Research Square

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

Early risk stratification for in-hospital mortality and 6-month unfavourable prognosis after injury in traumatic brain injury: machine learning models for clinical decision support

Yongfeng Wang
Nantong University
Jianjie Xu
Nantong University
Wei Shi
Affiliated Hospital of Nantong University
Ze Song
Tongzhou District People's Hospital of Nantong
Yixiang Guan
Hai'an People's Hospital
Haoliang Wu
Qidong People's Hospital
Hua Zhang
Qidong People's Hospital
Peizhe Jiang
Nantong University
Rui Jiang
Affiliated Hospital of Nantong University
Qianqian Liu
ntfyliuqianqian@163.com

Affiliated Hospital of Nantong University

Research Article

Keywords: traumatic brain injury, machine learning, prognostic prediction, external validation, clinical decision support

Posted Date: November 9th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11197260/v1

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

Additional Declarations: No competing interests reported.

Early risk stratification for in-hospital mortality and 6-month unfavourable prognosis after injury in traumatic brain injury: machine learning models for clinical decision support

Yongfeng Wang[a]†, Jianjie Xu[a]†, Wei Shi[b], Ze Song[c], Yixiang Guan[d], Haoliang Wu[e], Hua Zhang[e], Peizhe Jiang[a], Rui Jiang[b]*, Qianqian Liu[b]*

†Yongfeng Wang and Jianjie Xu contributed equally to this work.

*Correspondence: Qianqian Liu, ntfyliuqianqian@163.com; Rui Jiang, ntfyjiangrui@ntu.edu.cn

Correspondence address: Department of Neurosurgery, Affiliated Hospital of Nantong University, No. 20 Xisi Road, Nantong, Jiangsu 226001, China

[a] Nantong University, Nantong, Jiangsu, China

[b] Department of Neurosurgery, Affiliated Hospital of Nantong University, Nantong, Jiangsu, China

[c] Department of Neurosurgery, Tongzhou District People's Hospital of Nantong, Nantong, Jiangsu, China

[d] Department of Neurosurgery, Hai'an People's Hospital, Nantong, Jiangsu, China

[e] Department of Neurosurgery, Qidong People's Hospital, Nantong, Jiangsu, China

Acknowledgments

The authors thank the participating medical institutions and the clinical, research, and data management staff for their support in case data collection, patient follow-up, data organisation, and study implementation.

Author contributions

Yongfeng Wang and Qianqian Liu conceived and designed the study. Wei Shi, Hua Zhang, Haoliang Wu, Ze Song, and Yixiang Guan provided the study data. Peizhe Jiang checked the data. Yongfeng Wang, Qianqian Liu, Jianjie Xu, and Peizhe Jiang directly accessed and verified the underlying individual participant data used in this study. Patient follow-up was conducted by Qianqian Liu, Rui Jiang, and Yixiang Guan. Yongfeng Wang and Jianjie Xu were responsible for data analysis, model development, and figure preparation, and verified the consistency between the results reported in the manuscript and those generated by the analytical code. Yongfeng Wang and Jianjie Xu drafted the original manuscript. Qianqian Liu interpreted the results. Qianqian Liu and Rui Jiang supervised the study and reviewed and edited the manuscript. Qianqian Liu and Wei Shi were responsible for funding acquisition and project administration. All authors reviewed or critically revised the manuscript, approved the final version, and agreed to be accountable for the integrity and accuracy of the work.

Abstract

Purpose

To develop and externally validate machine-learning models using routinely available early admission data for in-hospital death and 6-month unfavourable prognosis after injury in traumatic brain injury (TBI).

Methods

This retrospective multicentre prognostic study included patients aged ≥15 years admitted within 12 h after TBI at four Chinese hospitals from November 2020 to October 2025. Two hospitals were used for development and two for independent external testing. Predictors included age, neurological assessments, laboratory measurements, and head computed tomography (CT) findings. Laboratory measurements were standardised using Value/upper limit of normal (ULN). Models were selected internally and evaluated using discrimination, classification, calibration, and decision curve analysis.

Results

The in-hospital death analysis included 2260 patients (523 deaths, 23.1%); the 6-month unfavourable prognosis after injury analysis included 1512 discharge survivors with available follow-up (254 unfavourable outcomes, 16.8%). For in-hospital death, CatBoost using features selected by Lasso-Logistic regression achieved external area under the receiver operating characteristic curve (AUROC) 0.903 (95% confidence interval [CI], 0.883–0.922), area under the precision–recall curve (AUPRC) 0.800 (95% CI, 0.760–0.844), sensitivity 82.0%, positive predictive value (PPV) 71.1%, and F1 score 76.2%. For 6-month unfavourable prognosis after injury, logistic regression using features selected by XGBoost-based recursive feature elimination with cross-validation (RFECV-XGBoost) achieved external AUROC 0.906 (95% CI, 0.873–0.936), AUPRC 0.727 (95% CI, 0.642–0.823), sensitivity 81.1%, PPV 72.5%, and F1 score 76.6%.

Conclusion

Both models maintained consistent cross-centre predictive performance and may support early risk stratification and treatment planning after TBI. Prospective multicentre evaluation is still needed before implementation.

Keywords: traumatic brain injury; machine learning; prognostic prediction; external validation; clinical decision support

Introduction

Traumatic brain injury (TBI) is a common traumatic condition mainly caused by traffic accidents, falls from height, and simple falls [7]. Despite advances in the clinical management of TBI and related research, mortality and disability remain high [13, 19]. The Glasgow Coma Scale (GCS) remains widely used for neurological assessment after TBI [19]. However, the GCS primarily reflects the level of consciousness and has limited prognostic value when used alone [21]. Therefore, a more comprehensive prognostic assessment should integrate neurological findings with additional clinical, laboratory, and imaging information [21].

In recent years, artificial intelligence has been widely applied to disease diagnosis and prognostic prediction [27]. Existing studies have shown that machine learning can integrate multidimensional information, including clinical characteristics, laboratory measurements, and imaging findings, thereby providing new approaches for disease diagnosis and prognostic prediction [11, 14, 16]. Meanwhile, the validation of prediction models in different populations and their clinical application value have also received increasing attention [3, 10, 28].

Nevertheless, some existing models still have limitations, including reliance on samples from a single centre, lack of external validation, and a focus on a single prediction outcome [10, 20]. In this study, we used routinely available early admission clinical data [31] to develop and validate prediction models for in-hospital death and 6-month unfavourable prognosis after injury in patients with TBI. Model performance and generalisability were evaluated in an external test set. Calibration curves [34] and decision curve analysis (DCA) [5, 35] were used to further assess the models. For each outcome, a final model was selected based on overall performance during internal evaluation, and model-specific interpretation methods were used to evaluate the contribution of key features to the prediction results.

Materials and Methods

Study design and participants

We retrospectively included patients with TBI admitted to the Affiliated Hospital of Nantong University (Centre A), Tongzhou District People's Hospital of Nantong (Centre B), Qidong People's Hospital (Centre C), and Hai'an People's Hospital (Centre D) between November 2020 and October 2025. Data from Centres A and B constituted the development set, whereas data from Centres C and D constituted the external test set for evaluation of model performance and generalisability.

All variables used for model development were obtained from the electronic medical record systems of the participating hospitals. We initially collected 76 variables, including age, sex, medical history, GCS score, Rotterdam CT score [18], and laboratory measurements (Tables S1 and S2 in Online Resource 1). Because clinical deterioration and unfavourable outcomes can occur in patients with GCS scores of 13 to 15, these patients were not excluded [29].

The outcomes of this study were in-hospital death and 6-month unfavourable prognosis after injury. In-hospital death was defined as death during hospitalisation or an irreversible moribund state characterised by absent spontaneous respiration and dependence on life support, with death judged to be imminent by the treating clinicians. The analysis of 6-month unfavourable prognosis after injury was restricted to patients who survived to hospital discharge. Prognosis was assessed using the Glasgow Outcome Scale (GOS) [12]. A GOS score of 4 or 5 was classified as favourable prognosis,

indicating the ability to live independently, whereas a GOS score of 1 to 3 was classified as unfavourable prognosis, indicating death or inability to live independently. Six-month prognosis after injury was obtained from telephone follow-up or electronic medical records. Patients without available 6-month outcome information were considered lost to follow-up and were excluded from the analysis.

The inclusion criteria were as follows: (1) admission within 12 h after TBI, including patients transferred from other hospitals; and (2) age 15 years or older. The general exclusion criteria were as follows: (1) acute infectious disease; (2) patients receiving chemotherapy for a malignant tumour; (3) active immune system disease; (4) haematologic disease that did not affect coagulation or haemostasis but might interfere with the interpretation of inflammation-related routine blood test indicators; (5) severe concomitant injuries to other body regions at the time of TBI that could directly lead to death or paralysis; (6) preexisting limb movement disorder before TBI; (7) refusal of treatment by the patient or family despite the presence of spontaneous respiration, heartbeat, or other vital signs and a reasonable opportunity for treatment; (8) no definite intracranial injury on head CT or poor-quality head CT images that prevented accurate assessment of TBI; (9) transfer from another hospital without available initial examination data; (10) pregnancy; and (11) absence of spontaneous respiration or heartbeat at admission. Additional outcome-specific exclusion criteria are described in Supplementary Method S1.

No formal sample-size calculation was performed; all eligible patients meeting the predefined criteria during the study period were included.

Ethics

This retrospective multicentre study was conducted in accordance with the Declaration of Helsinki and reported in accordance with the TRIPOD+AI statement [4]. The study protocol was approved by the Ethics Committee of the Affiliated Hospital of Nantong University (approval number: 2026-K041-01). Because of the retrospective nature of the study, the requirement for written informed consent was waived by the committee.

Data processing

Variables with more than 20% missing values were excluded (Supplementary Tables S1, S2). Additional variables were excluded based on availability, data accuracy, and redundancy, including selected demographic variables, baseline vital signs, medical history, injury mechanism, and redundant imaging variables (Supplementary Method S1).

The remaining variables were used for model development. To account for differences in laboratory reference ranges, we recorded the reference ranges corresponding to each patient's blood test results during data collection. Before model development, laboratory measurements were standardised using the Value/upper limit of normal (ULN) method [23, 26] (Supplementary Method S2, Supplementary Table S9). Because both the original measurements and their corresponding reference ranges were unavailable for missing laboratory values, these values could not be standardised using the Value/ULN method. Therefore, complete-case analysis was used, and only patients with complete data for variables retained for model development were included in the analysis [36]. To assess the potential impact of excluding patients with incomplete data, we compared event rates before and after exclusion in each centre and in the development and external test sets and calculated standardised mean differences (SMDs) [37]. The assessment showed no substantial imbalance (Supplementary Tables S3, S4); only

thrombin time (TT) in Centre B in the in-hospital death dataset showed a mild distributional shift (SMD = 0.130), slightly above the commonly used reference threshold of 0.10 (Supplementary Tables S5, S6).

Model development and performance evaluation

Fig. 1 shows the model development and evaluation workflow for the two outcomes. Feature subsets were selected using two methods: (1) XGBoost-based recursive feature elimination with cross-validation (RFECV-XGBoost) [2, 9] and (2) logistic regression with an L1 regularisation penalty (Lasso-Logistic regression) [6, 33]. Detailed parameter settings are provided in Supplementary Method S3.

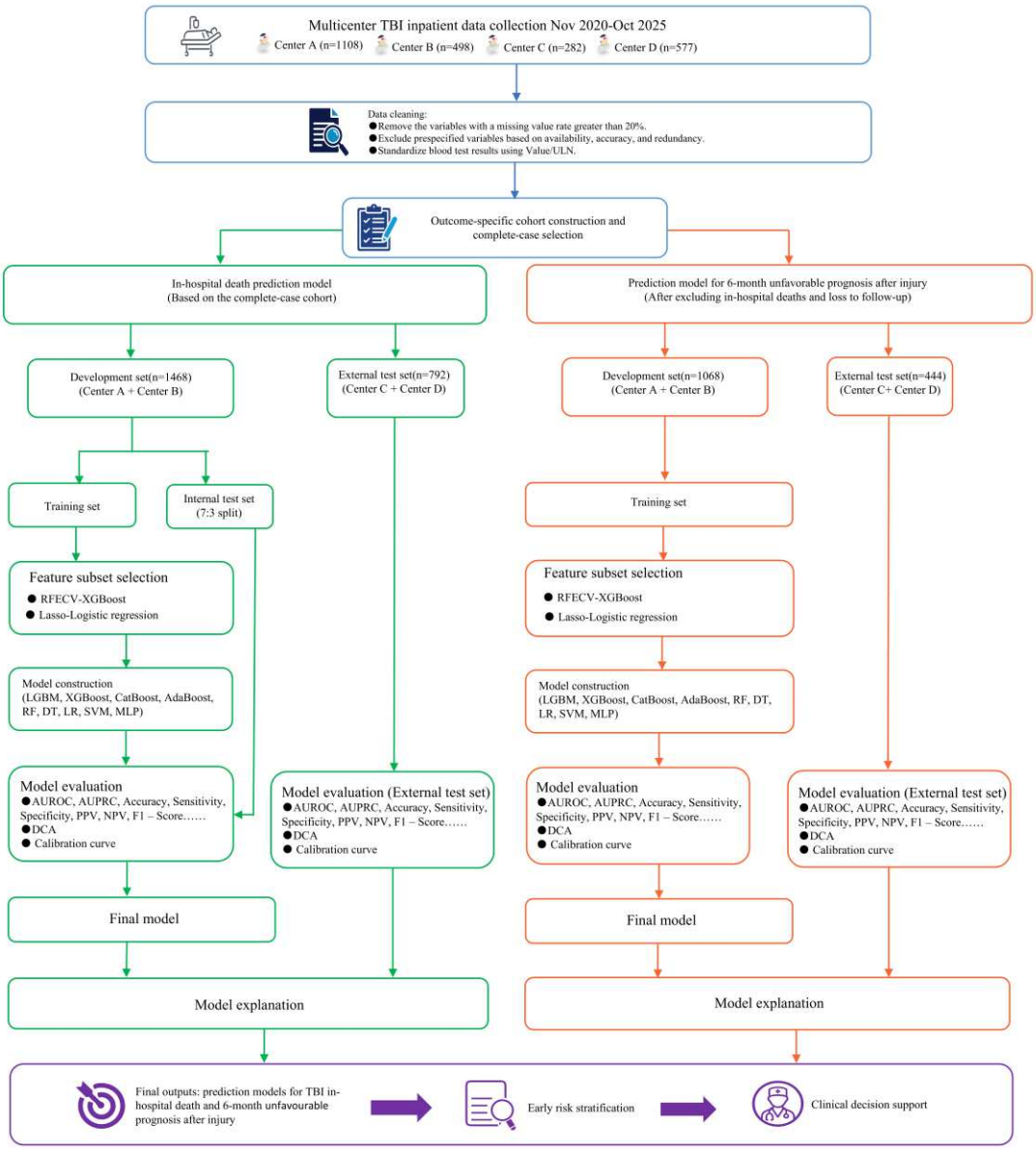


Fig. 1 Study workflow and technical framework for model development. Multicentre data from four medical institutions were processed using complete-case analysis after removing variables with more than 20% missing data and standardising blood test results using the Value/ULN method. Centres A and B formed the development dataset, whereas Centres C and D formed the external test dataset. For each outcome, feature subsets selected using RFECV-XGBoost and Lasso-Logistic regression were separately combined with nine machine-learning algorithms. Candidate models were evaluated using discrimination and classification metrics, DCA, and calibration curves, followed by final model selection and explanation. TBI: traumatic brain

injury; ULN: upper limit of normal; RFECV-XGBoost: XGBoost-based recursive feature elimination with cross-validation; Lasso-Logistic regression: logistic regression with an L1 regularisation penalty; LGBM: Light Gradient Boosting Machine; XGBoost: Extreme Gradient Boosting; CatBoost: Categorical Boosting; AdaBoost: Adaptive Boosting; RF: Random Forest; DT: Decision Tree; LR: Logistic Regression; SVM: Support Vector Machine; MLP: Multilayer Perceptron; AUROC: area under the receiver operating characteristic curve; AUPRC: area under the precision–recall curve; PPV: positive predictive value; NPV: negative predictive value; DCA: decision curve analysis

For the in-hospital death prediction model, because the number of positive events was relatively high (Table 1), the development set (Centres A and B) was divided into a training set and an internal test set at a ratio of 7:3 using stratified sampling; the training set was further divided into a sub-training set and an internal validation subset. Feature selection and hyperparameter optimisation were performed within the training set; the internal validation subset was used for early stopping, where applicable, and classification threshold determination, whereas the internal test set was reserved for internal evaluation. Bayesian optimisation [30] combined with 5-fold stratified cross-validation was used for hyperparameter tuning, with area under the receiver operating characteristic curve (AUROC) maximisation as the objective; the classification threshold was optimised rather than fixed at 0.5, with the optimal threshold determined by maximising the F1 score on the precision–recall curve [8, 32] (Supplementary Method S4).

For the prediction model for 6-month unfavourable prognosis after injury, patients who died in hospital and those without available follow-up information were excluded. Because of the relatively small number of positive events (Table 2), no separate internal test set was created; instead, the entire development set was used as the training set. Bayesian optimisation [30] combined with 5-fold stratified cross-validation was used for hyperparameter optimisation, with AUROC maximisation as the objective. After the optimal hyperparameters were determined, out-of-fold (OOF) predicted probabilities were generated using 10 repeats of 5-fold stratified cross-validation, and the optimal classification threshold was determined by maximising the F1 score on the precision–recall curve [8, 32] (Supplementary Method S4, Table S18).

Both models used class-weight-based cost-sensitive learning during training [1, 24], with class weights based on the ratio of negative to positive samples and adjusted during hyperparameter optimisation (Supplementary Method S4).

For model performance evaluation, the in-hospital death prediction model was internally validated using the internal test set, whereas the prediction model for 6-month unfavourable prognosis after injury was internally evaluated using out-of-fold predictions generated from repeated stratified cross-validation. Both models were externally validated using the independent external test set. Model performance was evaluated using AUROC, area under the precision–recall curve (AUPRC), accuracy, sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), and F1 score. Calibration curves [34] and DCA [5, 35] were used to assess calibration and clinical utility, respectively. Final model selection was based on internal evaluation only. For in-hospital death prediction, models with relatively high sensitivity were first considered, with greater emphasis placed on PPV; discrimination, overall classification performance, calibration, and decision-curve performance in the internal test set were used as supporting considerations. For the prediction model for 6-month unfavourable prognosis after injury, models with relatively high PPV values were first

considered, with greater emphasis placed on sensitivity; overall classification performance, calibration, and decision-curve performance based on out-of-fold evaluation were used as supporting considerations. The external test set was not used for model selection.

Statistical analysis

Statistical analyses were performed using the analytic datasets for model development. Continuous variables were reported as mean (standard deviation [SD]) or median (interquartile range [IQR]) according to their distributional and variance characteristics, and categorical variables were reported as counts (percentages). Between-group comparisons were performed using the independent-samples t test, Welch's t test, one-way analysis of variance, Mann–Whitney U test, Kruskal–Wallis rank-sum test, χ^2 test, or Fisher's exact test according to variable type, data distribution, number of groups, and homogeneity of variance (Supplementary Method S5). GCS was summarised and analysed as mean (SD). All tests were two-sided. P values were reported descriptively to characterise between-dataset differences and were not used as the sole basis for assessing clinical importance or for making model-development decisions.

All statistical analyses, data processing, and model development were performed in Python 3.12.10. The multilayer perceptron (MLP) model was implemented using the PyTorch framework.

Results

Baseline characteristics

A total of 2260 patients were included in the in-hospital death prediction analysis, of whom 523 (23.1%) had the in-hospital death outcome. A total of 1512 patients were included in the prediction analysis for 6-month unfavourable prognosis after injury, of whom 254 (16.8%) had an unfavourable prognosis. Table 1 shows the baseline characteristics and outcome event distributions of patients included in the in-hospital death prediction model: 1027 patients were included in the training set, 441 patients in the internal test set, and 792 patients in the external test set. Table 2 shows the baseline characteristics and outcome event distributions of patients included in the prediction model for 6-month unfavourable prognosis after injury: 1068 patients were included in the training set and 444 patients in the external test set. The external test sets for both models had higher proportions of in-hospital death and 6-month unfavourable prognosis after injury than the corresponding training sets (Tables 1, 2).

Feature subsets

After data preprocessing, 38 candidate variables were included in both the in-hospital death prediction task and the prediction task for 6-month unfavourable prognosis after injury. To reduce the risk of model overfitting caused by variable redundancy, feature selection was performed in the corresponding training data using the RFECV-XGBoost algorithm and Lasso-Logistic regression, and the corresponding feature subsets were ultimately determined (Fig. 2, Supplementary Fig. S1, Table S10, Table S11). Correlations among variables in the selected feature subsets are shown in Supplementary Fig. S2.

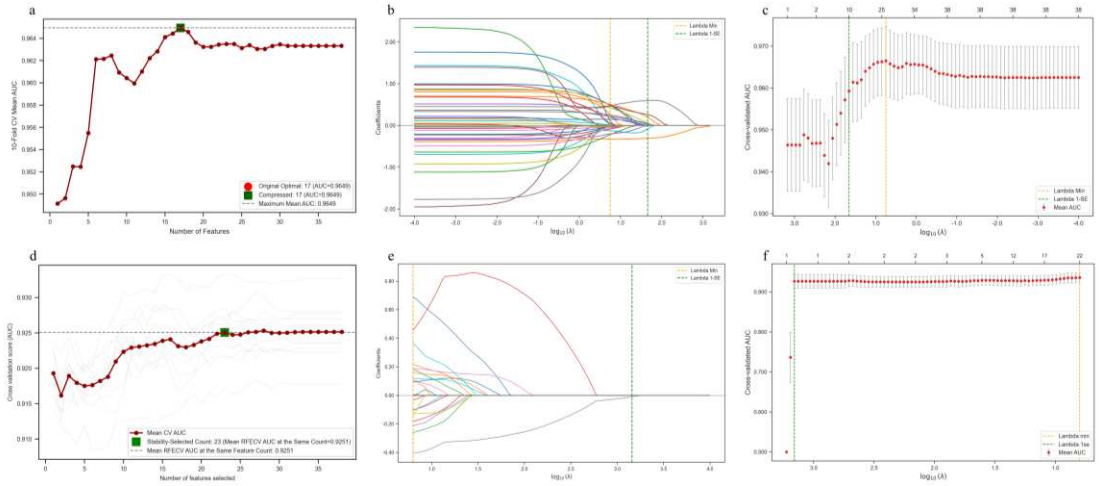


Fig. 2 Feature-selection processes for the two prediction outcomes. a RFECV-XGBoost feature-selection curve for in-hospital death prediction. The curve with circular markers shows the mean 10-fold cross-validation AUC across different numbers of retained features. The circle and square indicate the optimal and final selected subsets, respectively, which overlap at 17 features (AUC = 0.9649); the horizontal dashed line indicates the maximum mean AUC. b, c Lasso-Logistic regression coefficient path and 10-fold cross-validation AUC curve for in-hospital death prediction. d Stability-based feature-selection curve from repeated RFECV-XGBoost for 6-month unfavourable prognosis after injury. The light-grey curves represent individual RFECV runs, and the curve with circular markers represents the mean cross-validation AUC. The square indicates the stability-selected subset of 23 features; the horizontal dashed line indicates the mean RFECV AUC at the same feature count (0.9251). e, f Lasso-Logistic regression coefficient path and 10-fold cross-validation AUC curve for 6-month unfavourable prognosis after injury. In b, c, e, and f, the orange and green vertical dashed lines indicate Lambda Min and Lambda 1-SE, respectively; the upper axes in c and f indicate the numbers of non-zero coefficients, and error bars indicate standard errors. AUC: area under the curve; RFECV-XGBoost: XGBoost-based recursive feature elimination with cross-validation; Lasso-Logistic regression: logistic regression with an L1 regularisation penalty; Lambda Min: the λ value corresponding to the maximum mean cross-validation AUC; Lambda 1-SE: the more strongly regularised λ value with mean cross-validation

In-hospital death prediction models

The feature subsets selected using RFECV-XGBoost and Lasso-Logistic regression were separately combined with 9 machine-learning algorithms, resulting in 18 in-hospital death prediction models. Relative feature importance across the nine candidate models constructed using the RFECV-XGBoost and Lasso-Logistic regression feature subsets is shown in Supplementary Figs. S3 and S4, respectively. In the training set, AUROC ranged from 0.943 to 0.969 and AUPRC from 0.754 to 0.896 (Fig. 3, Supplementary Table S13). In the internal test set, AUROC ranged from 0.911 to 0.968 and AUPRC from 0.691 to 0.860 (Fig. 3, Supplementary Table S14).

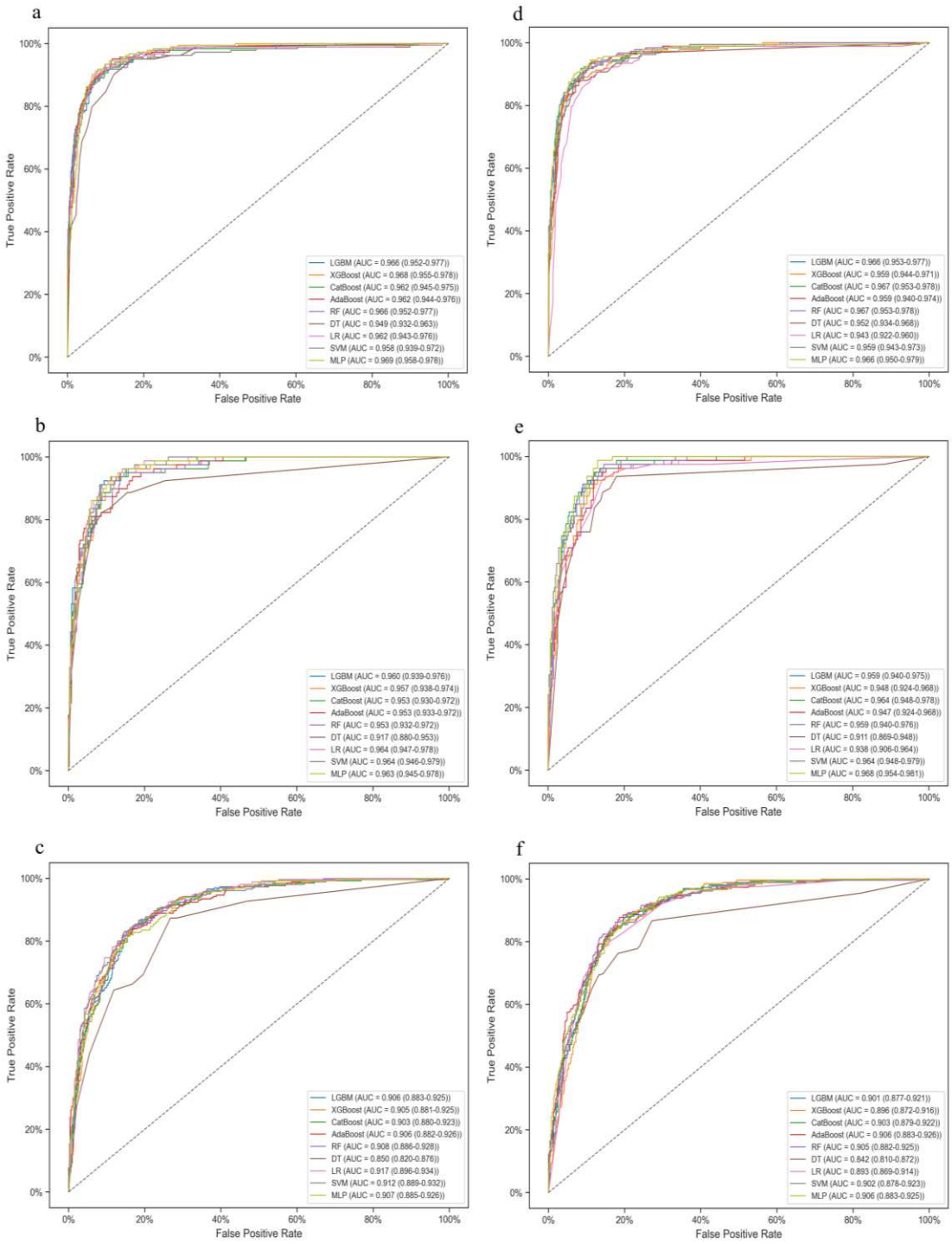


Fig. 3 ROC curves of the in-hospital death prediction models across different datasets. a–c Summary of ROC curves in the training set (a), internal test set (b), and external test set (c) for the nine models constructed using features selected by RFECV-XGBoost. d–f Summary of ROC curves in the training set (d), internal test set (e), and external test set (f) for the nine models constructed using features selected by Lasso-Logistic regression. The diagonal dashed line represents no discriminative ability. ROC: receiver operating characteristic; AUC: area under the curve; CI: confidence interval; LGBM: Light Gradient Boosting Machine; XGBoost: Extreme Gradient Boosting; CatBoost: Categorical Boosting; AdaBoost: Adaptive Boosting; RF: Random Forest; DT: Decision Tree; LR: Logistic Regression; SVM: Support Vector Machine; MLP: Multilayer Perceptron

In the internal test set, among models with relatively high sensitivity, the CatBoost model developed using features selected by Lasso-Logistic regression achieved the highest PPV (70.8%), with a sensitivity of 86.1%, an AUROC of 0.964 (95% CI, 0.947–0.979), an AUPRC of 0.849 (95% CI,

0.788–0.911), and an F1 score of 77.7% (Table 3). DCA showed positive net benefit across clinically relevant threshold probabilities, while calibration curves showed some deviation between predicted probabilities and observed event rates but overall acceptable agreement (Fig. 4, Supplementary Fig. S7). Taking these results together, this model was selected as the final in-hospital death prediction model.

In the external test set, the selected model maintained consistent cross-centre predictive performance, with an AUROC of 0.903 (95% CI, 0.883–0.922), AUPRC of 0.800 (95% CI, 0.760–0.844), sensitivity of 82.0%, PPV of 71.1%, and F1 score of 76.2% (Table 3, Supplementary Table S15). External decision curve and calibration analyses further supported the clinical utility and calibration performance of the selected model (Fig. 4, Supplementary Fig. S7). The final model was interpreted using SHapley Additive exPlanations (SHAP) analysis (Fig. 4) [17].

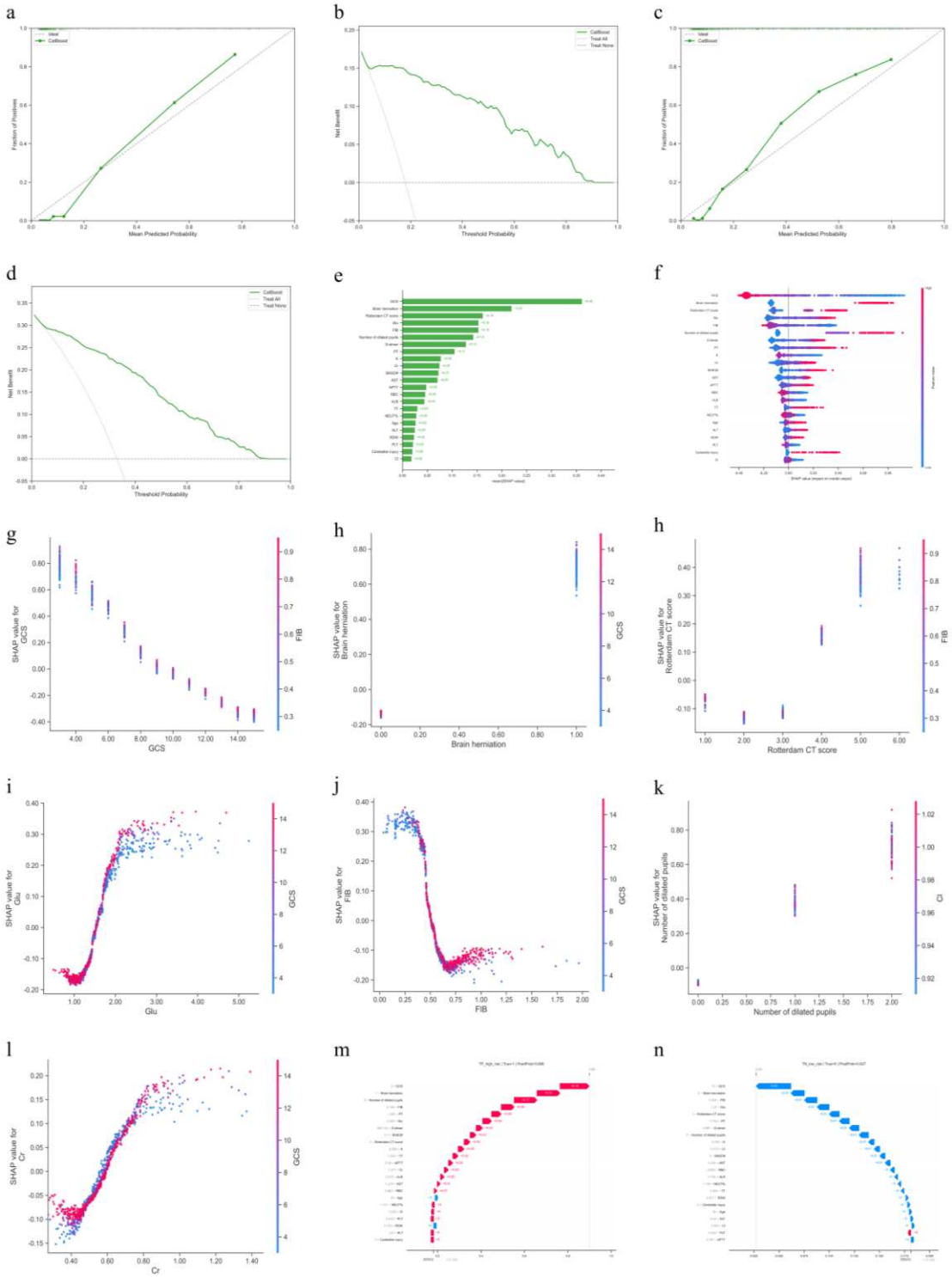


Fig. 4 Clinical net benefit, calibration performance, and SHAP interpretation of the final in-hospital death prediction model. The final model was CatBoost constructed using features selected by Lasso-Logistic regression. a Calibration curve in the internal test set. b Decision curve analysis in the internal test set. c Calibration curve in the external test set. d Decision curve analysis in the external test set. e SHAP feature-importance ranking based on mean absolute SHAP values. f SHAP summary dot plot. g–m SHAP dependence plots for GCS (g), brain herniation (h), Rotterdam CT score (i), Glu (j), FIB (k), number of dilated pupils (l), and Cr (m). In the Cr dependence plot, the horizontal axis was truncated at the 99th percentile; the complete plot is shown in Fig. S9. n, o Individual-level SHAP explanation plots. DCA: decision curve analysis; SHAP: SHapley Additive exPlanations; CatBoost: Categorical Boosting; Lasso-Logistic regression: logistic regression with an L1 regularisation penalty; GCS: Glasgow Coma Scale; Glu: blood glucose; FIB: fibrinogen; Cr: serum creatinine

Prediction models for 6-month unfavourable prognosis after injury

The feature subsets selected using RFECV-XGBoost and Lasso-Logistic regression were separately combined with 9 machine-learning algorithms, resulting in 18 prediction models for 6-month unfavourable prognosis after injury. Relative feature importance across the nine candidate models constructed using the RFECV-XGBoost and Lasso-Logistic regression feature subsets is shown in Supplementary Figs. S5 and S6, respectively. In the out-of-fold evaluation, AUROC ranged from 0.908 to 0.941 and AUPRC from 0.592 to 0.713 (Fig. 5, Supplementary Table S16). Among models with relatively high PPV values, the logistic regression model developed using features selected by RFECV-XGBoost achieved the highest sensitivity (80.3%), while maintaining a PPV of 62.2%, an AUROC of 0.937, an AUPRC of 0.690, and an F1 score of 70.1% (Table 4). DCA showed positive net benefit across a range of threshold probabilities, while the calibration curve showed some overprediction but overall acceptable agreement between predicted probabilities and observed event rates (Fig. 6, Supplementary Fig. S8). Taking these results together, this model was selected as the final prediction model for 6-month unfavourable prognosis after injury.

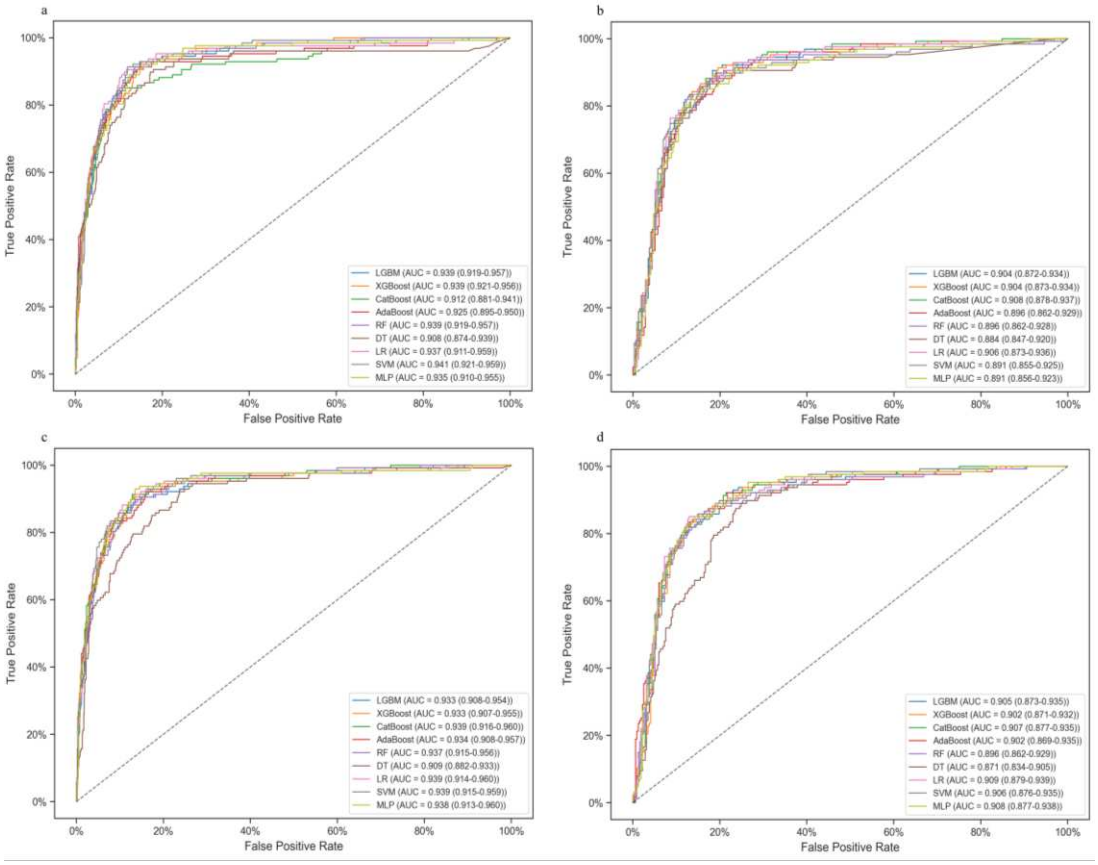


Fig. 5 ROC curves of the prediction models for 6-month unfavourable prognosis after injury across different datasets. a, b Summary of ROC curves based on OOF predictions in the training set (a) and predictions in the external test set (b) for the nine models constructed using features selected by RFECV-XGBoost. c, d Summary of ROC curves based on OOF predictions in the training set (c) and predictions in the external test set (d) for the nine models constructed using features selected by Lasso-Logistic regression. The diagonal dashed line represents no discriminative ability. ROC: receiver operating characteristic; AUC: area under the curve; CI: confidence interval; OOF: out-of-fold; LGBM: Light Gradient Boosting Machine; XGBoost: Extreme Gradient Boosting; CatBoost: Categorical Boosting; AdaBoost: Adaptive Boosting; RF: Random Forest; DT: Decision Tree; LR: Logistic Regression; SVM: Support Vector Machine; MLP: Multilayer Perceptron

In the external test set, the selected logistic regression model developed using features selected by RFECV-XGBoost maintained consistent cross-centre predictive performance, with an AUROC of

0.906 (95% CI, 0.873–0.936), an AUPRC of 0.727 (95% CI, 0.642–0.823), sensitivity of 81.1%, specificity of 87.7%, PPV of 72.5%, and an F1 score of 76.6% (Table 4, Supplementary Table S17). External DCA demonstrated positive net benefit across a range of threshold probabilities, and calibration analysis showed generally close agreement between predicted and observed risks, with some deviation at the extremes of the probability range. Model interpretation was performed using logistic regression coefficients and relative feature importance (Fig. 6, Supplementary Fig. S8).

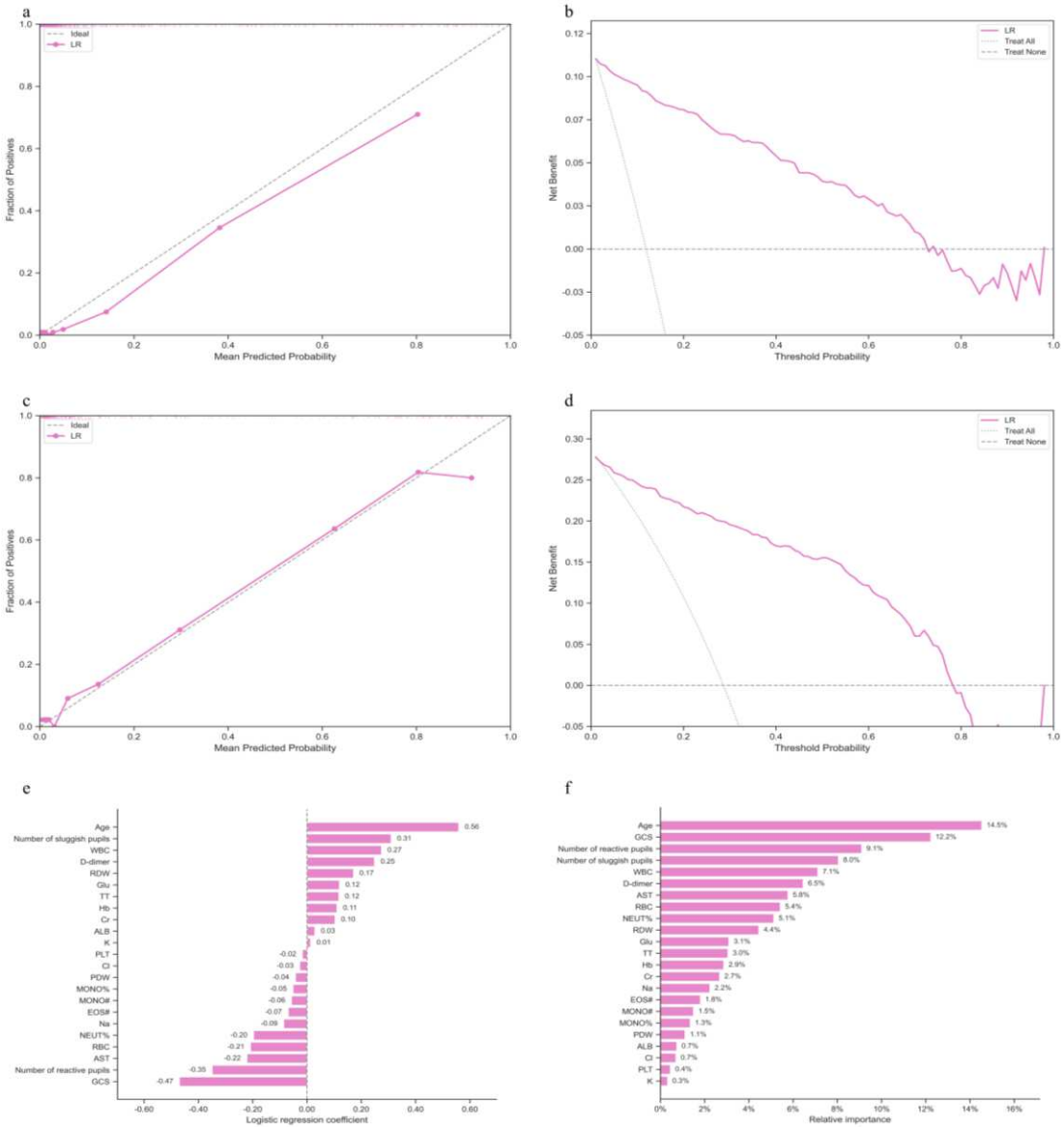


Fig. 6 Clinical net benefit, calibration performance, and interpretation of the final prediction model for 6-month unfavourable prognosis after injury. The final model was logistic regression constructed using features selected by RFECV-XGBoost. (a) Calibration curve based on OOF predictions in the training set. (b) Decision curve analysis based on OOF predictions in the training set. (c) Calibration curve in the external test set. (d) Decision curve analysis in the external test set. (e) Regression coefficients of the logistic regression model refitted on the complete development set using the selected feature set. (f) Corresponding relative feature importance based on the absolute regression coefficients normalised to sum to 100%. DCA: decision curve analysis; LR: Logistic Regression; RFECV-XGBoost: XGBoost-based recursive feature elimination with cross-validation; OOF: out-of-fold

Sensitivity analyses

Sensitivity analyses using raw laboratory values showed no consistent superiority of either representation across the two outcomes (Supplementary Tables S19 and S20; Supplementary Fig. S10).

Discussion

In this study, the selection of the final models was guided by the different clinical implications of the two outcomes rather than by the superiority of any single performance metric. For in-hospital death, models with relatively high sensitivity were first considered to support identification of patients at high risk, and greater emphasis was then placed on PPV to improve the reliability of positive predictions and reduce false-positive classifications [25]. For 6-month unfavourable prognosis after injury, models with relatively high PPV values were first considered, and greater emphasis was then placed on sensitivity to reduce missed identification of patients at risk of poor functional recovery [22] and support earlier rehabilitation planning [15] and follow-up. Therefore, although some candidate models performed better on individual metrics, the selected models were considered more appropriate for the intended clinical use of each outcome based on their overall performance and outcome-specific clinical priorities.

Compared with many previous TBI prognostic studies, the present study evaluated both outcomes and used an independent external test set to assess model performance [20]. In addition, model evaluation incorporated calibration and DCA alongside discrimination and classification metrics, providing a broader assessment of predictive reliability and potential clinical utility.

Notably, the model for 6-month unfavourable prognosis after injury showed some performance differences between OOF evaluation and external testing, particularly a higher PPV in the external test set. This difference may be mainly related to two factors. First, internal performance was estimated using repeated stratified cross-validation and aggregated out-of-fold predictions rather than direct resubstitution [3]. Second, the positive event rate was substantially higher in the external test set than in the development set (Table 2), and PPV is sensitive to outcome prevalence. Therefore, the higher positive event rate in the external test set likely increased the PPV. The model also showed favourable discrimination and sensitivity in the independent external test set, supporting its performance across centres.

This study has several limitations. First, although the models were evaluated in an independent external test set, differences in case composition, clinical workflows, and data recording practices across centres may still affect their generalisability. In addition, although comparisons before and after complete-case selection showed no substantial imbalance in event rates or variable distributions, residual selection bias cannot be completely excluded. Second, the models were developed using early admission data and did not incorporate serial neurological assessments, laboratory measurements, or follow-up imaging during hospitalisation; therefore, they primarily reflect risk during the early admission period. Third, the number of positive events for 6-month unfavourable prognosis after injury was relatively limited. Although cross-validation and regularisation were used to reduce the risk of overfitting, and cost-sensitive learning was used to address class imbalance, the limited number of events may still have affected estimate stability. Fourth, although laboratory measurements were standardised using the Value/ULN method to account for differences in laboratory reference ranges, sensitivity analyses using raw laboratory values did not demonstrate a clear performance advantage of the standardised representation. This may partly reflect the fact that all participating centres were located within the same geographical region, where between-centre heterogeneity in laboratory

reference ranges and testing practices may have been relatively limited. Therefore, the potential benefit of the Value/ULN approach in settings with greater inter-regional or inter-institutional variation remains to be evaluated. Larger prospective multicentre cohorts are needed to further assess model stability, calibration, generalisability, and clinical utility.

Overall, machine-learning models based on routinely available early admission data showed favourable predictive performance for both outcomes during internal evaluation and independent external validation, supporting their potential value for early risk stratification in patients with TBI.

References

1. Abdelhay O, Shatnawi A, Najadat H, Altamimi T (2025) Resampling methods for class imbalance in clinical prediction models: A scoping review protocol. PLoS ONE 20(11):e0330050. https://doi.org/10.1371/journal.pone.0330050
2. Chen T, Guestrin C (2016) XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining:785–794. https://doi.org/10.1145/2939672.2939785
3. Collins GS, Dhiman P, Ma J, Schlussel MM, Archer L, Van Calster B, Harrell FE, Jr., Martin GP, Moons KGM, van Smeden M, Sperrin M, Bullock GS, Riley RD (2024) Evaluation of clinical prediction models (part 1): from development to external validation. BMJ 384:e074819. https://doi.org/10.1136/bmj-2023-074819
4. Collins GS, Moons KGM, Dhiman P, Riley RD, Beam AL, Van Calster B, Ghassemi M, Liu X, Reitsma JB, van Smeden M, Boulesteix AL, Camaradou JC, Celi LA, Denaxas S, Denniston AK, Glocker B, Golub RM, Harvey H, Heinze G, Hoffman MM, Kengne AP, Lam E, Lee N, Loder EW, Maier-Hein L, Mateen BA, McCradden MD, Oakden-Rayner L, Ordish J, Parnell R, Rose S, Singh K, Wynants L, Logullo P (2024) TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ 385:e078378. https://doi.org/10.1136/bmj-2023-078378
5. Efthimiou O, Seo M, Chalkou K, Debray T, Egger M, Salanti G (2024) Developing clinical prediction models: a step-by-step guide. BMJ 386:e078276. https://doi.org/10.1136/bmj-2023-078276
6. Friedman JH, Hastie T, Tibshirani R (2010) Regularization Paths for Generalized Linear Models via Coordinate Descent. Journal of statistical software 33(1):1–22. https://doi.org/10.18637/jss.v033.i01
7. Gao G, Wu X, Feng J, Hui J, Mao Q, Lecky F, Lingsma H, Maas AIR, Jiang J (2020) Clinical characteristics and outcomes in patients with traumatic brain injury in China: a prospective, multicentre, longitudinal, observational study. Lancet Neurol 19(8):670–677. https://doi.org/10.1016/s1474-4422(20)30182-4
8. Guazzo A, Longato E, Fadini GP, Morieri ML, Sparacino G, Di Camillo B (2023) Deep-learning-based natural-language-processing models to identify cardiovascular disease hospitalisations of patients with diabetes from routine visits' text. Sci Rep 13(1):19132. https://doi.org/10.1038/s41598-023-45115-1
9. Guyon I, Weston J, Barnhill S, Vapnik V (2002) Gene Selection for Cancer Classification using Support Vector Machines. Machine Learning 46(1-3):389–422. https://doi.org/10.1023/A:1012487302797
10. Han R, Acosta JN, Shakeri Z, Ioannidis JPA, Topol EJ, Rajpurkar P (2024) Randomised controlled trials evaluating artificial intelligence in clinical practice: a scoping review. Lancet Digit Health 6(5):e367–e373. https://doi.org/10.1016/s2589-7500(24)00047-5
11. Huang SC, Pareek A, Seyyedi S, Banerjee I, Lungren MP (2020) Fusion of medical imaging and electronic health records using deep learning: a systematic review and implementation guidelines. NPJ Digit Med 3:136. https://doi.org/10.1038/s41746-020-00341-z

12. Jennett B, Bond M (1975) Assessment of outcome after severe brain damage. Lancet 1(7905):480–484. https://doi.org/10.1016/s0140-6736(75)92830-5
13. Jiang JY, Gao GY, Feng JF, Mao Q, Chen LG, Yang XF, Liu JF, Wang YH, Qiu BH, Huang XJ (2019) Traumatic brain injury in China. Lancet Neurol 18(3):286–295. https://doi.org/10.1016/s1474-4422(18)30469-1
14. Kline A, Wang H, Li Y, Dennis S, Hutch M, Xu Z, Wang F, Cheng F, Luo Y (2022) Multimodal machine learning in precision health: A scoping review. NPJ Digit Med 5(1):171. https://doi.org/10.1038/s41746-022-00712-8
15. Königs M, Beurskens EA, Snoep L, Scherder EJ, Oosterlaan J (2018) Effects of Timing and Intensity of Neurorehabilitation on Functional Outcome After Traumatic Brain Injury: A Systematic Review and Meta-Analysis. Arch Phys Med Rehabil 99(6):1149–1159.e1141. https://doi.org/10.1016/j.apmr.2018.01.013
16. Krones F, Marikkar U, Parsons G, Szmul A, Mahdi A (2025) Review of multimodal machine learning approaches in healthcare. Inf Fusion 114:102690. https://doi.org/10.1016/j.inffus.2024.102690
17. Lundberg S, Lee SI (2017) A Unified Approach to Interpreting Model Predictions. Advances in Neural Information Processing Systems 30:4765–4774.
18. Maas AI, Hukkelhoven CW, Marshall LF, Steyerberg EW (2005) Prediction of outcome in traumatic brain injury with computed tomographic characteristics: a comparison between the computed tomographic classification and combinations of computed tomographic predictors. Neurosurgery 57(6):1173–1182. https://doi.org/10.1227/01.neu.0000186013.63046.6b
19. Maas AIR, Menon DK, Manley GT, Abrams M, Åkerlund C, Andelic N, Aries M, Bashford T, Bell MJ, Bodien YG, Brett BL, Büki A, Chesnut RM, Citerio G, Clark D, Clasby B, Cooper DJ, Czeiter E, Czosnyka M, Dams-O'Connor K, De Keyser V, Diaz-Arrastia R, Ercole A, van Essen TA, Falvey É, Ferguson AR, Figaji A, Fitzgerald M, Foreman B, Gantner D, Gao G, Giacino J, Gravesteijn B, Guiza F, Gupta D, Gurnell M, Haagsma JA, Hammond FM, Hawryluk G, Hutchinson P, van der Jagt M, Jain S, Jain S, Jiang JY, Kent H, Kolias A, Kompanje EJO, Lecky F, Lingsma HF, Maegele M, Majdan M, Markowitz A, McCrea M, Meyfroidt G, Mikolić A, Mondello S, Mukherjee P, Nelson D, Nelson LD, Newcombe V, Okonkwo D, Orešič M, Peul W, Pisică D, Polinder S, Ponsford J, Puybasset L, Raj R, Robba C, Røe C, Rosand J, Schueler P, Sharp DJ, Smielewski P, Stein MB, von Steinbüchel N, Stewart W, Steyerberg EW, Stocchetti N, Temkin N, Tenovuo O, Theadom A, Thomas I, Espin AT, Turgeon AF, Unterberg A, Van Praag D, van Veen E, Verheyden J, Vyvere TV, Wang KKW, Wiegers EJA, Williams WH, Wilson L, Wisniewski SR, Younsi A, Yue JK, Yuh EL, Zeiler FA, Zeldovich M, Zemek R (2022) Traumatic brain injury: progress and challenges in prevention, clinical care, and research. Lancet Neurol 21(11):1004–1060. https://doi.org/10.1016/s1474-4422(22)00309-x
20. Malhotra AK, Shakil H, Smith CW, Huang YQ, Kwong JCC, Thorpe KE, Witiw CD, Kulkarni AV, Wilson JR, Nathens AB (2025) Predicting outcomes after moderate and severe traumatic brain injury using artificial intelligence: a systematic review. NPJ Digit Med 8(1):373. https://doi.org/10.1038/s41746-025-01714-y

21. Manley GT, Dams-O'Connor K, Alosco ML, Awwad HO, Bazarian JJ, Bragge P, Corrigan JD, Doperalski A, Ferguson AR, Mac Donald CL, Menon DK, McNett MM, van der Naalt J, Nelson LD, Pisică D, Silverberg ND, Umoh N, Wilson L, Yuh EL, Zetterberg H, Maas AIR, McCrea MA (2025) A new characterisation of acute traumatic brain injury: the NIH-NINDS TBI Classification and Nomenclature Initiative. Lancet Neurol 24(6):512–523. https://doi.org/10.1016/s1474-4422(25)00154-1
22. McCrea MA, Giacino JT, Barber J, Temkin NR, Nelson LD, Levin HS, Dikmen S, Stein M, Bodien YG, Boase K, Taylor SR, Vassar M, Mukherjee P, Robertson C, Diaz-Arrastia R, Okonkwo DO, Markowitz AJ, Manley GT, Adeoye O, Badjatia N, Bullock MR, Chesnut R, Corrigan JD, Crawford K, Duhaime AC, Ellenbogen R, Feeser VR, Ferguson AR, Foreman B, Gardner R, Gaudette E, Goldman D, Gonzalez L, Gopinath S, Gullapalli R, Hemphill JC, Hotz G, Jain S, Keene CD, Korley FK, Kramer J, Kreitzer N, Lindsell C, Machamer J, Madden C, Martin A, McAllister T, Merchant R, Ngwenya LB, Noel F, Nolan A, Palacios E, Perl D, Puccio A, Rabinowitz M, Rosand J, Sander A, Satris G, Schnyer D, Seabury S, Sherer M, Toga A, Valadka A, Wang K, Yue JK, Yuh E, Zafonte R (2021) Functional Outcomes Over the First Year After Moderate to Severe Traumatic Brain Injury in the Prospective, Longitudinal TRACK-TBI Study. JAMA Neurol 78(8):982–992. https://doi.org/10.1001/jamaneurol.2021.2043
23. Merz M, Lee KR, Kullak-Ublick GA, Brueckner A, Watkins PB (2014) Methodology to assess clinical liver safety data. Drug Saf 37 Suppl 1(Suppl 1):S33–45. https://doi.org/10.1007/s40264-014-0184-5
24. Pan X, Cheng J, Hou F, Lan R, Lu C, Li L, Feng Z, Wang H, Liang C, Liu Z, Chen X, Han C, Liu Z (2023) SMILE: Cost-sensitive multi-task learning for nuclear segmentation and classification with imbalanced annotations. Med Image Anal 88:102867. https://doi.org/10.1016/j.media.2023.102867
25. Raj R, Wennervirta JM, Tjerkaski J, Luoto TM, Posti JP, Nelson DW, Takala R, Bendel S, Thelin EP, Luostarinen T, Korja M (2022) Dynamic prediction of mortality after traumatic brain injury using a machine learning algorithm. NPJ Digit Med 5(1):96. https://doi.org/10.1038/s41746-022-00652-3
26. Raj SS, Fei T, Fried S, Ip A, Fein JA, Leslie LA, Alarcon Tomas A, Leithner D, Peled JU, Corona M, Dahi PB, Danylesko I, Epstein-Peterson Z, Funnell T, Giralt SA, Jacoby E, Kedmi M, Landego I, Lin RJ, Parascondola A, Pascual L, Orozco N, Park JH, Palomba ML, Salles G, Saldia A, Schöder H, Sdayoor I, Shah GL, Scordo M, Shem-Tov N, Shimoni A, Slingerland J, Yerushalmi R, Nagler A, Greenbaum BD, Vickers AJ, Suh HC, Avigdor A, Perales MA, van den Brink MRM, Shouval R (2025) An inflammatory biomarker signature of response to CAR-T cell therapy in non-Hodgkin lymphoma. Nat Med 31(4):1183–1194. https://doi.org/10.1038/s41591-025-03532-x
27. Rajpurkar P, Chen E, Banerjee O, Topol EJ (2022) AI in health and medicine. Nat Med 28(1):31–38. https://doi.org/10.1038/s41591-021-01614-0
28. Riley RD, Archer L, Snell KIE, Ensor J, Dhiman P, Martin GP, Bonnett LJ, Collins GS (2024) Evaluation of clinical prediction models (part 2): how to undertake an

external validation study. BMJ 384:e074820. https://doi.org/10.1136/bmj-2023-074820
29. Schellenberg M, Arase M, Wong MD, Demetriades D (2024) Mild traumatic brain injury: not always a mild injury. Eur J Trauma Emerg Surg 50(4):1229–1235. https://doi.org/10.1007/s00068-023-02365-y
30. Shahriari B, Swersky K, Wang Z, Adams RP, Freitas ND (2016) Taking the Human Out of the Loop: A Review of Bayesian Optimization. Proceedings of the IEEE 104(1):148–175. https://doi.org/10.1109/JPROC.2015.2494218
31. Steyerberg EW, Mushkudiani N, Perel P, Butcher I, Lu J, McHugh GS, Murray GD, Marmarou A, Roberts I, Habbema JD, Maas AI (2008) Predicting outcome after traumatic brain injury: development and international validation of prognostic scores based on admission characteristics. PLoS Med 5(8):e165. https://doi.org/10.1371/journal.pmed.0050165
32. Tian F, Zhou Y, Wang J, Feng P, Wang X, Ma M, Shangguan C, Zhu H (2026) Diagnosis and treatment of hepatic hilar lymph node metastasis in hepatic alveolar echinococcosis patients: a real-world single-center experience. Front Oncol 15:1708936. https://doi.org/10.3389/fonc.2025.1708936
33. Tibshirani R (2011) Regression shrinkage and selection via the lasso: a retrospective. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 73(3):273–282. https://doi.org/10.1111/j.1467-9868.2011.00771.x
34. Van Calster B, McLernon DJ, van Smeden M, Wynants L, Steyerberg EW (2019) Calibration: the Achilles heel of predictive analytics. BMC Med 17(1):230. https://doi.org/10.1186/s12916-019-1466-7
35. Vickers AJ, Elkin EB (2006) Decision curve analysis: a novel method for evaluating prediction models. Med Decis Making 26(6):565–574. https://doi.org/10.1177/0272989x06295361
36. White IR, Carlin JB (2010) Bias and efficiency of multiple imputation compared with complete-case analysis for missing covariate values. Stat Med 29(28):2920–2931. https://doi.org/10.1002/sim.3944
37. Zhang Z, Kim HJ, Lonjon G, Zhu Y (2019) Balance diagnostics after propensity score matching. Ann Transl Med 7(1):16. https://doi.org/10.21037/atm.2018.12.10

Statements and Declarations

Competing interests

All authors declare no competing interests.

Ethics approval

This retrospective multicentre study was conducted in accordance with the Declaration of Helsinki. The study protocol was approved by the Ethics Committee of the Affiliated Hospital of Nantong University (approval number: 2026-K041-01).

Consent to participate

Because of the retrospective nature of the study, the requirement for written informed consent was waived by the Ethics Committee of the Affiliated Hospital of Nantong University.

Funding

This study was supported by the Nantong Key Laboratory of Neurorestoration Medical-Engineering Transformation and the Postgraduate Research & Practice Innovation Program of Jiangsu Province (grant nos. KYCX25-3814 and KYCX25-3792). KYCX25-3814 was awarded to Peizhe Jiang; KYCX25-3792 was awarded to an individual who is not an author of this manuscript. The Nantong Key Laboratory provided institutional support without an individual author recipient or grant number. The funders had no role in study design, data collection, data analysis, data interpretation, preparation of the manuscript, or the decision to submit the manuscript for publication.

Data availability

The de-identified individual participant data and research code used in this study may be made available upon reasonable request after publication. Applicants will be required to obtain approval from the investigators and the relevant ethics committee, and to sign a data access agreement or other required documents before access is granted. The study protocol, statistical analysis plan, and relevant supplementary materials may also be provided upon reasonable request. Requests for further information should be directed to the corresponding author(ntfyliuqianqian@163.com) or the first author(2431310127@stmail.ntu.edu.cn).

Consent to publish

Not applicable.

Clinical trial number

Not applicable.

Table 1 Baseline characteristics of patients included in the in-hospital death prediction analysis, by dataset.

Variable	Training (N=1027)	Internal Test (N=441)	External Test (N=792)	p value
GCS, mean (SD)	11.0 (4.3)	10.9 (4.4)	8.7 (4.5)	<0.001
Age, Median (IQR)	64.0 (55.0, 71.0)	63.0 (56.0, 71.0)	66.0 (58.0, 72.0)	0.002
WBC, Median (IQR)	1.24 (0.91, 1.66)	1.24 (0.93, 1.63)	1.41 (1.04, 1.81)	<0.001
NEUT%, Median (IQR)	1.17 (1.08, 1.22)	1.17 (1.09, 1.22)	1.15 (1.02, 1.23)	0.074
MONO%, Median (IQR)	0.42 (0.33, 0.51)	0.42 (0.34, 0.52)	0.56 (0.42, 0.72)	<0.001
NEUT#, Median (IQR)	1.50 (1.05, 2.06)	1.52 (1.03, 2.05)	1.60 (1.11, 2.17)	0.107
MONO#, Median (IQR)	0.56 (0.40, 0.80)	0.56 (0.41, 0.75)	0.85 (0.62, 1.19)	<0.001
EOS#, Median (IQR)	0.06 (0.02, 0.15)	0.06 (0.02, 0.12)	0.08 (0.02, 0.20)	<0.001
BASO#, Median (IQR)	0.01 (0.00, 0.04)	0.01 (0.00, 0.10)	0.30 (0.17, 0.50)	<0.001
RBC, Median (IQR)	0.81 (0.74, 0.88)	0.81 (0.74, 0.89)	0.77 (0.70, 0.84)	<0.001
RDW-SD, Median (IQR)	0.81 (0.77, 0.85)	0.81 (0.77, 0.85)	0.82 (0.78, 0.87)	<0.001
PDW, Median (IQR)	1.01 (0.92, 1.04)	1.01 (0.91, 1.03)	0.67 (0.59, 0.84)	<0.001
PLT, Median (IQR)	0.57 (0.47, 0.70)	0.60 (0.49, 0.74)	0.61 (0.48, 0.73)	0.012
Hb, Median (IQR)	0.85 (0.77, 0.92)	0.85 (0.78, 0.92)	0.79 (0.72, 0.86)	<0.001
AST, Median (IQR)	0.81 (0.59, 1.19)	0.83 (0.61, 1.15)	0.78 (0.59, 1.12)	0.368
ALT, Median (IQR)	0.56 (0.40, 0.82)	0.58 (0.42, 0.83)	0.57 (0.42, 0.80)	0.629
ALB, Median (IQR)	0.76 (0.71, 0.81)	0.77 (0.71, 0.82)	0.80 (0.73, 0.87)	<0.001
Cr, Median (IQR)	0.55 (0.47, 0.65)	0.55 (0.47, 0.66)	0.58 (0.48, 0.69)	0.001
K, Median (IQR)	0.73 (0.66, 0.78)	0.72 (0.65, 0.78)	0.69 (0.61, 0.75)	<0.001
Na, Median (IQR)	0.95 (0.94, 0.97)	0.95 (0.94, 0.97)	0.96 (0.94, 0.97)	0.275
Cl, Median (IQR)	0.97 (0.94, 0.99)	0.97 (0.94, 0.99)	0.95 (0.93, 0.98)	<0.001
Glu, Median (IQR)	1.39 (1.14, 1.79)	1.37 (1.12, 1.75)	1.49 (1.20, 1.88)	0.001
INR, Median (IQR)	0.69 (0.65, 0.74)	0.70 (0.66, 0.75)	0.72 (0.66, 0.83)	<0.001
APTT, Median (IQR)	0.74 (0.68, 0.80)	0.74 (0.67, 0.81)	0.70 (0.63, 0.79)	<0.001
FIB, Median (IQR)	0.58 (0.46, 0.70)	0.58 (0.46, 0.71)	0.53 (0.40, 0.65)	<0.001
PT, Median (IQR)	0.86 (0.82, 0.94)	0.88 (0.84, 0.95)	0.88 (0.84, 0.95)	0.001
TT, Median (IQR)	0.83 (0.80, 0.87)	0.83 (0.79, 0.87)	0.87 (0.81, 0.95)	<0.001
D-dimer, Median (IQR)	20.18 (7.52, 60.14)	22.77 (7.38, 64.00)	30.88 (10.52, 78.32)	0.002
Number of reactive pupils				<0.001
0, n (%)	167 (16.3%)	73 (16.6%)	262 (33.1%)	
1, n (%)	28 (2.7%)	16 (3.6%)	40 (5.1%)	
2, n (%)	832 (81.0%)	352 (79.8%)	490 (61.9%)	
Number of sluggish pupils				0.004
0, n (%)	854 (83.2%)	362 (82.1%)	618 (78.0%)	
1, n (%)	118 (11.5%)	52 (11.8%)	78 (9.8%)	
2, n (%)	55 (5.4%)	27 (6.1%)	96 (12.1%)	
Number of dilated pupils				<0.001
0, n (%)	906 (88.2%)	389 (88.2%)	583 (73.6%)	
1, n (%)	55 (5.4%)	23 (5.2%)	96 (12.1%)	
2, n (%)	66 (6.4%)	29 (6.6%)	113 (14.3%)	
Number of constricted pupils				0.613
0, n (%)	1007 (98.1%)	431 (97.7%)	780 (98.5%)	
1, n (%)	7 (0.7%)	5 (1.1%)	8 (1.0%)	
2, n (%)	13 (1.3%)	5 (1.1%)	4 (0.5%)	
Rotterdam CT score				0.002

1, n (%)	32 (3.1%)	14 (3.2%)	11 (1.4%)	
2, n (%)	257 (25.0%)	120 (27.2%)	192 (24.2%)	
3, n (%)	443 (43.1%)	191 (43.3%)	310 (39.1%)	
4, n (%)	144 (14.0%)	59 (13.4%)	129 (16.3%)	
5, n (%)	138 (13.4%)	50 (11.3%)	135 (17.0%)	
6, n (%)	13 (1.3%)	7 (1.6%)	15 (1.9%)	
Anisocoria				0.003
NO, n (%)	843 (82.1%)	361 (81.9%)	602 (76.0%)	
YES, n (%)	184 (17.9%)	80 (18.1%)	190 (24.0%)	
Cerebral contusion and laceration				<0.001
NO, n (%)	297 (28.9%)	149 (33.8%)	318 (40.2%)	
YES, n (%)	730 (71.1%)	292 (66.2%)	474 (59.8%)	
PBSI				<0.001
NO, n (%)	986 (96.0%)	427 (96.8%)	692 (87.4%)	
YES, n (%)	41 (4.0%)	14 (3.2%)	100 (12.6%)	
Cerebellar injury				0.011
NO, n (%)	988 (96.2%)	413 (93.7%)	769 (97.1%)	
YES, n (%)	39 (3.8%)	28 (6.3%)	23 (2.9%)	
Brain herniation				<0.001
NO, n (%)	879 (85.6%)	362 (82.1%)	600 (75.8%)	
YES, n (%)	148 (14.4%)	79 (17.9%)	192 (24.2%)	
Death				<0.001
NO, n (%)	844 (82.2%)	362 (82.1%)	531 (67.0%)	
YES, n (%)	183 (17.8%)	79 (17.9%)	261 (33.0%)	

Data are mean (SD), median (IQR), or n (%). Laboratory variables are presented as Value/ULN ratios. P values compare the training, internal test, and external test sets and were interpreted descriptively; they were not used as the sole basis for determining clinical significance. Death: in-hospital death; GCS: Glasgow Coma Scale; SD: standard deviation; IQR: interquartile range; ULN: upper limit of normal; WBC: white blood cell count; NEUT%: neutrophil percentage; MONO%: monocyte percentage; NEUT#: neutrophil count; MONO#: monocyte count; EOS#: eosinophil count; BASO#: basophil count; RBC: red blood cell count; RDW-SD: red cell distribution width-standard deviation; PDW: platelet distribution width; PLT: platelet count; Hb: haemoglobin; AST: aspartate aminotransferase; ALT: alanine aminotransferase; ALB: albumin; Cr: serum creatinine; K: serum potassium; Na: serum sodium; Cl: serum chloride; Glu: blood glucose; INR: international normalised ratio; APTT: activated partial thromboplastin time; FIB: fibrinogen; PT: prothrombin time; TT: thrombin time; PBSI: primary brainstem injury.

Table 2 Baseline characteristics of patients included in the prediction analysis for 6-month unfavourable prognosis after injury, by dataset.

Variable	Training (N=1068)	External Test (N=444)	p value
GCS, mean (SD)	12.3 (3.4)	10.6 (4.0)	<0.001
Age, Median (IQR)	63.0 (55.0, 70.0)	64.0 (56.0, 71.0)	0.134
WBC, Median (IQR)	1.19 (0.89, 1.56)	1.34 (1.00, 1.73)	<0.001
NEUT%, Median (IQR)	1.18 (1.09, 1.22)	1.19 (1.07, 1.25)	0.058
MONO%, Median (IQR)	0.42 (0.34, 0.52)	0.61 (0.46, 0.76)	<0.001
NEUT#, Median (IQR)	1.45 (1.01, 1.96)	1.53 (1.10, 2.12)	0.013
MONO#, Median (IQR)	0.55 (0.40, 0.76)	0.85 (0.62, 1.14)	<0.001
EOS#, Median (IQR)	0.04 (0.02, 0.12)	0.06 (0.02, 0.18)	0.007
BASO#, Median (IQR)	0.01 (0.00, 0.03)	0.30 (0.10, 0.50)	<0.001
RBC, Median (IQR)	0.82 (0.76, 0.89)	0.77 (0.71, 0.84)	<0.001
RDW-SD, Median (IQR)	0.81 (0.77, 0.85)	0.82 (0.78, 0.87)	<0.001
PDW, Median (IQR)	1.01 (0.93, 1.03)	0.69 (0.59, 0.83)	<0.001
PLT, Median (IQR)	0.58 (0.48, 0.72)	0.62 (0.51, 0.76)	<0.001
Hb, Median (IQR)	0.86 (0.79, 0.93)	0.79 (0.73, 0.85)	<0.001
AST, Median (IQR)	0.75 (0.58, 1.06)	0.75 (0.56, 1.07)	0.906
ALT, Median (IQR)	0.57 (0.40, 0.80)	0.58 (0.42, 0.80)	0.207
ALB, Median (IQR)	0.77 (0.72, 0.82)	0.80 (0.74, 0.87)	<0.001
Cr, Median (IQR)	0.54 (0.45, 0.64)	0.56 (0.48, 0.67)	0.001
K, Median (IQR)	0.73 (0.67, 0.78)	0.69 (0.63, 0.76)	<0.001
Na, Median (IQR)	0.96 (0.94, 0.97)	0.96 (0.94, 0.97)	0.727
Cl, Median (IQR)	0.97 (0.94, 0.99)	0.95 (0.93, 0.98)	<0.001
Glu, Median (IQR)	1.31 (1.10, 1.61)	1.36 (1.08, 1.71)	0.213
INR, Median (IQR)	0.69 (0.65, 0.72)	0.68 (0.64, 0.76)	0.494
APTT, Median (IQR)	0.74 (0.67, 0.79)	0.67 (0.62, 0.73)	<0.001
FIB, Median (IQR)	0.61 (0.52, 0.72)	0.56 (0.47, 0.68)	<0.001
PT, Median (IQR)	0.86 (0.82, 0.91)	0.87 (0.83, 0.92)	0.002
TT, Median (IQR)	0.83 (0.79, 0.86)	0.86 (0.80, 0.92)	<0.001
D-dimer, Median (IQR)	17.36 (6.65, 41.37)	19.41 (7.73, 47.42)	0.221
Number of reactive pupils			<0.001
0, n (%)	58 (5.4%)	63 (14.2%)	
1, n (%)	31 (2.9%)	22 (5.0%)	
2, n (%)	979 (91.7%)	359 (80.9%)	
Number of sluggish pupils			<0.001
0, n (%)	894 (83.7%)	331 (74.5%)	
1, n (%)	120 (11.2%)	54 (12.2%)	
2, n (%)	54 (5.1%)	59 (13.3%)	
Number of dilated pupils			<0.001
0, n (%)	1025 (96.0%)	388 (87.4%)	
1, n (%)	37 (3.5%)	41 (9.2%)	
2, n (%)	6 (0.6%)	15 (3.4%)	
Number of constricted pupils			0.418
0, n (%)	1050 (98.3%)	439 (98.9%)	
1, n (%)	6 (0.6%)	2 (0.5%)	
2, n (%)	12 (1.1%)	3 (0.7%)	
Rotterdam CT score			0.147

1, n (%)	41 (3.8%)	10 (2.3%)	
2, n (%)	306 (28.7%)	136 (30.6%)	
3, n (%)	533 (49.9%)	196 (44.1%)	
4, n (%)	117 (11.0%)	50 (11.3%)	
5, n (%)	70 (6.6%)	50 (11.3%)	
6, n (%)	1 (0.1%)	2 (0.5%)	
Anisocoria			0.020
NO, n (%)	910 (85.2%)	356 (80.2%)	
YES, n (%)	158 (14.8%)	88 (19.8%)	
Cerebral contusion and laceration			<0.001
NO, n (%)	319 (29.9%)	184 (41.4%)	
YES, n (%)	749 (70.1%)	260 (58.6%)	
PBSI			<0.001
NO, n (%)	1052 (98.5%)	410 (92.3%)	
YES, n (%)	16 (1.5%)	34 (7.7%)	
Cerebellar injury			0.607
NO, n (%)	1032 (96.6%)	432 (97.3%)	
YES, n (%)	36 (3.4%)	12 (2.7%)	
Brain herniation			<0.001
NO, n (%)	1011 (94.7%)	379 (85.4%)	
YES, n (%)	57 (5.3%)	65 (14.6%)	
Unfavourable			<0.001
NO, n (%)	941 (88.1%)	317 (71.4%)	
YES, n (%)	127 (11.9%)	127 (28.6%)	

Data are mean (SD), median (IQR), or n (%). Laboratory variables are presented as Value/ULN ratios. P values compare the training and external test sets and were interpreted descriptively; they were not used as the sole basis for determining clinical significance. Unfavourable: 6-month unfavourable prognosis after injury; GCS: Glasgow Coma Scale; SD: standard deviation; IQR: interquartile range; ULN: upper limit of normal; WBC: white blood cell count; NEUT%: neutrophil percentage; MONO%: monocyte percentage; NEUT#: neutrophil count; MONO#: monocyte count; EOS#: eosinophil count; BASO#: basophil count; RBC: red blood cell count; RDW-SD: red cell distribution width-standard deviation; PDW: platelet distribution width; PLT: platelet count; Hb: haemoglobin; AST: aspartate aminotransferase; ALT: alanine aminotransferase; ALB: albumin; Cr: serum creatinine; K: serum potassium; Na: serum sodium; Cl: serum chloride; Glu: blood glucose; INR: international normalised ratio; APTT: activated partial thromboplastin time; FIB: fibrinogen; PT: prothrombin time; TT: thrombin time; PBSI: primary brainstem injury; CT: computed tomography.

Table 3 Performance of the final in-hospital death prediction model across datasets.

Metric	Training set	Internal test set	External test set
AUROC (95% CI)	0.967 (0.953–0.978)	0.964 (0.947–0.979)	0.903 (0.883–0.922)
AUPRC (95% CI)	0.896 (0.860–0.926)	0.849 (0.788–0.911)	0.800 (0.760–0.844)
Accuracy (95% CI)	92.8% (91.3%–94.3%)	91.2% (88.4%–93.7%)	83.1% (80.4%–85.7%)
Sensitivity (95% CI)	84.7% (79.2%–89.6%)	86.1% (78.5%–93.7%)	82.0% (77.4%–86.2%)
Specificity (95% CI)	94.5% (93.0%–96.0%)	92.3% (89.5%–95.0%)	83.6% (80.4%–86.6%)
PPV (95% CI)	77.1% (72.2%–82.3%)	70.8% (63.8%–78.9%)	71.1% (67.2%–75.4%)
NPV (95% CI)	96.6% (95.5%–97.7%)	96.8% (95.1%–98.5%)	90.4% (88.2%–92.5%)
F1 Score (95% CI)	80.7% (76.7%–84.7%)	77.7% (71.5%–83.6%)	76.2% (72.7%–79.5%)
Brier Score (95% CI)	N/A	0.061 (0.051–0.072)	0.124 (0.113–0.134)

Performance estimates are presented with 95% CIs; accuracy, sensitivity, specificity, PPV, NPV, and F1 score are expressed as percentages. Training-set performance was estimated from cross-validated out-of-fold predictions. AUROC: area under the receiver operating characteristic curve; AUPRC: area under the precision–recall curve; PPV: positive predictive value; NPV: negative predictive value; CI: confidence interval

Table 4 Performance of the final prediction model for 6-month unfavourable prognosis after injury across datasets.

Metric	**Training set (OOF)**	**External test set**
AUROC (95% CI)	0.937 (0.911–0.959)	0.906 (0.873–0.936)
AUPRC (95% CI)	0.690 (0.597–0.773)	0.727 (0.642–0.823)
Accuracy (95% CI)	91.9% (90.3%–93.4%)	85.8% (82.4%–89.2%)
Sensitivity (95% CI)	80.3% (73.0%–87.2%)	81.1% (73.7%–87.9%)
Specificity (95% CI)	93.4% (91.8%–94.9%)	87.7% (83.8%–91.3%)
PPV (95% CI)	62.2% (54.5%–69.6%)	72.5% (64.6%–79.7%)
NPV (95% CI)	97.2% (96.1%–98.3%)	92.1% (88.7%–95.0%)
F1 Score (95% CI)	70.1% (63.8%–75.8%)	76.6% (70.3%–82.2%)
Brier Score (95% CI)	N/A	0.105 (0.084–0.126)

Performance estimates are presented with 95% CIs; accuracy, sensitivity, specificity, PPV, NPV, and F1 score are expressed as percentages. Training-set performance was estimated from OOF predictions generated using repeated stratified cross-validation. OOF: out-of-fold; AUROC: area under the receiver operating characteristic curve; AUPRC: area under the precision–recall curve; PPV: positive predictive value; NPV: negative predictive value; CI: confidence interval

Supplementary Files

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

- Supplementarymaterial.pdf