

Development and Cross-Database Evaluation of 6-Hour Landmark Prediction Models for a Subsequent Packed Red Blood Cell Transfusion Record in Adult Trauma ICU Patients

gong zhang

zhgg02zsu@zmu.edu.cn

Fifth Affiliated Hospital of Zunyi Medical University

Cui Yang

Affiliated Hospital of Zunyi Medical University

Shirong Jiang

Fifth Affiliated Hospital of Zunyi Medical University

Xiaoyue Li

Fifth Affiliated Hospital of Zunyi Medical University

Haizhen Duan

Affiliated Hospital of Zunyi Medical University

Research Article

Keywords: trauma, packed red blood cells, machine learning, prediction model, landmark analysis, intensive care, electronic health records, SHAP

Posted Date: August 24th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10708009/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.

Most trauma transfusion-prediction studies address prehospital or arrival-phase decisions. We examined whether routinely documented information available six hours after ICU admission could stratify the risk of a subsequent packed red blood cell (pRBC) transfusion record.

Methods.

We developed five prediction models in MIMIC-IV v3.1 and evaluated them in a held-out MIMIC test set and eICU-CRD v2.0. Eligible adult trauma ICU stays had no qualifying pRBC record available by the six-hour landmark. The outcome was a subsequent qualifying pRBC event after six hours through the earlier of 24 hours or ICU exit; thus, ICU exit truncated ascertainment under routine disposition. Model development used 12 harmonized predictors, training-only preprocessing, five-fold out-of-fold procedures, and development-only logistic recalibration. Evaluation covered discrimination, calibration, Brier score, exploratory decision curves, TreeSHAP, competing events, operating points, and prespecified or post hoc robustness analyses.

Results.

The analysis included 6,779 MIMIC development stays (396 events), 2,270 held-out MIMIC stays (136 events), and 6,578 eICU stays (262 events) from 176 hospitals. No algorithm dominated across all metrics. Reference weighted XGBoost achieved AUROCs of 0.844 (95% CI, 0.810–0.876) in MIMIC and 0.851 (0.824–0.873) in eICU, with eICU AUPRC 0.235 (0.170–0.293), Brier score 0.0341 (0.0283–0.0399), calibration intercept – 0.139 (–0.507 to 0.224), and slope 1.045 (0.950–1.166). In eICU, AUROC was 0.869 (0.848–0.894) when early hemoglobin was measured and 0.727 (0.668–0.779) when it was not. At the development-locked Youden review operating point, sensitivity was 69.5%, specificity 85.0%, positive predictive value 16.1%, and 17.1% of stays were flagged.

Conclusions.

A six-hour landmark framework provided cross-database risk ranking for a subsequent documented pRBC-record event, but performance depended on the available information state and documentation workflow. The score may support targeted clinical review and reassessment; it does not diagnose bleeding or traumatic coagulopathy and is not a transfusion trigger.

Registration

This retrospective database study was not registered.

Background

Major traumatic bleeding is time-critical and requires rapid recognition, hemorrhage control, and coagulation-aware resuscitation [1]. After the initial resuscitation, ICU clinicians face a different task: identifying patients who may still require blood-product review, repeat testing, or preparation while physiology and documentation continue to evolve.

Published trauma transfusion-prediction studies use heterogeneous prediction times, predictor windows, and outcomes, and relatively few models have been evaluated in a distinct dataset [2]. Most focus on prehospital care, emergency-department arrival, or massive-transfusion activation rather than subsequent ICU management.

Schwinn et al. evaluated a three-hour ICU observation window followed by a two-hour transfusion-prediction window across five databases [3]. Sheikhalishahi et al. examined postoperative ICU information windows extending from one to six hours [4]. A related surgical-ICU study used three hours of vital signs and biomarkers to predict hemorrhage-control intervention within nine hours [5]. These studies are temporally close comparators, but their populations and endpoints do not reproduce a six-hour post-admission pRBC-record target in adult trauma ICU patients.

Routinely collected ICU data also encode clinical workflow. Whether and how often a laboratory test is measured can influence both missingness indicators and extrema, making the information state part of the transportability problem. Trauma-focused methodological commentary therefore emphasizes intended use and clinical relevance [6]. Evaluation in a new setting remains context dependent rather than permanently certifying a prediction model [7].

We developed and evaluated five models for a subsequent qualifying pRBC-record event among adult trauma ICU patients with no qualifying pRBC record available by six hours. MIMIC-IV supported development and held-out testing; eICU-CRD supported cross-database evaluation. We further examined calibration, hypothetical review utility, SHAP behavior, information-state boundaries, competing events, fixed operating points, and robustness to alternative model, feature, documentation, follow-up, and landmark contracts.

Methods

Study design, data sources, and reporting

This retrospective clinical prediction-model study used two deidentified public ICU databases. MIMIC-IV v3.1 covers 2008–2022 and was used for development and held-out testing [8, 9]. Because exact unshifted ICU admission dates are unavailable, we report the database coverage and use `anchor_year_group` only as a deidentification-era proxy; included MIMIC trauma stays spanned all five groups from 2008–2010 through 2020–2022. The included eICU-CRD v2.0 stays occurred during 2014–2015 and were used for cross-database evaluation across multiple US hospitals [10, 11]. We used all qualifying records in these fixed releases; the source cohorts were extracted on 7 August 2026, and no later clinical records can enter the releases. No patient linkage was attempted between databases, and no eICU information was used for feature selection, hyperparameter specification, model fitting, or recalibration.

The intended users are ICU clinicians responsible for post-landmark reassessment of adult trauma patients. The model is intended to support targeted review of current physiology, laboratory testing, ongoing procedures or bleeding concerns, and blood-product readiness; it is not intended for autonomous use, diagnosis, or automatic transfusion decisions.

Reporting followed TRIPOD [12]. RECORD extends STROBE for observational research using routinely collected health data [13], and STROBE recommendations were also followed [14]. PROBAST domains were considered when describing quality, risk of bias, and applicability, but this was not presented as a formal independent assessment [15].

The MIMIC-IV research resource was approved for data sharing by the Institutional Review Board of Beth Israel Deaconess Medical Center, with a waiver of informed consent [8]. eICU-CRD was released as a deidentified research resource [10]. Gong Zhang obtained credentialed access under the applicable database data-use agreements. Under the applicable database governance and the authors' institutional requirements, this secondary analysis required no additional local ethics application or individual informed consent.

Study population and six-hour landmark

We identified adults aged 18 years or older with a trauma admission and retained the first qualifying ICU stay per patient. Time zero was ICU admission. Stays ending at or before six hours were excluded. The primary cohort required no qualifying pRBC record available by the six-hour landmark (`early_prbc_recorded_by_landmark` = 0), a rule that could be determined from the recorded information state at prediction time. MIMIC contained 9,769 first qualifying adult trauma ICU stays; 98 ended by six

hours and 622 had a qualifying pRBC record available by the landmark, leaving 9,049 stays. eICU contained 7,081 first qualifying stays; 210 ended by six hours and 293 had an early record available, leaving 6,578 stays (Fig. 1).

Study size was determined by all qualifying stays available in the specified database releases after application of the locked eligibility criteria. No a priori sample-size or post hoc power calculation, outcome-based sample reduction, or discretionary subsampling was performed. The MIMIC development set contained 396 events for 22 retained model columns (approximately 18 events per retained column), providing a pragmatic basis for the fixed-complexity candidate models, development-only out-of-fold procedures, and subsequent held-out evaluation; this event-to-column summary is not a substitute for a formal model-specific sample-size calculation. The eligible MIMIC cohort retained its locked 75:25 development/test split, and all eligible eICU stays were used for cross-database evaluation.

The MIMIC cohort retained its locked outcome-stratified 75:25 split: 6,779 stays for development and 2,270 for held-out testing. The primary eICU cohort represented 176 hospitals. The prediction window was the first six hours after ICU admission. Follow-up began immediately after the six-hour landmark and ended at the first qualifying pRBC-record event, ICU exit, or 24 hours after ICU admission, whichever occurred first. ICU discharge or death before a qualifying pRBC event was treated as a competing event in the prespecified sensitivity analyses.

pRBC phenotype, eligibility, and outcome

In MIMIC-IV, qualifying standard ICU pRBC records were identified from inpatient events using the locked product item set, positive recorded volume, and non-rewritten records. Occurrence of a qualifying pRBC event was timed by starttime. For real-time eligibility, an event starting or continuing during the first six hours counted as an early record only when its storetime was no later than the landmark; storetime was not used to time the subsequent outcome. Laboratory availability similarly required charttime within the predictor window and storetime no later than the landmark.

In eICU, qualifying records were positive numeric intakeoutput entries carrying one of the locked pRBC labels. Event occurrence was timed by intakeoutputOffset, and real-time early-record availability additionally required intakeoutputEntryOffset no later than 360 minutes. Subsequent outcomes were timed by intakeoutputOffset in both the primary and recorded-by-horizon sensitivity definitions. For laboratory predictors, measurement time was labResultOffset; an otherwise eligible result was available only when $\max(\text{labResultOffset}, \text{labResultRevisedOffset})$ was no later than 360 minutes, with the original offset used when no revised offset existed. Vital signs used observationOffset because eICU provides no separate entry or store timestamp for these tables. The eICU table does not provide an independent infusion end time, administration status, or transfusion episode identifier. Consequently, the outcome is reported as a subsequent documented pRBC event and not as biological transfusion need or a confirmed new transfusion episode.

Source-specific timing differences were preserved explicitly. A retrospective event-time eligibility rule ($\text{early_prbc} = 0$) was evaluated separately because it can use early events entered after the landmark.

Cohort eligibility and the outcome were derived programmatically from structured database fields using locked, database-specific code, product-label, value, and timestamp rules. No clinician adjudication, image interpretation, chart review, or other subjective outcome assessment was performed.

Primary eligibility excluded stays with a qualifying pRBC record available by the six-hour landmark; the retrospective event-time exclusion was examined only as a sensitivity analysis. Treatments and procedures other than pRBC records were not included as predictors or adjustment variables, and the post-landmark qualifying pRBC record constituted the outcome. No treatment-effect analysis was performed.

Predictors and preprocessing

The 12 initial predictors were specified on the basis of clinical relevance to post-trauma physiology and their ability to be harmonized across both databases. No univariable outcome screening, assessment-dataset feature selection, or eICU-guided predictor selection was used. The common predictors were age, female sex, minimum hemoglobin, minimum platelet count, maximum lactate, minimum pH, minimum base excess, maximum international normalized ratio, maximum heart rate, minimum systolic blood pressure, maximum respiratory rate, and minimum peripheral oxygen saturation. Dynamic values were restricted to audited plausible ranges and summarized from information available during the first six hours. Each variable with observed missingness contributed a missingness indicator. In critical-care prediction, missingness can encode observation and workflow patterns and should not automatically be interpreted as a biological mechanism [16].

Predictor values and their availability indicators were derived programmatically from structured demographic, laboratory, and vital-sign fields using the locked six-hour window and plausible-value rules; no subjective assessor interpretation was involved.

Within each training fold, numeric imputation parameters were estimated from training observations only and applied unchanged to assessment observations. SVM-RBF additionally used training-fold means and standard deviations. Constant and exactly duplicated columns were removed within the training split, and ordinary logistic regression underwent QR full-rank filtering. No assessment-set feature selection, outcome oversampling, or undersampling was performed. The primary specification retained 22 model columns after removal of constant age- and sex-missing indicators (Supplementary Table S15).

Model development, recalibration, and reference model

Five stratified development folds generated out-of-fold predictions for logistic regression; elastic net (alpha 0.5; lambda 0.01); probability random forest (350 trees, minimum node size 20, and $mtry = \max[2, \text{floor}(\sqrt{p})]$); weighted XGBoost (depth 2, learning rate 0.05, 200 rounds, minimum child weight 5, row and column subsampling 0.8, L2 penalty 1, and development class-ratio weighting); and SVM-RBF (cost 1; gamma $1/p$; training-fold-only centering and scaling). Fixed configurations were selected for reproducibility and were not exhaustively tuned against eICU.

Because the outcome was uncommon, weighted XGBoost assigned the positive class a training-split weight equal to the ratio of non-events to events. This weighting addressed class imbalance during fitting without over- or undersampling and did not alter the observed outcome prevalence; probability distortion was addressed by the development-only logistic recalibration described below.

For each algorithm, a two-parameter logistic recalibration model was fitted only to raw MIMIC development out-of-fold predictions. Each final model was then refitted on the full development set and the frozen recalibration map was applied unchanged to MIMIC test and eICU probabilities. Calibration was assessed because discrimination alone does not establish agreement between predicted and observed risk [17]. Weighted XGBoost was retained post hoc as the reference model for exploratory DCA, SHAP, competing-risk, and sensitivity analyses. This designation was not intended to rank algorithms as universally better.

Performance, uncertainty, and clinical-utility analyses

We reported AUROC, AUPRC, Brier score, calibration intercept, calibration slope, and equal-count calibration plots. Primary performance intervals used 1,000 individual-stay bootstrap resamples for MIMIC and 1,000 hospital-cluster resamples for eICU. Extended and Fine-Gray sensitivities used 1,000 resamples with the same database-specific resampling units; washout differences used 2,000 resamples. Candidate-versus-reference AUROC comparisons used 10,000 paired resamples on identical rows, a two-sided + 1 corrected sign-tail P value, and Holm adjustment across four comparisons within each dataset. P values were not calculated for between-model differences in AUPRC, Brier score, or calibration.

Resampling eICU hospitals with replacement preserved within-hospital dependence when estimating uncertainty for pooled cross-database performance. This procedure was not an analysis of between-hospital heterogeneity, and no hospital-specific

performance or calibration parameters, random-effects variance, or formal heterogeneity statistic was estimated.

Decision-curve analysis evaluates net benefit across threshold probabilities and depends on the consequences assigned to false-positive and false-negative decisions [18]. We interpreted this exploratory DCA only as a hypothetical additional clinical review or reassessment strategy because no implemented workflow or utility weights were evaluated. The prespecified analysis covered 1%-20% thresholds; Fig. 3 displays 1%-15% to preserve legibility in the clinically relevant low-risk range. Review all and review none were descriptive comparators; model thresholds were not transfusion triggers.

Model interpretation, competing events, and operating points

TreeSHAP was calculated for the uncalibrated reference-model log-odds margin. SHAP values describe fitted-model behavior rather than causal or physiological effects [19]. Global importance used mean absolute SHAP values, and dependence displays used observed predictor values only. Margin additivity and observed-value provenance were audited.

ICU discharge or death before a subsequent pRBC event was treated as a competing event. Aalen-Johansen cumulative incidence was displayed by score tertiles fixed from MIMIC development out-of-fold predictions. Fine-Gray subdistribution hazards quantified association per one development-standard-deviation increase in the standardized score. MIMIC inference included individual-stay bootstrap intervals; eICU inference used hospital-cluster bootstrap intervals only, and ordinary model-based crr P values were not used for eICU. These analyses assessed score association and risk-group ordering, not full competing-risk prediction validation.

Three review/reassessment thresholds were selected exclusively from calibrated MIMIC development out-of-fold predictions: maximum Youden index, the highest threshold retaining at least 80% development sensitivity, and the highest threshold retaining at least 90% development sensitivity. Thresholds were applied unchanged to MIMIC test and eICU, with sensitivity, specificity, positive and negative predictive values, and flagged fraction reported with bootstrap intervals.

Robustness and sensitivity analyses

Model-form analyses compared unweighted XGBoost, three-degree-of-freedom natural-spline logistic regression, and hemoglobin-only logistic regression with the reference. Feature-form analyses evaluated values without explicit missingness indicators, missingness indicators alone, measurement counts alone, and first- or last-value summaries. Information-state analyses stratified performance by availability of early hemoglobin and compared the full model with a measured-only hemoglobin logistic model on identical rows.

Additional analyses restricted to complete 24-hour ICU follow-up; used a MIMIC-only expanded pRBC-source screen; assessed MIMIC FinishedRunning records; used a later anchor-year-group proxy; required eICU outcome documentation by the horizon; retained the complete landmark cohort and added early-pRBC-recorded-by-six-hours as a predictor; reinstated the retrospective event-time eligibility definition; and imposed one- and two-hour washouts without model refitting. Analyses changing population, outcome, or information state were treated as different estimands rather than head-to-head performance comparisons.

Race, ethnicity, socioeconomic status, and other sociodemographic attributes were not available in a consistently harmonized form across the two databases. We therefore performed no fairness-oriented model constraints, reweighting, post-processing, formal health-inequality analysis, or assessment of differential performance across those groups. Inclusion of age and recorded sex as predictors should not be interpreted as evidence of equitable performance.

Software and reproducibility

Analyses used R 4.5.0 with data.table 1.18.2.1, glmnet 5.0, ranger 0.18.0, xgboost 3.2.1.1, e1071 1.7–17, pROC 1.19.0.1, PRROC 1.4, survival 3.8-6, and cmprsk 2.2–12. Each formal run wrote a session record, input hashes, preprocessing audits,

prediction-level outputs, and completion checks. Additional file 2 provides aggregate table and figure source data plus an executable, privacy-audited reference weighted-XGBoost object, predictor contract, scoring function, synthetic example, and numerical verification. The fitted reference object reproduced 100 locked held-out MIMIC raw and recalibrated probabilities with maximum absolute differences below $1e-12$ after serialization. Comparator specifications are reported in Supplementary Table S15. Restricted patient-level database records, prediction-level outputs, patient-level SHAP values, and fitted comparator objects that embed row-level training features are not redistributed; the analysis scripts and retained run-provenance materials are available from the corresponding authors on reasonable request, subject to the database agreements.

Results

Cohort derivation and characteristics

The primary MIMIC cohort comprised 9,049 stays with 532 subsequent pRBC events (5.9%); 6,779 development stays with 396 events and 2,270 held-out test stays with 136 events. The eICU cohort included 6,578 stays with 262 events (4.0%) from 176 hospitals (Figure 1). Complete ICU follow-up through 24 hours was available for 75.9% of MIMIC and 72.1% of eICU stays. Compared with MIMIC, the eICU population was younger and had substantially greater missingness for several laboratory predictors (Table 1). Descriptive between-database P values were not used for model selection.

Five-model benchmark

No single algorithm dominated both datasets and all metrics (Table 2; Figure 2). In the MIMIC held-out test set, reference weighted XGBoost had the highest AUROC point estimate, 0.844 (95% CI, 0.810-0.876), whereas random forest had the highest AUPRC point estimate, 0.294 (0.216-0.378). The reference XGBoost AUPRC was 0.275 (0.214-0.355), and its Brier score was 0.0489 (0.0425-0.0559). Logistic regression, elastic net, and random forest did not differ significantly from reference XGBoost in Holm-adjusted AUROC comparisons ($P=0.307$, 0.209 , and 0.095); SVM-RBF performed lower (adjusted $P<0.001$).

In eICU, elastic net had a marginally higher AUROC point estimate than reference XGBoost (0.852 vs 0.851), whereas reference XGBoost had the highest AUPRC point estimate (0.235; 95% CI, 0.170-0.293). The adjusted AUROC P value for elastic net versus reference XGBoost was 0.819. Logistic regression and random forest also showed no adjusted AUROC difference ($P=0.302$ and 0.134), whereas SVM-RBF was lower ($P<0.001$). Thus, XGBoost was used as a common reference for downstream analyses, not described as the best algorithm.

Calibration and hypothetical review utility

For reference XGBoost, the MIMIC calibration intercept was -0.051 (95% CI, -0.384 to 0.351) and slope 0.958 (0.824-1.112). In eICU, Brier score was 0.0341 (0.0283-0.0399), calibration intercept -0.139 (-0.507 to 0.224), and slope 1.045 (0.950-1.166) (Table 2; Figure 3). Equal-count plots showed broadly preserved probability ordering in the low-risk range. Across clinically illustrative thresholds, several models had positive net benefit over review-none; these curves quantify a hypothetical additional review strategy and not an automatic transfusion action.

Reference-model interpretation and information state

Minimum hemoglobin had the largest eICU mean absolute TreeSHAP value (0.971), followed by minimum systolic blood pressure (0.438), maximum heart rate (0.243), age (0.157), minimum oxygen saturation (0.127), maximum lactate (0.113), minimum base excess (0.111), maximum INR (0.106), minimum pH (0.104), and minimum platelet count (0.062) (Figure 4; Supplementary Table S14). Dependence displays were restricted to observed hemoglobin and systolic-pressure values and should not be read as causal thresholds.

Early hemoglobin availability marked an important performance boundary. In eICU, reference-model AUROC was 0.869 (95% CI, 0.848-0.894) among 2,975 stays with hemoglobin measured (191 events) and 0.727 (0.668-0.779) among 3,603 stays without an early measurement (71 events); the conditional difference was 0.142 (0.092-0.201). On measured-hemoglobin rows, reference XGBoost exceeded a strict hemoglobin-only logistic comparator by 0.020 AUROC (0.008-0.029) (Table 3; Supplementary Table S7).

Model-form, feature-form, and washout robustness

Unweighted XGBoost remained close to the reference in eICU (AUROC difference -0.002; 95% CI, -0.007 to 0.003), as did spline logistic regression (-0.009; -0.020 to 0.000); hemoglobin-only logistic regression was lower (-0.060; -0.079 to -0.043). Values-only XGBoost retained AUROC 0.849 (0.826-0.870), with a paired difference of -0.002 (-0.005 to 0.002). This similarity does not show that missingness was inconsequential, because values-only preprocessing, measurement availability, and extrema can still reflect documentation workflow. Missingness-only and measurement-count-only models retained some discrimination, 0.667 (0.613-0.706) and 0.693 (0.651-0.723), but remained substantially lower than the reference; workflow-only signal therefore did not account for full-model performance. First-value AUROC was 0.836 (0.811-0.862) and last-value AUROC 0.848 (0.826-0.867) (Table 3; Figure 5).

After a two-hour washout, the unchanged reference model retained AUROC 0.833 (0.794-0.867) among 2,233 MIMIC stays with 115 events and 0.840 (0.814-0.861) among 6,435 eICU stays with 219 events. Paired differences from no washout were -0.012 (-0.025 to 0.004) and -0.011 (-0.019 to -0.002), respectively. Because washout changed the risk set and prevalence, accompanying AUPRC and Brier changes were not interpreted as pure model improvement or deterioration (Figure 5; Supplementary Table S13).

Competing events, risk groups, and fixed operating points

Per one development-standard-deviation increase in the score, the Fine-Gray subdistribution hazard ratio was 4.14 (individual-stay bootstrap 95% CI, 3.47-5.12) in MIMIC and 4.75 (hospital-cluster bootstrap 95% CI, 4.19-5.50) in eICU; no ordinary model-based crr P value was reported for eICU. Using development-fixed score tertiles, 18-hour Aalen-Johansen cumulative incidence was 0.87%, 2.09%, and 14.83% in MIMIC low-, intermediate-, and high-risk groups and 0.38%, 1.52%, and 10.78% in eICU (Figure 6; Supplementary Table S9). These results demonstrate ordered cumulative incidence, not individual competing-risk calibration.

At the development-locked Youden threshold of 0.06636, eICU sensitivity was 69.5% (95% CI, 62.6%-74.6%), specificity 85.0% (83.1%-86.8%), positive predictive value 16.1% (12.7%-19.5%), negative predictive value 98.5% (98.1%-98.9%), and flagged fraction 17.1% (15.3%-19.2%). Lower locked thresholds increased eICU sensitivity to 76.7% and 90.5% but flagged 24.2% and 44.1% of stays, respectively (Figure 6; Supplementary Table S10). These estimates display the sensitivity-workload trade-off for review, not a transfusion indication.

Alternative follow-up, documentation, and landmark contracts

When the locked model was restricted to complete 24-hour follow-up, AUROC was 0.833 (95% CI, 0.798-0.868) in MIMIC and 0.842 (0.815-0.862) in eICU. A MIMIC-only expanded early-source sensitivity yielded AUROC 0.845 (0.808-0.881). A later MIMIC anchor-year-group proxy test yielded AUROC 0.892 among 1,114 stays but only 24 events; this was treated as a temporal proxy rather than exact calendar-time validation. FinishedRunning-only and documentation-by-horizon sensitivities also preserved similar point discrimination (Supplementary Tables S6, S8, and S16; Supplementary Figure S2).

The secondary full-landmark operational model retained all six-hour landmark stays and added early pRBC recorded by six hours as a predictor. It achieved AUROC 0.873 (0.848-0.897) among 2,419 MIMIC test stays with 201 events and 0.878 (0.860-

0.896) among 6,871 eICU stays with 380 events. This model answers a broader operational question and is not directly comparable with the primary no-early-record estimand (Table 3; Supplementary Table S11 and Figure S5).

Recorded-by-landmark and retrospective event-time early-pRBC states were discordant for 28 MIMIC development stays, 14 MIMIC test stays, and 32 eICU stays; 14, 8, and 11 subsequently met the outcome. Under retrospective event-time eligibility, AUROC/AUPRC were 0.845 (0.812-0.879)/0.268 (0.208-0.347) in MIMIC and 0.851 (0.826-0.872)/0.227 (0.166-0.291) in eICU. Similar discrimination does not remove the real-time information problem, so this definition remained a sensitivity analysis (Supplementary Table S12).

Discussion

Principal findings

In two public ICU databases, a prediction framework anchored to information available six hours after ICU admission discriminated a subsequent documented pRBC-record event among adult trauma patients without an early pRBC record available by that landmark. Four of five candidate algorithms achieved broadly similar discrimination; elastic net had a marginally higher eICU AUROC point estimate than reference XGBoost, while XGBoost retained the highest eICU AUPRC point estimate. This pattern supports a model-family conclusion rather than a claim that one complex algorithm was uniformly better.

The study's most clinically important boundary was not the difference between logistic regression and XGBoost, but the information state. Performance was substantially higher when early hemoglobin had been measured. Values-only performance remained close to the full model, whereas missingness-only and measurement-count-only models were weaker but non-random. These results indicate that physiology and documentation workflow jointly shape transportability.

Comparison with related studies

Most recent trauma models target early massive transfusion at or near presentation [20]. Benjamin et al. developed an early massive-transfusion model in traumatic hemorrhage [21], while De La Rosa et al. examined early transfusion requirements [22] and Safari et al. early pRBC receipt [23]. Sigle et al. studied prehospital transfusion need [24], Richards et al. acute traumatic coagulopathy [25], and Yu et al. on-scene massive-transfusion prediction [26]. Each addresses a different decision time or outcome; their AUROCs should not be ranked directly against a post-ICU landmark estimand.

A meta-model in non-traumatic ICU patients addresses a broader critical-care population [27]. The closest temporal comparisons are the five-database post-ICU analysis [3], postoperative ICU work [4], and surgical-ICU hemorrhage-control model [5]. Our design differs by fixing real-time recorded eligibility at six hours, limiting the population to trauma ICU stays, separating the later pRBC window, and auditing both event time and documentation availability.

Prospective ShockMatrix work has examined workflow feasibility [28] and compared model predictions with clinicians [29]. Our retrospective database analysis cannot provide either type of evidence. A 2024 systematic review emphasized substantial heterogeneity in trauma transfusion prediction [2]. A subsequent systematic review and meta-analysis similarly showed why evidence synthesis must account for differences in intended use and study design [30]. Together, these findings reinforce the need to compare landmark, phenotype, prevalence, and information availability before comparing performance statistics.

Clinical and informatics interpretation

The intended action is targeted review or reassessment: for example, verifying current physiology, repeating selected laboratory tests, reviewing ongoing bleeding or procedures, and checking blood-product readiness. It is not automatic transfusion. At the locked Youden operating point, approximately 17 reviews would be generated per 100 eligible eICU stays

while identifying about 70% of subsequent events; higher-sensitivity strategies substantially increased workload. Any local use would therefore need an explicit alert pathway and prospective measurement of burden, timeliness, safety, and clinician response.

TreeSHAP ranked hemoglobin and hemodynamic features highly, but these attributions describe model behavior, not pathophysiological causation or treatment thresholds. Extrema may partly reflect monitoring intensity because repeatedly tested patients have more opportunities to generate extreme observations. The small difference between values-only and full models argues against explicit missingness indicators being the sole driver, but the missingness- and measurement-count-only results confirm that workflow carries predictive information.

Clinical adoption of transfusion prediction systems requires workflow and safety evaluation [31]. Deployed prediction tools also require monitoring for dataset shift and performance deterioration [32]. Local implementation should first reproduce the real-time pRBC phenotype and feature timestamps, perform silent prospective evaluation, recalibrate probabilities, inspect information-state and subgroup performance, compare predictions with clinicians or transparent bedside rules, and measure alert workload before any decision support is activated.

Strengths

Strengths include a time-anchored eligibility rule based on records available at prediction, a locked held-out MIMIC test set, evaluation across 176 eICU hospitals without eICU refitting, common preprocessing across five algorithms, development-only recalibration, hospital-cluster uncertainty, and evaluation beyond AUROC through AUPRC, Brier score, calibration, decision curves, SHAP, competing events, and fixed operating points. The internal analysis archive retained prediction-level audit files, content hashes, and completion checks. The submission provides aggregate table and figure source data and a privacy-audited executable reference weighted-XGBoost package; versioned analysis code and retained run-provenance materials are available from the corresponding authors on reasonable request, subject to the database agreements.

Limitations

First, the outcome is a qualifying documented ICU pRBC event, not biological transfusion need, active hemorrhage, traumatic coagulopathy, or adjudicated appropriateness. In eICU, limited administration timing and status fields prevent confirmation of a new transfusion episode.

Second, early-pRBC and timestamp contracts were necessarily database specific. MIMIC has charttime, storetime, starttime, and status information that eICU does not fully reproduce. The primary recorded-by-landmark rule improves real-time interpretability but cannot eliminate source asymmetry.

Third, approximately one quarter of stays left ICU observation before 24 hours. Complete-follow-up and competing-event analyses were reassuring for score ordering, but the primary estimand remains documented transfusion under routine ICU disposition processes rather than transfusion need irrespective of disposition.

Fourth, laboratory missingness was substantial, particularly in eICU, and the hemoglobin-unmeasured stratum performed less well. Median imputation and missingness indicators do not recover unobserved physiology, and extrema can encode measurement intensity.

Fifth, eICU reflects 2014-2015 practice and cannot establish transportability to contemporary transfusion workflows. The later MIMIC anchor-year-group analysis was only a deidentification-era proxy with few events, and the primary MIMIC split was random rather than temporal.

Sixth, weighted XGBoost became the reference model post hoc, model configurations were fixed rather than exhaustively tuned, and earlier project iterations had examined held-out and cross-database results. The study should therefore be read as

development with held-out and cross-database evaluation, not pristine confirmatory evidence.

Seventh, the hospital-cluster intervals accounted for dependence but should not be interpreted as evidence of consistent performance across individual hospitals; between-hospital performance heterogeneity was not estimated.

Finally, there was no prospective deployment, comparison with clinician judgment in the same workflow, causal analysis, fairness mitigation, formal fairness or health-inequality analysis, or treatment-effect assessment. Sociodemographic attributes beyond age and recorded sex were not consistently harmonized across databases. Threshold results are retrospective workload illustrations only.

Conclusions

A six-hour landmark framework using routinely documented ICU information provided cross-database risk ranking for a subsequent qualifying pRBC-record event among adult trauma patients with no early pRBC record available at prediction. Several models performed similarly; reference weighted XGBoost achieved eICU AUROC 0.851 and AUPRC 0.235, but performance was strongly conditioned by hemoglobin availability and documentation workflow. The score may support targeted review after local recalibration and prospective silent evaluation. It does not diagnose bleeding or traumatic coagulopathy, define transfusion indication, or replace clinical judgment.

Abbreviations

AUPRC
area under the precision-recall curve
AUROC
area under the receiver operating characteristic curve
CI
confidence interval
DCA
decision-curve analysis
eICU-CRD
eICU Collaborative Research Database
ICU
intensive care unit
INR
international normalized ratio
NPV
negative predictive value
OOF
out-of-fold
pRBC
packed red blood cell
PPV
positive predictive value
SHAP
Shapley additive explanations
SHR
subdistribution hazard ratio
SVM-RBF
support vector machine with radial basis function kernel
XGBoost

extreme gradient boosting.

Declarations

Ethics approval and consent to participate

Gong Zhang obtained credentialed access to MIMIC-IV v3.1 and eICU-CRD v2.0 under the applicable data-use agreements. This secondary analysis used deidentified public research data and involved no direct participant contact or intervention. Under the applicable database governance and institutional requirements, no additional local ethics application or informed consent was required.

Consent for publication

Not applicable.

Availability of data and materials

MIMIC-IV v3.1 and eICU-CRD v2.0 are available through PhysioNet to qualified users who independently complete the required credentialing and data-use agreements. Additional file 2 contains aggregate table and figure source data and a privacy-audited executable reference weighted-XGBoost package with its predictor contract, R scoring function, synthetic example, and numerical verification. The analysis scripts and retained run-provenance materials are available from the corresponding authors on reasonable request, subject to the database agreements. Patient-level records, derived patient-level datasets, prediction-level outputs, patient-level SHAP values, and fitted comparator objects that embed row-level training features cannot be redistributed.

Competing interests

The authors declare that they have no competing interests.

Funding

This work was supported by the National Natural Science Foundation of China (Nos. 82460375 and 81960361); the National Natural Science Foundation of China-China Academy of General Technology Joint Fund for Basic Research (No. 82260385); the Science and Technology Support Program of Guizhou Province (Qian Ke He Support [2025] General 121); the Basic Research Program of Guizhou Province (Qian Ke He Foundation-ZK [2024] General 341 and Qian Ke He Foundation MS [2026] 937); the Zhuhai Municipal Science and Technology Program for Social Development (Nos. 2420004000300, 2420004000107, and 2520004003291); the Guizhou Postgraduate Innovation Program (No. 2024YJSKYJJ323); the Science and Technology Fund of the Health Commission of Guizhou Province (Nos. gzwkj2025-011, gzwkj2025-015, gzwkj2025-208, and gzwkj2025-345); the Health Commission of Guizhou Province (No. gzwkj2022-103); the Guizhou Clinical Research Center for Emergency Medicine (No. LCZX[2025]004); the Science and Technology Program of Guizhou Province (No. ZK [2023]580); the Medical Research Union Fund for High-quality Health Development of Guizhou Province (No. 2024GZYXKYJJXM0035); the National Clinical Key Specialty Construction Project (No. XM04212); and the Collaborative Innovation Center of Chinese Ministry of Education (No. 2020-39). Qian Ke He Foundation MS [2026] 937 was awarded to Gong Zhang. Nos. 82260385, ZK [2023]580, gzwkj2022-103, and 2024GZYXKYJJXM0035 were awarded to Haizhen Duan. The funders had no role in the study design, data collection, data analysis, interpretation of the results, preparation of the manuscript, or the decision to submit the manuscript for publication.

Study protocol

No formal study protocol was prepared for this retrospective database study.

Registration

This retrospective database study was not registered.

Patient and public involvement

Patients and/or the public were not involved in the design, conduct, reporting, interpretation, or dissemination plans of this study.

Authors' contributions

GZ: Conceptualization, Methodology, Data curation, Formal analysis, Software, Visualization, Funding acquisition, and Writing - original draft. CY: Validation, Investigation, and Writing - review & editing. SJ and XL: Clinical interpretation, Supervision, and Writing - review & editing. HD: Conceptualization, Methodology, Supervision, Project administration, Funding acquisition, and Writing - review & editing. All authors read and approved the final manuscript.

Acknowledgements

Not applicable.

References

1. Rossaint R, Afshari A, Bouillon B, Cerny V, Cimpoesu D, Curry N, et al. The European guideline on management of major bleeding and coagulopathy following trauma: sixth edition. *Critical care* (London, England). 2023;27(1):80. doi:10.1186/s13054-023-04327-7. PMID: 36859355.
2. Oakley W, Tandle S, Perkins Z, Marsden M. Predicting blood transfusion following traumatic injury using machine learning models: A systematic review and narrative synthesis. *The journal of trauma and acute care surgery*. 2024;97(4):651-659. doi:10.1097/TA.0000000000004385. PMID: 38720200.
3. Schwinn J, Sheikhalishahi S, Morhart M, Soto Rey I, Simon P, Kaspar M, et al. Predicting blood transfusion after ICU admission in five databases: A comparison of three machine learning paradigms. *Digital health*. 2026;12:20552076261428383. doi:10.1177/20552076261428383. PMID: 42078831.
4. Sheikhalishahi S, Goss S, Seidlmayer LK, Zaghdoudi S, Hinske LC, Kaspar M. Predicting blood transfusion demand in intensive care patients after surgery by comparative analysis of temporally extended data selection. *BMC medical informatics and decision making*. 2024;24(1):397. doi:10.1186/s12911-024-02800-z. PMID: 39695617.
5. Forrester JP, Del Rio MB, Meyer CH, Paci SPR, Rastegar ER, Li T, et al. A Combined Model of Vital Signs and Serum Biomarkers Outperforms Shock Index in the Prediction of Hemorrhage Control Interventions in Surgical Intensive Care Unit Patients. *Journal of intensive care medicine*. 2025;40(6):632-641. doi:10.1177/08850666241312614. PMID: 39924935.
6. Tandle S, Wohlgemut JM, Marsden MER, Pisirir E, Kyrimi E, Stoner RS, et al. Enhancing the clinical relevance of haemorrhage prediction models in trauma. *Military Medical Research*. 2023;10(1):43. doi:10.1186/s40779-023-00476-6. PMID: 37726859.
7. Van Calster B, Steyerberg EW, Wynants L, van Smeden M. There is no such thing as a validated prediction model. *BMC medicine*. 2023;21(1):70. doi:10.1186/s12916-023-02779-w. PMID: 36829188.
8. Johnson AEW, Bulgarelli L, Shen L, Gayles A, Shammout A, Horng S, et al. MIMIC-IV, a freely accessible electronic health record dataset. *Scientific data*. 2023;10(1):1. doi:10.1038/s41597-022-01899-x. PMID: 36596836.
9. Johnson A, Bulgarelli L, Pollard T, Gow B, Moody B, Horng S, et al. MIMIC-IV (version 3.1). *PhysioNet*. 2024. doi:10.13026/kpb9-mt58.
10. Pollard TJ, Johnson AEW, Raffa JD, Celi LA, Mark RG, Badawi O. The eICU Collaborative Research Database, a freely available multi-center database for critical care research. *Scientific data*. 2018;5:180178. doi:10.1038/sdata.2018.178. PMID: 30204154.
11. Pollard T, Johnson A, Raffa J, Celi LA, Badawi O, Mark R. eICU Collaborative Research Database (version 2.0). *PhysioNet*. 2019. doi:10.13026/C2WM1R.

12. 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 (Clinical research ed.)*. 2024;385:e078378. doi:10.1136/bmj-2023-078378. PMID: 38626948.
13. Benchimol EI, Smeeth L, Guttman A, Harron K, Moher D, Petersen I, et al. The REporting of studies Conducted using Observational Routinely-collected health Data (RECORD) statement. *PLoS medicine*. 2015;12(10):e1001885. doi:10.1371/journal.pmed.1001885. PMID: 26440803.
14. von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP, et al. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. *Journal of clinical epidemiology*. 2008;61(4):344-9. doi:10.1016/j.jclinepi.2007.11.008. PMID: 18313558.
15. 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 (Clinical research ed.)*. 2025;388:e082505. doi:10.1136/bmj-2024-082505. PMID: 40127903.
16. Singh J, Sato M, Ohkuma T. On Missingness Features in Machine Learning Models for Critical Care: Observational Study. *JMIR medical informatics*. 2021;9(12):e25022. doi:10.2196/25022. PMID: 34889756.
17. 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 medicine*. 2019;17(1):230. doi:10.1186/s12916-019-1466-7. PMID: 31842878.
18. Zhao L, Leng Y, Hu Y, Xiao J, Li Q, Liu C, et al. Understanding decision curve analysis in clinical prediction model research. *Postgraduate medical journal*. 2024;100(1185):512-515. doi:10.1093/postmj/qgae027. PMID: 38453146.
19. Ponce-Bobadilla AV, Schmitt V, Maier CS, Mensing S, Stodtmann S. Practical guide to SHAP analysis: Explaining supervised machine learning model predictions in drug development. *Clinical and translational science*. 2024;17(11):e70056. doi:10.1111/cts.70056. PMID: 39463176.
20. Nikouline A, Feng J, Rudzicz F, Nathens A, Nolan B. Machine learning in the prediction of massive transfusion in trauma: a retrospective analysis as a proof-of-concept. *European journal of trauma and emergency surgery : official publication of the European Trauma Society*. 2024;50(3):1073-1081. doi:10.1007/s00068-023-02423-5. PMID: 38265444.
21. Benjamin AJ, Young AJ, Holcomb JB, Fox EE, Wade CE, Meador C, et al. Early Prediction of Massive Transfusion for Patients With Traumatic Hemorrhage: Development of a Multivariable Machine Learning Model. *Annals of surgery open : perspectives of surgical history, education, and clinical approaches*. 2023;4(3):e314. doi:10.1097/AS9.0000000000000314. PMID: 37746616.
22. De La Rosa MR, Bowers A, Eastridge B, Wang M, Jenkins D. Early prediction of transfusion requirements in trauma patients using explainable machine learning. *Transfusion*. 2026;66 Suppl 1(Suppl 1):S42-S52. doi:10.1111/trf.70106. PMID: 41693248.
23. Safari S, Zarei H, Zare K, Aghili SH, Saadatipour N, Vazirizadeh-Mahabadi M, et al. Machine Learning Models for Predicting the Need for Early Packed Red Blood Cell Transfusion in Multiple Trauma Patients. *Archives of academic emergency medicine*. 2026;14(1):e1. doi:10.22037/aaem.v14i1.2820. PMID: 41664679.
24. Sigle M, Boos M, Weiss T, van Iersel M, Nwafor-Okoli C, McBeth P, et al. AI-enabled forecasting of prehospital transfusion needs in patients with trauma: a multinational, registry-based, retrospective, machine learning development and validation study. *The Lancet. Digital health*. 2026;8(1):100945. doi:10.1016/j.landig.2025.100945. PMID: 41620321.
25. Richards JE, Yang S, Kozar RA, Scalea TM, Hu P. A machine learning-based Coagulation Risk Index predicts acute traumatic coagulopathy in bleeding trauma patients. *The journal of trauma and acute care surgery*. 2025;98(4):614-620. doi:10.1097/TA.0000000000004463. PMID: 39330762.
26. Yu B, Cho J, Kim H, Hwang SH, Hwang J, Kim S, et al. On-scene machine learning prediction model for massive transfusion in trauma and its association with in-hospital mortality. *BJS open*. 2025;10(1). doi:10.1093/bjsopen/zraf167. PMID: 41705531.
27. Rafiei A, Moore R, Choudhary T, Marshall C, Smith G, Roback JD, et al. Robust Meta-Model for Predicting the Likelihood of Receiving Blood Transfusion in Non-traumatic Intensive Care Unit Patients. *Health data science*. 2024;4:0197.

doi:10.34133/hds.0197. PMID: 39507297.

28. Gauss T, Moyer JD, Colas C, Pichon M, Delhay N, Werner M, et al. Pilot deployment of a machine-learning enhanced prediction of need for hemorrhage resuscitation after trauma - the ShockMatrix pilot study. BMC medical informatics and decision making. 2024;24(1):315. doi:10.1186/s12911-024-02723-9. PMID: 39468585.

29. Gauss T, James A, Colas C, Delhay N, Holleville M, Bijok B, et al. Comparison of machine learning and human prediction to identify trauma patients in need of hemorrhage control resuscitation (ShockMatrix study): a prospective observational study. The Lancet regional health. Europe. 2025;55:101340. doi:10.1016/j.lanepe.2025.101340. PMID: 40584589.

30. Han C, Zhang G, Gao H, Zhang Y. Predicting Massive Transfusion Needs in Trauma Using Machine Learning: Systematic Review and Meta-Analysis. Journal of trauma nursing : the official journal of the Society of Trauma Nurses. 2026;33(4):228-240. doi:10.1097/JTN.0000000000000921. PMID: 42228445.

31. Maynard S, Farrington J, Raza S, Stanworth SJ. Artificial Intelligence Implementation in Transfusion Medicine: Addressing the Challenges of Clinical Adoption. Transfusion medicine reviews. 2026;40(1):150961. doi:10.1016/j.tmr.2026.150961. PMID: 41723890.

32. Davis SE, Embí PJ, Matheny ME. Sustainable deployment of clinical prediction tools-a 360° approach to model maintenance. Journal of the American Medical Informatics Association : JAMIA. 2024;31(5):1195-1198. doi:10.1093/jamia/ocae036. PMID: 38422379.

Tables

Table 1. Characteristics, predictor availability, follow-up, and outcome frequency in the primary cohorts

Characteristic	MIMIC-IV primary cohort (N=9,049)	MIMIC-IV missing, n (%)	eICU cross-database evaluation (N=6,578)	eICU missing, n (%)	P value
Age, years	66 [49, 80]	0 (0.0%)	59 [39, 77]	0 (0.0%)	<0.001
Female sex	3,464/9,049 (38.3%)	0 (0.0%)	2,467/6,564 (37.6%)	14 (0.2%)	0.376
Minimum hemoglobin, g/dL	11.3 [9.6, 12.8]	3,240 (35.8%)	11.5 [9.9, 13.1]	3,603 (54.8%)	<0.001
Minimum platelet count, 10 ⁹ /L	194 [147, 245]	3,226 (35.7%)	185 [143, 230]	3,895 (59.2%)	<0.001
Maximum lactate, mmol/L	2.1 [1.4, 3.2]	5,832 (64.4%)	2.3 [1.4, 3.5]	5,639 (85.7%)	0.009
Minimum pH	7.34 [7.28, 7.40]	5,613 (62.0%)	7.34 [7.28, 7.40]	5,370 (81.6%)	0.162
Minimum base excess, mmol/L	-2.0 [-5.0, 0.0]	5,723 (63.2%)	-3.5 [-6.4, -0.7]	5,434 (82.6%)	<0.001
Maximum INR	1.20 [1.10, 1.40]	4,385 (48.5%)	1.19 [1.10, 1.38]	5,115 (77.8%)	<0.001
Maximum heart rate, beats/min	93 [80, 107]	111 (1.2%)	97 [85, 112]	253 (3.8%)	<0.001
Minimum systolic blood pressure, mmHg	105 [92, 119]	122 (1.3%)	104 [90, 119]	256 (3.9%)	<0.001
Maximum respiratory rate, breaths/min	23 [20, 27]	142 (1.6%)	24 [21, 29]	608 (9.2%)	<0.001
Minimum SpO ₂ , %	95 [93, 98]	135 (1.5%)	94 [91, 97]	316 (4.8%)	<0.001
pRBC recorded from ICU admission through the 6-hour landmark	0/9,049 (0.0%)	0 (0.0%)	0/6,578 (0.0%)	0 (0.0%)	Not applicable
Complete ICU follow-up through 24 hours	6,865/9,049 (75.9%)	0 (0.0%)	4,744/6,578 (72.1%)	0 (0.0%)	<0.001
Subsequent qualifying pRBC-record event from >6 to 24 hours	532/9,049 (5.9%)	0 (0.0%)	262/6,578 (4.0%)	0 (0.0%)	<0.001

Values are median [interquartile range] or n/N (%). P values compare databases using two-sided Wilcoxon rank-sum tests for continuous variables and Pearson chi-square tests for binary variables on nonmissing observations. They are descriptive and were not used for predictor screening or model selection. pRBC, packed red blood cell; INR, international normalized ratio; SpO₂, peripheral oxygen saturation.

Table 2. Five-model performance in the MIMIC held-out test and eICU cross-database evaluation

Dataset	Model	n (events)	AUROC (95% CI)	AUPRC (95% CI)	Brier score (95% CI)	Calibration intercept (95% CI)	Calibration slope (95% CI)	Holm- adjusted P value vs reference
MIMIC held-out test	Logistic regression	2,270 (136)	0.838 (0.799 to 0.870)	0.273 (0.212 to 0.344)	0.0493 (0.0420 to 0.0566)	0.127 (-0.277 to 0.535)	1.019 (0.880 to 1.180)	0.307
MIMIC held-out test	Elastic net	2,270 (136)	0.834 (0.798 to 0.869)	0.278 (0.214 to 0.360)	0.0489 (0.0422 to 0.0556)	0.151 (-0.250 to 0.573)	1.028 (0.883 to 1.194)	0.209
MIMIC held-out test	Random forest	2,270 (136)	0.824 (0.786 to 0.859)	0.294 (0.216 to 0.378)	0.0485 (0.0413 to 0.0557)	-0.133 (-0.530 to 0.239)	0.902 (0.758 to 1.074)	0.095
MIMIC held-out test	Reference weighted XGBoost	2,270 (136)	0.844 (0.810 to 0.876)	0.275 (0.214 to 0.355)	0.0489 (0.0425 to 0.0559)	-0.051 (-0.384 to 0.351)	0.958 (0.824 to 1.112)	Reference
MIMIC held-out test	SVM-RBF	2,270 (136)	0.715 (0.665 to 0.769)	0.188 (0.136 to 0.250)	0.0541 (0.0457 to 0.0626)	1.333 (0.386 to 2.432)	1.463 (1.134 to 1.859)	<0.001
eICU cross- database evaluation	Logistic regression	6,578 (262)	0.843 (0.821 to 0.862)	0.216 (0.163 to 0.271)	0.0353 (0.0294 to 0.0420)	-0.414 (-0.812 to 0.118)	0.921 (0.816 to 1.077)	0.302
eICU cross- database evaluation	Elastic net	6,578 (262)	0.852 (0.825 to 0.871)	0.227 (0.173 to 0.286)	0.0350 (0.0286 to 0.0415)	-0.406 (-0.810 to 0.122)	0.925 (0.819 to 1.070)	0.819
eICU cross- database evaluation	Random forest	6,578 (262)	0.836 (0.806 to 0.862)	0.211 (0.153 to 0.267)	0.0343 (0.0288 to 0.0406)	-0.340 (-0.743 to 0.027)	0.949 (0.838 to 1.064)	0.134
eICU cross- database evaluation	Reference weighted XGBoost	6,578 (262)	0.851 (0.824 to 0.873)	0.235 (0.170 to 0.293)	0.0341 (0.0283 to 0.0399)	-0.139 (-0.507 to 0.224)	1.045 (0.950 to 1.166)	Reference
eICU cross- database evaluation	SVM-RBF	6,578 (262)	0.718 (0.677 to 0.757)	0.138 (0.099 to 0.188)	0.0380 (0.0304 to 0.0459)	-0.276 (-1.407 to 1.697)	1.065 (0.652 to 1.803)	<0.001

All metrics use two-parameter logistic recalibration learned from MIMIC development out-of-fold predictions. Marginal 95% confidence intervals use 1,000 bootstrap resamples (individual stays in MIMIC and hospital clusters in eICU). Adjusted P values compare candidate-model AUROC with reference weighted XGBoost using 10,000 paired resamples and Holm adjustment within each dataset. They do not test AUPRC, Brier score, or calibration measures. Weighted XGBoost is a post hoc reference model and is not intended to rank algorithms.

Table 3. Selected sensitivity analyses and deployment-boundary evaluations

Section	Analysis	Dataset	Model	n (events)	AUROC (95% CI)	Paired contrast or interpretation	Estimand
Model form	Post hoc model-form sensitivity	elCU cross- database evaluation	Unweighted XGBoost	6,578 (262)	0.849	-0.002 (-0.007 to 0.003)	Same primary population and outcome; alternative model form
Model form	Post hoc model-form sensitivity	elCU cross- database evaluation	3-df natural spline logistic	6,578 (262)	0.842	-0.009 (-0.020 to 0.000)	Same primary population and outcome; alternative model form
Model form	Post hoc model-form sensitivity	elCU cross- database evaluation	Hemoglobin- only logistic	6,578 (262)	0.791	-0.060 (-0.079 to -0.043)	Same primary population and outcome; alternative model form
Feature form	Extended feature-form sensitivity	elCU cross- database evaluation	Values-only XGBoost	6,578 (262)	0.849 (0.826 to 0.870)	-0.002 (-0.005 to 0.002)	Same primary population and outcome; alternative feature information form
Feature form	Extended feature-form sensitivity	elCU cross- database evaluation	Missingness- only XGBoost	6,578 (262)	0.667 (0.613 to 0.706)	-0.184 (-0.234 to -0.148)	Same primary population and outcome; alternative feature information form
Feature form	Extended feature-form sensitivity	elCU cross- database evaluation	First-value XGBoost	6,578 (262)	0.836 (0.811 to 0.862)	-0.015 (-0.028 to 0.000)	Same primary population and outcome; alternative feature information form
Feature form	Extended feature-form sensitivity	elCU cross- database evaluation	Last-value XGBoost	6,578 (262)	0.848 (0.826 to 0.867)	-0.003 (-0.013 to 0.007)	Same primary population and outcome; alternative feature information form

Section	Analysis	Dataset	Model	n (events)	AUROC (95% CI)	Paired contrast or interpretation	Estimand
Feature form	Extended feature-form sensitivity	eICU cross-database evaluation	Measurement-count-only XGBoost	6,578 (262)	0.693 (0.651 to 0.723)	-0.158 (-0.198 to -0.126)	Same primary population and outcome; alternative feature information form
Information state	Hemoglobin information-state sensitivity	eICU cross-database evaluation: Early hemoglobin measured	Reference weighted XGBoost: early hemoglobin measured	2,975 (191)	0.869 (0.848 to 0.894)	0.142 (0.092 to 0.201)	Performance conditional on whether early hemoglobin was measured
Information state	Hemoglobin information-state sensitivity	eICU cross-database evaluation: Early hemoglobin not measured	Reference weighted XGBoost: early hemoglobin not measured	3,603 (71)	0.727 (0.668 to 0.779)	0.142 (0.092 to 0.201)	Performance conditional on whether early hemoglobin was measured
Information state	Strict measured-only hemoglobin comparator	eICU cross-database evaluation: early hemoglobin measured	Strict measured-only Hb logistic	2,975 (191)	0.850 (0.830 to 0.873)	0.020 (0.008 to 0.029)	Measured-Hb subgroup only; conditional same-row comparator
Follow-up	Complete 24-hour follow-up sensitivity	MIMIC held-out test: complete 24-hour follow-up	Reference weighted XGBoost applied unchanged	1,735 (126)	0.833 (0.798 to 0.868)	Not directly comparable	Conditioned on complete ICU follow-up through 24 hours
Follow-up	Complete 24-hour follow-up sensitivity	eICU cross-database evaluation: complete 24-hour follow-up	Reference weighted XGBoost applied unchanged	4,744 (242)	0.842 (0.815 to 0.862)	Not directly comparable	Conditioned on complete ICU follow-up through 24 hours
Definition	Expanded-source sensitivity	MIMIC held-out test	Expanded-source sensitivity weighted XGBoost	2,182 (110)	0.845 (0.808 to 0.881)	Not directly comparable	MIMIC-only expanded early-pRBC source exclusion; primary outcome window retained
Definition	Documentation availability sensitivity	MIMIC held-out test: event-time outcome; storetime-aware landmark eligibility	Cross-outcome-definition documentation-recording weighted XGBoost	2,270 (136)	0.844	Not directly comparable	event-time outcome; MIMIC event-time component

Section	Analysis	Dataset	Model	n (events)	AUROC (95% CI)	Paired contrast or interpretation	Estimand
Definition	Documentation availability sensitivity	eICU cross-outcome-definition evaluation: event-and-recorded-by-horizon outcome; not the same estimand	Cross-outcome-definition documentation-recording weighted XGBoost	6,578 (244)	0.851	Not directly comparable	event-and-recorded-by-horizon outcome; Not same-estimand cross-database evaluation
Operational	Recorded-by-6h operational model	MIMIC held-out full-landmark test	Operational recorded-by-6h weighted XGBoost	2,419 (201)	0.873 (0.848 to 0.897)	Not directly comparable	Full landmark cohort; early-pRBC recorded-by-6h status is a predictor, not an exclusion

Table 3 (continued). Selected sensitivity analyses and deployment-boundary evaluations

Section	Analysis	Dataset	Model	n (events)	AUROC (95% CI)	Paired contrast or interpretation	Estimand
Operational	Recorded-by-6h operational model	eICU full-landmark cross-database evaluation	Operational recorded-by-6h weighted XGBoost	6,871 (380)	0.878 (0.860 to 0.896)	Not directly comparable	Full landmark cohort; early-pRBC recorded-by-6h status is a predictor, not an exclusion
Eligibility	Retrospective event-time eligibility	MIMIC held-out event-time-eligibility test	Retrospective event-time eligibility weighted XGBoost	2,256 (128)	0.845 (0.812 to 0.879)	Not directly comparable	Retrospective event-time early-pRBC exclusion; not the real-time recorded-by-landmark eligibility rule
Eligibility	Retrospective event-time eligibility	eICU event-time-eligibility cross-database evaluation	Retrospective event-time eligibility weighted XGBoost	6,546 (251)	0.851 (0.826 to 0.872)	Not directly comparable	Retrospective event-time early-pRBC exclusion; not the real-time recorded-by-landmark eligibility rule
Washout	Two-hour washout sensitivity	MIMIC-IV held-out test	Reference weighted XGBoost after 2-hour washout	2,233 (115)	0.833 (0.794 to 0.867)	-0.012 (-0.025 to 0.004)	Landmark cohort conditioned on remaining free of outcome/competing event for 2 hours after 6-hour landmark
Washout	Two-hour washout sensitivity	eICU-CRD cross-database evaluation	Reference weighted XGBoost after 2-hour washout	6,435 (219)	0.840 (0.814 to 0.861)	-0.011 (-0.019 to -0.002)	Landmark cohort conditioned on remaining free of outcome/competing event for 2 hours after 6-hour landmark

Rows deliberately include distinct sensitivity estimands and must not be interpreted as one head-to-head comparison. Model- and feature-form contrasts preserve the primary population and outcome; information-state, complete-follow-up, operational full-landmark, retrospective event-time, documentation, source-screen, and washout rows alter conditioning or estimand. Where only a formal point estimate was available, no marginal confidence interval was invented.

Figures

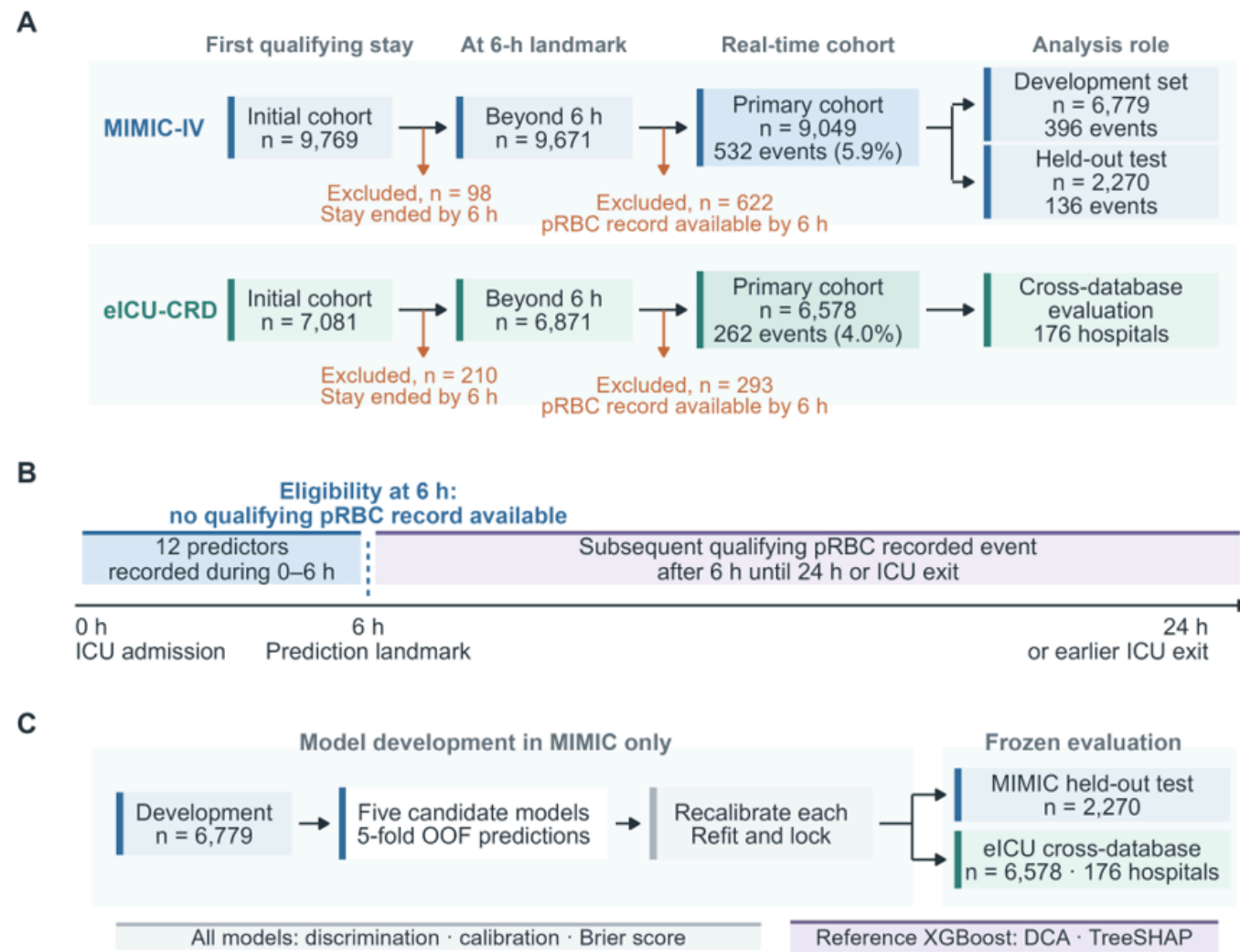


Figure 1

Study design, real-time landmark eligibility, and modelling workflow. (A) Derivation of the parallel MIMIC-IV and eICU-CRD trauma ICU cohorts. The primary cohort required no qualifying pRBC record available by the 6-hour landmark. MIMIC-IV retained its outcome-stratified 75:25 development/test split; eICU-CRD was used for cross-database evaluation. (B) Predictors were summarized from information recorded during the first 6 hours after ICU admission. The outcome was a subsequent qualifying pRBC event after the landmark through the earlier of 24 hours or ICU exit. (C) Five candidate algorithms were developed with five-fold out-of-fold procedures. A separate two-parameter logistic recalibration map was learned for each model from MIMIC development out-of-fold predictions. Each model was then refitted, locked, and evaluated without further fitting in the MIMIC held-out test and eICU cohorts. Decision-curve analysis and TreeSHAP were restricted to the post hoc reference weighted XGBoost model.

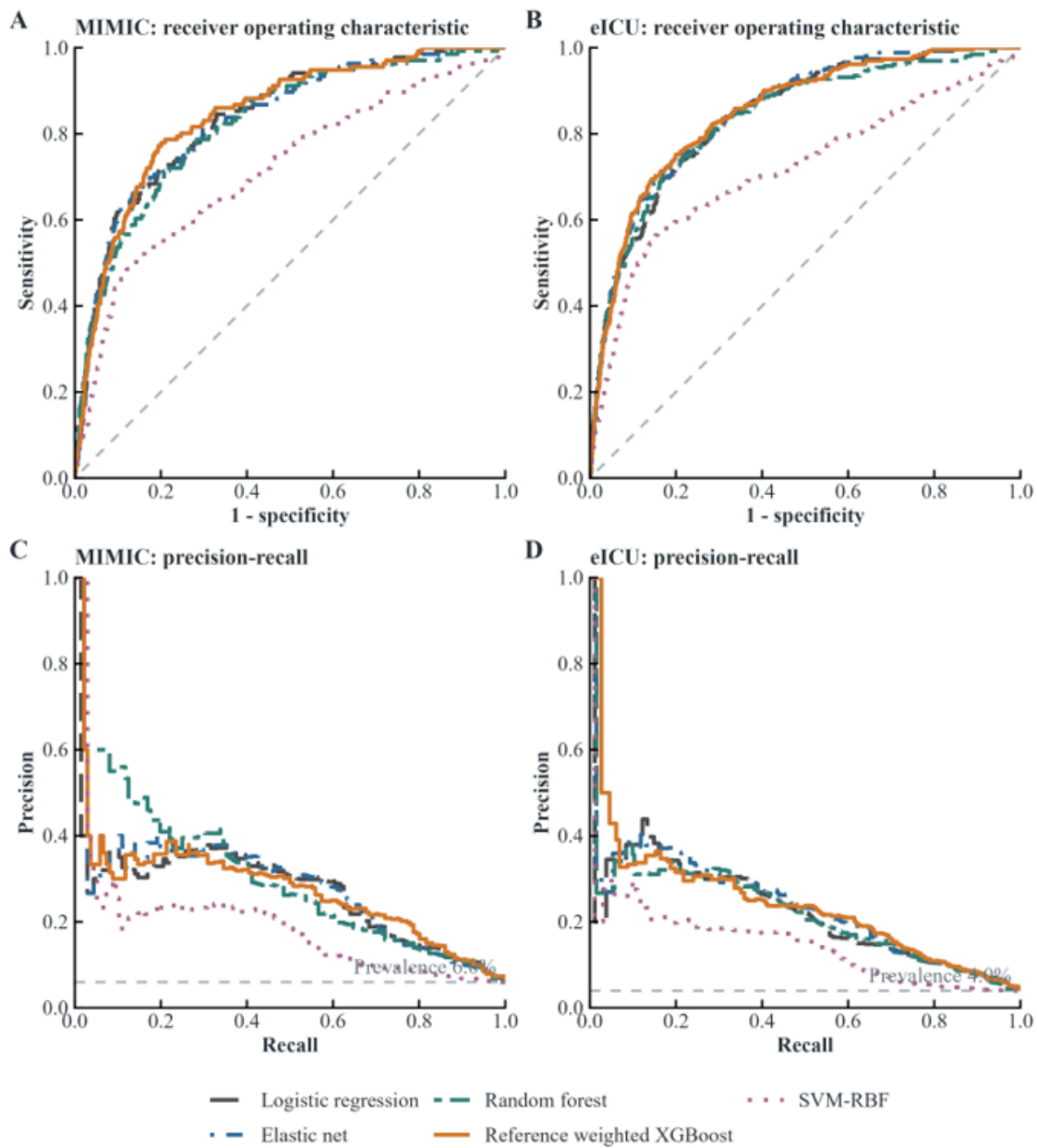


Figure 2

Discrimination of five candidate models in MIMIC-IV and eICU-CRD. Receiver operating characteristic curves in the MIMIC held-out test (A) and eICU cross-database evaluation (B), and precision-recall curves in the MIMIC held-out test (C) and eICU cross-database evaluation (D). Curves use calibrated probabilities; monotonic recalibration does not alter rank-based discrimination. Dashed precision-recall reference lines indicate outcome prevalence.

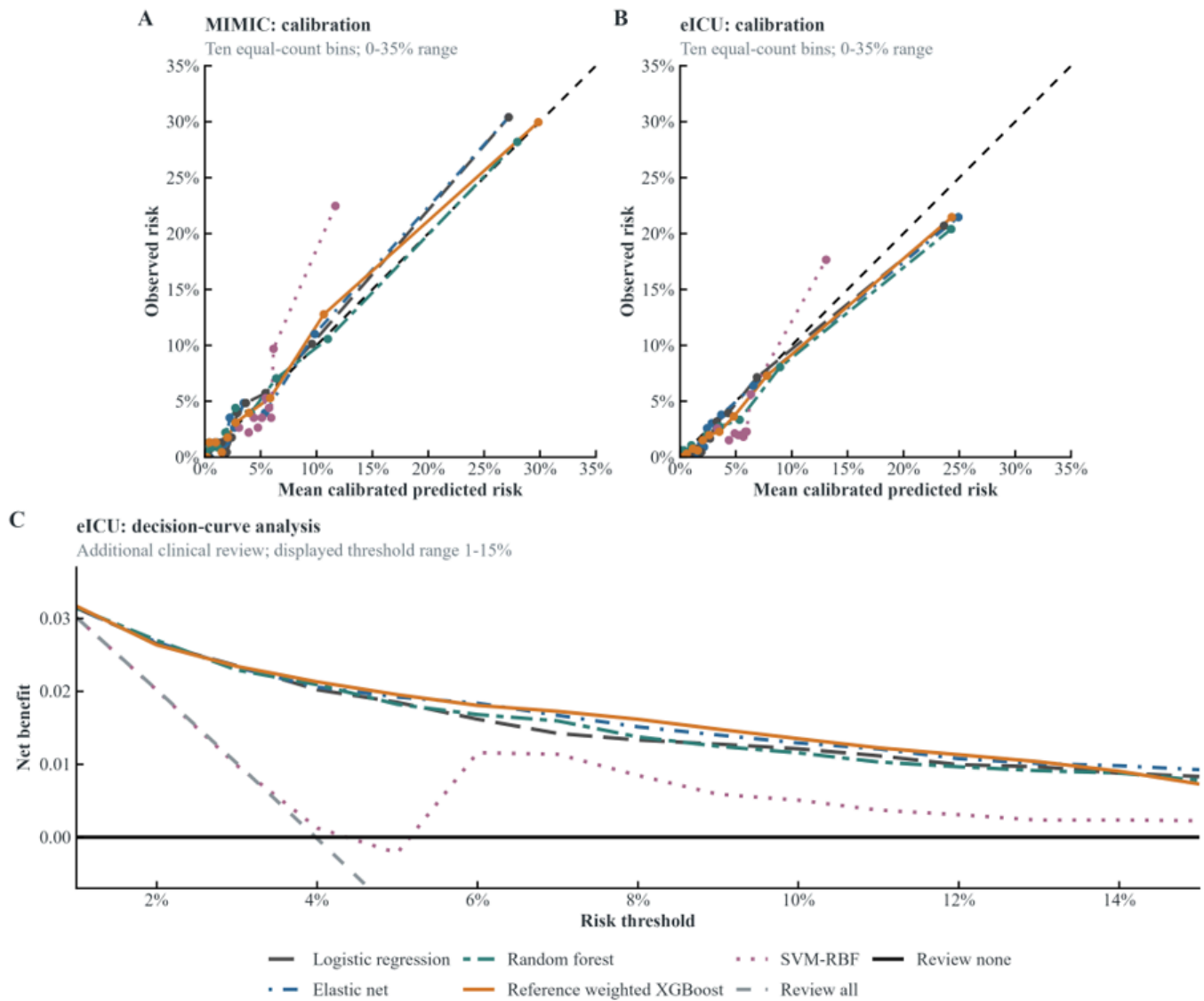


Figure 3

Calibration and clinical utility. Descriptive equal-count calibration plots for the MIMIC held-out test (A) and eICU cross-database evaluation (B). The dashed diagonal denotes ideal calibration; axes display the clinically relevant low-risk range. (C) Exploratory decision curves in eICU; the analysis covered 1%-20% risk thresholds and the panel displays 1%-15% for legibility. Net benefit refers to a hypothetical additional clinical review or reassessment strategy and is not an automatic transfusion trigger.

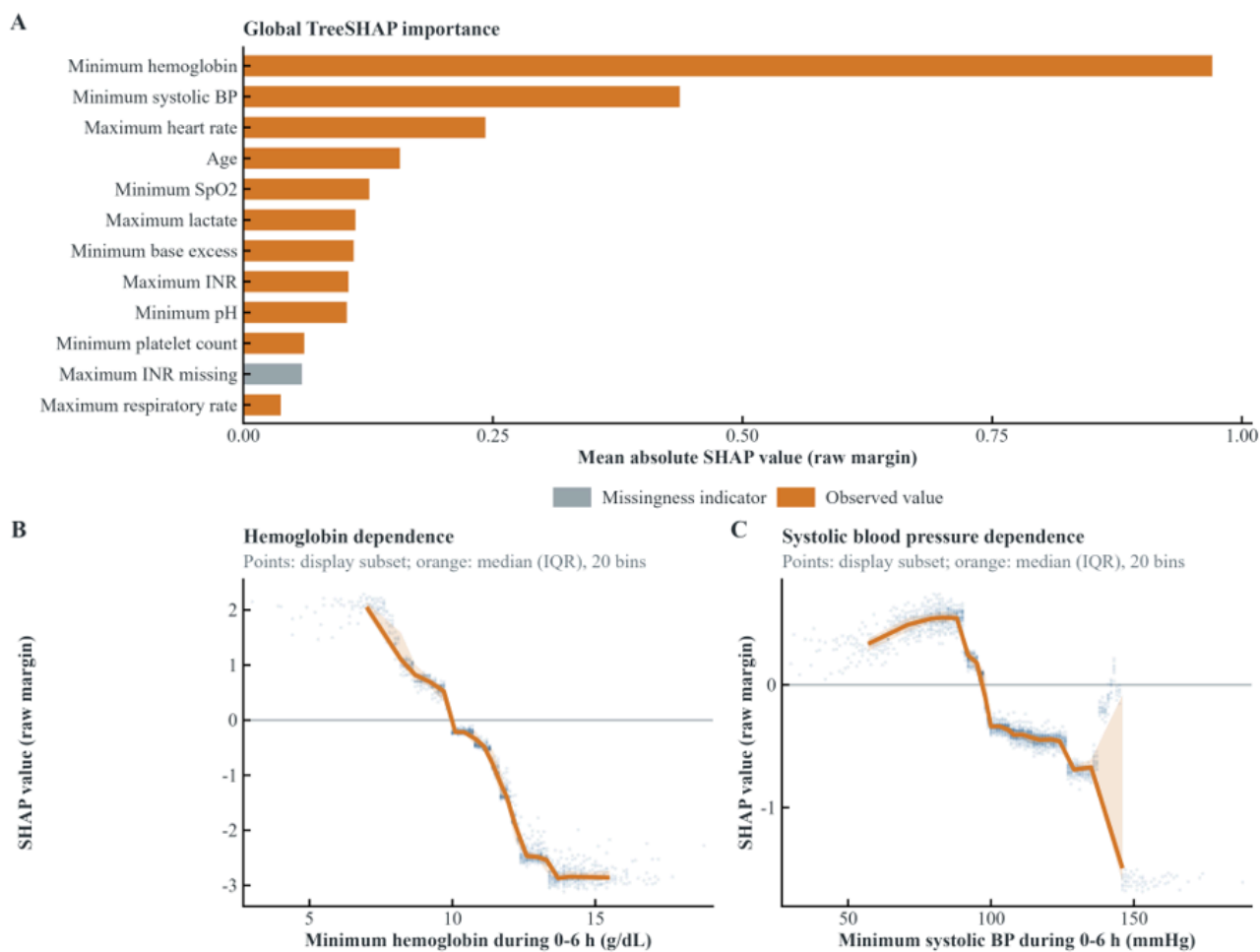


Figure 4

TreeSHAP interpretation of the post hoc reference weighted XGBoost model in eICU. (A) Global importance quantified as mean absolute SHAP value on the raw model margin; the 12 highest-ranked of 22 model columns are displayed. Dependence of raw-margin SHAP values on observed minimum hemoglobin (B) and minimum systolic blood pressure (C). Light points are a deterministic display subset; orange lines and ribbons are median and interquartile range within 20 equal-count bins from all observed values. SHAP describes model behavior and has no causal interpretation.

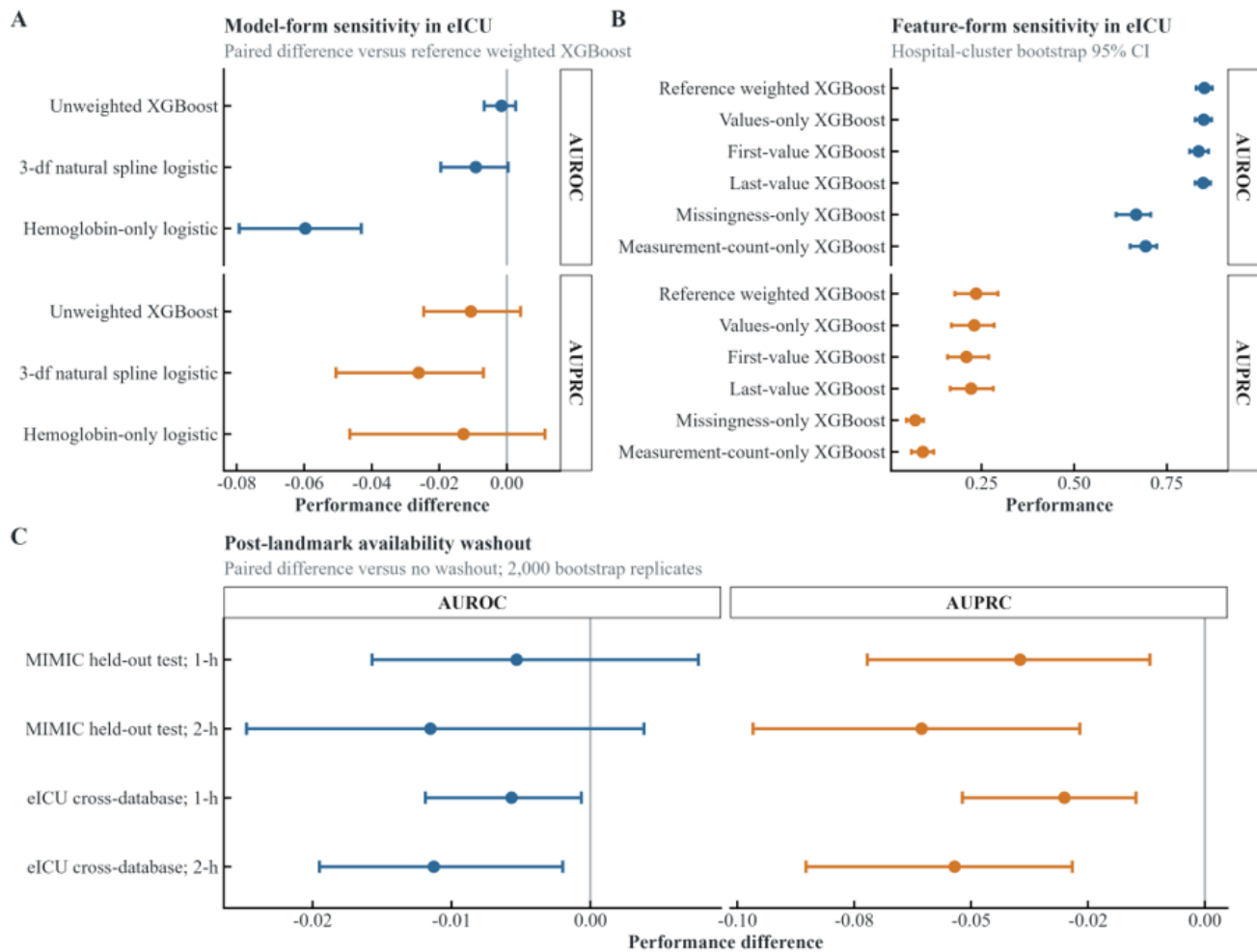


Figure 5

Model-form, feature-form, and post-landmark availability robustness. (A) Paired eICU AUROC and AUPRC differences for post hoc model-form alternatives versus reference weighted XGBoost. (B) Absolute eICU performance for feature-form sensitivities with hospital-cluster bootstrap 95% confidence intervals. (C) Paired AUROC and AUPRC differences after excluding events within 1 or 2 hours after the landmark; intervals used 2,000 individual-stay bootstrap samples for MIMIC and hospital-cluster samples for eICU.

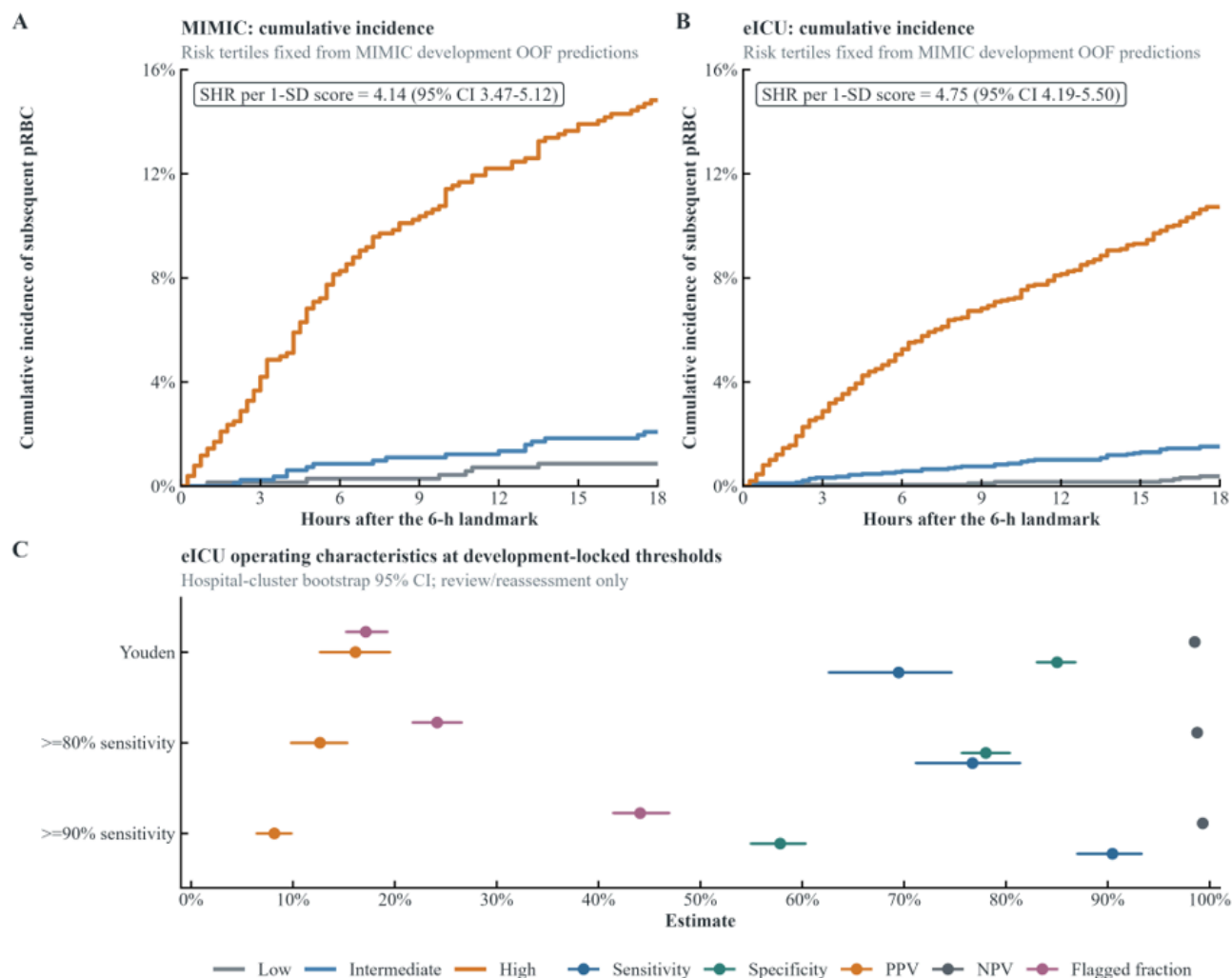


Figure 6

Risk stratification and operating characteristics. Aalen-Johansen cumulative incidence of subsequent pRBC events across development-fixed risk tertiles in MIMIC (A) and eICU (B). Competing events were ICU discharge or death before pRBC. The subdistribution hazard ratio is the association per one development-standard-deviation increase in the fixed score and is not full competing-risk prediction validation. (C) eICU operating characteristics at three development-locked thresholds. These are review/reassessment operating points, not transfusion triggers.

Supplementary Files

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

- [Supplementarymaterials.docx](#)