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.

How calibration fails when prediction models cross hospitals: logistic regression versus tree-based models for septic shock mortality

Honghong Duan
983 Hospital of Joint Logistic Support Force
Ying Zhang
983 Hospital of Joint Logistic Support Force
Haijin Suo
983 Hospital of Joint Logistic Support Force
Jingyi Ren
Xijing Hospital, The Fourth Military Medical University
Junjie Li
Xijing Hospital, The Fourth Military Medical University
Jing Yang
Xijing Hospital, The Fourth Military Medical University
Zelong Yang
youngderl@hotmail.com

Xijing Hospital, The Fourth Military Medical University

Research Article

Keywords: Mortality prediction, Logistic regression, External validation, Model calibration, Model selection bias

Posted Date: November 12th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11105693/v1

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

Additional Declarations: No competing interests reported.

How calibration fails when prediction models cross hospitals: logistic regression versus tree-based models for septic shock mortality

Honghong Duan[1#], Ying Zhang[1#], Haijin Suo[1], Jingyi Ren[2], Junjie Li[2*], Jing Yang[2*], Zelong Yang[2*]

1 Department of Emergency Medicine, 983 Hospital of Joint Logistic Support Force, Tianjin 300142, China

2 Department of Emergency Medicine, Xijing Hospital, The Fourth Military Medical University, Xi'an 710032, China

# These authors contributed equally to this work.

* Corresponding authors: Junjie Li, E-mail: lijunjie@fmmu.edu.cn; Jing Yang, E-mail: dryangjing@qq.com; Zelong Yang, E-mail: youngderl@hotmail.com

Abstract

Background: Clinical prediction models developed at single institutions rarely undergo rigorous external validation. Whether model calibration, the agreement between predicted and observed event rates, generalizes across populations with substantially different baseline risk remains poorly characterized.

Methods: We developed prediction models in MIMIC-IV (n=17,516) and externally validated them in eICU-CRD (n=2,819) using a pragmatic septic shock definition (suspected infection, organ dysfunction, and vasopressor use). Logistic regression (LR), random forest (RF), and XGBoost were compared on native probabilities using 72 predictors. We assessed discrimination (AUC), calibration (Brier score), and calibratability with small local samples (intercept correction and Platt scaling).

Results: Internal AUC ranking did not predict external performance. LR ranked second internally (0.820) but achieved the highest external AUC (0.813 vs. XGBoost 0.800, RF 0.759). LR's calibration slope stayed near 1.0, and its external error was a pure location shift that 50 local patients sufficed to correct. XGBoost's raw Brier of 0.185 reflected a coincidental match between its class-weight-centered predictions and eICU's 42.4% mortality, and its uncorrected calibration was prevalence-dependent in simulation. RF's error was a compression of predicted risks that a single intercept could not repair; only full recalibration reduced its Brier score (0.211 to 0.196).

Conclusions: When local recalibration data are scarce, the structure of calibration failure should guide model selection. LR's failure is visible and cheap to repair, XGBoost's is prevalence-dependent and may be invisible without local data, and RF's requires full recalibration rather than a single intercept.

Keywords: Mortality prediction, Logistic regression, External validation, Model calibration, Model selection bias

Background

Clinical prediction models are increasingly used to support decision-making in intensive care, yet most are developed and validated at single institutions [1,2], and even widely deployed commercial models have performed poorly under external validation [3]. External validation studies have repeatedly shown that discrimination, a model's ability to rank patients by risk, often generalizes across populations [4,5,6]. Calibration, the agreement between predicted and observed event rates, is less transportable [7,8].

These failures matter because clinical decisions depend on absolute risk. Treatment thresholds, resource allocation, and prognostic communication all draw on predicted probabilities, so a model that ranks patients well but systematically underestimates risk can guide decisions in a consistently wrong direction. Machine learning models are known to produce poorly calibrated probabilities even when discrimination is strong, and calibration can drift further after deployment. How such errors behave when a model crosses institutions with very different baseline risk, and whether they can be repaired with the small local samples a receiving hospital can realistically assemble, remains unclear. The Medical Information Mart for Intensive Care IV (MIMIC-IV) [1] and the eICU Collaborative Research Database (eICU-CRD) [9] provide a natural experiment for studying this phenomenon. Among patients identified with septic shock, MIMIC-IV has approximately 23% in-hospital mortality, while eICU-CRD has approximately 42%, a 1.8-fold difference.

We addressed three questions. First, does model discrimination persist across this baseline risk gap, and is internal AUC ranking a reliable guide to external performance? Second, what is the structure of calibration shift, and does it differ between model classes? Third, how sensitive is model selection to methodological choices in the evaluation pipeline?

Methods

Study population

This retrospective cohort study used two publicly available databases: MIMIC-IV (single tertiary academic medical center, Boston, USA; 2008-2019) and eICU-CRD (208 US hospitals; 2014-2015). As both databases are de-identified and publicly available, this study was exempt from additional ethics review. This study was conducted in accordance with the Declaration of Helsinki. We included adult patients (≥18 years) who developed septic shock during the ICU stay, defined by a pragmatic, database-specific operationalization approximating, but not fully implementing, the Sepsis-3 criteria: suspected infection with organ dysfunction, vasopressor therapy to maintain mean arterial pressure (MAP) ≥ 65 mm Hg, and serum lactate > 2 mmol/L [10]. We operationalized this as suspected infection with organ dysfunction plus vasopressor administration (vasopressor use as a proxy for the MAP criterion), without requiring serum lactate: the MIMIC-IV sepsis3 definition (suspected infection with SOFA ≥ 2), and infection diagnoses with organ dysfunction (APACHE IV acute physiology score ≥ 20) in eICU-CRD. Shock onset (T0) was the start of the first vasopressor infusion. ICU stays shorter than 24 hours were excluded, and each qualifying ICU admission was analyzed as a separate episode. The primary outcome was in-hospital mortality. The final cohort consisted of 17,516 ICU stays in MIMIC-IV and 2,819 in eICU-CRD (**Figure 1A**).

Predictor variables

We identified 87 candidate predictors from the 48 hours preceding shock onset (T0). In MIMIC-IV, T0 was confined to the ICU stay (patients who received vasopressors before ICU admission were excluded). After excluding redundant variables and those with high missingness, 72 predictors were retained, covering seven domains: demographics, comorbidities, infection sites, illness severity score, vital signs, laboratory parameters, and first-6-hour ICU trajectory variables.

To reduce multicollinearity, the 13 redundant candidate predictors were removed. Diastolic blood pressure was excluded because of 86.3% missingness in eICU-CRD. The SOFA score was omitted because its cardiovascular component overlaps with the cohort definition and its remaining components were already included as individual predictors.

Among the 30 first-6-hour trajectory variables, 26 were available in both databases. The four blood-pressure variables were 100% missing in eICU-CRD and were imputed to the MIMIC-IV training-set median. They were retained for their independent predictive signal in the development cohort. With constant imputation in eICU-CRD, they are inert rather than biased in external validation. The full definitions of the 72 predictors are given in **Table S1**.

Table S1. Complete 72-variable feature dictionary

#	Variable	Category	Type	Description
1	age	Demographics	Continuous	Patient age at ICU admission (years)
2	aids	Comorbidity	Binary	HIV/AIDS
3	apsiii_score	Severity score	Continuous	Acute Physiology Score III
4	bicarbonate_max	Lab value	Continuous	Maximum bicarbonate (mEq/L)
5	bilirubin_max	Lab value	Continuous	Maximum bilirubin (mg/dL)
6	bun_max	Lab value	Continuous	Maximum blood urea nitrogen (mg/dL)
7	chf	Comorbidity	Binary	Congestive heart failure
8	ckd	Comorbidity	Binary	Chronic kidney disease
9	copd	Comorbidity	Binary	Chronic pulmonary disease
10	cvd	Comorbidity	Binary	Cerebrovascular disease
11	dementia	Comorbidity	Binary	Dementia
12	diabetes	Comorbidity	Binary	Diabetes mellitus
13	glucose_max	Lab value	Continuous	Maximum glucose (mg/dL)
14	heart_rate_max	Vital sign	Continuous	Maximum heart rate (bpm)

15	height_cm	Demographics	Continuous	Height in centimeters
16	hemoglobin_max	Lab value	Continuous	Maximum hemoglobin (g/dL)
17	icu6h_bicarb_max	6-h window lab	Continuous	First-6-h maximum bicarbonate (mEq/L)
18	icu6h_bicarb_min	6-h window lab	Continuous	First-6-h minimum bicarbonate (mEq/L)
19	icu6h_bili_max	6-h window lab	Continuous	First-6-h maximum bilirubin (mg/dL)
20	icu6h_bun_max	6-h window lab	Continuous	First-6-h maximum blood urea nitrogen (mg/dL)
21	icu6h_bun_min	6-h window lab	Continuous	First-6-h minimum blood urea nitrogen (mg/dL)
22	icu6h_cr_max	6-h window lab	Continuous	First-6-h maximum creatinine (mg/dL)
23	icu6h_cr_min	6-h window lab	Continuous	First-6-h minimum creatinine (mg/dL)
24	icu6h_dbp_max	6-h window vital	Continuous	First-6-h maximum diastolic blood pressure (mmHg)
25	icu6h_dbp_min	6-h window vital	Continuous	First-6-h minimum diastolic blood pressure (mmHg)
26	icu6h_glu_max	6-h window lab	Continuous	First-6-h maximum glucose (mg/dL)
27	icu6h_glu_min	6-h window lab	Continuous	First-6-h minimum glucose (mg/dL)
28	icu6h_hgb_max	6-h window lab	Continuous	First-6-h maximum hemoglobin (g/dL)
29	icu6h_hgb_min	6-h window lab	Continuous	First-6-h minimum hemoglobin (g/dL)
30	icu6h_hr_max	6-h window vital	Continuous	First-6-h maximum heart rate (bpm)
31	icu6h_hr_min	6-h window vital	Continuous	First-6-h minimum heart rate (bpm)

32	icu6h_k_max	6-h window lab	Continuous	First-6-h maximum potassium (mEq/L)
33	icu6h_k_min	6-h window lab	Continuous	First-6-h minimum potassium (mEq/L)
34	icu6h_na_max	6-h window lab	Continuous	First-6-h maximum sodium (mEq/L)
35	icu6h_na_min	6-h window lab	Continuous	First-6-h minimum sodium (mEq/L)
36	icu6h_plt_max	6-h window lab	Continuous	First-6-h maximum platelets (×10^3/μL)
37	icu6h_plt_min	6-h window lab	Continuous	First-6-h minimum platelets (×10^3/μL)
38	icu6h_rr_max	6-h window vital	Continuous	First-6-h maximum respiratory rate (breaths/min)
39	icu6h_rr_min	6-h window vital	Continuous	First-6-h minimum respiratory rate (breaths/min)
40	icu6h_sbp_max	6-h window vital	Continuous	First-6-h maximum systolic blood pressure (mmHg)
41	icu6h_sbp_min	6-h window vital	Continuous	First-6-h minimum systolic blood pressure (mmHg)
42	icu6h_spo2_min	6-h window vital	Continuous	First-6-h minimum SpO2 (%)
43	icu6h_tempc_max	6-h window vital	Continuous	First-6-h maximum temperature (Celsius)
44	icu6h_tempc_min	6-h window vital	Continuous	First-6-h minimum temperature (Celsius)
45	icu6h_wbc_max	6-h window lab	Continuous	First-6-h maximum white blood cell count (×10^3/μL)
46	icu6h_wbc_min	6-h window lab	Continuous	First-6-h minimum white blood cell count (×10^3/μL)
47	inf_abdomen	Infection site	Binary	Abdominal infection

48	inf_bacteremia	Infection site	Binary	Bacteremia
49	inf_pneumonia	Infection site	Binary	Pneumonia infection
50	inf_uti	Infection site	Binary	Urinary tract infection
51	initial_lactate	Lab value	Continuous	Initial lactate (mmol/L)
52	liver_disease	Comorbidity	Binary	Liver disease
53	malignancy	Comorbidity	Binary	Malignancy (leukemia, lymphoma, metastatic solid tumor)
54	mechvent_at_t0	Vital sign	Binary	Mechanical ventilation at shock onset
55	mi	Comorbidity	Binary	Myocardial infarction
56	ne_aminoglycosides	Medication	Binary	Aminoglycoside exposure pre-shock
57	ne_contrast	Medication	Binary	Iodinated contrast exposure pre-shock
58	ne_vancomycin	Medication	Binary	Vancomycin exposure pre-shock
59	oasis_score	Severity score	Continuous	OASIS severity score
60	platelets_max	Lab value	Continuous	Maximum platelet count (×10^3/μL)
61	potassium_max	Lab value	Continuous	Maximum potassium (mEq/L)
62	pre_t0_max_creat	Lab value	Continuous	Pre-shock maximum creatinine (mg/dL)
63	pud	Comorbidity	Binary	Peptic ulcer disease
64	pvd	Comorbidity	Binary	Peripheral vascular disease
65	resp_rate_max	Vital sign	Continuous	Maximum respiratory rate (breaths/min)

66	rheumd	Comorbidity	Binary	Rheumatologic disease
67	sbp_max	Vital sign	Continuous	Maximum systolic blood pressure (mmHg)
68	sodium_max	Lab value	Continuous	Maximum sodium (mEq/L)
69	spo2_min	Vital sign	Continuous	Minimum SpO2 (%)
70	temp_c_max	Vital sign	Continuous	Maximum temperature (Celsius)
71	wbc_max	Lab value	Continuous	Maximum white blood cell count (×10^3/μL)
72	weight_kg	Demographics	Continuous	Weight in kilograms

Data cleaning

Respiratory rate was capped at 60 breaths per minute, and MIMIC-IV temperatures were converted from Fahrenheit to Celsius. Continuous predictors were rescaled to zero mean and unit standard deviation using the MIMIC training data, with the same scaling applied to all cohorts. OASIS is not comparable across the two databases, because MIMIC-IV derives it from continuous monitoring data and eICU-CRD from nursing shift summaries. Systolic blood pressure maximum was missing in 57.4% of MIMIC and 68.6% of eICU patients, and recorded eICU values reflected admission triage readings rather than true pre-shock maxima. We therefore added a sensitivity analysis excluding SBP max and OASIS, and external performance was essentially unchanged.

Missing data and model architecture

All 72 predictor variables were retained in the model. Missing values were imputed using the median from the MIMIC training set. Missingness rates by database are given in **Table S2**. Variables with high missingness and their cross-database comparability are addressed in the Limitations section. As a robustness check, models were retrained after excluding the two variables with known cross-database comparability problems (OASIS and SBP max).

Table S2. Missingness across all 72 variables

#	Variable	MIMIC missing (%)	eICU missing (%)
1	age	0.0	0.0
2-14	Comorbidities (13 variables)	0.0	0.0
15-18	Infection sites (4 variables)	0.0	0.0
19	height_cm	29.7	1.2
20	weight_kg	1.0	1.8
21	mechvent_at_t0	0.0	0.0
22	apsiii_score	0.0	8.5
23	oasis_score	0.0	0.0
24	hemoglobin_max	22.7	5.7
25	platelets_max	25.3	6.3
26	wbc_max	25.9	6.4
27	bicarbonate_max	33.9	5.1
28	bun_max	33.9	5.4
29	pre_t0_max_creat	33.9	5.4
30	glucose_max	58.4	5.5
31	sodium_max	22.4	4.8
32	potassium_max	21.9	4.9
33	bilirubin_max	60.7	20.3
34	initial_lactate	32.1	27.6
35	heart_rate_max	0.0	2.6
36	sbp_max	57.4	68.6
37	resp_rate_max	0.1	4.4
38	temp_c_max	15.6	7.9
39	spo2_min	0.0	19.3
40	ne_vancomycin	0.0	0.0

41	ne_aminoglycosides	0.0	0.0
42	ne_contrast	0.0	0.0
43-46	ICU first-6h SBP/DBP (4 variables)	49.3-78.2	100.0
47-72	ICU first-6h labs/vitals (26 variables)	3.0-64.3	1.5-65.2

The MIMIC-IV dataset was split into training (80%) and test (20%) sets using a fixed random seed of 42, with stratification on the mortality outcome. Three models were trained on the same training set.

Logistic regression with L1 regularization [11]. The regularization parameter C=0.0464 was selected by 5-fold stratified cross-validation on the MIMIC training set using ROC AUC as the scoring metric. No class_weight adjustment was applied.

Random forest [12]. Configured with 200 trees, maximum depth 6, minimum samples per leaf 10. class_weight="balanced_subsample" was applied within the random forest algorithm.

XGBoost [19]. Configured with 200 trees, maximum depth 4, learning rate 0.05. scale_pos_weight was set to the ratio of negative to positive training samples (3.3:1). RF and XGBoost hyperparameters were fixed at defaults, so neither tree-based model received a tuning advantage over logistic regression.

All three models were evaluated on their native predict_proba outputs. No additional calibration steps (Platt scaling, isotonic regression) were applied to any model in the primary comparison. To assess calibratability, intercept correction was applied to both LR and XGBoost by fitting a logistic regression with only the intercept term on randomly sampled eICU patients (n=30, 50, 75, 100, 200; 200 repetitions each). For XGBoost, local Platt scaling [13] was also evaluated using the same sample sizes. These experiments test whether both models are similarly calibratable when given the same amount of local calibration data.

Statistical analysis

External validation methodology followed published recommendations [14]. Discrimination was assessed using the area under the receiver operating characteristic curve (AUC) [15]. Calibration was evaluated using calibration curves (10 bins), the Brier score, and the Van Calster decomposition [8], which separates prediction error into calibration-in-the-large, calibration slope, calibration loss, refinement, and uncertainty. Calibration-in-the-large is the difference between observed and mean predicted mortality. The calibration slope was estimated by regressing the binary outcome on the logit of the predicted probability. Calibration loss is the part of the Brier score left after accounting for refinement and uncertainty.

Optimal thresholds were derived using Youden's index in each cohort. Decision curve analysis followed Vickers and Elkin [16] across thresholds of 1%–50%, computed on native predicted probabilities. For LR, feature importance was assessed with SHAP using LinearExplainer [17] on the MIMIC test set. Bootstrap resampling (500 iterations) was used for confidence interval estimation on key performance metrics. Subgroup analysis was performed across age, mechanical ventilation, and diabetes groups.

To isolate the effect of outcome prevalence on calibration, we created reweighted versions of the full eICU cohort (n=2,819) with target mortality rates from 10% to 70%. Weights were based on the outcome only, set to π/π_obs for deaths and $(1-\pi)/(1-\pi_obs)$ for survivors, where π is the target rate and π_obs the observed rate of 0.42. Characteristics within each outcome group were unchanged; only the death proportion varied. The effective sample size ranged from 1,971 to 2,819, and the maximum weight ratio was 6.6. Each scenario was repeated 200 times with bootstrap resampling.

Model selection can be swayed by choices in the analysis pipeline. We therefore tested the four choices made in the original analysis one at a time: trying multiple random seeds and keeping the best run; selecting the LR regularization parameter C on eICU data rather than by MIMIC-only cross-validation; applying class_weight="balanced" to LR; and calibrating RF and XGBoost with MIMIC-fitted Platt scaling while leaving LR uncalibrated. Each choice was added in turn to a fair

baseline in which all models were compared on native probabilities, and its individual and cumulative effect on the external model ranking was examined.

Following Van Calster et al. [8], the Brier score was decomposed into three components, namely calibration loss, refinement, and uncertainty. Calibration loss represents the component of prediction error attributable to miscalibration. Uncertainty is determined by outcome prevalence and represents the minimum achievable Brier score for a naive model predicting the population mean. Refinement captures the improvement from risk stratification beyond the population average. This decomposition shows whether changes in Brier score across datasets reflect genuine model degradation or are simply driven by baseline risk differences.

Software

All analyses were conducted in Python 3.12.3 using scikit-learn 1.6.1 [18], XGBoost 3.3.0 [19], SHAP 0.52.0 [17], pandas 2.3.3 [20], NumPy 2.4.6 [21], SciPy 1.18.0 [22], and Matplotlib 3.11.0 [23].

Results

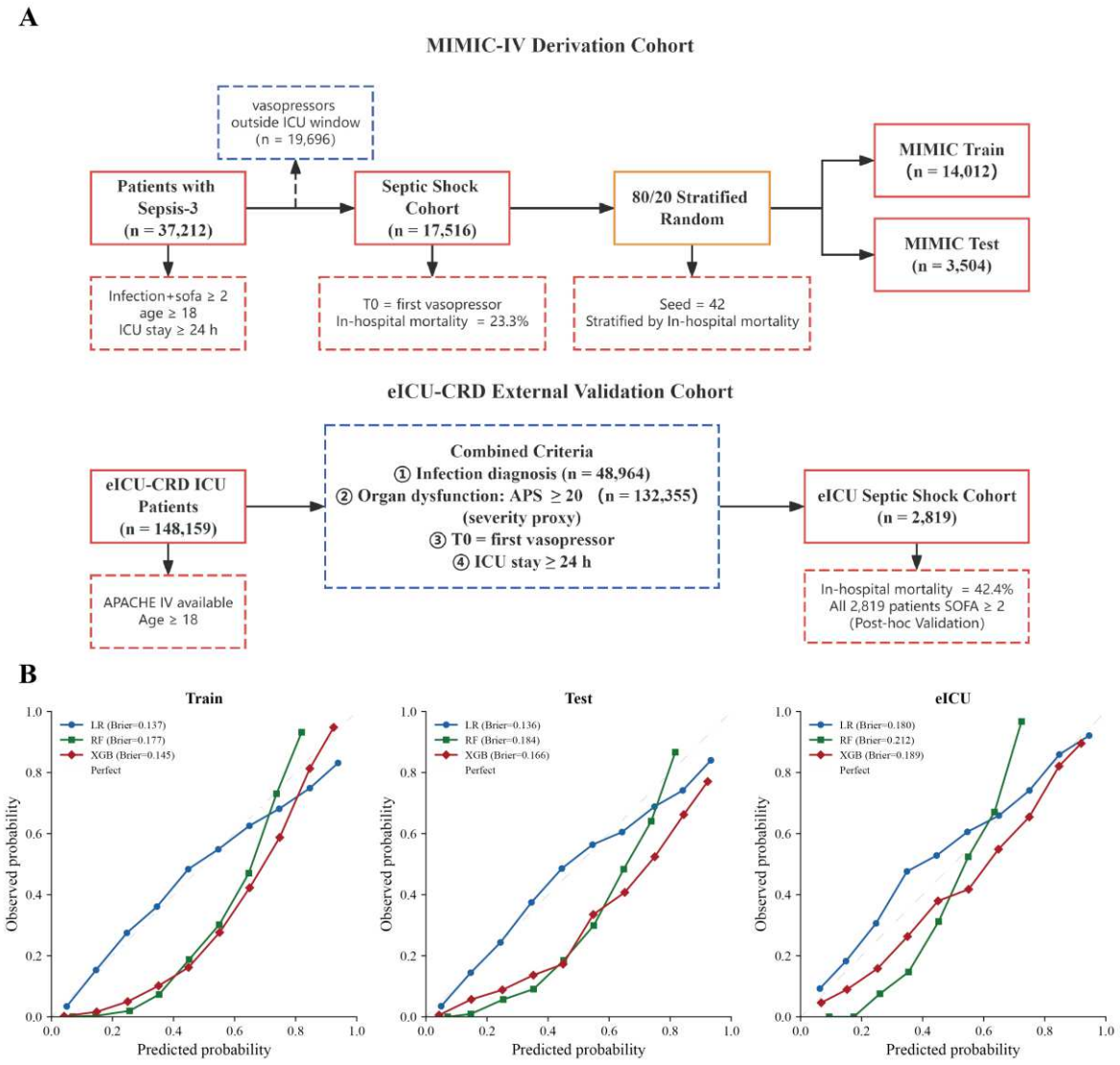


Figure 1. Study design and calibration. A, patient selection flowcharts for the MIMIC-IV development cohort and the eICU-CRD external validation cohort. B, calibration curves for logistic regression, random forest, and XGBoost on the MIMIC training, MIMIC test, and eICU cohorts, computed on native uncalibrated predicted probabilities with 10 bins. The dotted line is perfect calibration, and the Brier score is shown for each model.

Cohort characteristics

Baseline characteristics differed substantially between the two cohorts (**Table 1**). MIMIC patients had lower in-hospital mortality (23.3% vs. 42.4% in eICU), lower median APS scores (52 vs. 71), higher rates of pneumonia (41.3% vs. 6.7%) and mechanical ventilation at shock onset (53.7% vs. 28.2%). The most extreme differences appeared in variables dependent on diagnosis coding, with

liver disease recorded in 17.1% of MIMIC patients but essentially none in eICU, and bacteremia in 20.3% vs. 0.4%. This pattern is consistent with more intensive comorbidity and infection-site coding in MIMIC's tertiary academic setting.

Table 1. Baseline characteristics of the study cohorts

Characteristic	MIMIC-IV (n=17 516)	eICU-CRD (n=2 819)
Age (years)	68 (58-78)	68 (58-79)
Male	60.0%	51.5%
APS (APACHE IV acute physiology score)	52 (38-70)	71 (51-97)
OASIS	36 (30-42)	15 (12-21)
SOFA (computed, first 24h)	8 (6-11)	8 (6-10)
Heart rate (max, bpm)	81 (66-99)	116 (99-134)
SBP (max, mmHg)	121 (106-141)	99 (98-99)
Respiratory rate (max, /min)	62 (25-82)	30 (25-36)
Temperature (max, °C)	37.1 (36.7-37.6)	37.3 (36.8-38.2)
SpO2 (min, %)	95 (91-99)	91 (84-95)
Hemoglobin (min, g/dL)	8.9 (7.6-10.5)	10.1 (8.5-12.1)
Hemoglobin (max, g/dL)	11.2 (9.5-12.8)	11.6 (9.9-13.3)
Platelets (min, $\times 10^9$/L)	160 (110-226)	168 (107-239)
WBC (max, $\times 10^9$/L)	13 (9-18)	16 (10-23)
Bicarbonate (min, mmol/L)	22 (19-25)	18 (14-22)
BUN (max, mg/dL)	26 (17-44)	34 (22-52)
Creatinine (max, mg/dL)	1.2 (0.8-2)	1.9 (1.2-2.9)
Bilirubin (max, mg/dL)	0.7 (0.4-1.8)	0.9 (0.5-1.6)
Glucose (max, mg/dL)	167 (138-200)	162 (121-226)
Sodium (min, mmol/L)	136 (133-138)	136 (132-140)
Potassium (max, mmol/L)	4.7 (4.2-5.4)	4.5 (4-5.1)
Initial lactate (mmol/L)	2.2 (1.6-3.3)	3.3 (2.2-6.1)
Pneumonia	41.3%	6.7%

Urinary tract infection	16.9%	3.6%
Abdominal infection	4.9%	1.2%
Bacteremia	20.3%	0.4%
Myocardial infarction	22.1%	8.1%
Congestive heart failure	35.8%	17.1%
Peripheral vascular disease	14.7%	6.2%
Cerebrovascular disease	13.6%	10.0%
Dementia	3.7%	6.3%
COPD	27.5%	16.9%
Rheumatic disease	3.8%	2.2%
Peptic ulcer disease	3.6%	3.6%
Liver disease	17.1%	0.0%
Diabetes	32.9%	30.5%
Chronic kidney disease	25.5%	10.5%
AIDS/HIV	0.7%	0.8%
Malignancy	14.1%	21.2%
Mechanical ventilation at shock onset	53.7%	28.2%
Vancomycin exposure	24.0%	32.5%
ICU length of stay (days)	4.3 (2.2-9.2)	2 (1-4.1)
Hospital length of stay (days)	10.9 (6.2-20.6)	4.6 (1.4-8.8)
In-hospital mortality	23.3%	42.4%

Data are median (IQR) or percentage. APS = acute physiology score; OASIS = Oxford Acute Severity of Illness Score; SBP = systolic blood pressure; WBC = white blood cell count; BUN = blood urea nitrogen.

Model performance

Internal AUC ranking did not predict external performance (**Table 2**). XGBoost had the highest internal AUC at 0.836 on the MIMIC test set, against 0.820 for LR and 0.814 for RF, but on eICU the order reversed, with LR highest at 0.813, XGBoost at 0.800, and RF at 0.759. Model selection based on internal validation alone would have chosen XGBoost over LR. Paired bootstrap tests on

eICU confirmed that these differences exceeded sampling variation, with a ΔAUC of 0.013 for LR versus XGBoost and 0.054 for LR versus RF. The Brier scores told a different story. LR's was the highest of the three on eICU at 0.195, above XGBoost at 0.185, and the decomposition shows that this gap was largely a location shift driven by the 1.8-fold difference in baseline mortality rather than a loss of discrimination.

Table 2. Model performance on internal and external validation

Model	Test AUC	Test Brier	eICU AUC	eICU Brier	eICU AUPRC
LR	0.820	0.136	0.813	0.195	0.760
RF	0.814	0.184	0.759	0.211	0.683
XGB	0.836	0.166	0.800	0.185	0.730

All models evaluated on native (uncalibrated) probabilities. eICU bootstrap 95% CIs (500 iterations): LR AUC 0.813 [0.798-0.829], Brier 0.195 [0.186-0.205]; RF AUC 0.759 [0.740-0.777], Brier 0.211 [0.208-0.215]; XGB AUC 0.800 [0.785-0.816], Brier 0.185 [0.178-0.192].

Discrimination and risk stratification

Risk stratification by predicted-risk quartiles was preserved externally for all three models, although the gradients were compressed (**Table 3**). For LR, observed mortality rose monotonically from 10.6% in the lowest predicted quartile to 78.6% in the highest on eICU, a 7.4-fold gradient, compared with 2.6% to 54.3% (20.9-fold) on MIMIC test. RF showed the same pattern (12.8% to 70.2% on eICU, 5.5-fold; 2.3% to 52.1% on MIMIC test, 22.7-fold), as did XGBoost (9.2% to 75.7% on eICU, 8.2-fold; 1.3% to 55.0% on MIMIC test, 42.3-fold). All three models therefore produced monotonically increasing mortality across quartiles in both cohorts, indicating that graded risk stratification transported across cohorts for all three model families. The compression is consistent with the higher eICU baseline mortality, which lifts the floor in the lowest quartile and leaves little headroom in the highest.

Table 3. Observed mortality by predicted risk quartile across models

Model	Quartile	MIMIC Test	eICU
LR	Q1 (lowest risk)	2.6%	10.6%

LR	Q2	11.3%	27.0%
LR	Q3	24.8%	53.6%
LR	Q4 (highest risk)	54.3%	78.6%
RF	Q1 (lowest risk)	2.3%	12.8%
RF	Q2	11.9%	35.2%
RF	Q3	26.8%	51.6%
RF	Q4 (highest risk)	52.1%	70.2%
XGBoost	Q1 (lowest risk)	1.3%	9.2%
XGBoost	Q2	10.2%	32.3%
XGBoost	Q3	26.6%	52.4%
XGBoost	Q4 (highest risk)	55.0%	75.7%

Risk quartiles defined by each model's predicted-probability thresholds.

Clinical thresholds

For LR, the best externally performing model, the Youden-optimal threshold was nearly identical in the two cohorts (0.211 on MIMIC test; 0.208 on eICU) despite the 1.8-fold difference in baseline mortality (**Table S3**). This stability is expected. Youden's index is read off the ROC curve, which LR preserved across cohorts, so the operating point (sensitivity/specificity) is invariant to any monotone rescaling of the predicted probabilities. Positive predictive value rose from 0.434 on MIMIC test to 0.662 on eICU, reflecting eICU's higher baseline risk. Because LR systematically underestimated risk in eICU (calibration-in-the-large +0.130; Table 4), local intercept recalibration would raise the numeric threshold to approximately 0.37 while leaving sensitivity and specificity unchanged. For RF and XGBoost, Youden-optimal thresholds were higher and likewise similar across cohorts. For RF this reflects the compression of its native probabilities, with a calibration slope of 2.261 on eICU, whereas for XGBoost it reflects the class-weight-driven centering of its predictions near 0.5 rather than a distortion of their spread.

Table 4. Calibration decomposition of the three models across cohorts

Model	Dataset	Cal-in-large	Slope	Recal Brier
LR	MIMIC Train	-0.000	1.051	0.137
LR	MIMIC Test	-0.003	1.073	0.136
LR	eICU	+0.130	0.994	0.166
RF	MIMIC Train	-0.201	2.108	0.110
RF	MIMIC Test	-0.201	1.811	0.126
RF	eICU	-0.072	2.261	0.196
XGB	MIMIC Train	-0.165	1.591	0.100
XGB	MIMIC Test	-0.168	1.114	0.122
XGB	eICU	−0.071	1.086	0.179

Cal-in-large = observed minus mean predicted mortality; Slope = logistic regression coefficient of outcome on logit(predicted probability).

Table S3. Youden-optimal operating thresholds and associated test characteristics

Model	Cohort	Threshold	Sensitivity	Specificity	PPV	NPV
Logistic regression	MIMIC test	0.211	0.811	0.679	0.434	0.922
Logistic regression	eICU	0.208	0.778	0.707	0.662	0.812
Random forest	MIMIC test	0.533	0.725	0.752	0.470	0.900
Random forest	eICU	0.518	0.706	0.688	0.625	0.760
XGBoost	MIMIC test	0.492	0.807	0.724	0.470	0.925
XGBoost	eICU	0.525	0.739	0.714	0.656	0.788

Thresholds are Youden-optimal on native (uncalibrated) probabilities. MIMIC test = held-out 20% MIMIC-IV split; eICU = external validation cohort (n = 2,819).

Calibration decomposition

LR was well calibrated on the MIMIC test set, and its calibration slope was near 1 in all three cohorts (1.051, 1.073, and 0.994; **Table 4**). On eICU it underestimated mortality (mean predicted 29.4% versus observed 42.4%), but the calibration curve stayed parallel to the diagonal, so the error

was a location shift only (**Figure 1B**, eICU panel). The decomposition assigned almost all of this error to calibration-in-the-large (0.130), with negligible calibration loss. XGBoost had a smaller offset in the other direction (-0.071). Its scale_pos_weight of about 3.3 pulls predictions toward 40–50% in every cohort, so it overestimated mortality on MIMIC (-0.165) but landed within about 7 percentage points of eICU's baseline risk. Its slightly lower native Brier (0.185 versus 0.195; **Table 2**) came from stronger risk stratification (refinement -0.061 versus -0.051), not from better calibration. RF showed a different failure. Its eICU slope of 2.261 (95% CI 2.062–2.455) means the predicted probabilities were compressed toward the mean, a pattern already visible in the training cohort (**Figure 1B**, training panel), so it needs full logistic recalibration rather than a simple intercept shift. For all three models the eICU Brier score was dominated by uncertainty, which is fixed by the outcome prevalence.

Local recalibration of LR, RF, and XGBoost

Calibration functions do not transfer across prevalence differences (**Table 5**). Because LR's slope was preserved, a single intercept parameter should suffice. Fitted on 50 randomly sampled eICU patients, it reduced the mean Brier from 0.195 to 0.178 (95% range 0.174–0.188), reaching 0.175 by n = 200. AUC and AUPRC were unchanged, as expected for a monotonic transformation. The same correction changed XGBoost little at n = 50 and 200, consistent with its small native offset; local Platt scaling gave similar results (0.186 at n = 50; 0.182 at n = 200). Given the same local data, the two models were similarly calibratable, with XGBoost consistently about 0.006 Brier higher. RF differed. An intercept shift left RF’s Brier essentially unchanged (0.212 at n = 50 and 0.208 at n = 200), whereas full logistic recalibration reduced it to 0.196, consistent with its slope distortion requiring more than a location correction. The comparison also has a practical implication. XGBoost's native Brier was already close to its corrected value because its small offset (−0.071) leaves little bias to remove, whereas LR's larger offset (+0.130) made even a noisy n = 50 estimate worthwhile. A model that appears to need less correction may simply have a less visible

need. Finally, Platt scaling fitted on MIMIC and transported to eICU degraded XGBoost (Brier 0.185 to 0.198).

Table 5. Recalibration effects on eICU performance for LR, RF, and XGBoost

Model	Calibration	N local	Brier	95% range	Δ from native
LR	Native	0	0.195	-	-
LR	Intercept	50	0.178	0.174-0.188	-0.017
LR	Intercept	200	0.175	0.174-0.177	-0.020
XGBoost	Native	0	0.185	-	-
XGBoost	Intercept	50	0.184	0.180-0.195	-0.001
XGBoost	Intercept	200	0.181	0.180-0.183	-0.004
XGBoost	Platt	50	0.186	0.181-0.200	+0.001
XGBoost	Platt	200	0.182	0.180-0.185	-0.003
XGBoost	Platt (MIMIC-fitted, transported)	0	0.198	–	+0.013
RF	Native	0	0.211	-	-
RF	Intercept	50	0.212	0.207-0.225	+0.001
RF	Intercept	200	0.208	0.207-0.211	-0.003
RF	Full logistic recalibration	all eICU	0.196	-	-0.015

Brier means and 95% ranges (5th–95th percentiles) over 1000 paired random samples. Δ from native = Brier (corrected) − Brier (native); negative values indicate improvement. All corrections preserve AUC and AUPRC. RF full logistic recalibration was fitted on the full eICU cohort.

Variable contributions across model families

SHAP analysis across the three model families (**Figure 2**) identified illness severity measured by APS, early physiological derangements such as maximum hemoglobin, maximum respiratory rate, and initial lactate, age, infection site with pneumonia leading, and the comorbidity malignancy as prominent predictors, consistent with shock severity driving prognosis in this cohort. APS ranked first in all three models, and first-6-hour minimum hemoglobin ranked second in the LR model.

Top-15 overlap between model families was 7 of 15 between LR and XGBoost, 4 of 15 between LR and random forest, and 8 of 15 between XGBoost and random forest, supporting the interpretation that the LR single-index structure is not capturing spurious associations.

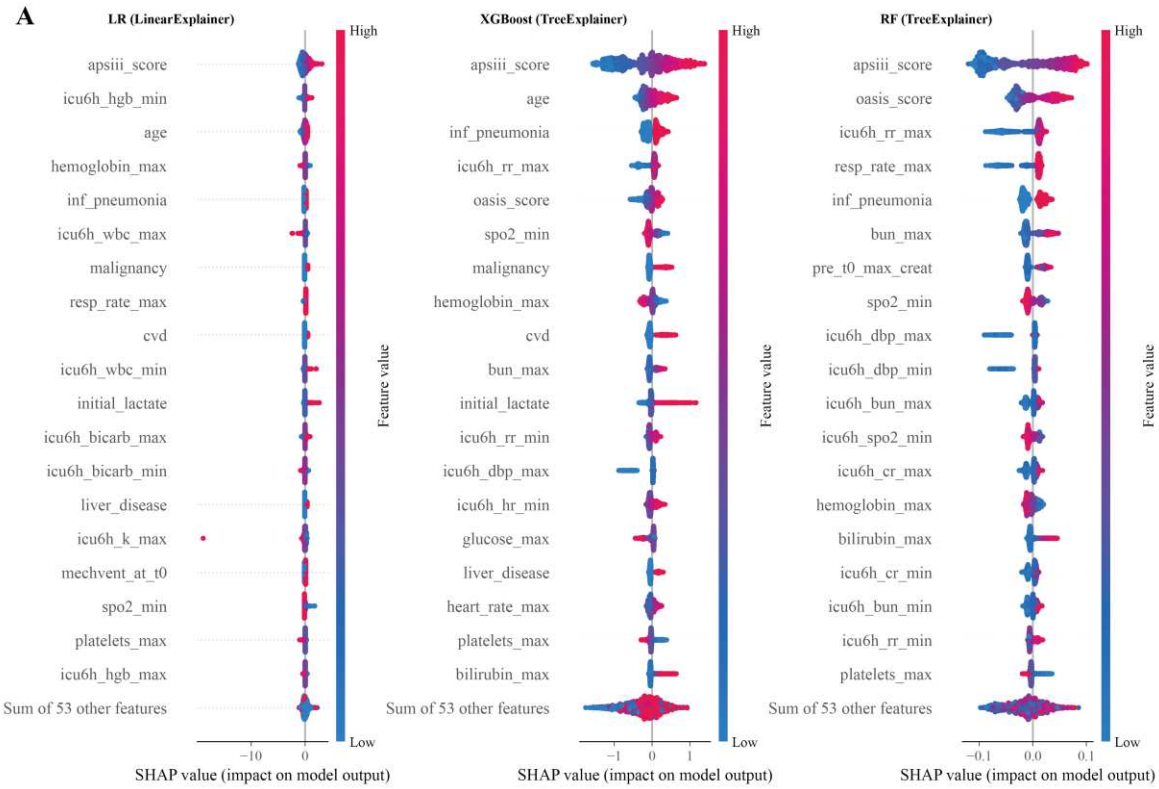


Figure 2. SHAP summary plots of feature importance for the logistic regression, XGBoost, and random forest models on the MIMIC test set, computed with the LinearExplainer for logistic regression and the TreeExplainer for the tree-based models.

Prevalence-dependent XGBoost calibration

The reweighting design (Figure 3) revealed a mirror-image pattern. As the target mortality rate rose from 10% to 70%, XGBoost's average predicted risk moved only from 0.41 to 0.57, far less than the target itself, so its native weighted Brier score stayed nearly flat. Intercept correction with 75 local samples helped mainly when the target rate was far from this center, lowering the Brier from 0.197 to 0.080 at 10% prevalence. LR showed the opposite behavior, with correction helping most when the target was far from the 23.3% rate seen in the development cohort, lowering the Brier from 0.279 to 0.156 at 70% prevalence. Random forest's native Brier was also flat but consistently

higher than XGBoost's, and even after intercept correction it remained the least well calibrated of the three models at every prevalence level, for example 0.085 versus 0.076 and 0.078 at 10% prevalence, which is consistent with its slope distortion of 2.261 on eICU. Intercept-corrected Brier values across the full prevalence grid, for local samples of n = 30, 75, and 200, are given in **Table S4**.

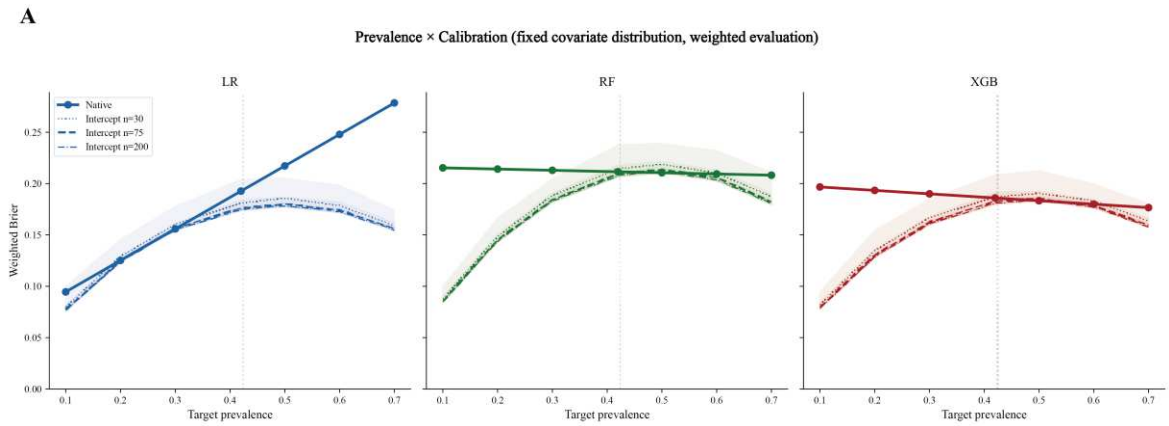


Figure 3. Prevalence-dependent calibration. Native and intercept-corrected weighted Brier scores for logistic regression, random forest, and XGBoost across target mortality rates of 10-70% in the fixed-covariate reweighted eICU cohort.

Table S4. Intercept correction across simulated prevalence levels (LR, RF, XGBoost)

Target mortality	Model	Native Brier	Intercept n=30	Intercept n=75	Intercept n=200
10%	LR	0.095	0.080 [0.076–0.100]	0.077 [0.076–0.081]	0.076 [0.076–0.078]
10%	RF	0.215	0.088 [0.084–0.100]	0.086 [0.084–0.090]	0.085 [0.084–0.087]
10%	XGB	0.197	0.082 [0.078–0.094]	0.080 [0.078–0.085]	0.078 [0.078–0.080]
20%	LR	0.125	0.129 [0.123–0.145]	0.125 [0.123–0.129]	0.124 [0.123–0.127]
20%	RF	0.214	0.149 [0.143–0.167]	0.145 [0.143–0.151]	0.144 [0.143–0.146]
20%	XGB	0.193	0.135 [0.128–0.154]	0.130 [0.128–0.136]	0.129 [0.128–0.131]

30%	LR	0.156	0.160 [0.154–0.178]	0.156 [0.154–0.161]	0.155 [0.154–0.158]
30%	RF	0.213	0.188 [0.182–0.205]	0.184 [0.182–0.191]	0.183 [0.182–0.187]
30%	XGB	0.190	0.167 [0.160–0.185]	0.162 [0.160–0.168]	0.161 [0.160–0.164]
42%	LR	0.193	0.181 [0.174–0.204]	0.176 [0.174–0.184]	0.175 [0.174–0.178]
42%	RF	0.212	0.215 [0.206–0.238]	0.210 [0.206–0.218]	0.208 [0.206–0.211]
42%	XGB	0.186	0.187 [0.180–0.208]	0.183 [0.180–0.192]	0.181 [0.180–0.184]
50%	LR	0.217	0.186 [0.177–0.205]	0.180 [0.177–0.187]	0.178 [0.177–0.181]
50%	RF	0.211	0.219 [0.211–0.239]	0.214 [0.211–0.221]	0.212 [0.211–0.215]
50%	XGB	0.183	0.191 [0.183–0.213]	0.185 [0.183–0.192]	0.184 [0.183–0.187]
60%	LR	0.248	0.179 [0.172–0.198]	0.174 [0.172–0.182]	0.173 [0.172–0.175]
60%	RF	0.209	0.210 [0.203–0.232]	0.206 [0.203–0.213]	0.204 [0.202–0.208]
60%	XGB	0.180	0.183 [0.176–0.200]	0.179 [0.176–0.184]	0.177 [0.176–0.181]
70%	LR	0.279	0.159 [0.154–0.174]	0.156 [0.154–0.162]	0.155 [0.154–0.157]
70%	RF	0.208	0.187 [0.179–0.211]	0.182 [0.179–0.191]	0.181 [0.179–0.184]
70%	XGB	0.177	0.164 [0.157–0.180]	0.160 [0.157–0.166]	0.158 [0.157–0.161]

Outcome-based marginal reweighting on the full eICU cohort (n=2,819); 200 bootstrap repetitions. Brier reported as mean [5th-95th percentile].

Methodological sensitivity analysis

The original model selection pipeline made several methodological choices that each tilted the comparison in a different direction, and together they decided which model a developer following that pipeline would have picked (**Table 6**). Each choice on its own barely moved the external AUC, and the external ranking stayed LR > XGBoost > RF no matter how they were combined. What did change was which model looked best on the internal MIMIC data. At every step the tree models came out on top internally, while on eICU the best model was actually logistic regression. The clearest example is the transported Platt scaling. It was fitted on MIMIC and then applied to eICU, and it worsened XGBoost's calibration (Brier 0.185 to 0.198; random forest 0.212 to 0.229). This was not because the tree models are hard to calibrate. It was because they were denied the local eICU data that logistic regression's correction had received (**Table 5**).

Table 6. Impact of methodological choices on model ranking

Condition	Best Model (eICU AUC)	Key Issue
Fair comparison (all native)	LR (0.813)	-
+ Seed tuning (multiple seeds)	XGB/RF may appear best	Inflates complex model apparent performance
+ C selected using eICU	LR benefits (data leakage)	Exaggerates LR
+ class_weight="balanced"	Shifts LR probabilities	Unfair probability change
+ Platt scaling trained on MIMIC, applied to eICU	XGB (Brier 0.198, worse)	Unfair: XGB denied local data

Robustness analyses

We checked whether the findings held when we changed predictor definitions, feature timing, class imbalance handling, and the patient split. After removing OASIS and maximum systolic blood pressure, the two predictors not measured comparably across databases, and after restricting eICU to the 1,600 patients meeting the Sepsis-3 lactate criterion (50.2% mortality), the model ranking and the calibration pattern did not change (**Tables S5 and S6**). Thirty of the 72 predictors summarized the first six hours of the ICU stay, a window that could include information recorded after shock had begun. Retraining without these features made no difference to any conclusion, and

the largest AUC change was 0.006 (**Table S7**). For XGBoost, removing the class weighting barely changed discrimination or the calibration slope, but its predictions moved back toward the development prevalence. The calibration offset went from −0.165 to −0.000 on the development set, from −0.168 to −0.001 on the MIMIC test set, and from −0.071 to +0.087 on eICU. Its apparently good external calibration in the main analysis thus came from the class weighting, not from the model itself (**Table S8**). Finally, when all ICU stays from the same hospitalization were placed on the same side of the split, internal AUC changed by at most 0.015 and external AUC by at most 0.003, and the ranking and calibration pattern stayed the same (**Table S9**).

Table S5. Sensitivity analysis excluding OASIS and systolic blood pressure maximum

Model	AUC (full set)	AUC (70 features)	ΔAUC	Brier (full set)	Brier (70 features)	ΔBrier
LR	0.813	0.812	-0.001	0.195	0.188	-0.008
RF	0.759	0.767	+0.007	0.211	0.215	+0.004
XGBoost	0.800	0.809	+0.009	0.185	0.197	+0.012

Models retrained on the 70-feature set (72 minus OASIS and SBP max); external AUC ranking LR > XGBoost > RF preserved; max |dAUC| = 0.009.

Table S6. Sensitivity analysis restricted to admissions with lactate > 2 mmol/L

Dataset	Lactate subgroup	Model	N	Mortality %	AUC	Brier	Cal-in-large	Cal slope
MIMIC Test	All	LR	3504	23.3	0.820	0.136	−0.003	1.073
MIMIC Test	All	RF	3504	23.3	0.814	0.184	−0.201	1.811
MIMIC Test	All	XGB	3504	23.3	0.836	0.166	−0.168	1.114
MIMIC Test	> 2	LR	1366	24.3	0.857	0.127	−0.004	1.095
MIMIC Test	> 2	RF	1366	24.3	0.858	0.157	−0.158	1.817

MIMIC Test	> 2	XGB	1366	24.3	0.867	0.148	−0.136	1.067
eICU	All	LR	2819	42.4	0.813	0.195	+0.130	0.994
eICU	All	RF	2819	42.4	0.759	0.211	−0.072	2.261
eICU	All	XGB	2819	42.4	0.800	0.185	−0.071	1.086
eICU	> 2	LR	1600	50.2	0.816	0.203	+0.148	0.951
eICU	> 2	RF	1600	50.2	0.762	0.210	−0.013	2.251
eICU	> 2	XGB	1600	50.2	0.801	0.184	−0.049	1.048
eICU	≤ 2	LR	442	19.9	0.809	0.126	+0.030	1.129
eICU	≤ 2	RF	442	19.9	0.705	0.211	−0.255	1.736
eICU	≤ 2	XGB	442	19.9	0.753	0.173	−0.176	1.008

Subgroups by initial serum lactate at T0: All = full cohort; > 2 = Sepsis-3 lactate criterion; <= 2 = reference.

Table S7. Sensitivity analysis excluding ICU first-6-hour features

Model	**Set**	**AUC (full set)**	**AUC (42 features)**	**ΔAUC**	**Brier (full set)**	**Brier (42 features)**	**ΔBrier**	**Cal-in-large (full set)**	**Cal-in-large (42 features)**	**Cal slope (full set)**	**Cal slope (42 features)**
LR	MIMIC Test	0.820	0.814	−0.006	0.136	0.137	+0.001	−0.003	−0.004	1.073	1.025
LR	eICU	0.813	0.813	−0.001	0.195	0.208	+0.012	+0.130	+0.164	0.994	1.014
RF	MIMIC Test	0.814	0.816	+0.002	0.184	0.182	−0.001	−0.201	−0.202	1.811	1.755
RF	eICU	0.759	0.763	+0.004	0.211	0.208	−0.004	−0.072	−0.067	2.261	2.039
XGB	MIMIC Test	0.836	0.832	−0.004	0.166	0.169	+0.003	−0.168	−0.172	1.114	1.070
XGB	eICU	0.800	0.804	+0.004	0.185	0.182	−0.003	−0.071	−0.059	1.086	1.051

All 30 ICU first-6-hour (icu6h) predictors excluded (42 features retained); external ranking and calibration pattern preserved; max |dAUC| = 0.006.

Table S8. Sensitivity analysis of XGBoost class weighting (scale_pos_weight = 1.0 vs. 3.3)

Set	N	Prevalence	AUC (w=3.3)	AUC (w=1.0)	ΔAUC	Brier (w=3.3)	Brier (w=1.0)	ΔBrier	Cal-in-large (w=3.3)	Cal-in-large (w=1.0)	Cal slope (w=3.3)	Cal slope (w=1.0)
Train	14012	23.3%	0.891	0.888	−0.003	0.145	0.112	−0.033	−0.165	−0.000	1.591	1.562
Test	3504	23.3%	0.836	0.835	−0.001	0.166	0.132	−0.035	−0.168	−0.001	1.114	1.118
eICU	2819	42.4%	0.800	0.807	+0.008	0.185	0.186	+0.000	−0.071	+0.087	1.086	1.087

XGBoost retrained with scale_pos_weight = 1.0 (primary analysis: 3.3).

Table S9. Sensitivity analysis using a hospitalization-level split

Model	Set	N	Prevalence	AUC (episode split)	AUC (patient split)	ΔAUC	Brier (episode split)	Brier (patient split)	Cal-in-large (episode split)	Cal-in-large (patient split)	Cal slope (episode split)	Cal slope (patient split)
LR	Train	13990	23.2%	0.816	0.821	+0.005	0.138	0.135	−0.000	−0.000	1.051	1.009
LR	Test	3526	23.4%	0.820	0.805	−0.015	0.136	0.144	−0.003	+0.004	1.073	0.862
LR	eICU	2819	42.4%	0.813	0.814	+0.001	0.195	0.189	+0.130	+0.111	0.994	0.901
RF	Train	13990	23.2%	0.839	0.842	+0.003	0.177	0.176	−0.201	−0.200	2.108	2.141
RF	Test	3526	23.4%	0.814	0.810	−0.004	0.184	0.183	−0.201	−0.198	1.811	1.729
RF	eICU	2819	42.4%	0.759	0.763	+0.003	0.211	0.209	−0.072	−0.061	2.261	2.223
XGB	Train	13990	23.2%	0.891	0.893	+0.002	0.145	0.143	−0.165	−0.164	1.591	1.575

XGB	Test	3526	23.4%	0.836	0.827	−0.009	0.166	0.166	−0.168	−0.158	1.114	1.012
XGB	eICU	2819	42.4%	0.800	0.798	−0.002	0.185	0.185	−0.071	−0.065	1.086	1.123

Hospitalization-level split: 13,990 training / 3,526 test stays (primary analysis: 14,012 / 3,504).

Decision curve and subgroup analyses

Decision curves computed on native probabilities showed positive net benefit across clinically relevant thresholds of 1–30% for all three models in all three cohorts. LR and XGBoost were broadly overlapping, RF was slightly lower on eICU but approached zero by 30% on the MIMIC cohorts (**Figure 4**). LR stayed positive up to the highest evaluated threshold in every cohort. RF net benefit turned negative above approximately 30% on the MIMIC cohorts, and XGBoost above approximately 41% on the MIMIC test set. On eICU all three models stayed positive across the entire range, and XGBoost had the highest net benefit at every threshold. Because these curves use native probabilities, they show the models as deployed, and net benefit itself depends on calibrated risk. LR's location shift on eICU means its nominal thresholds correspond to somewhat higher true risks, so the curves should be read together with the recalibration analysis. LR and XGBoost had near-equivalent net benefit, so the curves confirm decision value for both but do not choose between them. That choice rests on the calibration analyses (**Tables 4–5**). In subgroup analyses on eICU, the ranking LR over XGBoost over RF held within every age, ventilation, and diabetes stratum, and all subgroup confidence intervals overlapped the overall confidence interval (**Figure 5**). The clearest difference was by age. LR performed best in patients under 65 (AUC 0.847, 95% CI 0.823–0.869), compared with 0.783 (95% CI 0.761–0.805) in those 65 or older.

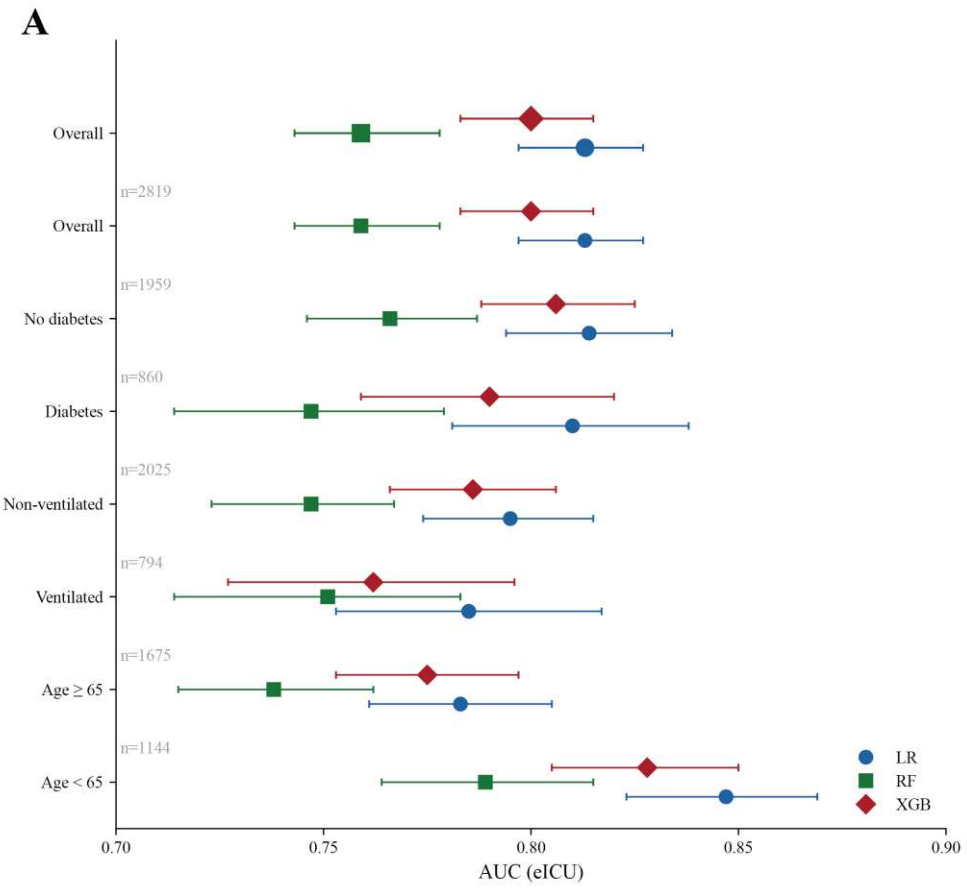


Figure 4. Decision curves showing net benefit against threshold probability for logistic regression, random forest, and XGBoost on the MIMIC training cohort A, the MIMIC test cohort B, and the eICU cohort C, computed on native predicted probabilities. Grey lines show the treat-all and treat-none strategies.

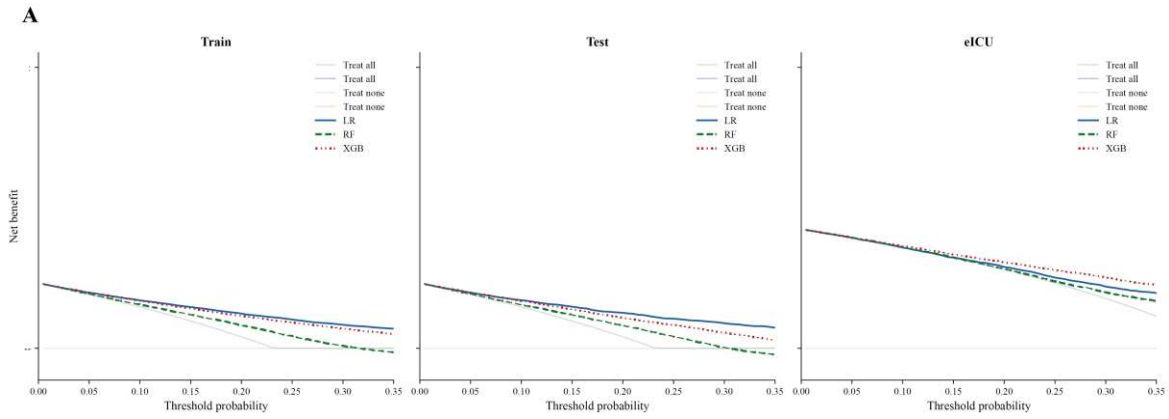


Figure 5. Subgroup analysis on the eICU cohort, showing AUC with 95% bootstrap confidence intervals for logistic regression, random forest, and XGBoost within age, mechanical ventilation, and diabetes strata, together with the overall estimate.

Discussion

Prediction models rarely perform identically when moved from one hospital to another, but exactly how they fail, and what fixes them, is poorly understood. Comparing three model classes across two critical care databases, we report three findings.

First, LR and XGBoost were both miscalibrated after transport, but the errors looked very different. LR systematically underestimated mortality on eICU, predicting 29.4% on average against an observed 42.4%. Every prediction was roughly 13 points too low, yet the ordering of patients by risk remained correct, and the error was large enough to be obvious at a glance. XGBoost's error was of the same kind but only −0.071, and this owed to luck rather than to the model. A class weight applied during training had pulled its predictions toward 50%, and eICU's mortality happened to be 42.4%. When we removed the class weight and retrained, discrimination was unchanged ($\Delta\text{AUC} \leq 0.008$ in all cohorts), predictions moved back to the development mortality rate, and the small eICU offset reversed to +0.087. The practical lesson is that a model can look well calibrated for reasons that have nothing to do with how well it travels. Prevalence-driven miscalibration of this kind has also been documented over time, as in the AKI prediction models of Davis et al., where calibration drifted while discrimination persisted [6].

Second, how well a model can be calibrated and how predictable its calibration needs are are different properties, and only the latter distinguishes these models. Given 50 local patients, LR and XGBoost reached similar Brier scores, 0.178 versus 0.184, so tree-based models are not inherently harder to calibrate. The difference lies in what happens before any local data are available. In our simulation, patient characteristics were held fixed and only the death rate varied, and XGBoost's uncorrected accuracy stayed nearly flat across death rates from 10% to 70%. Intercept correction with 75 local samples improved it by 59% at a death rate of 10% but did almost nothing between 42% and 60%, where its predictions already sat. Whether a tree-based model needs correction, and whether a simple intercept suffices, therefore depends on how far the local death rate is from the

level its predictions are centered on, and this cannot be known without local data. RF showed the same unpredictability in a different form. Its failure was not a uniform shift but a compression of predicted probabilities that left too little spread between risks, so a single intercept could not in principle repair it and full recalibration was required. LR's decision rule is simpler and, in this setting, sufficient. Recalibrate when the local death rate departs substantially from the development rate.

Third, the model ranking on internal data was a poor guide to performance on external data. XGBoost had the highest internal AUC at 0.836 but fell to 0.800 on eICU, while LR, second internally at 0.820, ranked first externally at 0.813, with bootstrap confidence intervals excluding zero for both pairwise differences. Selecting on internal AUC alone would have deployed the wrong model. This gap is not unique to our setting. In the largest systematic review of COVID-19 prediction models, nearly all were at high risk of bias and almost none had been externally validated [24], and as strongly validated real-time machine learning models proliferate in critical care [25], understanding the gap becomes a practical rather than academic concern.

Beyond these findings, the study carries a methodological warning. The original pipeline's ranking was overturned not by any single error but by the accumulation of four choices that were each minor on its own, namely seed tuning, hyperparameter selection on external data, class weighting applied to only one model, and calibration data given to only one model. Reproducible comparison therefore requires fixed and pre-specified seeds, hyperparameter selection on training data only, deliberate rather than default use of class weights, and identical evaluation conditions for all candidates. If one model receives calibration post-processing, all must. These pitfalls are not hypothetical. A systematic review found deficient validation and reporting to be the norm in machine learning prediction studies [26], which motivates updated reporting and risk-of-bias standards [27,28].

These findings also translate into different deployment pathways for the two model classes. For LR, deployment with scarce local data follows a predictable sequence. Deploy as-is for risk

ranking, collect 50 to 100 local outcomes, and fit a single intercept. This is a pragmatic minimum, below the roughly 100 events recommended for full external validation, but adequate for estimating a single parameter [29]. No retraining, tuning, or diagnostic step is required. Tree-based models carry an additional burden under the same sequence. Before choosing between intercept-only and full recalibration, one must first determine whether the calibration error is of the compressed-probability type, a determination that itself consumes local data. A sample large enough to fit LR's single parameter may be too small to make it reliably. This difference in diagnostic burden, rather than any difference in achievable calibration, is the practical argument for simpler models when local outcome data are scarce. Reporting of such early-stage local evaluations should follow emerging standards for clinical AI decision support [30].

Our results agree with and extend prior work. Earlier external validations of ICU and sepsis mortality models found that discrimination was preserved while the accuracy of predicted probabilities degraded [4,5]. We confirm this pattern across model classes and show that the structure of calibration failure differs. LR was off only in its overall level, an offset correctable with 50 patients, whereas tree-based models failed less predictably. The failure of transported Platt scaling echoes Niculescu-Mizil and Caruana's observation that good calibration is data-dependent and cannot be assumed to transfer across datasets [31]. We add that the problem lies in transporting the calibration function rather than in the model's ability to be calibrated, since locally fitted Platt scaling performed adequately once given target-population data. The measurement-definition problems we encountered are instances of the definitional variability documented in a recent scoping review [32] and reinforce the call for standardized, transparent variable reporting.

Limitations

Several limitations qualify these findings. Both databases are American, and we do not know whether the results would hold in other healthcare systems. The external cohort of 2,819 patients is adequate for validation but small compared with the development set. Patients within each hospital in eICU are correlated, and a small fraction of MIMIC records came from hospitalizations

with more than one ICU stay, yet our bootstrap intervals treated every record as independent, so they may be slightly too narrow by the standards of clustered prediction research [33]. The between-model differences are large enough that this is unlikely to change any conclusion. We deliberately kept the pre-specified set of 72 variables in the main analysis, even though some proved uninformative externally, because dropping variables based on the validation data would have tuned the analysis to eICU. Keeping the pre-specified set with imputation, together with sensitivity analyses, is the recommended way to handle systematically missing predictors in external validation [14,28]. One variable deserves note. Maximum systolic blood pressure was missing for 68.6% of eICU patients, mostly replaced by triage-level readings, but its coefficient was shrunk to exactly zero by L1 regularization, and removing it together with OASIS changed nothing. Maximum respiratory rate and heart rate also differed between databases in ways that suggest different measurement windows rather than true physiology, patterns that tree-based models may exploit in training and lose after transport more readily than LR, although this remains speculative. Our septic shock definition was broader than strict Sepsis-3 because it did not require the lactate threshold of 2 mmol/L [10], but a sensitivity analysis restricted to lactate-positive admissions reproduced the main results in direction and magnitude. Finally, the finding that LR fails only through its overall level comes from one disease, one outcome, and one pair of databases. Whether it holds under stronger distortion or outside the United States is an open question.

Conclusions

In this cross-hospital septic shock cohort, LR and XGBoost had calibration errors that were similarly easy to repair, but the errors differed in how visible they were. LR underestimated mortality on eICU by about 13 points on average, a gap large enough to be obvious, whereas XGBoost's predictions sat close to eICU's death rate by coincidence, hiding a need that our simulation showed to depend on the local death rate. The two tree-based models also failed differently, one repairable with a single intercept and one requiring full recalibration, and only local

data can reveal which repair is needed. When local outcome data are scarce, understanding how a model is expected to fail can complement internal discrimination as a guide to model selection.

Declarations

Ethics approval and consent to participate

This study was a secondary analysis of two de-identified, publicly available critical care databases. Approval for the establishment and sharing of MIMIC-IV was granted by the Institutional Review Boards of the Massachusetts Institute of Technology (No. 0403000206) and Beth Israel Deaconess Medical Center, with a waiver of informed consent. eICU-CRD was certified as meeting HIPAA Safe Harbor de-identification standards by an independent privacy expert (Privacert, Cambridge, MA; Certification No. 1031219-2). Both databases were accessed under the PhysioNet Credentialed Health Data License after completion of the required training in human subjects' research. No additional institutional review was required for this analysis. The study was conducted in accordance with the Declaration of Helsinki.

Clinical trial number: not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Funding

This work was supported by the Xijing Hospital 2025 Youth Boost Program (grant number XJZT25QN36). The funders had no role in study design, data collection and analysis, interpretation of results, or manuscript preparation.

Authors' contributions

Honghong Duan: Conceptualization, Formal analysis, Software, Writing – original draft. Ying Zhang: Conceptualization, Formal analysis, Writing – original draft. Haijin Suo and Jingyi Ren:

Data curation, Software. Junjie Li: Methodology, Validation. Jing Yang and Junjie Li: Supervision, Writing – review & editing. Zelong Yang: Conceptualization, Funding acquisition, Supervision. All authors reviewed and approved the final manuscript.

Acknowledgements

Not applicable.

Availability of data and materials

MIMIC-IV is available through PhysioNet (https://physionet.org/content/mimiciv/) and eICU-CRD through PhysioNet (https://physionet.org/content/eicu-crd/) to credentialed researchers. The complete analysis pipeline, including cohort extraction, model training, and all validation and sensitivity analyses, is available from the corresponding author on reasonable request.

References

[1] Johnson AEW, Bulgarelli L, Shen L, Gayles A, Shammout A, Horng S, et al. MIMIC-IV, a freely accessible electronic health record database. Sci Data. 2023;10(1):1. doi:10.1038/s41597-022-01899-x.

[2] Collins GS, Dhiman P, Ma J, Schlussel MM, Archer L, Van Calster B, et al. Evaluation of clinical prediction models (part 1): from development to external validation. BMJ. 2024;384:e074819. doi:10.1136/bmj-2023-074819.

[3] 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 Intern Med. 2021;181(8):1065-70. doi:10.1001/jamainternmed.2021.2626.

[4] Cox EGM, Wiersema R, Eck RJ, Kaufmann T, Granholm A, Vaara ST, et al. External validation of mortality prediction models for critical illness reveals preserved discrimination but poor calibration. Crit Care Med. 2023;51(1):80-90. doi:10.1097/CCM.0000000000005712.

[5] Hargovan S, Simpson C, Sivalingam S, Carter A, Gunnarsson R. External validation of eight different models to predict sepsis mortality in intensive care units. J Crit Care. 2025;90:155174. doi:10.1016/j.jcrc.2025.155174.

[6] Davis SE, Lasko TA, Chen G, Siew ED, Matheny ME. Calibration drift in regression and machine learning models for acute kidney injury. J Am Med Inform Assoc. 2017;24(6):1052-61. doi:10.1093/jamia/ocx030.

[7] 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(1):230. doi:10.1186/s12916-019-1466-7.

[8] 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. J Clin Epidemiol. 2016;74:167-76. doi:10.1016/j.jclinepi.2015.12.005.

[9] 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. doi:10.1038/sdata.2018.178.

[10] Singer M, Deutschman CS, Seymour CW, Shankar-Hari M, Annane D, Bauer M, et al. The Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3). JAMA. 2016;315(8):801-10. doi:10.1001/jama.2016.0287.

[11] Tibshirani R. Regression shrinkage and selection via the lasso. J R Stat Soc Series B. 1996;58(1):267-88. doi:10.1111/j.2517-6161.1996.tb02080.x.

[12] Breiman L. Random forests. Mach Learn. 2001;45(1):5-32. doi:10.1023/A:1010933404324.

[13] Platt JC. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. Adv Large Margin Classif. 1999;10(3):61-74.

[14] 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. doi:10.1136/bmj-2023-074820.

[15] 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(1):128-38. doi:10.1097/EDE.0b013e3181c30fb2.

[16] Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. Med Decis Making. 2006;26(6):565-74. doi:10.1177/0272989X06295361.

[17] Lundberg SM, Lee SI. A unified approach to interpreting model predictions. In: Advances in Neural Information Processing Systems 30. 2017. p. 4768-77. doi:10.5555/3295222.3295230.

[18] 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.

[19] 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. 2016. p. 785-94. doi:10.1145/2939672.2939785.

[20] McKinney W. Data structures for statistical computing in Python. In: Proceedings of the 9th Python in Science Conference. 2010. p. 56-61. doi:10.25080/Majora-92bf1922-00a.

[21] Harris CR, Millman KJ, van der Walt SJ, Gommers R, Virtanen P, Cournapeau D, et al. Array programming with NumPy. Nature. 2020;585(7825):357-62. doi:10.1038/s41586-020-2649-2.

[22] Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat Methods. 2020;17(3):261-72. doi:10.1038/s41592-019-0686-2.

[23] Hunter JD. Matplotlib: a 2D graphics environment. Comput Sci Eng. 2007;9(3):90-95. doi:10.1109/MCSE.2007.55.

[24] Wynants L, Van Calster B, Collins GS, Riley RD, Heinze G, Schuit E, et al. Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal. BMJ. 2020;369:m1328. doi:10.1136/bmj.m1328.

[25] Lim L, Gim U, Cho K, Yoo D, Ryu HG, Lee HC. Real-time machine learning model to predict short-term mortality in critically ill patients: development and international validation. Crit Care. 2024;28(1):76. doi:10.1186/s13054-024-04866-7.

[26] Andaur Navarro CL, Damen JAA, van Smeden M, Takada T, Nijman SWJ, Dhiman P, et al. Systematic review identifies the design and methodological conduct of studies on machine learning-based prediction models. J Clin Epidemiol. 2023;154:8-22. doi:10.1016/j.jclinepi.2022.11.015.

[27] 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. doi:10.1136/bmj-2023-078378.

[28] Moons KGM, Damen JAA, Kaul T, Hooft L, Andaur Navarro C, Dhiman P, et al. PROBAST+AI: an updated quality, risk of bias, and applicability assessment tool for prediction models using regression or artificial intelligence methods. BMJ. 2025;388:e082505. doi:10.1136/bmj-2024-082505.

[29] Riley RD, Snell KIE, Archer L, Ensor J, Debray TPA, van Calster B, et al. Evaluation of clinical prediction models (part 3): calculating the sample size required for an external validation study. BMJ. 2024;384:e074821. doi:10.1136/bmj-2023-074821.

[30] Vasey B, Nagendran M, Campbell B, Clifton DA, Collins GS, Denaxas S, et al. Reporting guideline for the early-stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. Nat Med. 2022;28(5):924-33. doi:10.1038/s41591-022-01772-9.

[31] Niculescu-Mizil A, Caruana R. Predicting good probabilities with supervised learning. In: Proceedings of the 22nd International Conference on Machine Learning (ICML). 2005. p. 625-32. doi:10.1145/1102351.1102430.

[32] Cox EGM, Meijs DAM, Wynants L, Sels JEM, Koeze J, Keus F, et al. The definition of predictor and outcome variables in mortality prediction models: a scoping review and quality of reporting study. J Clin Epidemiol. 2025;178:111605. doi:10.1016/j.jclinepi.2024.111605.

[33] Debray TPA, Collins GS, Riley RD, Snell KIE, Van Calster B, Reitsma JB, et al. Transparent reporting of multivariable prediction models developed or validated using clustered data: TRIPOD-Cluster checklist. BMJ. 2023;380:e071018. doi:10.1136/bmj-2022-071018.