and refitted models evaluated in identical external populations. Points and horizontal intervals show AUROC and 95% hospital-cluster bootstrap confidence intervals at 24 and 72 hours. Frozen models were trained in the corresponding diagnostic cohorts and refitted models in the expanded cohorts. Paired differences are reported in Table 3 and Additional file 1. Each panel represents the population eligible at its specified landmark.

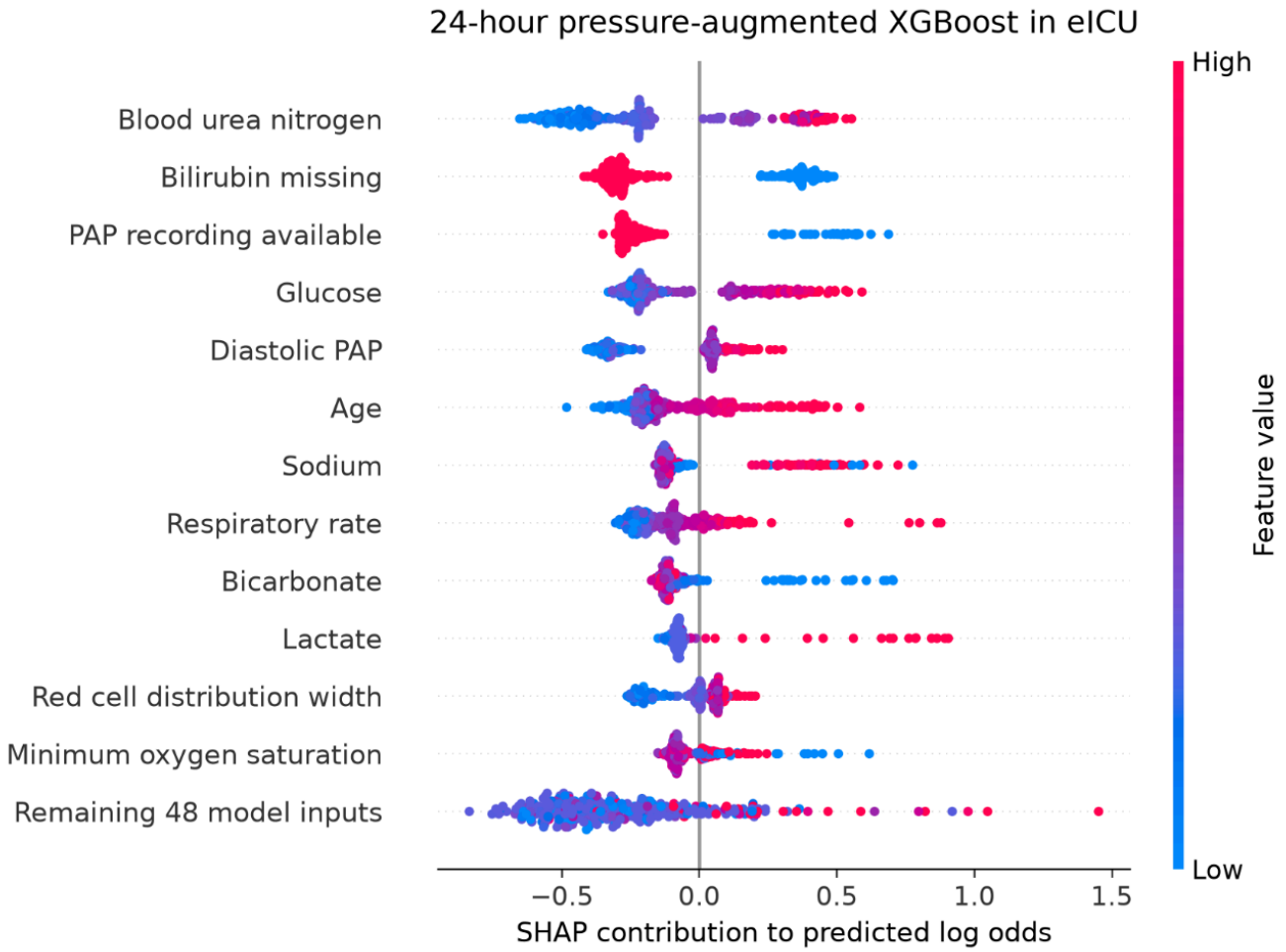

Figure 5

SHAP explanations for the 24-hour pressure-augmented XGBoost model. The analysis used a reproducible random sample of 300 expanded-cohort eICU patients. Each point represents a contribution to model log odds; positive values increase the predicted probability of death. Colour indicates feature value. Binary indicators describe pressure availability or laboratory missingness. BUN denotes blood urea nitrogen; PAPd, diastolic pulmonary artery pressure; RDW, red cell distribution width; and minimum SpO2, the minimum hourly oxygen saturation. The final row combines remaining contributions.

Supplementary Files

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

- Additionalfile1.docx
- Additionalfile3.docx
- Additionalfile4.docx
- Additionalfile6.docx

- Additionalfile5.docx
- Additionalfile2.zip