

Temporal evaluation of machine-learning models for early in-hospital mortality prediction after acute trauma using routinely collected emergency department data: a multicentre retrospective cohort study

Ji Sun Lee

Daejeon University

Joung Hwan Back

back2000@dju.kr

Daejeon University

Hyun Jin Kim

Far East University

Research Article

Keywords: clinical decision support, emergency department, machine learning, mortality, prediction model, trauma, temporal transportability

Posted Date: October 7th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11079806/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

Clinical prediction models are increasingly considered for decision support in emergency care, but high discrimination alone does not establish temporal reliability. We evaluated whether models based only on data available at initial emergency department assessment retained calibrated and threshold-specific performance over time and across clinically important subgroups.

Methods

We conducted a retrospective multicentre cohort study of 626,233 adults with acute trauma recorded in the Korean Emergency Department-based Injury In-depth Surveillance database from 2019 to 2023. Logistic regression, random forest, Extreme Gradient Boosting (XGBoost), and Light Gradient Boosting Machine (LightGBM) models using 15 initial-assessment variables were developed in 2019–2022 and compared in a temporally separated 2023 cohort. Evaluation included AUROC, AUPRC, Brier score, probability calibration, derivation-defined operating thresholds, decision-curve analysis, and age- and injury-mechanism subgroup performance. Because 2023 performance informed designation of XGBoost for extended characterisation, the 2023 cohort was not treated as independent external validation of a prespecified final model.

Results

In 2023, 1,653 of 119,234 patients (1.39%) died in hospital. AUROCs were 0.977 for logistic regression, 0.980 for random forest, 0.978 for XGBoost, and 0.978 for LightGBM; corresponding AUPRCs were 0.630, 0.708, 0.714, and 0.716. For XGBoost, derivation-fitted isotonic calibration improved the calibration slope from 0.919 to 1.021 and reduced the Brier score from 0.0367 to 0.0065. A sensitivity-constrained operating threshold that achieved 90.01% sensitivity during derivation achieved 79.49% when applied unchanged in 2023. AUROC was 0.993 among adults aged 19–64 years and 0.930 among adults aged 65 years or older.

Conclusions

Routinely collected initial emergency department data carried substantial prognostic information, but strong aggregate discrimination coexisted with temporal instability of fixed operating points and lower discrimination in older adults. These findings support distinguishing ranking performance, calibrated risk estimation, and threshold-specific classification when prediction models are evaluated for clinical decision-support use, with attention to temporal transportability and subgroup reliability. Independent validation is required before clinical implementation.

Background

Trauma remains a major cause of mortality and disability worldwide and continues to create substantial demand for emergency and hospital care [1, 2]. During the first phase of trauma assessment, clinicians must identify patients at increased risk before many laboratory, imaging, and downstream treatment data are available. Routinely collected demographic, injury, prehospital, physiological, and neurological data therefore offer a potential information substrate for prediction-supported clinical decision support. Such systems are intended to augment, rather than replace, clinician decision making, and their value depends not only on predictive accuracy but also on fit with the intended use, workflow, and prediction time point [3].

Conventional physiological trauma scores provide simple approaches to early risk stratification. The Revised Trauma Score (RTS) combines Glasgow Coma Scale (GCS), systolic blood pressure, and respiratory rate [4], whereas the reverse Shock Index multiplied by Glasgow Coma Scale (rSIG) combines haemodynamic and neurological information [5, 6]. Shock-index approaches have also been evaluated across trauma populations [7]. These scores are readily calculated but have fixed mathematical structures and use a limited subset of the information available during early assessment. Machine-learning models can accommodate a broader set of routinely collected variables and nonlinear relationships, but greater algorithmic flexibility does not by itself establish clinical reliability [8–11].

For prediction-supported decision making, discrimination is only one dimension of model performance. A high area under the receiver operating characteristic curve (AUROC) indicates that a model ranks patients effectively across possible thresholds, but it does not establish whether predicted probabilities correspond to observed risks [12, 13]. Calibration and prediction error matter when outputs are communicated as patient-level risks, while decision-curve analysis can examine potential net benefit under specified threshold probabilities [14]. Fixed operating thresholds raise a separate question: whether the sensitivity and specificity attached to a decision rule remain stable when the same rule is transported to a later population.

Temporal reliability and subgroup performance are therefore central to evaluating translation from a prediction model to a clinical decision-support component. Changes in case mix, referral pathways, data completeness, clinical practice, or documentation can alter calibration and threshold-specific operating characteristics even when aggregate discrimination remains high. Aggregate metrics may also obscure clinically important heterogeneity. Older adults are particularly relevant in trauma because mortality risk is higher and early physiological responses may differ from those of younger adults. A model intended for broad emergency-care use should therefore be examined for temporal transportability and subgroup reliability before implementation [15].

The aim of this study was to develop and temporally evaluate models for early prediction of in-hospital mortality among adults with acute trauma using information available during initial emergency department assessment.

Methods

Study design, reporting standard, and data source

This retrospective multicentre cohort study used the Emergency Department-based Injury In-depth Surveillance (EDIIS) database administered by the Korea Disease Control and Prevention Agency. EDIIS prospectively collects standardised demographic, injury-related, prehospital, physiological, emergency-department, and hospital-outcome information from participating emergency departments across South Korea. All injury-related emergency-department records from 1 January 2019 through 31 December 2023 were assessed for eligibility. The study was reported with reference to the TRIPOD + AI reporting guideline for clinical prediction models using regression or machine-learning methods [12]. The EDIIS data used in this study were accessed through the authorised data-use process of the Korea Disease Control and Prevention Agency under research management number KDCA-12-02-EI-2026-000126.

Calendar time was used to separate model development from later evaluation. Data from 2019–2022 constituted the derivation cohort and 2023 data constituted a temporally separated model-comparison cohort. All preprocessing, imputation, model fitting, and probability-calibration procedures requiring parameter estimation were fitted using derivation data and then applied unchanged to 2023. The primary model-level analysis was the comparison of candidate algorithms rather than a claim that one algorithm was universally superior. Because performance in the 2023 cohort contributed to reference-model designation, performance estimates for the selected model were interpreted as temporally held-out comparative estimates rather than independent external validation of a prespecified final model; selection-related optimism may therefore remain.

Participants

Adults aged 19 years or older who presented to an emergency department with an acute traumatic injury were eligible. Acute trauma was defined a priori as presentation within 24 hours of injury. Patients were excluded if they were younger than 19 years; had non-traumatic injuries such as poisoning, thermal injury, or drowning; were dead on arrival; had no observable definitive outcome because of direct transfer from the emergency department or unknown disposition; or presented more than 24 hours after injury. Patients transferred to another institution after hospital admission were retained when a definitive hospital outcome was available. Sensitivity analyses examined exclusion of these patients and alternative acute-trauma windows of 12 hours and 6 hours.

All eligible records available during the predefined study period were included. No formal sample-size calculation was performed because the analysis used the complete eligible surveillance cohort.

Intended use, prediction time point, outcome, and candidate predictors

The intended use context was early prognostic decision support after completion of the initial emergency department assessment: to estimate the risk of in-hospital mortality using only information available by that point. This study evaluated the predictive information and reliability of such a potential decision-support component; it did not deploy a clinical decision-support system, display model outputs to clinicians, or assign care according to predicted risk. Predictors were therefore restricted to information available at or before the prediction time point. Physiological and neurological measures were the first documented values at emergency-department presentation, and laboratory results, imaging findings, procedures, treatments, and other downstream information were excluded to reduce temporal information leakage.

The primary outcome was in-hospital mortality, defined as death during the index emergency-department stay or subsequent hospitalisation associated with the index presentation. Patients with an observable definitive hospital outcome who did not die were classified as survivors.

Fifteen candidate predictors were specified based on availability during initial assessment and representation in EDIIS: age, sex, injury mechanism, injury intentionality, time from injury to emergency-department arrival, systolic blood pressure, diastolic blood pressure, heart rate, respiratory rate, body temperature, GCS, Alert/Voice/Pain/Unresponsive (AVPU) level of consciousness, mode of transportation to the emergency department, nighttime presentation, and alcohol-related status. In the analytical pipeline, transport mode was represented as ambulance transportation (yes/no). Total GCS was calculated from the initial eye, verbal, and motor components. Operational definitions and analytical representations are provided in Additional file 1: Table S1.

Data quality, missing data, and preprocessing

Source-data codes representing unavailable or unknown information, physiologically implausible values, and internally inconsistent measurements were treated as missing according to prespecified data-quality rules. Predictor-specific missingness and validation rules are provided in Additional file 1: Table S2.

Continuous and binary predictors were imputed using median values estimated from the 2019–2022 derivation cohort and were standardised using derivation-cohort parameters. Multicategory predictors underwent mode imputation followed by one-hot encoding, with unseen categories in 2023 handled without refitting the encoder. All fitted preprocessing parameters were estimated within the derivation pipeline and applied unchanged to 2023. Robustness to the primary missing-data strategy was examined using an alternative iterative multivariable imputation procedure fitted within derivation data and applied without using 2023 outcome information; this was treated as an alternative single imputation strategy rather than formal multiple imputation with pooled inference.

Model development and algorithm comparison

Four candidate algorithms were fitted using the complete 2019–2022 derivation cohort: logistic regression, random forest, Extreme Gradient Boosting (XGBoost), and Light Gradient Boosting Machine (LightGBM) [16–18]. Class imbalance was addressed using algorithm-specific inverse-frequency weighting. For XGBoost, `scale_pos_weight` was set to the derivation-cohort ratio of survivors to deaths ($500,334/6,665 = 75.0689$). Logistic regression, random forest, and LightGBM used their implementations of `class_weight='balanced'`. These approaches targeted similar imbalance correction but were not assumed to be mathematically identical across algorithms. Synthetic minority oversampling was not used. A separate five-variable parsimonious logistic model was fitted without class weighting so that its predicted probabilities remained on the observed-risk scale.

Candidate algorithms were compared in the 2023 cohort using discrimination, prediction error, and calibration. Before model-specific calibration slopes were calculated, eligibility for extended characterisation was defined as an unrecalibrated calibration-slope point estimate between 0.90 and 1.10; among eligible algorithms, the model with the highest AUROC was designated as the reference model for extended temporal, threshold, decision-curve, subgroup, and explainability analyses. Preliminary discrimination results had been examined before the calibration-slope criterion was finalised. Because the 2023 cohort contributed to reference-model designation, subsequent performance estimates for the selected model were interpreted as temporally held-out comparative estimates rather than independent external-validation estimates. The reference-model designation criterion and candidate-algorithm comparison are summarised in Additional file 1: Table S3. Supplementary analyses of derivation sample size and alternative reference-model designation strategies restricted to the 2019–2022 data are reported in Additional file 1: Table S4.

Computational environment and model specification

Analyses were performed in Python 3.14.6 using NumPy 2.4.6, pandas 3.0.3, SciPy 1.18.0, scikit-learn 1.9.0, XGBoost 3.3.0, LightGBM 4.6.0, statsmodels 0.14.6, SHAP 0.52.0, and Matplotlib 3.11.0. Primary XGBoost settings were `n_estimators = 400`, `max_depth = 6`, `learning_rate = 0.05`, `subsample = 0.8`, `colsample_bytree = 0.8`, `scale_pos_weight = 75.0689`, `eval_metric='logloss'`, and `random_state = 42`. The random forest used `n_estimators = 200`, `min_samples_leaf = 50`, `class_weight='balanced'`, and `random_state = 42`; LightGBM used `n_estimators = 400`, `num_leaves = 63`, `learning_rate = 0.05`, `subsample = 0.8`, `colsample_bytree = 0.8`, `class_weight='balanced'`, and `random_state = 42`; logistic regression used `max_iter = 2000` and `class_weight='balanced'`. Parameters not explicitly specified used package-version defaults. Full computational and analytical specifications are provided in Additional file 1: Supplementary Methods S1.

Performance measures and conventional trauma-score comparators

Discrimination was assessed using AUROC and area under the precision–recall curve (AUPRC). Ninety-five percent confidence intervals for principal discrimination estimates were obtained from 1,000 percentile bootstrap resamples of the 2023 cohort using random seed 42. AUPRC was included because

in-hospital mortality was uncommon and precision–recall performance is informative when evaluating detection of the minority outcome class.

RTS and rSIG were evaluated as conventional physiological comparators. RTS was recalculated using the established scoring algorithm [4]. rSIG was calculated as $GCS \times (\text{systolic blood pressure} / \text{heart rate})$ [5]. Comparisons with each score were performed in the corresponding paired complete-score subset. An alternative mean arterial pressure-based formulation of rSIG was examined in sensitivity analysis.

Probability calibration

Probability calibration was performed without using 2023 outcomes for estimation or model updating. Stratified five-fold cross-validation within the derivation cohort (shuffle=True, random_state = 42) generated out-of-fold XGBoost predictions that were used to fit an isotonic regression mapping with out-of-bounds predictions clipped to the fitted range. This mapping was then applied unchanged to 2023 predictions. Calibration was evaluated using calibration slope, recalibration intercept, calibration-in-the-large, and Brier score [13, 19]. Ninety-five percent confidence intervals were obtained from 1,000 percentile bootstrap resamples of the fixed prediction–outcome pairs in 2023. Discrimination before and after calibration was also examined.

Operating thresholds and temporal stability

Two operating thresholds were determined exclusively from the class-weighted, uncalibrated XGBoost output in the derivation cohort and applied unchanged to 2023. The first maximised the Youden index. The second was a sensitivity-constrained operating threshold, defined as the highest model-output cut-point retaining at least 90% sensitivity for in-hospital mortality during derivation. No threshold was re-optimised using 2023 outcomes. Sensitivity, specificity, positive predictive value, negative predictive value, and F1 score were calculated at each threshold. Because these thresholds were derived from class-weighted, uncalibrated outputs, they were interpreted as model-specific operating cut-points rather than absolute mortality probabilities. Their purpose was to test transportability of a fixed operating point over calendar time, not to define a clinical action threshold. Calibration and decision-curve analyses used the separate isotonic-calibrated probability scale; therefore, the uncalibrated operating cut-points are not interchangeable with calibrated probability thresholds and would require redefinition for a calibrated clinical implementation.

Decision-curve, subgroup, and explainability analyses

Potential clinical utility was examined using decision-curve analysis based on isotonic-calibrated mortality probabilities in the full 2023 cohort. Net benefit was compared with treat-all and treat-none strategies across the displayed threshold-probability range of 0 to 0.20 [14]. Decision-curve analysis was interpreted as decision-analytic evidence under specified threshold preferences and not as evidence that implementation improves clinician decisions, workflow, or patient outcomes.

Predictive performance was evaluated among adults aged 19–64 years and 65 years or older and across major injury-mechanism groups. The same fitted XGBoost model was applied without subgroup-specific refitting or recalibration. Mortality prevalence and AUROC were estimated within each subgroup, with 95% confidence intervals from 1,000 bootstrap resamples. Findings were treated as predictive-performance heterogeneity rather than causal differences.

Model behaviour was examined using SHapley Additive exPlanations (SHAP) [20]. Global feature contributions were summarised using mean absolute SHAP values, and dependence patterns were examined for selected predictors. Age-stratified SHAP analyses used the same model without subgroup-specific refitting and were interpreted as descriptions of model behaviour rather than causal explanations.

Parsimonious model and sensitivity analyses

A separate unpenalised logistic regression model incorporating age, systolic blood pressure, heart rate, respiratory rate, and GCS was fitted in the derivation cohort without class weighting and evaluated in 2023. AUROC, AUPRC, calibration slope, and mean predicted probability were assessed. A nomogram was used to visualise this reduced model (Additional file 1: Figure S1); it was not treated as a validated bedside score.

Sensitivity and robustness analyses examined alternative acute-trauma windows (12 and 6 hours), exclusion of patients transferred after hospital admission, an alternative mean arterial pressure-based rSIG formulation, iterative multivariable imputation, random-search hyperparameter optimisation, and predictor-block ablation (Additional file 1: Table S5). The XGBoost random search was conducted only as a robustness analysis using 15 iterations of three-fold cross-validation on a random subsample of 150,000 derivation patients, with AUROC as the optimisation objective, random seed 42, and no early stopping; the tuned configuration was not adopted for the primary model. A supplementary multivariable logistic model containing all 15 candidate predictors characterised adjusted associations and collinearity but was not used for predictor selection or primary model development (Additional file 1: Table S6).

Ethics approval and AI-assisted language editing

The study was approved by the Institutional Review Board of Daejeon University (approval no. 1040647-202607-HR-015-01). The requirement for individual informed consent was waived because the study used de-identified secondary data. Analyses were conducted through the authorised EDIIS data-access process and under the applicable conditions of data use. During manuscript preparation, ChatGPT (OpenAI) was used solely for English-language editing and structural refinement to improve grammar, readability, linguistic clarity, and alignment with journal formatting. It was not used for study design, data collection, statistical analysis, interpretation of results, or generation of scientific conclusions. All scientific content was reviewed and approved by the authors, who take responsibility for the manuscript.

Results

Study population and temporal cohort characteristics

Among 1,071,424 injury-related emergency-department records identified from 2019 through 2023, 56 complete duplicate records were removed. After sequential application of eligibility criteria, 626,233 adults with acute trauma were included (Fig. 1): 506,999 in the 2019–2022 derivation cohort and 119,234 in the 2023 temporal model-comparison cohort. Overall, 8,318 patients (1.33%) died during the index emergency-department stay or subsequent hospitalisation. Mortality was 1.31% in the derivation cohort and 1.39% in 2023.

Patients who died were older than survivors (median age 65 versus 50 years) and were more often male (69.7% versus 57.4%). Ambulance transportation was recorded in 93.7% of deaths and 37.7% of survivors. Median GCS was 3 among patients who died and 15 among survivors; 55.1% of patients who died were unresponsive on AVPU compared with 0.2% of survivors (Table 1).

Table 1
Baseline characteristics of the acute trauma cohort according to in-hospital mortality

Characteristic	Survivors (n = 617,915)	In-hospital deaths (n = 8,318)
Age, years, median (IQR)	50 (33–64)	65 (52–78)
Sex, n (%)		
Male	354,406 (57.4)	5,801 (69.7)
Female	263,509 (42.6)	2,517 (30.3)
Systolic blood pressure, mmHg, median (IQR)	139 (123–155)	134 (104–161)
Diastolic blood pressure, mmHg, median (IQR)	82 (73–92)	80 (61–97)
Heart rate, beats/min, median (IQR)	83 (74–94)	90 (75–109)
Respiratory rate, breaths/min, median (IQR)	20 (18–20)	20 (18–22)
Body temperature, °C, median (IQR)	37 (36–37)	36 (36–37)
Glasgow Coma Scale, median (IQR)	15 (15–15)	3 (3–12)
Level of consciousness (AVPU), n (%)		
Alert	599,617 (97.0)	1,681 (20.2)
Responsive to voice	11,822 (1.9)	576 (6.9)
Responsive to pain	5,201 (0.8)	1,475 (17.7)
Unresponsive	1,275 (0.2)	4,586 (55.1)
Time from injury to ED arrival, hours, median (IQR)	1 (1–3)	1 (0–2)
Ambulance transportation, yes, n (%)	233,230 (37.7)	7,793 (93.7)
Injury mechanism, n (%)		
Fall or slip	214,585 (34.7)	4,152 (49.9)
Transportation-related injury	115,163 (18.6)	3,380 (40.6)
Blunt injury	111,748 (18.1)	353 (4.2)
Penetrating injury	91,611 (14.8)	195 (2.3)
Overexertion	24,409 (4.0)	24 (0.3)
Machinery	8,681 (1.4)	45 (0.5)
Note. Continuous variables are presented as median (IQR), and categorical variables as n (%). Percentages for subcategories are calculated within each outcome group. AVPU, Alert, Voice, Pain, Unresponsive; ED, emergency department; IQR, interquartile range.		

Characteristic	Survivors (n = 617,915)	In-hospital deaths (n = 8,318)
Natural disaster	31 (0.0)	1 (0.0)
Other	51,687 (8.4)	168 (2.0)
Injury intentionality, n (%)		
Unintentional	571,082 (92.4)	6,871 (82.6)
Self-harm or suicide	11,604 (1.9)	891 (10.7)
Assault or violence	34,315 (5.6)	111 (1.3)
Unknown (C98)	63 (0.0)	3 (0.0)
Unknown (C99)	851 (0.1)	442 (5.3)
Nighttime ED presentation, n (%)	173,348 (28.1)	1,816 (21.8)
Alcohol-related status, yes, n (%)	83,462 (13.5)	519 (6.2)
Note. Continuous variables are presented as median (IQR), and categorical variables as n (%). Percentages for subcategories are calculated within each outcome group. AVPU, Alert, Voice, Pain, Unresponsive; ED, emergency department; IQR, interquartile range.		

Missingness ranged from 0.60% for body temperature to 1.92% for GCS; alcohol-related status had 6.22% missing or unavailable information (Additional file 1: Table S2). Absolute standardised mean differences were 0.10 or less for all measured characteristics, with the largest values observed for age (– 0.100) and fall/slip mechanism (– 0.095) (Additional file 1: Table S7).

Algorithm comparison and predictive performance

AUROC in the 2023 cohort ranged from 0.977 to 0.980 across the four algorithms (Table 2; Fig. 2). AUROC was 0.977 (95% CI 0.974–0.980) for logistic regression, 0.980 (0.977–0.982) for random forest, 0.978 (0.976–0.981) for XGBoost, and 0.978 (0.975–0.981) for LightGBM. Corresponding AUPRC values were 0.630, 0.708, 0.714, and 0.716. RTS and rSIG had AUROC of 0.814 and 0.805, respectively. In paired complete-score subsets, the XGBoost AUROC exceeded RTS by 0.149 (95% CI 0.135–0.163) and rSIG by 0.158 (0.140–0.176).

Table 2

Predictive performance of prediction models and conventional trauma scores for in-hospital mortality in the 2023 temporal model-comparison cohort

Model	AUROC (95% CI)	AUPRC (95% CI)
rSIG	0.805 (0.786–0.825)	0.229 (0.198–0.264)
RTS	0.814 (0.799–0.831)	0.278 (0.251–0.313)
Logistic regression	0.977 (0.974–0.980)	0.630 (0.603–0.657)
Random forest	0.980 (0.977–0.982)	0.708 (0.687–0.729)
XGBoost	0.978 (0.976–0.981)	0.714 (0.693–0.735)
LightGBM	0.978 (0.975–0.981)	0.716 (0.695–0.737)

Note. Values are AUROC or AUPRC with 95% confidence intervals obtained from 1,000 percentile bootstrap resamples of the 2023 temporal model-comparison cohort. Conventional trauma scores were evaluated among patients with all components required for score calculation (rSIG, n = 1,411 missing; RTS, n = 1,456 missing). Comparisons between XGBoost and each conventional score were performed in the corresponding paired complete-score subset. AUPRC, area under the precision–recall curve; AUROC, area under the receiver operating characteristic curve; CI, confidence interval; RTS, Revised Trauma Score; rSIG, reverse Shock Index multiplied by Glasgow Coma Scale. LightGBM, Light Gradient Boosting Machine; XGBoost, Extreme Gradient Boosting.

Unrecalibrated calibration-slope point estimates were 0.785 for logistic regression, 1.117 for random forest, 0.919 for XGBoost, and 0.875 for LightGBM (Additional file 1: Table S3). XGBoost was the only algorithm within the prespecified 0.90–1.10 band and was therefore used as the reference model for extended characterisation. When derivation size increased from 2019 alone to 2019–2022, the XGBoost calibration slope in 2023 increased from 0.786 to 0.919 while AUROC varied by less than 0.006. When reference-model designation was restricted to 2019–2022 using either an internal temporal split or five-fold cross-validation, none of the candidate algorithms met the specified calibration-slope criterion (Additional file 1: Table S4).

Probability calibration

Observed mortality in 2023 was 1.39%, whereas the mean uncalibrated XGBoost output was 9.09%. Before recalibration, the calibration slope was 0.919 (95% CI 0.892–0.951) and the Brier score was 0.0367 (0.0361–0.0374). After application of the derivation-fitted isotonic mapping, mean predicted mortality was 1.47%; the calibration slope was 1.021 (0.990–1.057) and the Brier score was 0.0065 (0.0061–0.0069) (Fig. 3A).

Temporal stability of fixed operating thresholds

The Youden-index operating cut-point achieved 97.18% sensitivity during derivation and 86.57% in 2023. The sensitivity-constrained operating cut-point achieved 90.01% sensitivity during derivation and 79.49% in 2023, an absolute decrease of 10.52 percentage points (Fig. 4; Additional file 1: Table S8).

At the sensitivity-constrained operating cut-point in 2023, 1,314 of 1,653 patients who died were identified and 339 were not identified. Specificity was 98.26%, positive predictive value 39.05%, negative predictive value 99.71%, and F1 score 0.524. At the Youden-index cut-point, specificity changed from 96.22% in derivation to 95.49% in 2023, while sensitivity decreased from 97.18% to 86.57%. At the sensitivity-constrained cut-point, specificity changed from 98.51% to 98.26%, while sensitivity decreased from 90.01% to 79.49%.

Subgroup performance, decision-curve analysis, and explainability

Performance differed by age (Table 3). Among 85,792 adults aged 19–64 years, 745 died (0.87%) and AUROC was 0.993 (95% CI 0.990–0.995). Among 33,442 adults aged 65 years or older, 908 died (2.72%) and AUROC was 0.930 (0.920–0.938). Injury-mechanism AUROCs ranged from 0.953 for falls or slips to 0.999 for penetrating injury; estimates in penetrating and other-mechanism groups were based on 33 and 29 deaths (Additional file 1: Table S9).

Table 3
Age-stratified discrimination of the XGBoost model for in-hospital mortality in the 2023 temporal model-comparison cohort

Age group	N	Mortality prevalence, n (%)	AUROC (95% CI)
19–64 years	85,792	745 (0.87)	0.993 (0.990–0.995)
≥ 65 years	33,442	908 (2.72)	0.930 (0.920–0.938)

Note. The XGBoost model fitted using the complete 2019–2022 derivation cohort was applied unchanged to each age group without subgroup-specific refitting or recalibration. Mortality prevalence and AUROC were estimated separately within each subgroup. Ninety-five per cent confidence intervals for AUROC were obtained from 1,000 percentile bootstrap resamples within the corresponding subgroup. CI, confidence interval; AUROC, area under the receiver operating characteristic curve. XGBoost, Extreme Gradient Boosting.

Decision-curve analysis using isotonic-calibrated probabilities showed positive net benefit for XGBoost relative to treat-all and treat-none strategies across the examined threshold-probability range (Fig. 3B).

Global SHAP analysis identified AVPU level of consciousness, age, ambulance transportation, GCS, and initial physiological measures among the variables with the largest contributions to model output (Fig. 5). Age-stratified SHAP analyses showed differences in the distribution and relative contribution of predictor information between younger and older adults (Additional file 1: Figure S2), but these patterns were not interpreted as causal mechanisms.

Parsimonious model and robustness analyses

The five-variable unweighted logistic model achieved an AUROC of 0.932 (95% CI 0.925–0.940) and an AUPRC of 0.530 (95% CI 0.502–0.557) in 2023. Its calibration slope was 0.977 and its mean predicted

probability was 1.36%, compared with an observed mortality of 1.39%. The corresponding nomogram is shown in Additional file 1: Figure S1.

XGBoost AUROC was 0.978 under the primary 24-hour definition, 0.980 with a 12-hour definition, and 0.981 with a 6-hour definition. Excluding 29,561 post-admission transfers increased AUROC to 0.983. AUROC was 0.9783 with median imputation and 0.9787 with iterative multivariable imputation. Random-search tuning changed AUROC from 0.9783 to 0.9802. In contrast, removal of neurological and physiological predictor groups reduced AUROC to 0.898 (Additional file 1: Table S5).

Discussion

This large multicentre study evaluated early mortality prediction after acute trauma as a clinical prediction and decision-support problem rather than as a discrimination-only modelling task. Routinely collected information available during initial emergency-department assessment provided substantial prognostic information, and all four regression and machine-learning algorithms demonstrated substantially greater discrimination than the conventional physiological trauma scores evaluated. However, several findings limit the extent to which these models can be considered ready for clinical decision-support implementation.

First, differences in discrimination across the four candidate algorithms were small relative to the loss of performance observed when neurological and physiological predictor groups were removed. Logistic regression and the tree-based models had AUROCs within approximately 0.003 of one another, whereas ablation of neurological and physiological predictors reduced XGBoost AUROC by approximately 0.08. This pattern suggests that predictor content and alignment with the intended prediction time point may be at least as influential as algorithmic complexity for early trauma-risk prediction.

Second, probability calibration and fixed-threshold operating performance showed different temporal behaviour. Derivation-fitted isotonic calibration produced predicted mortality probabilities that were closely aligned with observed mortality in the 2023 cohort, whereas sensitivity at fixed derivation-defined operating cut-points decreased when those cut-points were applied unchanged over time. Third, aggregate performance concealed important subgroup heterogeneity. Adults aged 65 years or older had a higher mortality prevalence but lower discrimination than younger adults. Taken together, these findings support evaluation of the complete prediction pipeline across multiple dimensions rather than reliance on a single aggregate discrimination measure.

Implications for prediction-supported clinical decision making

The findings reinforce the distinction among ranking performance, probability estimation, and classification at a specified operating point. AUROC summarises the ability of a model to rank patients across all possible thresholds but does not establish that predicted probabilities are accurate or that

sensitivity and specificity at a fixed cut-point will remain stable over time. Calibration addresses whether predicted probabilities correspond to observed risk [13], whereas operating-threshold performance reflects the consequences of using a specific cut-point for classification.

These distinctions are particularly relevant when prediction outputs are considered for clinical decision support. A displayed patient-level risk probability requires adequate calibration, whereas an alerting or escalation rule requires acceptable and stable threshold-specific operating characteristics. Neither property can be inferred from AUROC alone. The present findings add a temporal dimension to this distinction: strong rank discrimination may persist while the sensitivity associated with a previously defined operating cut-point decreases in a later cohort.

This observation is consistent with broader medical-informatics concerns regarding secondary use of health data, class-imbalance handling, threshold adjustment, emergency-department prediction, and uncertainty in predictive systems [21–24]. For clinical implementation, performance monitoring should therefore extend beyond discrimination to include probability calibration, threshold-specific operating characteristics, data quality, and subgroup performance.

The class-weighted XGBoost model also illustrates why raw model outputs should not automatically be interpreted as patient-level probabilities. Mean uncalibrated output exceeded observed mortality, consistent with the effect of inverse-frequency class weighting on the model-output scale. Application of the isotonic calibration mapping fitted within the derivation data reduced this discrepancy in the 2023 cohort. If model outputs are intended to communicate absolute risk, calibration and subsequent monitoring of calibration stability are therefore necessary. If a model is instead used to trigger an alert at a fixed score, the temporal stability of the associated sensitivity and specificity becomes the more relevant operating property.

Algorithm complexity, predictor timing, and interpretability

The small differences in discrimination among logistic regression, random forest, XGBoost, and LightGBM suggest that predictor availability and prediction-time alignment may be as important as algorithmic complexity for this task. Predictors were intentionally restricted to information available during initial emergency-department assessment. Laboratory results, imaging findings, procedures, and subsequent treatments were excluded because they were not available at the intended prediction time point. This approach reduced the potential for temporal information leakage and maintained correspondence between model inputs and the proposed early decision-support context.

The five-variable parsimonious logistic model further illustrated the prognostic information contained in routinely available variables. Using age, systolic blood pressure, heart rate, respiratory rate, and GCS, the model achieved an AUROC of 0.932 in 2023, with a calibration slope of 0.977 and a mean predicted probability close to the observed mortality prevalence. Although its discrimination was lower than that of the full models, the findings indicate that substantial prognostic information was retained using a limited set of immediately available variables. The nomogram should nevertheless be regarded as a

transparency and visualisation tool rather than a validated bedside score; independent validation and prospective workflow evaluation would be required before clinical use.

SHAP analyses characterised how individual predictors contributed to fitted XGBoost outputs and how those contributions varied across predictor values. These analyses may facilitate model inspection and interpretation but do not establish causal effects or clinical mechanisms [20]. In prediction-supported decision systems, explainability should therefore complement, rather than substitute for, evaluation of calibration, threshold behaviour, subgroup performance, and real-world usability.

Subgroup reliability and older adults

The lower discrimination observed among adults aged 65 years or older warrants particular attention because this subgroup also had a higher mortality prevalence. Age-related differences in physiological reserve, frailty, multimorbidity, and medication exposure may contribute to greater heterogeneity in early presentation after trauma [25]. However, these factors were not available in EDIIS, and the observed performance difference cannot be attributed to a specific mechanism.

The subgroup findings instead demonstrate that performance estimated in an overall population should not automatically be assumed to apply uniformly across clinically important patient groups. Future validation studies should prespecify subgroup analyses and evaluate discrimination, calibration, and relevant operating characteristics within higher-risk populations. Additional information such as frailty, comorbidity, medication exposure, or serial physiological measurements may improve prediction among older adults, although any additional predictors should remain compatible with the intended early prediction time point.

Implementation and future work

The models evaluated in this study were not deployed in clinical practice, clinicians did not receive model outputs, and care decisions were not assigned according to predicted risk. Accordingly, decision-curve net benefit should be interpreted as a decision-analytic property rather than evidence that model implementation improves clinical decisions, workflow, or patient outcomes.

Before independent validation, model-selection procedures should ideally be prespecified and fixed so that performance estimates are not influenced by the same data used to select among candidate models. The complete prediction pipeline should subsequently undergo independent temporal or geographic validation in data not used for algorithm selection [15]. Prospective studies would then be required to evaluate how model outputs are displayed, interpreted, and integrated into clinical workflow, as well as their effects on clinician judgement, alert burden, automation bias, safety, and patient outcomes.

Any future implementation strategy should include ongoing evaluation of the complete prediction pipeline. Monitoring should distinguish changes in outcome prevalence, data completeness, calibration, discrimination, threshold-specific sensitivity and specificity, and subgroup performance. Where class

weighting or other imbalance-handling approaches alter the model-output scale, raw outputs should not be presented as absolute risks unless an appropriate calibration procedure has been independently validated. Operating thresholds intended to trigger clinical action should preferably be defined on an interpretable probability scale and reassessed when patient characteristics, documentation practices, or care processes change.

Strengths and limitations

This study has several strengths. It used a large multicentre cohort with a substantial number of outcome events and separated model development from later evaluation by calendar time. Predictors were restricted to information available at the intended early-assessment time point, and preprocessing, imputation, model fitting, and calibration procedures requiring parameter estimation were performed using derivation data before being applied unchanged to the 2023 cohort. Model evaluation extended beyond AUROC to include AUPRC, calibration, Brier score, fixed-threshold transportability, decision-curve analysis, subgroup performance, explainability, and multiple sensitivity and robustness analyses. Computational environment details and analytical specifications were also reported to support reproducibility.

Several limitations should be considered. First, the retrospective observational design is susceptible to measurement error, missing data, and temporal or institutional variation in documentation. For example, GCS missingness decreased from 2.31% in the derivation period to 0.27% in 2023, indicating temporal changes in data completeness.

Second, although EDIIS includes multiple emergency departments across South Korea, transportability to other countries, trauma systems, institutional workflows, and documentation environments requires independent evaluation. In addition, 2023 outcomes contributed to designation of XGBoost as the reference model. Its reported 2023 performance therefore represents temporally held-out model characterisation rather than independent external validation of a prespecified final model and may retain selection-related optimism.

Third, inverse-frequency class weighting altered the raw model-output scale, limiting interpretation of uncalibrated outputs as absolute mortality probabilities. Although derivation-fitted isotonic calibration improved calibration in 2023, the stability and transportability of calibrated probabilities require independent validation.

Fourth, available predictors did not include frailty, detailed comorbidity, medication exposure, anatomical injury severity, laboratory or imaging findings, or serial physiological measurements. These omissions may partly account for residual subgroup heterogeneity, particularly among older adults; however, incorporating downstream information would also alter the intended early prediction time point.

Fifth, some subgroup and comparator analyses require cautious interpretation. Very high AUROC estimates for penetrating injury and other mechanisms were based on only 33 and 29 deaths, respectively, and should be regarded as exploratory. Comparisons with RTS and rSIG were restricted to

paired complete-score subsets; although component missingness was limited, these subsets may not fully represent the entire cohort.

Finally, SHAP analyses describe model behaviour rather than causal relationships, and this study did not evaluate real-world implementation. Usability, workflow integration, clinician behaviour, alert burden, safety, and patient outcomes therefore require prospective evaluation.

Conclusions

Routinely collected information available during initial emergency-department assessment provided substantial prognostic information for in-hospital mortality after acute trauma. Regression and machine-learning algorithms demonstrated high discrimination, but high aggregate AUROC did not ensure temporal stability of fixed operating-threshold performance or uniform discrimination across age groups.

For XGBoost, derivation-fitted isotonic calibration improved probability calibration in the 2023 cohort, whereas fixed derivation-defined operating cut-points showed reduced sensitivity when applied unchanged over time. These findings highlight the importance of distinguishing rank discrimination, calibrated risk estimation, and threshold-specific classification when prediction models are evaluated for clinical decision-support use.

Evaluation should therefore consider temporal transportability and subgroup reliability in addition to aggregate predictive performance. Independent validation in data not used for algorithm selection, followed by prospective assessment of usability, workflow integration, safety, and clinical impact, is required before routine implementation.

Abbreviations

AUPRC

area under the precision–recall curve

AUROC

area under the receiver operating characteristic curve

AVPU

Alert/Voice/Pain/Unresponsive

CI

confidence interval

ED

emergency department

EDIIS

Emergency Department-based Injury In-depth Surveillance

GCS

Glasgow Coma Scale

LightGBM

Light Gradient Boosting Machine
NPV
negative predictive value
PPV
positive predictive value
rSIG
reverse Shock Index multiplied by Glasgow Coma Scale
RTS
Revised Trauma Score
SHAP
SHapley Additive exPlanations
XGBoost
Extreme Gradient Boosting.

Declarations

Ethics approval and consent to participate: The study was approved by the Institutional Review Board of Daejeon University (approval no. 1040647-202607-HR-015-01). The requirement for individual informed consent was waived because the study involved secondary analysis of de-identified EDIIS data. Data were analysed through the authorised Korea Disease Control and Prevention Agency EDIIS access process and in accordance with the applicable data-use conditions.

Consent for publication: Not applicable.

Availability of data and materials: The data that support the findings of this study are available from the Korea Disease Control and Prevention Agency through the Emergency Department-based Injury In-depth Surveillance (EDIIS) data-access process. Restrictions apply to the availability of these data, and the authors are not permitted to redistribute the dataset publicly or directly to third parties. Researchers may apply to the Korea Disease Control and Prevention Agency for access, subject to the agency's eligibility criteria, approval procedures, and applicable data-use conditions.

Competing interests: The authors declare no competing interests.

Funding: This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (RS-2026-25490096).

Authors' contributions: JSL contributed to conceptualisation, methodology, supervision, investigation, project administration, data curation, funding acquisition, writing—original draft, and writing—review and editing. JHB contributed to data curation, formal analysis, validation, and writing—review and editing. HJK contributed to writing—review and editing. All authors reviewed and approved the final manuscript.

Acknowledgements: This study used data provided by the Korea Disease Control and Prevention Agency (KDCA; research management number KDCA-12-02-EI-2026-000126). The findings and interpretations presented in this article are those of the authors and do not represent the official views of the KDCA.

References

1. World Health Organization. Injuries and violence. World Health Organization; 2024. <https://www.who.int/news-room/fact-sheets/detail/injuries-and-violence>. Accessed 17 July 2026.
2. GBD 2021 Diseases and Injuries Collaborators. Global incidence, prevalence, years lived with disability (YLDs), disability-adjusted life-years (DALYs), and healthy life expectancy (HALE) for 371 diseases and injuries in 204 countries and territories and 811 subnational locations, 1990–2021: a systematic analysis for the Global Burden of Disease Study 2021. *Lancet*. 2024;403(10440):2133–2161. doi:10.1016/S0140-6736(24)00757-8.
3. Sutton RT, Pincock D, Baumgart DC, Sadowski DC, Fedorak RN, Kroeker KI. An overview of clinical decision support systems: benefits, risks, and strategies for success. *NPJ Digit Med*. 2020;3:17. doi:10.1038/s41746-020-0221-y.
4. Champion HR, Sacco WJ, Copes WS, Gann DS, Gennarelli TA, Flanagan ME. A revision of the Trauma Score. *J Trauma*. 1989;29(5):623–629. doi:10.1097/00005373-198905000-00017.
5. Kimura A, Tanaka N. Reverse shock index multiplied by Glasgow Coma Scale score (rSIG) is a simple measure with high discriminant ability for mortality risk in trauma patients: an analysis of the Japan Trauma Data Bank. *Crit Care*. 2018;22(1):87. doi:10.1186/s13054-018-2014-0.
6. Park JS, Choi SJ, Kim MJ, Choi SY, Kim HY, Park YS, et al. Cutoff of the reverse shock index multiplied by the Glasgow Coma Scale for predicting in-hospital mortality in adult patients with trauma: a retrospective cohort study. *BMC Emerg Med*. 2024;24(1):55. doi:10.1186/s12873-024-00978-z.
7. Vang M, Østberg M, Steinmetz J, Rasmussen LS. Shock index as a predictor for mortality in trauma patients: a systematic review and meta-analysis. *Eur J Trauma Emerg Surg*. 2022;48(4):2559–2566. doi:10.1007/s00068-022-01932-z.
8. Cardosi JD, Shen H, Groner JI, Armstrong M, Xiang H. Machine learning for outcome predictions of patients with trauma during emergency department care. *BMJ Health Care Inform*. 2021;28(1):e100407. doi:10.1136/bmjhci-2021-100407.
9. Look CSJ, Teixayavong S, Djärv T, Ho AFW, Tan KBK, Ong MEH. Improved interpretable machine learning emergency department triage tool addressing class imbalance. *Digit Health*. 2024;10:20552076241240910. doi:10.1177/20552076241240910.
10. Ling L, Ma J, Xia X, Yuan H, Liu S, Zhu Y. Early mortality prediction after severe trauma using ensemble machine learning: a single-center retrospective study. *Front Public Health*. 2025;13:1716910. doi:10.3389/fpubh.2025.1716910.
11. Oh NE, Oh TYC, Hsu J, et al. Prehospital real-time AI for trauma mortality prediction: a multi-institutional and multi-national validation study. *Nat Commun*. 2026;17(1):1464.

doi:10.1038/s41467-025-68198-y.

12. Collins GS, Moons KGM, Dhiman P, 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.
13. Van Calster B, McLernon DJ, van Smeden M, Wynants L, Steyerberg EW. Calibration: the Achilles heel of predictive analytics. *BMC Med*. 2019;17(1):230. doi:10.1186/s12916-019-1466-7.
14. Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. *Med Decis Making*. 2006;26(6):565–574. doi:10.1177/0272989X06295361.
15. Pennestrì F, Cabitza F, Picerno N, Banfi G. Sharing reliable information worldwide: healthcare strategies based on artificial intelligence need external validation. Position paper. *BMC Med Inform Decis Mak*. 2025;25(1):56. doi:10.1186/s12911-025-02883-2.
16. Breiman L. Random forests. *Mach Learn*. 2001;45(1):5–32. doi:10.1023/A:1010933404324.
17. 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*. Association for Computing Machinery; 2016:785–794. doi:10.1145/2939672.2939785.
18. Ke G, Meng Q, Finley T, et al. LightGBM: a highly efficient gradient boosting decision tree. *Adv Neural Inf Process Syst*. 2017;30:3146–3154.
19. Brier GW. Verification of forecasts expressed in terms of probability. *Mon Weather Rev*. 1950;78(1):1–3. doi:10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2.
20. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. *Adv Neural Inf Process Syst*. 2017;30:4765–4774.
21. Pant D, Nytrø Ø, Leventhal BL, Clausen C, Koochakpour K, Stien L, et al. Secondary use of health records for prediction, detection, and treatment planning in the clinical decision support system: a systematic review. *BMC Med Inform Decis Mak*. 2025;25(1):190. doi:10.1186/s12911-025-03021-8.
22. Nikbakht-Fard R, Tapak L, Khazei M, et al. Optimizing stroke mortality prediction: a comparative study of sampling, cost-sensitive learning, and threshold adjustment. *BMC Med Inform Decis Mak*. 2026. doi:10.1186/s12911-026-03707-7.
23. Zhu Y, Feng J, Liu R, Ren W, Ma X, Zhang C, et al. Development and validation of machine learning-based prediction models for adult bowel necrosis for patients presenting with acute abdominal pain at the emergency department: a multicenter retrospective cohort study. *BMC Med Inform Decis Mak*. 2026. doi:10.1186/s12911-026-03628-5.
24. Bayly H, Tripodis Y, Lenio S, Patil P; Alzheimer's Disease Neuroimaging Initiative. Reducing overconfident errors in clinical prediction models. *BMC Med Inform Decis Mak*. 2026. doi:10.1186/s12911-026-03681-0.
25. Kim SY, Hong KJ, Shin SD, et al. Validation of the shock index, modified shock index, and age shock index for predicting mortality of geriatric trauma patients in emergency departments. *J Korean Med Sci*. 2016;31(12):2026–2032. doi:10.3346/jkms.2016.31.12.2026.

Figures

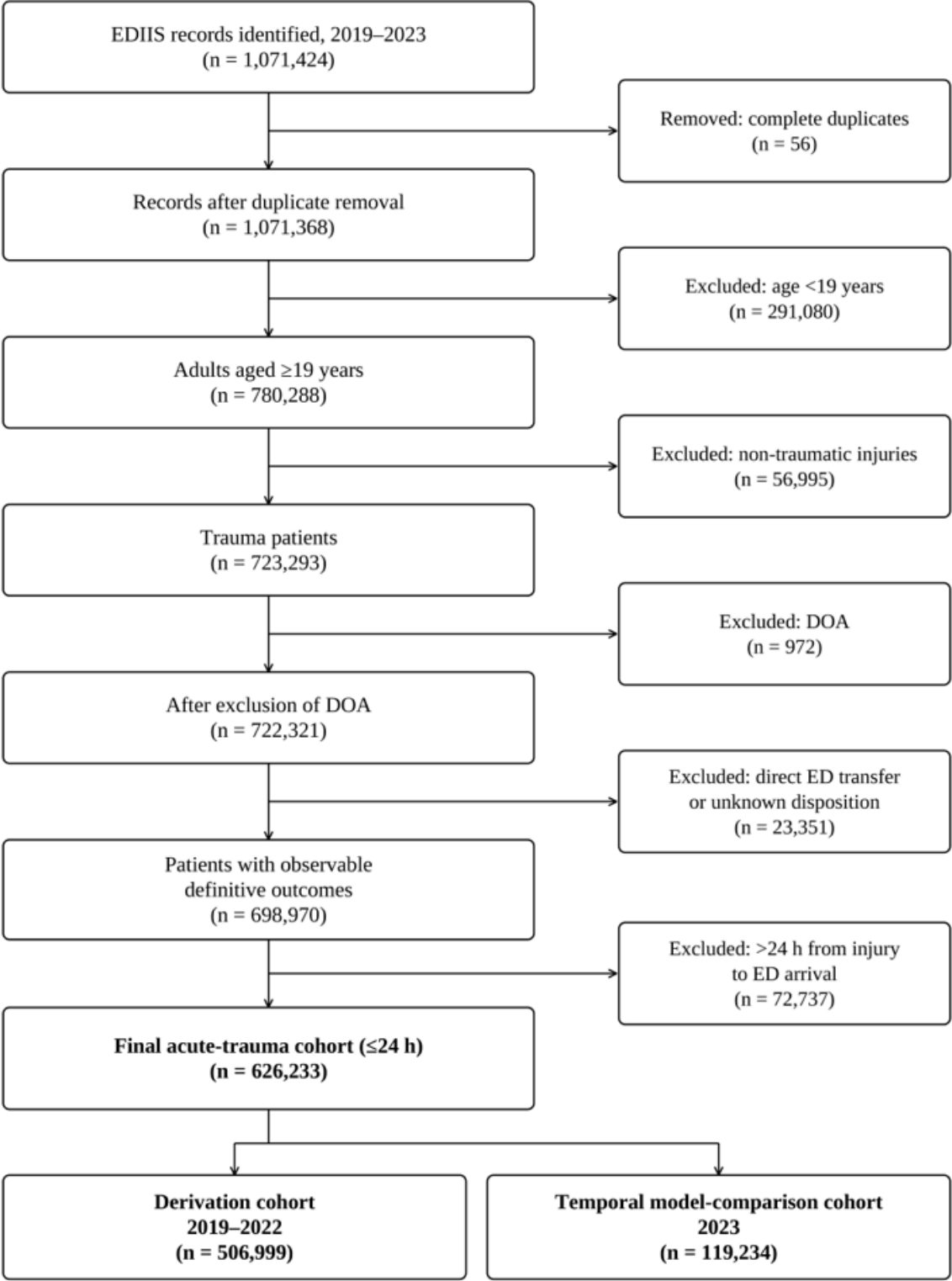


Figure 1

Flow diagram of study population selection.

Complete duplicate records were removed before application of the eligibility criteria. DOA, dead on arrival; ED, emergency department; EDIIS, Emergency Department-based Injury In-depth Surveillance.

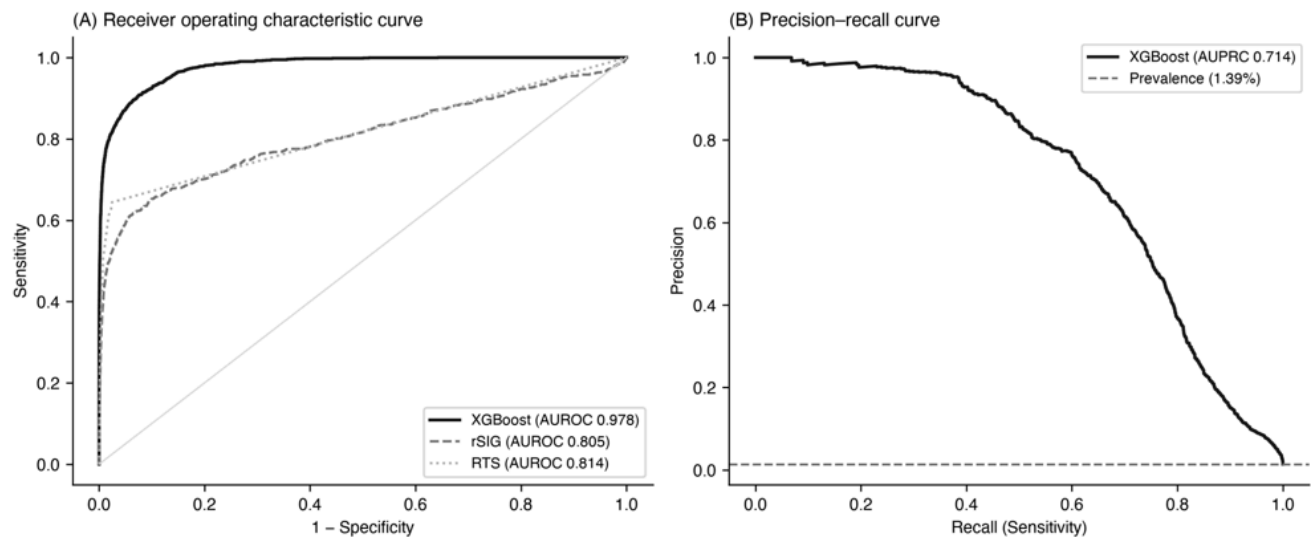


Figure 2

Discrimination of prediction models and conventional trauma scores for in-hospital mortality in the 2023 temporal model-comparison cohort.

(A) Receiver operating characteristic curves for the selected XGBoost model and the conventional physiological trauma scores (Revised Trauma Score and reverse Shock Index multiplied by Glasgow Coma Scale). (B) Precision-recall curve for the XGBoost model. The horizontal reference line represents the observed prevalence of in-hospital mortality (1.39%) in the 2023 cohort. AUROC, area under the receiver operating characteristic curve; AUPRC, area under the precision-recall curve; RTS, Revised Trauma Score; rSIG, reverse Shock Index multiplied by Glasgow Coma Scale.

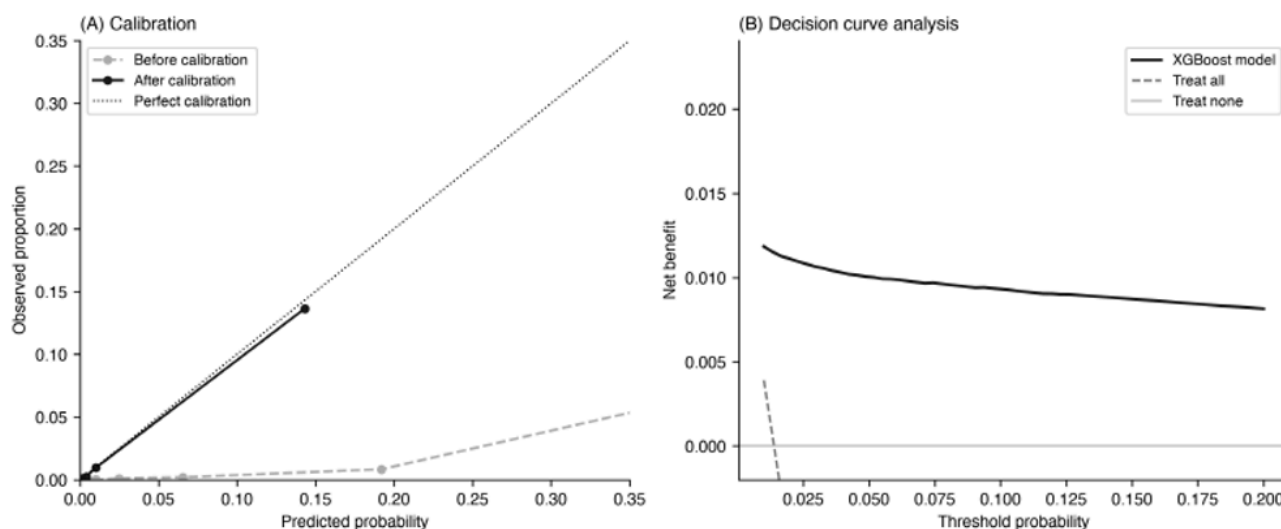


Figure 3

Probability calibration and decision-curve analysis of XGBoost in the 2023 temporal model-comparison cohort.

(A) Calibration plot before and after application of the isotonic calibration mapping estimated from out-of-fold predictions generated exclusively within the 2019–2022 derivation cohort. The diagonal line represents perfect calibration. (B) Decision-curve analysis based on isotonic-calibrated mortality probabilities, with net benefit compared with treat-all and treat-none strategies across the examined threshold-probability range. The calibration mapping was applied unchanged to the 2023 cohort; no 2023 outcomes were used to estimate or update the calibration function. Decision-curve analysis represents decision-analytic evaluation and should not be interpreted as evidence of clinical effectiveness.

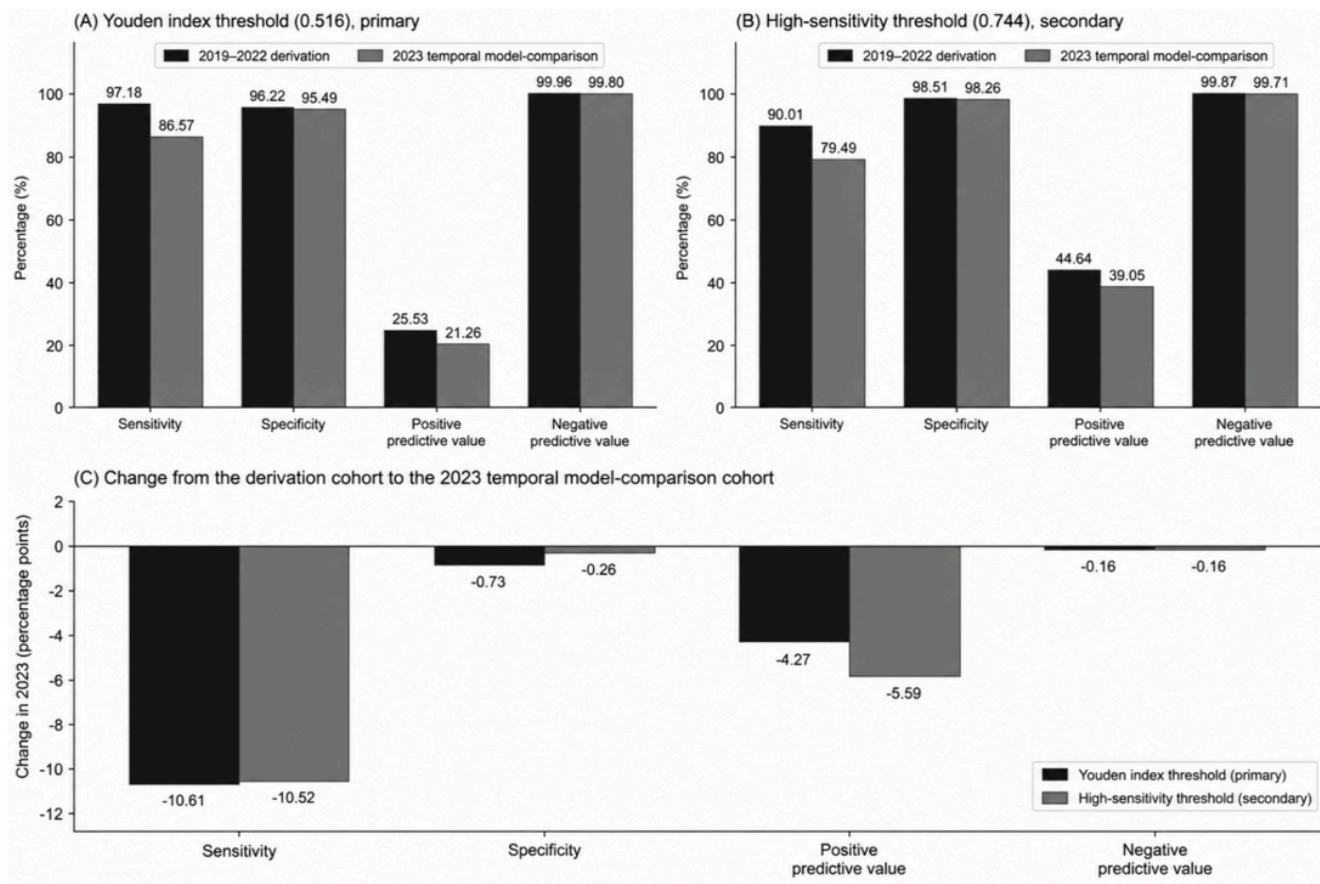


Figure 4

Temporal change in operating-threshold performance between the derivation and 2023 model-comparison cohorts.

(A) Classification performance at the model-output threshold maximising the Youden index in the 2019–2022 derivation cohort and subsequently applied unchanged to the 2023 temporal model-comparison cohort. (B) Classification performance at the model-output threshold defined as the highest threshold retaining at least 90% sensitivity in the 2019–2022 derivation cohort and subsequently applied unchanged to the 2023 temporal model-comparison cohort. (C) Changes in sensitivity, specificity, positive predictive value and negative predictive value between the derivation and 2023 cohorts at both operating thresholds. Thresholds were determined exclusively from class-weighted, uncalibrated model outputs in the derivation cohort and were not re-optimised using 2023 outcomes. Accordingly, the thresholds represent model-specific operating cut-points rather than absolute mortality probabilities. NPV, negative predictive value; PPV, positive predictive value.

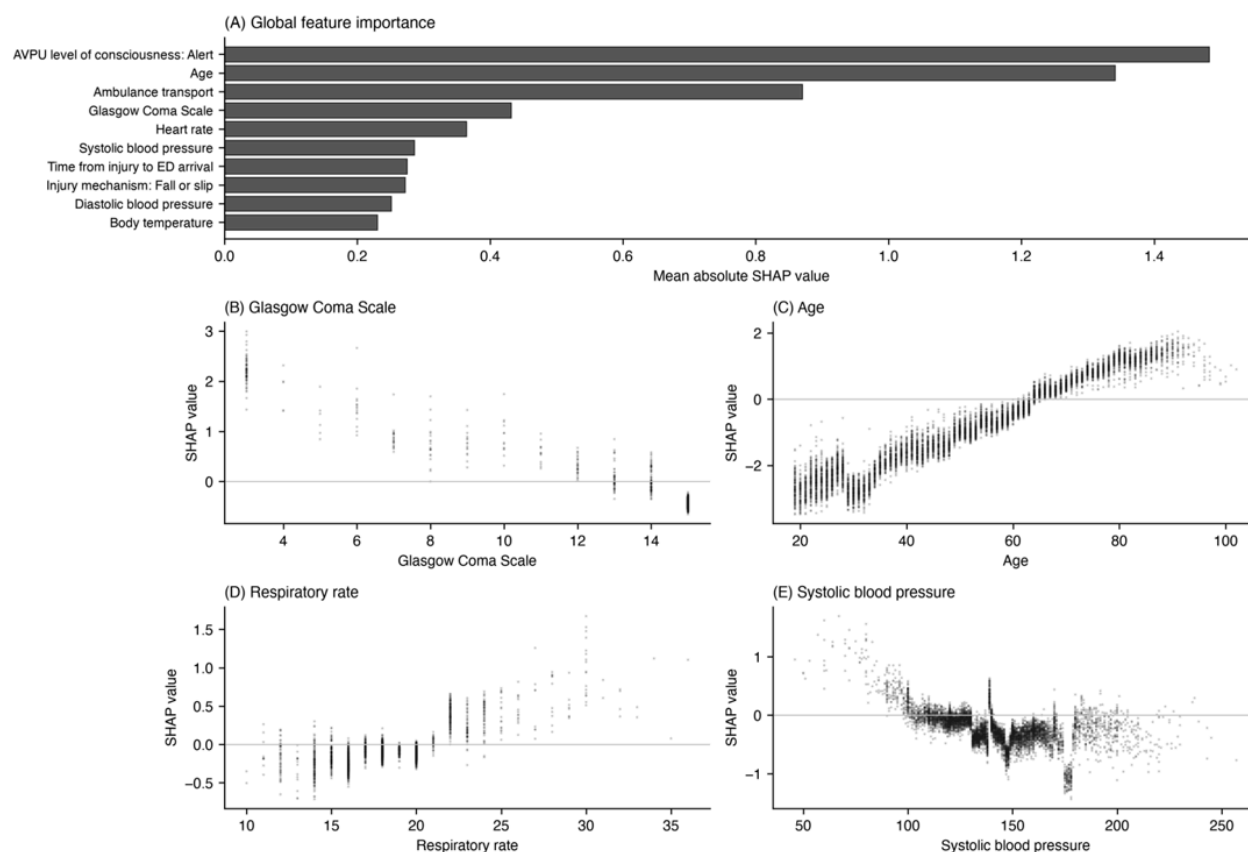


Figure 5

SHAP-based characterisation of XGBoost model behaviour.

(A) Global feature contributions based on mean absolute SHapley Additive exPlanations (SHAP) values. Higher mean absolute SHAP values indicate greater contributions to model output. Dependence plots show relationships between predictor values and model contributions for (B) Glasgow Coma Scale, (C) age, (D) respiratory rate and (E) systolic blood pressure. Positive SHAP values indicate contributions towards higher model output for in-hospital mortality and negative values indicate contributions towards lower model output. SHAP values characterise model behaviour and should not be interpreted as causal effects.

Supplementary Files

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

- [BMCAAdditionalFile0918.docx](#)