Development of a machine learning-based predictive model for postoperative delirium in elderly intensive care unit patients: Retrospective Study

Houfeng Li, Qinglai Zang, Qi Li, Yanchen Lin, Jintao Duan, Jing Huang, Huixiu Hu, Ying Zhang, Dengyun Xia, Miao Zhou

Submitted to: Journal of Medical Internet Research
on: October 08, 2024

Table of Contents

Development of a machine learning-based predictive model for postoperative delirium in elderly intensive care unit patients: Retrospective Study

Houfeng Li[1*] BMed; Qinglai Zang[2, 3*] MSc; Qi Li[2, 4*] MMed; Yanchen Lin[1] MMed; Jintao Duan[5] BSc; Jing Huang[6] BMed; Huixiu Hu[1] BMed; Ying Zhang[1] BMed; Dengyun Xia[7*] MMed; Miao Zhou[8*] MD

[1]Graduate School Hebei North University Zhangjiakou CN
[2]School of Anesthesiology Naval Medical University Shanghai CN
[3]Information Center The Second Affiliated Hospital of Naval Medical University Shanghai CN
[4]Department of Anesthesiology Shanghai Ninth People's Hospital Shanghai Jiao Tong University School of Medicine Shanghai CN
[5]School of Health Science and Engineering University of Shanghai for Science and Technology Shanghai CN
[6]Graduate School Wannan Medical College Wuhu CN
[7]Department of Anesthesiology The First Affiliated Hospital of Hebei North University Zhangjiakou CN
[8]Department of Anesthesiology The Affiliated Cancer Hospital of Nanjing Medical University, Jiangsu Cancer Hospital, Jiangsu Institute of Cancer Research Nanjing Medical University Nanjing CN
[*]these authors contributed equally

Corresponding Author:
Miao Zhou MD
Department of Anesthesiology
The Affiliated Cancer Hospital of Nanjing Medical University, Jiangsu Cancer Hospital, Jiangsu Institute of Cancer Research
Nanjing Medical University
No. 42, Baiziting Community, Xuanwu District
Nanjing
CN

Abstract

Background: The occurrence of delirium is a prevalent phenomenon among patients admitted to the geriatric intensive care unit (ICU), with the potential to adversely impact prognosis and augment the risk of complications.

Objective: This study aimed to construct and validate a predictive model for postoperative delirium in elderly patients in ICUs, providing timely and effective early identification of high-risk individuals and assisting clinicians in decision-making.

Methods: The data from patients admitted to the ICU for over 24 hours were extracted from the Medical Information Marketplace for Intensive Care IV (MIMIC-IV) database and the eICU Collaborative Research Database (eICU-CRD). The MIMIC-IV data were split into a training set and an internal validation set (7:3 ratio), while the eICU-CRD data served as an external validation set. A delirium prediction was conducted for the subsequent prediction windows (12h, 24h, 48h, and whole stay time) utilising data from the first 24 hours post-admission. The corresponding feature variables were subjected to Boruta feature selection, and the prediction models were constructed using logistic regression, support vector classifier, random forest classifier, and extreme gradient boosting (XGB). Subsequently, the model performance was evaluated using receiver operating characteristic curves, calibration curves, decision curve analysis, and external validation.

Results: The MIMIC-IV and eICU-CRD datasets comprised 5897 and 618 patients, respectively, who were included in the analysis. A total of 57 features were selected for the construction of the predictive model. In the context of internal validation, the XGB model demonstrated the most effective prediction of delirium across different prediction windows. The Area under the curve values for the four prediction windows (12h, 24h, 48h, and whole stay time) were 0.860(95% CI: 0.839-0.880), 0.871(95% CI: 0.850-0.889), 0.851(95% CI: 0.829-0.871), and 0.846(95% CI: 0.827-0.867), respectively. The Area under the curve values for the external validation set were 0.828(95% CI: 0.768-0.880), 0.811(95% CI: 0.762-0.855), 0.756(95% CI: 0.705-0.803), and 0.750(95% CI: 0.701-0.795). Furthermore, the XGB model demonstrated the most accurate calibration across all prediction windows, with values of 0.115, 0.119, 0.136, and 0.144, respectively. Additionally, the decision curve analysis revealed that the XGB model outperformed the other models in terms of net gain for the majority of threshold probability values. The five most

significant predictive features identified were the first day's delirium assessment results, invasive ventilation, Sequential Organ Failure Assessment score, minimum Glasgow Coma Scale score, and type of first care unit.

Conclusions: The high-performance XGB model for predicting postoperative delirium in elderly ICU patients has been successfully developed and validated. The model predicts the incidence of delirium at 12h, 24h, 48h, and whole stay time after the first day of hospitalisation within ICU. This enables physicians to identify high-risk patients early, thus facilitating the optimisation of personalised management strategies and care plans.

(JMIR Preprints 08/10/2024:67258)
DOI: https://doi.org/10.2196/preprints.67258

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✓ **Please make my preprint PDF available to anyone at any time (recommended).**
Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.
Only make the preprint title and abstract visible.
No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**
Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v
Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http

Original Manuscript

Development of a machine learning-based predictive model for postoperative delirium in elderly intensive care unit patients: Retrospective Study

Houfeng Li[1*], BMed; Qinglai Zang[2,3*], MSc; Qi Li[2,4*], MMed; Yanchen Lin[1], MMed; Jintao Duan[5], BSc; Jing Huang[6], BMed; Huixiu Hu[1], BMed; Ying Zhang[1], BMed; Dengyun Xia[7*], MMed; Miao Zhou[8*], MD

[1]Graduate School, Hebei North University, Zhangjiakou, Hebei, China.
[2]School of Anesthesiology, Naval Medical University, Shanghai, China.
[3]Information Center, The Second Affiliated Hospital of Naval Medical University, Shanghai, China.
[4]Department of Anesthesiology, Shanghai Ninth People's Hospital, Shanghai Jiao Tong University School of Medicine, Shanghai, China.
[5]School of Health Science and Engineering, University of Shanghai for Science and Technology, Shanghai, China.
[6]Graduate School, Wannan Medical College, Wuhu, Anhui, China.
[7]Department of Anesthesiology, The First Affiliated Hospital of Hebei North University, Zhangjiakou, Hebei, China.
[8]Department of Anesthesiology, The Affiliated Cancer Hospital of Nanjing Medical University, Jiangsu Cancer Hospital, Jiangsu Institute of Cancer Research, Nanjing Medical University, Nanjing, Jiangsu, China.

*these authors contributed equally

Corresponding Author:

Miao Zhou, MD
Department of Anesthesiology
The Affiliated Cancer Hospital of Nanjing Medical University, Jiangsu Cancer Hospital, Jiangsu Institute of Cancer Research
Nanjing Medical University
No. 42, Baiziting Community, Xuanwu District
Nanjing, 210009
China
Phone: 86 182-1756-7295
Email: zhoumiao@jszlyy.com.cn

Abstract

Background

The occurrence of delirium is a prevalent phenomenon among patients admitted to the geriatric intensive care unit (ICU), with the potential to adversely impact prognosis and augment the risk of complications.

Objective

This study aimed to construct and validate a predictive model for postoperative delirium in elderly patients in ICUs, providing timely and effective early identification of high-risk individuals and assisting clinicians in decision-making.

Methods

The data from patients admitted to the ICU for over 24 hours were extracted from the Medical Information Marketplace for Intensive Care IV (MIMIC-IV) database and the eICU Collaborative Research Database (eICU-CRD). The MIMIC-IV data were split into a training set and an internal validation set (7:3

ratio), while the eICU-CRD data served as an external validation set. A delirium prediction was conducted for the subsequent prediction windows (12h, 24h, 48h, and whole stay time) utilising data from the first 24 hours post-admission. The corresponding feature variables were subjected to Boruta feature selection, and the prediction models were constructed using logistic regression, support vector classifier, random forest classifier, and extreme gradient boosting (XGB). Subsequently, the model performance was evaluated using receiver operating characteristic curves, calibration curves, decision curve analysis, and external validation.

Results

The MIMIC-IV and eICU-CRD datasets comprised 5897 and 618 patients, respectively, who were included in the analysis. A total of 57 features were selected for the construction of the predictive model. In the context of internal validation, the XGB model demonstrated the most effective prediction of delirium across different prediction windows. The Area under the curve values for the four prediction windows (12h, 24h, 48h, and whole stay time) were 0.860(95% CI: 0.839-0.880), 0.871(95% CI: 0.850-0.889), 0.851(95% CI: 0.829-0.871), and 0.846(95% CI: 0.827-0.867), respectively. The Area under the curve values for the external validation set were 0.828(95% CI: 0.768-0.880), 0.811(95% CI: 0.762-0.855), 0.756(95% CI: 0.705-0.803), and 0.750(95% CI: 0.701-0.795). Furthermore, the XGB model demonstrated the most accurate calibration across all prediction windows, with values of 0.115, 0.119, 0.136, and 0.144, respectively. Additionally, the decision curve analysis revealed that the XGB model outperformed the other models in terms of net gain for the majority of threshold probability values. The five most significant predictive features identified were the first day's delirium assessment results, invasive ventilation, Sequential Organ Failure Assessment score, minimum Glasgow Coma Scale score, and type of first care unit.

Conclusions

The high-performance XGB model for predicting postoperative delirium in elderly ICU patients has been successfully developed and validated. The model predicts the incidence of delirium at 12h, 24h, 48h, and whole stay time after the first day of hospitalisation within ICU. This enables physicians to identify high-risk patients early, thus facilitating the optimisation of personalised management strategies and care plans.

Keywords: Elderly; delirium; machine learning; artificial intelligence; delirium assessment; predictive modeling

Introduction

Delirium is an acute neuropsychiatric syndrome for which the pathogenesis remains incompletely understood. It is a type of cerebral dysfunction caused by a combination of precipitating factors and external stresses, accompanied by impairment of cognition, consciousness, attention, and mindset [1, 2]. Delirium can result in prolonged hospital stays, increased healthcare costs, adverse effects on surgical prognosis, and may even lead to long-term cognitive impairment and a decline in daily living standards outside the intensive care unit (ICU) [3-5]. Furthermore, delirium is associated with an elevated risk of postoperative mortality [6-8]. The diagnosis of delirium is a common occurrence in critically ill patients, with a prevalence of up to 82% [9], and prevention is complicated by the multiplicity and interplay of postoperative delirium triggers. However, the highly preventable nature of delirium suggests that early intervention will reduce the incidence of delirium in high-risk patients [10-12]. Fortunately, approximately 30% to 40% of delirium cases can benefit from delirium reduction strategies [13]. Consequently, the early prediction of delirium is of particular importance, it supports clinicians to implement timely interventions and targeted treatments that can maximize the benefits of early preventative measures.

Machine learning is an artificial intelligence technique that can process a substantial number of variables in a non-linear and highly interactive manner [14]. It enables computers to learn from data and make predictions or decisions without being explicitly programmed, thereby overcoming complex problems while exhibiting good predictive performance [15, 16]. Machine learning has been applied in the medical field for diagnosing diseases, recognizing medical images, providing treatment strategies and predicting outcomes [17]. To date, several delirium prediction models have been developed that are more effective in predicting postoperative delirium than traditional clinician-based regression models [18]. However, the challenge of achieving generalised replication across populations, differences in the inclusion of delirium factors across model groups, lack of reliability due to limited and partially missing retrospective data, and the dispersed and non-targeted nature of the populations covered remain to be addressed [19-22].

In this study, based on retrospective target data for the elderly population comprehensively collected from the Medical Information Marketplace for Intensive Care IV (MIMIC-IV) database and the eICU Collaborative Research Database (eICU-CRD), we developed an early prediction model for delirium using

machine learning algorithms. Furthermore, the independent variables were ranked according to their predictive importance, thus enhancing the interpretability. Notably, this study included delirium within 24 hours as a predictor in the model for the first time, retained patients who already had delirium on the first day, and did not neglect this elderly population with a not-so-low prevalence, which could help to comprehensively predict postoperative delirium in the elderly. More interestingly, Furthermore, this study evaluated the model's performance across a range of observational and predictive timeframes, addressing the issue of temporal variability and preventive effects. This enabled the prediction of postoperative delirium in elderly patients at both short- and long-term intervals, thereby reducing the likelihood of misestimating the risk of postoperative delirium due to errors in cognitive attention.

Our ultimate goal is to provide clinicians with a tool to identify high-risk patients faster and more comprehensively, and to be able to implement accurate and uniquely personalised risk prevention for older patients earlier based on prediction, thereby adjusting pretreatment strategies and care plans and ultimately improving prognosis.

Methods

Ethical review

The MIMIC-IV database is approved by the Institutional Review Boards of Beth Israel Deaconess Medical Centre and Massachusetts Institute of Technology. Access to the eICU-CRD was approved by the Massachusetts Institute of Technology Institutional Review Board. Due to the retrospective design, lack of direct patient intervention and safety structure, and the fact that all protected health information in the database is de-identified. Our study followed the Transparent Reporting of Multivariate Predictive Models for Individual Prognosis or Diagnosis statement [23].

Study population

The MIMIC-IV database comprises electronic health record data for 76,943 ICU admissions at Beth Israel Deaconess Medical Centre between 2008 and 2019 [24]. The eICU-CRD is a multicentre telemedicine database comprising data on over 200,000 patients admitted to 335 ICUs at 208 hospitals across the United States between 2014 and 2015 [25]. The study population comprised patients aged 65 years and above who were admitted to the ICU for the first time following surgery. The inclusion criteria also required that the ICU duration be at least 24 hours and that at least one validated Confusion Assessment Method for the ICU (CAM-ICU) be conducted during the initial 24-hour observation window and during subsequent prediction windows. Furthermore, patients above 89 years of age were excluded from the study, as the database for this age group was restricted to protect the privacy of the ultra-high elderly population. Additionally, physiological changes in this age group vary significantly, and their inclusion would not have enhanced the reliability or reference value of the data. In total, 5897 and 618 patients from the two databases were included in this study. The detailed flowchart illustrating the inclusion and exclusion criteria is presented in Figure 1.

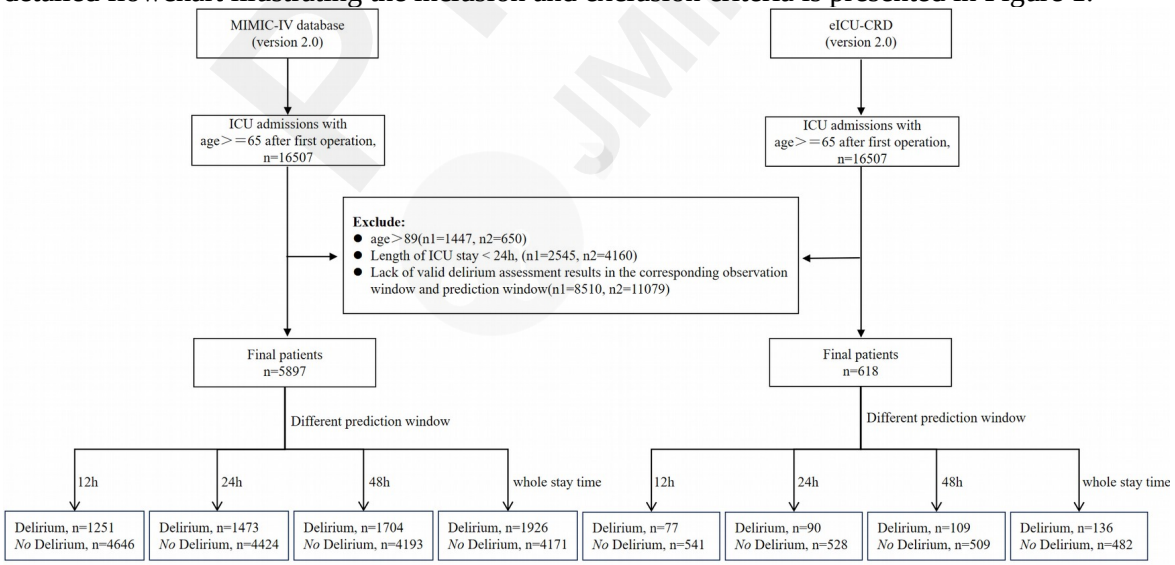


Figure 1. Cohort selection schema. MIMIC-IV: Medical Information Marketplace for Intensive Care IV; eICU-CRD: eICU Collaborative Research Database; ICU: intensive care unit

Delirium assessment

To identify instances of delirium, all intensive care unit patients received a CAM-ICU. The observation

window in the study refers to the time period during which patient data were collected and models were derived. The prediction window in the study refers to the time period between the end of the observation window and the artificially set deadline. We observed the patients' condition on the first day after their admission to the ICU. The incidence of delirium was predicted for the subsequent 12, 24 and 48 hours and whole ICU stay time. Delirium was diagnosed if at least one valid CAM-ICU value was positive in different prediction windows.

Data extraction and processing

In light of the existing literature on delirium prediction, the availability of data in relevant databases, the ease of data extraction and monitoring in a clinical setting, the total of 57 categorical or numerical variables were identified that met the aforementioned criteria and were subsequently categorized into the following domains: demographic data, vital signs, laboratory values, scores, comorbidities, and treatment measures. Furthermore, the first 24 hours of delirium assessment results and the first care unit type was documented in this study, thus providing the initial overall picture of the patient's condition. As the admission diagnosis was not consistent across the application dataset, similarly, downstream variables such as outcome were not available in real time and were therefore all excluded from the study. The first 24 hours of data were extracted for the valid variables, and in addition, where patients underwent multiple vital sign measurements or laboratory tests on the first day of admission, averages were calculated and extracted to ensure reliability for subsequent analyses. A list of all the variables used can be found in Textbox 1.

Textbox 1. Variables included in the prediction models.

Demographic data
- Age, gender, race, weight

Vital signs
- Heart rate, systolic blood pressure, diastolic blood pressure, mean blood pressure, temperature, respiratory rate, oxygen saturation

Laboratory results
- Hematocrit, hemoglobin, platelets, white blood cell, anion gap, bicarbonate, blood urea nitrogen, chloride, creatinine, glucose, sodium, potassium, international normalized ratio, prothrombin time, partial thromboplastin time, urine output

Comorbidity
- Myocardial infarction, congestive heart failure, peripheral vascular disease, cerebrovascular disease, dementia, chronic pulmonary disease, peptic ulcer disease, renal disease, Acquired Immunodeficiency Syndrome, liver disease, diabetes, tumor

Score
- Sequential Organ Failure Assessment (SOFA), Glasgow Coma Scale (GCS), Charlson Comorbidity Index

Treatment measures
- Invasive ventilation, renal replacement therapy, acetaminophen, anticholinergics, anticoagulants, antihistamines, antipsychotics, benzodiazepines, diuretics, general anesthetics, Nonsteroidal Antiinflammatory Drugs, opioids, vasopressors

Other information
- First 24h delirium assessment, first care unit type

Furthermore, to minimize the impact of missing data on the results, variables with more than 15% missing values were excluded from the final cohort. Missing values were then compensated for using multiple imputation [26]. Additionally, the data were shuffled to adjust the order of the samples.

Feature Selection

The training data of elderly patients admitted to the ICU following surgical procedures were analyzed using the pandas data analysis tool. Subsequently, the feature data is normalized and scaled to a standard normal distribution with a mean of 0 and a standard deviation of 1. The Boruta algorithm identifies the most salient features by comparing the Z-value of each feature with that of the 'shadow feature'. The Z-value of each attribute is obtained from the Random Forest model at each iteration by copying all the real features and shuffling them sequentially. In contrast, the Z-value of the shadow is generated by randomly shuffling the real features. In multiple independent trials, if the Z-value of a real feature exceeds the maximum Z-value of a shaded feature, the real feature is deemed to be "important." In this process, the Random Forest model is trained several times on the dataset, and the most important features are selected to predict the target variable [27]. This approach helps to ensure that the strongest predictive feature factors are retained while maintaining the performance of the model.

Parameter tuning and model development

The MIMIC-IV dataset was divided into a training set and a test set, with the former accounting for 70% and the latter for 30% of the total data. The eICU-CRD dataset was used as an external validation set. Four algorithms, namely logistic regression (LR), support vector classifier (SVC), random forest classifier (RFC) and extreme gradient boosting (XGB), were employed to develop the prediction model for delirium. Through Bayesian optimisation, the optimal combination of hyper-parameters for LR, SVR, RFC and XGB was automatically identified and incorporated into the corresponding model, which was then trained to achieve a high level of prediction performance.

Model Performance Evaluation

In order to facilitate the prediction of results, auxiliary functions were created and Programming Language Theory library functions were employed for the generation of receiver operating characteristic (ROC) curves and confusion matrix plots. The performance of the various models was then compared using the area under the curve (AUC) values, accuracy, precision, sensitivity and specificity.

A calibration curve is a visual tool that facilitates comprehension of the reliability and precision of a model by plotting the correlation between the predicted probability of the model and the actual frequency of observations [28, 29]. Brier score of 0 signifies optimal calibration, with a closer value to 0 indicating superior calibration [30]. A decision curve is a tool used to assess the performance of a predictive model under different thresholds, and it enables users to comprehend the impact of utilising the model in diverse decision-making scenarios by plotting the model's prediction curves under varying decision thresholds [31, 32]. Consequently, this study employs calibration curves and Brier scores to evaluate the model's reliability. Decision curve analysis was employed to assess the net clinical benefit. Shapley Additive Explanations (SHAP) was utilised to investigate the interpretability of the final predictive model.

Finally, in order to assess the generalisation ability of the model and the ability of the model to predict new samples, the applicability performance of the model predictions was assessed using external validation.

Statistical analyses

Stata version 17.0, SPSS version 27.0.1 and Python version 3.9 were applied for data processing, statistical analysis, and the development and validation of machine learning algorithms. Categorical variables were expressed as frequency and percentage and were compared using the $\chi 2$ test. Normally distributed continuous variables were expressed as mean and were compared using t-test. Non-normally distributed continuous variables, shown as median and interquartile distance, were compared using a rank sum test. P <0.05 indicates a statistically significant difference, and all tests were two-tailed.

Results

Baseline Characteristics

The final study cohort comprised 5897 patients from the MIMIC-IV dataset, of whom 1926 (32.7%) were assessed as delirium during the remaining stay after the first day in the ICU. Additionally, 618 patients from the eICU-CRD database were included, of whom 136 (22.0%) were assessed as delirium during the remaining stay after the first day in the ICU. Table 1 presents the characteristics of patients who were delirious and non-delirious in the prediction window of whole stay time. The characteristics of patients in other prediction windows are presented in Multimedia Appendix 1, Multimedia Appendix 2, and Multimedia Appendix 3.

Table 1. Baseline Characteristics of delirium and Non-delirium Patients in the prediction window of whole stay time.

Patients Characteristics	MIMIC-IV[a] cohort			eICU-CRD[b] cohort		
	No Delirium (n=3971)	Delirium (n=1926)	P Value	No Delirium (n=482)	Delirium (n=136)	P Value
Demographic data						
Age, year, median (IQR)	74.0(69.0,80.0)	76.0(70.0,82.0)	<0.001	74.0(69.0,80.0)	75.5(71.0,80.0)	0.140
Gender, male, n (%)	2263(57.0)	1037(53.8)	0.022	244(50.6)	81(59.6)	0.065
Weight, kg, median (IQR)	79.5(67.9,92.0)	77.0(65.2,91.0)	<0.001	77.8(65.5,91.4)	77.3(66.1,92.1)	0.777
Race, n (%)			<0.001			0.186
Black	253(6.4)	176(9.1)		51(10.6)	22(16.2)	
White	2857(71.9)	1220(63.5)		382(79.3)	99(72.8)	
Asian	111(2.8)	30(1.6)		2(0.4)	1(0.7)	
Hispanic	87(2.2)	47(2.4)		23(4.8)	4(2.9)	
Other/Unknown	663(16.7)	453(23.5)		24(5.0)	10(7.4)	
First care unit type, n (%)			<0.001			0.048
Cardiovascular ICU[c]	1861(46.9)	481(25.0)		103(21.4)	40(29.4)	
Neurological ICU	251(6.3)	240(12.5)		79(16.4)	27(19.9)	
Other ICU	1859(46.8)	1205(62.6)		300(62.2)	69(50.7)	

First 24h delirium assessment, n (%)			<0.001			<0.001
Unable to assess	251(6.3)	233(12.1)		8(1.7)	3(2.2)	
Negative	3182(80.1)	636(33.0)		439(91.1)	73(53.7)	
Positive	538(13.5)	1057(54.9)		35(7.3)	60(44.1)	
Vital signs, median (IQR)						
Heart rate, beats/min	79.8(72.2,88.4)	82.5(73.8,93.2)	<0.001	83.2(74.6,91.6)	86.6(77.5,97.0)	0.006
Systolic blood pressure, mmHg	115.1(106.9,125.2)	115.2(106.6,125.9)	0.600	119.0(108.2,133.4)	121.0(109.2,130.3)	0.768
Diastolic blood pressure, mmHg	58.3(53.1,64.3)	59.4(53.5,65.2)	0.002	61.0(55.9,67.3)	62.0(56.4,68.6)	0.274
Mean blood pressure, mmHg	74.8(69.9,80.8)	75.3(70.3,81.6)	0.011	78.1(71.2,85.1)	77.8(71.5,85.6)	0.750
Respiratory rate, beats/min	18.2(16.5,20.2)	18.8(16.9,21.1)	<0.001	17.5(15.7,19.6)	17.8(15.6,20.8)	0.402
Temperature, ℃	36.8(36.6,37.0)	36.9(36.6,37.2)	<0.001	36.8(36.6,37.1)	36.9(36.6,37.1)	0.722
Oxygen saturation, %	97.2(95.9,98.3)	97.7(96.3,98.8)	<0.001	97.3(95.8,98.4)	97.4(96.1,98.6)	0.375
Laboratory results, median (IQR)						
Hematocrit, %	31.8(28.4,35.5)	32.4(28.6,36.6)	<0.001	31.5(28.0,35.0)	29.9(25.2,33.4)	<0.001
Hemoglobin, g/dL	10.4(9.3,11.7)	10.5(9.3,11.9)	0.157	10.5(9.3,11.7)	9.6(8.5,10.9)	<0.001
Platelet, 10^9/L	175.0(133.0,225.5)	180.7(133.5,238.7)	0.021	191.8(144.1,243.0)	177.0(129.8,247.6)	0.429
White blood cell, 10^9/L	11.4(8.6,14.7)	12.2(9.2,15.7)	<0.001	11.9(9.1,15.2)	11.6(9.4,16.2)	0.685
Anion gap, mmol/L	13.8(12.0,16.0)	15.0(13.0,17.3)	<0.001	11.0(8.0,14.0)	11.2(8.0,15.0)	0.468
Bicarbonate, mmol/L	23.0(21.0,25.0)	22.0(19.7,24.3)	<0.001	24.0(22.0,25.6)	23.5(21.0,26.0)	0.400
Blood urea nitrogen, mg/dL	19.5(14.5,29.0)	23.4(16.5,37.0)	<0.001	18.5(13.0,26.4)	20.4(14.6,34.8)	0.003
Chloride, mmol/L	104.9(101.5,107.5)	104.5(101.0,108.0)	0.534	104.5(102.0,108.0)	105.9(102.5,109.6)	0.043
Creatinine, mg/dL	1.0(0.8,1.40)	1.10(0.8,1.7)	<0.001	1.0(0.8,1.40)	1.10(0.9,1.6)	<0.001
Glucose, mg/dL	128.5(110.5,153.5)	138.7(114.6,174.0)	<0.001	143.0(121.6,166.0)	139.3(120.2,165.7)	0.622
Sodium, mmol/L	138.5(136.5,140.3)	139.0(136.5,141.3)	<0.001	138.3(136.0,140.5)	139.0(136.6,142.0)	0.012
Potassium, mmol/L	4.2(3.9,4.5)	4.2(3.9,4.6)	0.857	4.2(3.9,4.6)	4.2(3.9,4.5)	0.964
International normalized ratio	1.3(1.1,1.4)	1.3(1.1,1.5)	<0.001	1.3(1.2,1.5)	1.5(1.3,1.7)	<0.001
Prothrombin time, s	14.0(12.4,15.4)	14.2(12.3,16.0)	<0.001	15.7(14.1,17.2)	17.0(14.8,19.5)	<0.001
partial thromboplastin time, s	31.8(28.0,38.4)	31.6(27.7,39.2)	0.551	34.9(34.2,35.2)	34.9(32.5,35.2)	0.478
Urine output, ml	1550.0(1050.0,2160.0)	1310.0(805.0,1875.0)	<0.001	1450.8(943.5,1876.5)	1360.0(770.5,1993.8)	0.268
Comorbidity, n (%)						
Myocardial infarct	907(22.8)	455(23.6)	0.503	18(3.7)	3(2.2)	0.548
Congestive heart failure	1301(32.8)	733(38.1)	<0.001	34(7.1)	5(3.7)	0.153
Peripheral vascular disease	621(15.6)	297(15.4)	0.829	11(2.3)	4(2.9)	0.900
Cerebrovascular disease	597(15.0)	520(27.0)	<0.001	26(5.4)	16(11.8)	0.009
Dementia	95(2.4)	218(11.3)	<0.001	5(1.0)	4(2.9)	0.218
Chronic pulmonary disease	1061(26.7)	601(31.2)	<0.001	36(7.5)	11(8.1)	0.810
Peptic ulcer disease	110(2.8)	83(4.3)	0.002	2(0.4)	2(1.5)	0.453
Renal disease	955(24.0)	556(28.9)	<0.001	82(17.0)	38(27.9)	0.004
AIDS[d]	1(0.0)	3(0.2)	0.105	1(0.2)	0(0)	1.000
Liver disease	289(7.3)	178(9.2)	0.009	5(1.0)	3(2.2)	0.525
Diabetes	1332(33.5)	705(36.6)	0.020	64(13.3)	21(15.4)	0.518
Tumor	680(17.1)	272(14.1)	0.003	104(21.6)	26(19.1)	0.535
Score, median (IQR)						
GCS[e]	15.00(14.0,15.0)	15.00(13.0,15.0)	<0.001	14.0(11.0,15.0)	13.0(8.0,14.0)	<0.001
SOFA[f]	4.00(2.00,6.00)	6.00(4.0,9.0)	<0.001	5.0(4.0,7.0)	7.0(6.0,9.0)	<0.001
Charlson Comorbidity Index	6.00(5.00,8.00)	7.00(5.0,9.0)	<0.001	5.0(3.0,6.0)	5.0(4.0,6.0)	0.025
Treatment measures, n (%)						
Renal replacement therapy	104(2.6)	99(5.1)	<0.001	19(3.9)	6(4.4)	0.806
Invasive ventilation	1882(47.4)	1446(75.1)	<0.001	178(36.9)	64(47.1)	0.033
Acetaminophen	3076(77.5)	1234(64.1)	<0.001	258(53.5)	87(64.0)	0.030
Anticholinergics	1364(34.3)	647(33.6)	0.566	80(16.6)	32(23.5)	0.064
Anticoagulants	2600(65.5)	1321(70.1)	<0.001	206(42.7)	57(41.9)	0.863
Antihistamines	234(5.1)	70(3.6)	<0.001	118(24.5)	17(12.5)	0.003
Antipsychotics	153(3.9)	193(10.0)	<0.001	5(1.0)	5(3.7)	0.077
Benzodiazepines	677(17.0)	310(16.1)	0.358	87(18.0)	39(28.7)	0.007
Diuretics	1675(42.2)	747(38.8)	0.013	134(27.8)	39(28.7)	0.841
General anesthetics	1660(41.8)	1171(60.8)	<0.001	74(15.4)	30(22.1)	0.065
NSAIDs[g]	2204(55.5)	755(39.2)	<0.001	114(23.7)	40(29.4)	0.171
Opioids	3417(86.0)	1736(90.1)	<0.001	282(58.5)	87(64.0)	0.252
Vasopressors	2022(50.9)	1083(56.2)	<0.001	90(18.7)	50(36.8)	<0.001

[a]MIMIC-IV: Medical Information Marketplace for Intensive Care IV.
[b]eICU-CRD: eICU Collaborative Research Database.
[c]ICU: intensive care unit.
[d]AIDS: Acquired Immunodeficiency Syndrome.
[e]GCS: Glasgow Coma Scale.
[f]SOFA: Sequential Organ Failure Assessment.
[g]NSAIDs: Nonsteroidal Antiinflammatory Drugs.

Evaluation of Model Performance

Four machine learning algorithms were employed in the construction of prediction models for the occurrence of delirium in elderly ICU patients following surgery. Figure 2 illustrates the discriminative performance of the ROC curves of the four models across different prediction windows. The XGB model

demonstrated the best prediction of postoperative delirium in elderly patients. The AUC values for the four prediction windows (12h, 24h, 48h, and whole stay time) were 0.860(95% CI: 0.839-0.880), 0.871(95% CI: 0.850-0.889), 0.851(95% CI: 0.829-0.871), and 0.846(95% CI: 0.827-0.867), respectively. The RFC model also exhibits satisfactory prediction performance, although it is slightly inferior to that of the XGB model in general. The corresponding AUC values for the four prediction windows of the RFC model are 0.854(95% CI: 0.832-0.872), 0.864(95% CI: 0.845-0.884), 0.847(95% CI: 0.826-0.867), and 0.841(95% CI: 0.821-0.860), respectively. Overall, both models exhibited a certain degree of decline in predicting delirium within the long-term prediction window compared to the short-term prediction window, which is consistent with our expectations. The SVC and LR models demonstrated significantly inferior performance compared to the first two models. Furthermore, the best performing XGB models were validated using the following metrics: accuracy, sensitivity, specificity, positive predictive value, and negative predictive value, as illustrated in Table 2. The confusion matrices associated with these evaluation metrics are presented in Supplemental Figure 1A.

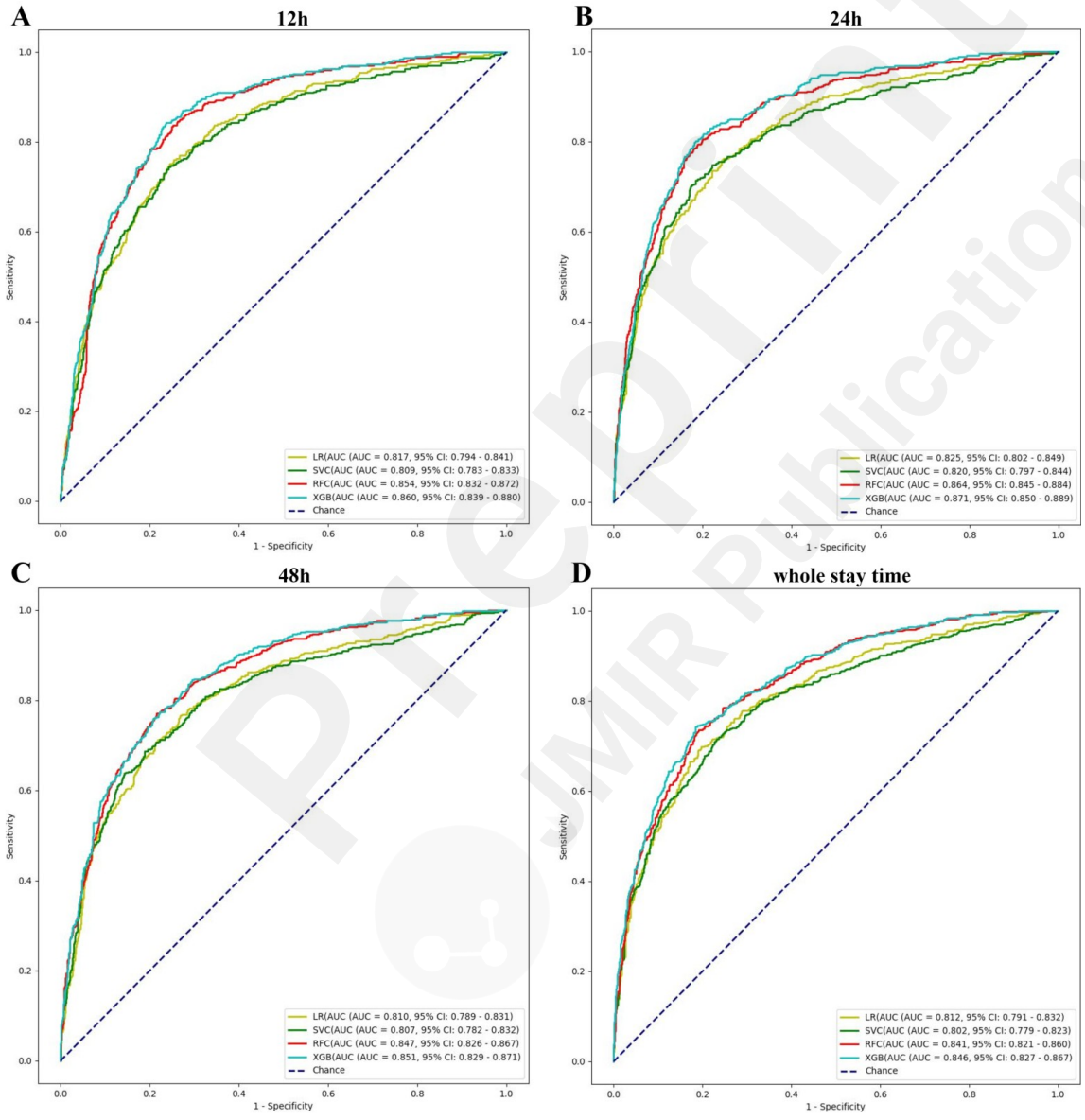


Figure 2. Receiver operating characteristic curves for all machine learning models in different prediction windows in the internal validation set. LR: logistic regression; SVC: support vector classifier; RFC: random forest classifier; XGB: extreme gradient boosting.

Table 2. The prediction performance of extreme gradient boosting models in different prediction windows in the internal validation set.

Prediction window	Accuracy, mean(95% CI)	Sensitivity, mean(95%CI)	Specificity, mean(95%CI)	PPV[a], mean(95%CI)	NPV[b], mean(95%CI)	AUC[c], mean(95%CI)

12h	0.828 (0.811-0.846)	0.456 (0.406-0.506)	0.928 (0.915-0.942)	0.631 (0.574-0.688)	0.866 (0.846-0.886)	0.860 (0.839-0.880)
24h	0.828 (0.811-0.846)	0.507 (0.460-0.553)	0.935 (0.922-0.948)	0.723 (0.673-0.772)	0.851 (0.832-0.869)	0.871 (0.850-0.889)
48h	0.811 (0.793-0.830)	0.585 (0.542-0.628)	0.903 (0.887-0.919)	0.710 (0.667-0.754)	0.843 (0.823-0.862)	0.851 (0.829-0.871)
whole stay time	0.797 (0.778-0.815)	0.647 (0.608-0.686)	0.869 (0.850-0.888)	0.706 (0.667-0.744)	0.835 (0.815-0.856)	0.846 (0.827-0.867)

[a]PPV: positive predictive value.

[b]NPV: negative predictive value.

[c]AUC: area under the curve.

In order to enhance the accuracy and precision of the model predictions, the model was calibrated utilising Brier scores and calibration curves. As illustrated in Figure 3, the XGB model demonstrates the optimal fit between the observed and predicted probabilities across diverse prediction windows, indicative of superior calibration. The Brier scores for the XGB model in predicting delirium across different windows are 0.115, 0.119, 0.136, and 0.144, respectively, substantiating the reliability of our model. Concurrently, the model demonstrates a comparatively higher degree of predictive precision in windows that are relatively early in time. A decision curve is a tool used to evaluate the performance of a predictive model under different thresholds. As illustrated in Figure 4, for the internal validation dataset, the XGB model exhibits superior performance compared to other machine learning models across a range of thresholds for diverse prediction windows, with the RFC model exhibiting a marginal advantage in a few instances. When multiple evaluation metrics are considered, the XGB model emerges as the best algorithm.

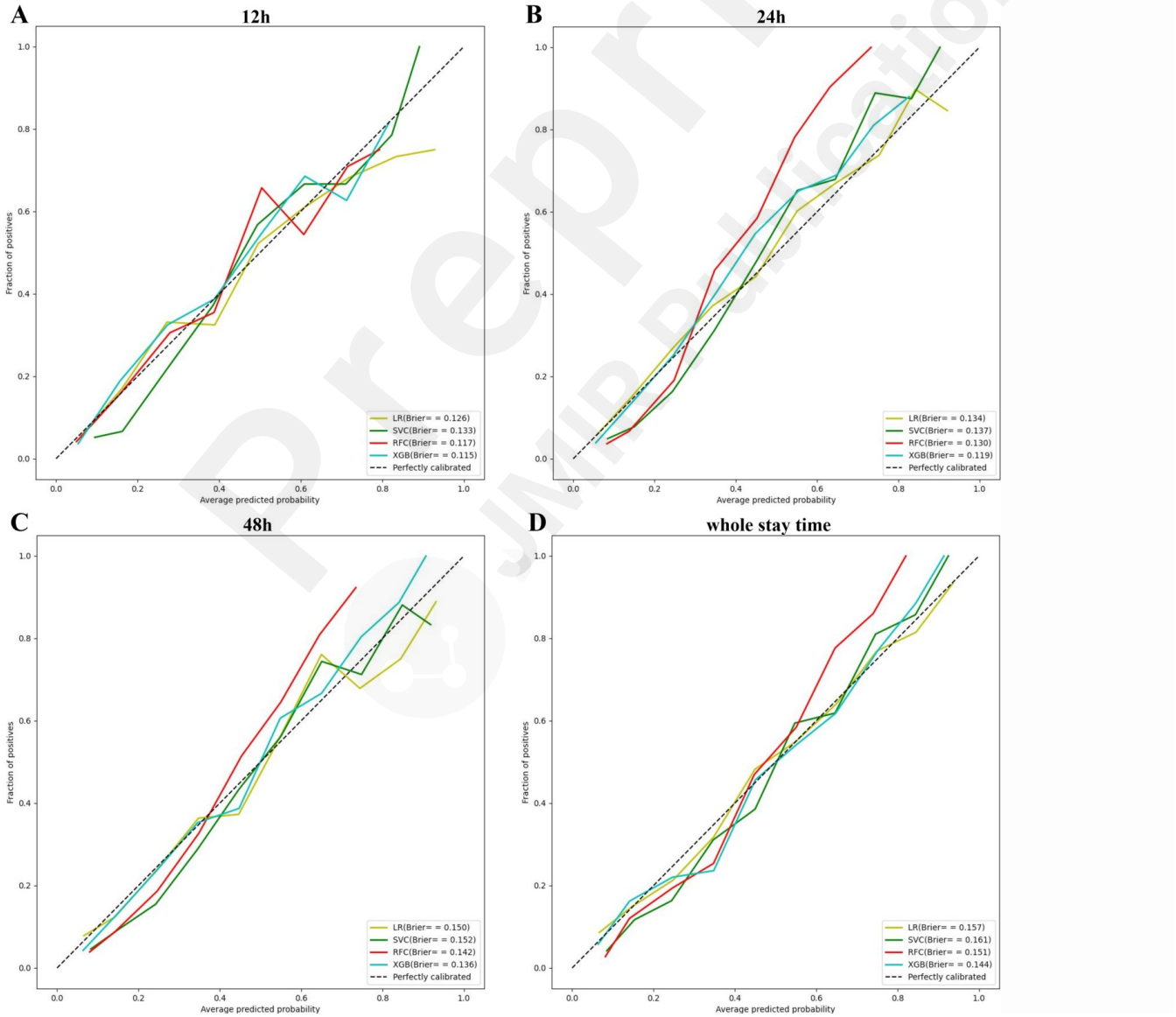


Figure 3. Calibration curves for all machine learning models in different prediction windows in the internal

validation set. LR: logistic regression; SVC: support vector classifier; RFC: random forest classifier; XGB: extreme gradient boosting. A Brier score of 0 indicates perfect calibration, and the closer the value is to 0, the better the model calibration, and XGB has the best Brier score.

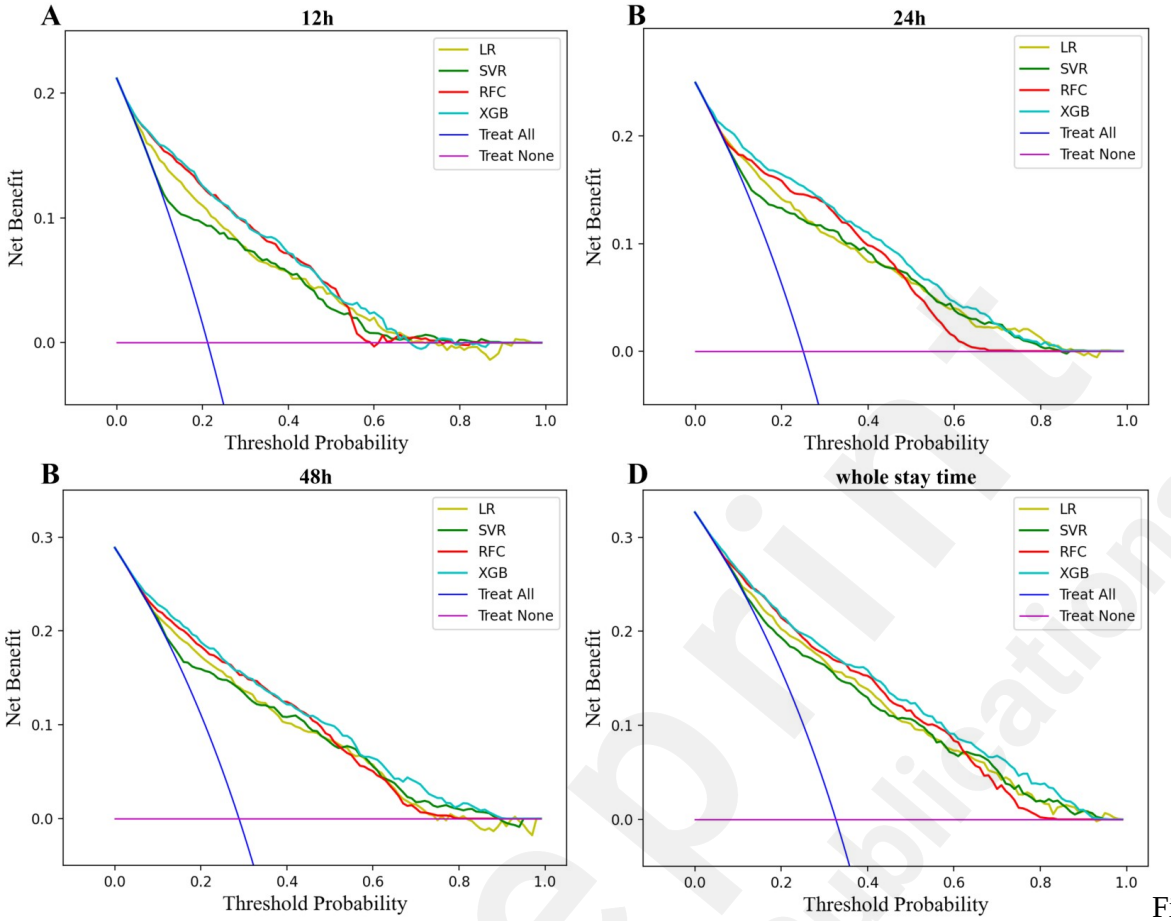


Figure 4. Decision curves for all machine learning models in different prediction windows in the internal validation set. LR: logistic regression; SVC: support vector classifier; RFC: random forest classifier; XGB: extreme gradient boosting.

To evaluate the model's capacity for generalisation and its ability to make predictions on new samples, an external validation of the XGB model was conducted using the eICU-CRD dataset from 208 different hospitals. With regard to the AUC values (Figure 5), the XGB model continues to demonstrate robust performance. The AUC values for the four prediction windows were 0.828(95% CI: 0.768-0.880), 0.811(95% CI: 0.762-0.855), 0.756(95% CI: 0.705-0.803), and 0.750(95% CI: 0.701-0.795), respectively. The comprehensive performance of the XGB model on the external validation set is presented in Table 3, and the associated confusion matrix plots are provided in Supplemental Figure 1B.

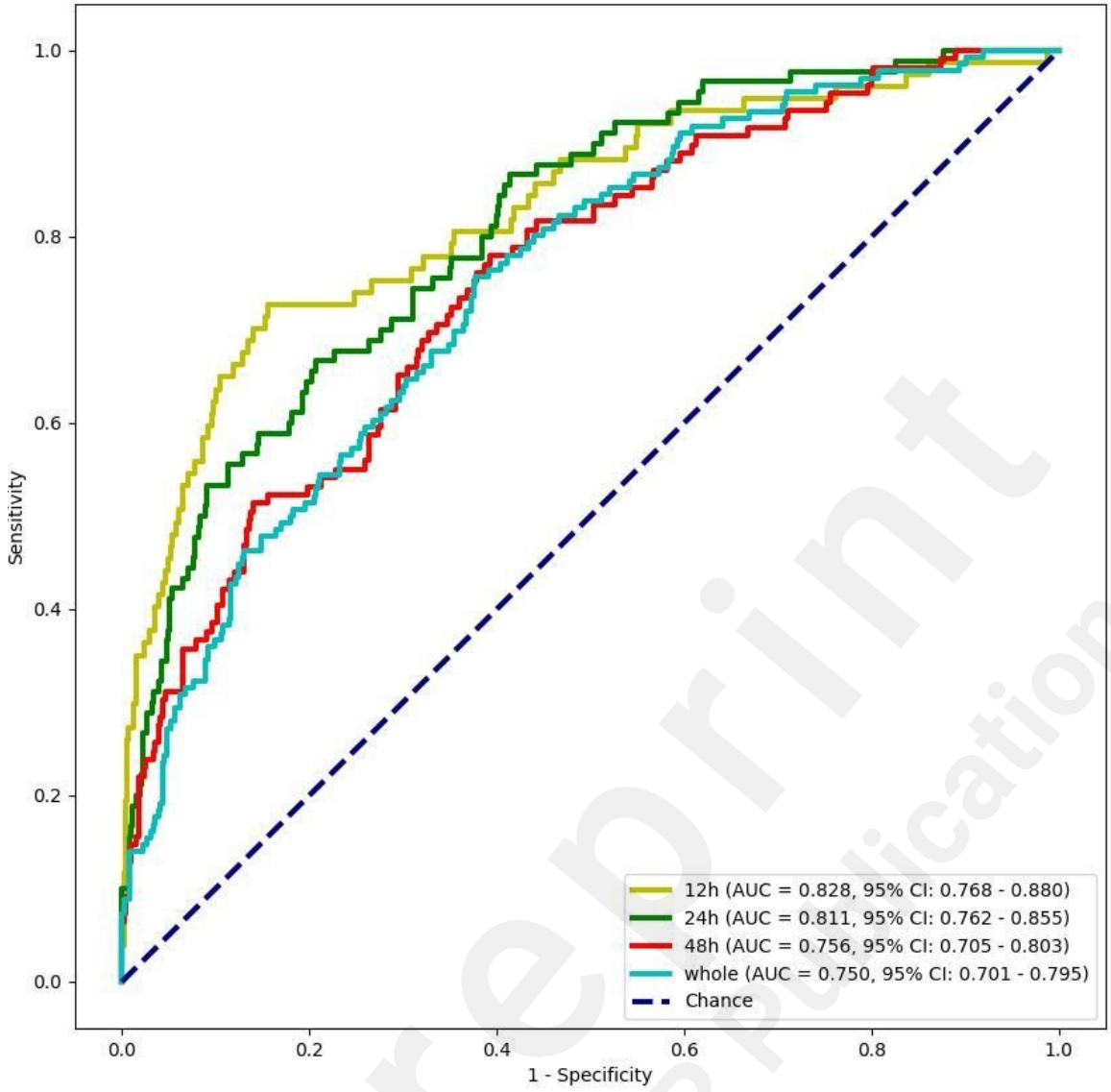


Figure 5.Receiver operating characteristic curves for extreme gradient boosting models in different prediction windows in the external validation set.

Table 3. The prediction performance of extreme gradient boosting models in different prediction windows in the external validation set.

Prediction window	Accuracy, mean(95%CI)	Sensitivity, mean(95%CI)	Specificity, mean(95%CI)	PPV[a], mean(95%CI)	NPV[b], mean(95%CI)	AUC[c], mean(95%CI)
12h	0.903 (0.880-0.926)	0.325 (0.220-0.429)	0.985 (0.975-0.995)	0.758 (0.611-0.904)	0.911 (0.888-0.934)	0.828 (0.768-0.880)
24h	0.867 (0.841-0.894)	0.333 (0.236-0.431)	0.958 (0.941-0.975)	0.577 (0.443-0.711)	0.894 (0.869-0.919)	0.811 (0.762-0.855)
48h	0.811 (0.780-0.842)	0.385 (0.294-0.477)	0.902 (0.876-0.928)	0.457 (0.355-0.558)	0.873 (0.844-0.901)	0.756 (0.705-0.803)
whole stay time	0.780 (0.747-0.813)	0.463 (0.379-0.547)	0.869 (0.839-0.899)	0.500 (0.413-0.587)	0.852 (0.820-0.883)	0.750 (0.701-0.795)

[a]PPV: positive predictive value.

[b]NPV: negative predictive value.

[c]AUC: area under the curve.

Variable Importance

The results of the study demonstrated that each variable had a distinct predictive value with respect to the occurrence of delirium in elderly ICU patients who had undergone surgery. In order to identify the most influential features in the model, we plotted the feature importance rankings of the XGB model for different prediction windows, comprising the top 20 features (as illustrated in Figure 6). The ranking of features

exhibits minor fluctuations across different prediction windows. In general, the most significant features were the first day's delirium assessment results, invasive ventilation, SOFA score, minimum GCS score, and type of first care unit. Furthermore, certain general anesthetics with concomitant sedation, mean body temperature, age, body weight and select laboratory metrics were also identified as relatively high-ranking features. The SHAP summary plot (Figure 7) complements the above ranking by illustrating the impact of each feature on the model output.

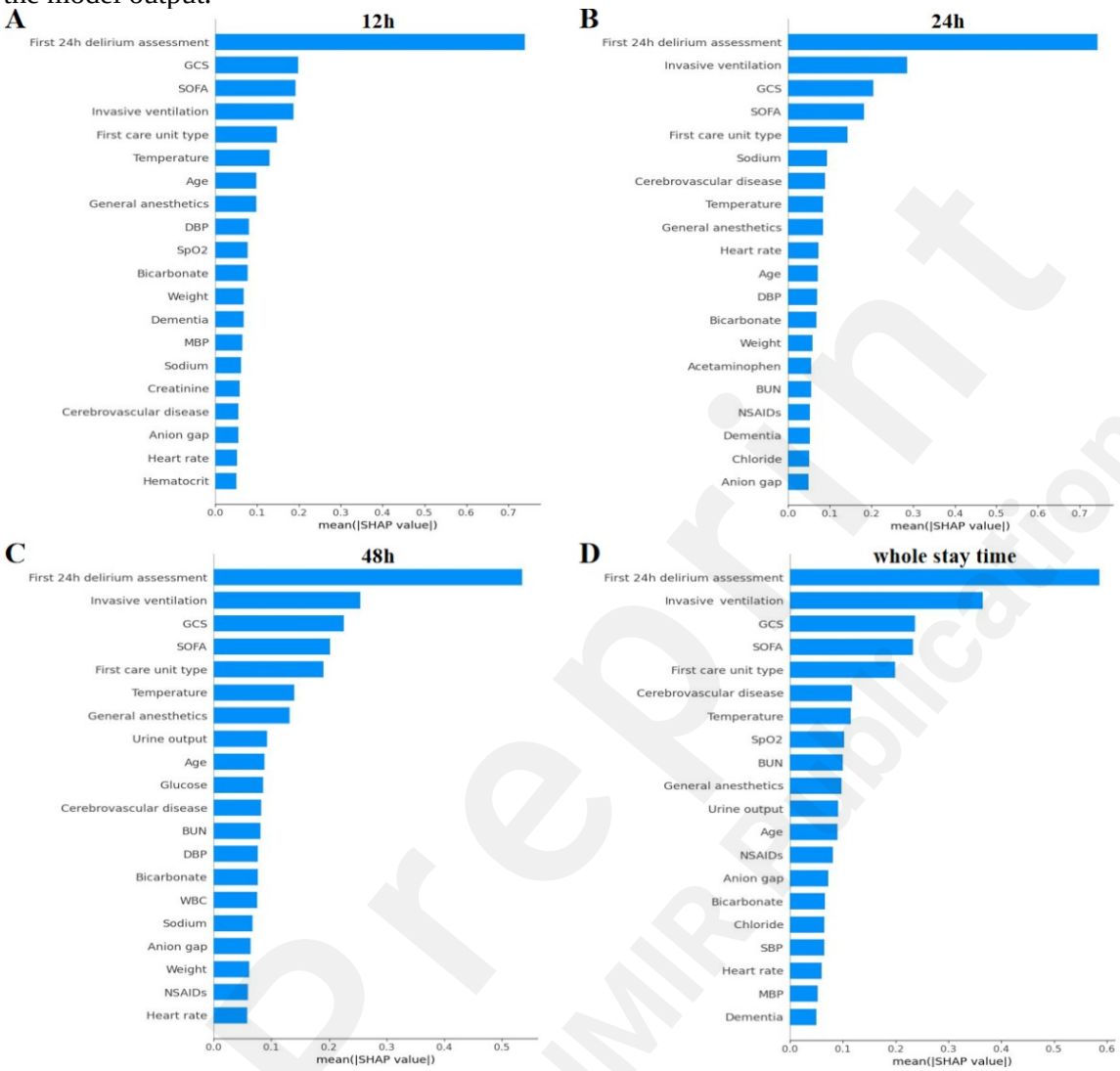


Figure 6. Feature importance ranking plot of the XGB machine learning models in different prediction windows (top 20 features). A, B, C, and D correspond to prediction windows of 12h, 24h, 48h, and whole stay time, respectively. GCS: Glasgow Coma Scale; SOFA: Sequential Organ Failure Assessment; SBP: systolic blood pressure; DBP: diastolic blood pressure; MBP: mean blood pressure; BUN: blood urea nitrogen; NSAIDS: Nonsteroidal Antiinflammatory Drugs; WBC: white blood cell; SpO2: oxygen saturation.

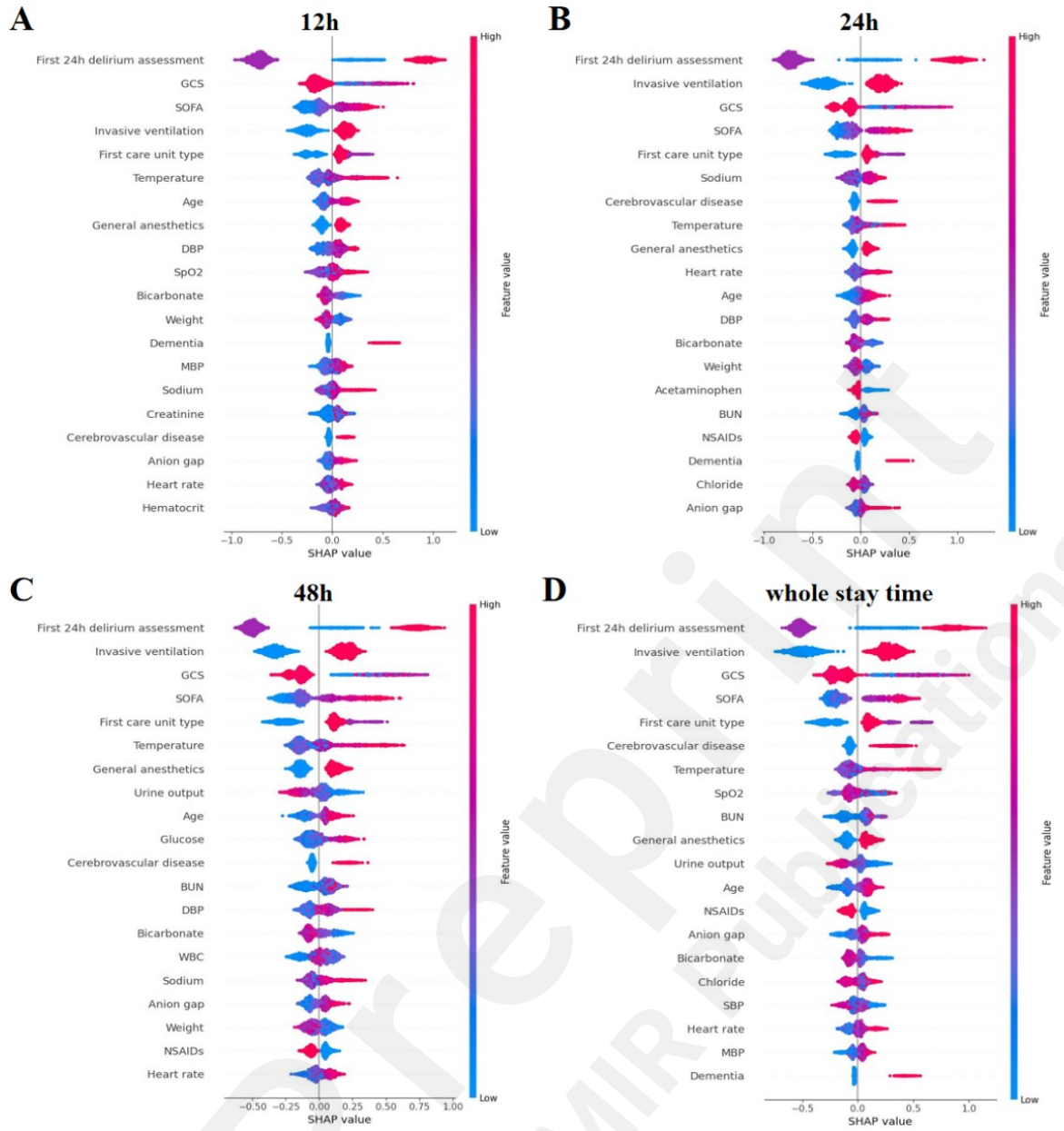


Figure 7. Shapley Additive Explanations (SHAP) summary plots of the XGB machine learning models in different prediction windows (top 20 features). A, B, C, and D correspond to prediction windows of 12h, 24h, 48h, and whole stay time, respectively. Each point in the plot in a given case corresponds to the SHAP value of the element. The y-axis represents the feature, and the x-axis position indicates the SHAP value or the extent of the feature's impact on the prediction. The color of the points represents the actual values of the features, with purple indicating low values and red indicating high values. GCS: Glasgow Coma Scale; SOFA: Sequential Organ Failure Assessment; SBP: systolic blood pressure; DBP: diastolic blood pressure; MBP: mean blood pressure; BUN: blood urea nitrogen; NSAIDS: Nonsteroidal Antiinflammatory Drugs; WBC: white blood cell; SpO2: oxygen saturation.

Discussion

Principal Results

This study presents the findings of a large-scale retrospective analysis conducted of elderly ICU population with a high prevalence of postoperative delirium. An early prediction model for delirium in elderly ICU patients was developed using four mainstream machine learning algorithms. The most effective XGB models for identifying high-risk delirium early were then selected. In addition, satisfactory discrimination and generalisation abilities are demonstrated in both internal and external validation.

To the best of our knowledge, this is the inaugural predictive model to forecast delirium episodes in elderly patients in a windowed manner. The model is capable of predicting the incidence of delirium at the next 12, 24, and 48 hours, as well as throughout the subsequent ICU stay, utilising clinical data obtained

within 24 hours of ICU admission. This addresses the issue of temporal variability and the preventive effect. Notably, delirium within 24 hours was included as a predictor in the model for the first time in this study, which helped to comprehensively predict postoperative delirium in the elderly and did not overlook the elderly population with repeated delirium episodes. In addition, the first day's delirium assessment results, invasive ventilation, SOFA score, minimum GCS score, and type of first care unit were the five most significant predictive features.

Comparison With Prior Work

Of all hospital departments, ICU have the highest incidence of delirium, with rates of up to 80% in ICUs compared to 14-29% in non-ICU inpatients [33]. This is attributable to the high prevalence of ICU delirium resulting from physical and medication disorders commonly observed in critically ill patients [34, 35]. In addition to the fact that patients' susceptibility to postoperative general anesthesia in the elderly population is significantly associated with an increased risk of ICU delirium [36, 37]. It is therefore imperative that delirium is predicted at the earliest opportunity upon admission to the ICU, particularly in elderly patients. At present, the CAM-ICU score is the most frequently employed method for diagnosing delirium. However, it necessitates the administration of multiple assessments to ascertain a positive result [38]. Prior research has demonstrated that clinicians' forecasts of delirium progression are less precise than those of ICU delirium prediction models. This discrepancy may be attributed to various factors, including the lack of clinical experience among ICU personnel, the volume and intricacy of delirium assessment, and the dearth of attention devoted to delirium [39-42]. Despite the development of multiple models for the assessment of delirium risk in the ICU, these models exhibit limitations in their scope and focus. Many models encompass a broad age range, while others concentrate on the post-surgical recovery period, neglecting the distinctive attributes of elderly ICU patients [43-45]. In contrast, our machine-learning predictive model demonstrated the capacity to anticipate delirium onset at an earlier stage, based on data from elderly ICU admissions within 24 hours. Interestingly, the prevalence of postoperative delirium was 41.8% in elderly patients in the MIMIC-IV database and 27.7% in elderly patients in the eICU-CRD, which is higher than in previous similar studies [21, 25, 46, 47]. This may be due to the fact that we retained patients who were diagnosed with delirium on the first day in order to focus on their subsequent progress, which resulted in an increased incidence of delirium in the overall data. Previous studies have demonstrated that patients with persistent or recurrent delirium tend to have longer hospital stays and higher mortality rates [48]. Consequently, the value of clinical research and prevention in this population cannot be overlooked. In conclusion, this study may assist clinicians in making optimal clinical decisions and providing preventive risk monitoring and personalised care plans for high-risk patients.

The present study identified five key factors most strongly associated with the onset of delirium in elderly patients in the ICU. These were the first day's delirium assessment results, invasive ventilation, SOFA score, minimum GCS score, and type of first care unit. This means that certain highly predictive features identified in previous studies (such as age, invasive ventilation, SOFA score, GCS score, and type of first care unit), were corroborated in the present study [24, 25, 49, 50]. In the present study, the factor of delirium on the first day was consistently identified as the most important characteristic. It seems plausible to suggest that this is due to the fact that delirium is a persistent illness from which patients are unlikely to recover in the short term. This has a cascading effect on subsequent delirium assessments. It is noteworthy that in the MIMIC-IV dataset of this study, 1595 individuals were identified as having delirium on day one. By day two, 660 (41.3%) of these had moved to a non-delirious state. This suggests that although the result of the delirium-first day was an important predictor of this study, it was not the sole determining factor. In the meantime, this study employed SHAP to elucidate the intrinsic information of the XGB model, thereby offering a transparent rationale for personalised risk prediction of delirium. This facilitates a more intuitive comprehension of the influence of pivotal features and provides guidance for clinical decision-making.

In addition, cognitive and behavioural functions in humans are related to neurotransmitter transmission, and certain anaesthetic drugs may induce delirium by affecting the balance of transmitter transmission and leading to neurological dysfunction, although the exact mechanism remains unknown [51]. The data from this study indicates that the anaesthetic drugs administered on the first day of admission to the ICU are a significant contributing factor to the development of delirium. Over 98% of these drugs are propofol and ketamine, which are primarily employed to facilitate the sedation of critically ill patients in the ICU, thereby enabling more effective monitoring and management of their condition. This finding aligns with the observations reported by Zhang Yang et al [22]. Concurrently, anaesthetic drugs are metabolised at a

diminished rate in elderly patients, thereby increasing the probability of adverse effects [52]. Conversely, specific anaesthetic drugs (e.g. dopamine D2 antagonists) play a pivotal role in the prevention and alleviation of delirium, which is commonly treated with the use of such drugs in ICU [53].

Several postoperative complications were included in this study to investigate risk factors for delirium. It was found that there was no significant correlation between postoperative complications and delirium and the only notable complication was cerebrovascular disease, which may be related to organic brain disease. Other predictors such as body temperature, anion gap, blood sodium, oxygen saturation, blood urea nitrogen, blood pressure, urine output, blood glucose, biarbonate, and platelets have been validated by similar studies or predictive models [22, 54, 55].

Strengths and Limitations

It is important to note that our study has several notable contributions and strengths. Firstly, this study differs from previous research in that it constructed four predictive models, rather than a single model. The optimal predictive XGB model was selected based on the utilisation of conventional clinical feature variables. The demonstration of favorable predictive performance in both internal and external validation enhances the clinical utility of the XGB model and provides compelling evidence for its popularity. Secondly, in order to ensure the quality and quantity of the data, two widely recognised high-quality databases were used: MIMIC-IV database and eICU-CRD. These databases are characterised by a large sample size and rich clinical data. The data set used in this study addresses data from an elderly population with a high prevalence of delirium and is therefore of a higher quality than some studies that use clinically collected data. Last but not least, the model was constructed using data that were readily available and collected within 24 hours of the patient's admission to the ICU. This is the inaugural instance in which the characteristics of first 24h delirium assessment have been incorporated, taking into account the recurrence and persistence of delirium and prognosis. Furthermore, different 12h, 24h, 48h, and whole stay time delirium prediction windows were constructed. The model considers both short- and long-term prediction of delirium, which is crucial for improving continuity of care and more effectively planning resource allocation in resource-limited settings. Additionally, early and accurate prediction of delirium allows clinicians to adjust treatment strategies with greater time efficiency.

It is important to acknowledge that our study has certain limitations. Firstly, our study was implemented and validated retrospectively, and therefore further prospective intervention studies are required to validate the performance of the model. Secondly, there are currently no clear diagnostic criteria for delirium. Although the CAM-ICU tool is considered highly sensitive and specific for the detection of delirium in the ICU, misdiagnosis and underdiagnosis are still inevitable and do not reflect the degree of deterioration of delirium. Thirdly, there is a possibility of selection bias and interpretive bias, as only variables that were available in all cohorts and easily extracted from the database were selected in order to ensure the accuracy and validity of the data. Furthermore, patients who did not have sufficiently valid CAM-ICU data (59% of total ICU admissions after the initial screening in the MIMIC-IV dataset) were excluded. Fourthly, the European and United States databases were sourced for this study, due to the inherent limitations of genetically distinct populations with unique attributes that prevent the predictive models derived from these databases from being generalised to other populations. Fifthly, state-of-the-art approaches to model interpretation, including SHAP and its alternatives, fail to account for dependencies between features, which inevitably introduces correlation bias [56, 57].

Conclusions

In this study, we constructed and validated a high-performance prediction model for delirium in elderly ICU patients. This model can predict the incidence of delirium in the subsequent 12 hours, 24 hours, 48 hours, and whole stay time using clinical data obtained within 24 hours of ICU admission. It enables clinicians to promptly identify elderly patients at elevated risk of delirium, thus facilitating the implementation of targeted and individualised interventions to enhance prognosis and optimise management strategies, while rationalising healthcare resources.

Acknowledgements

This work was supported the National Science Foundation for Young Scientists of China (NO. 82002122).

MZ and DX are the corresponding authors and take responsibility for the integrity of the whole work. HL drafted the first version of the manuscript. QL, YL, and HH revised the manuscript. QZ, HL and JD were responsible for data cleaning and algorithm implementation. JH and YZ were responsible for data access and privacy management. All authors have read and approved the final manuscript.

Conflicts of Interest
None declared.
Abbreviations
AIDS: Acquired Immunodeficiency Syndrome
AUC: area under the curve
BUN: blood urea nitrogen
CAM-ICU: Confusion Assessment Method for the Intensive Care Unit
DBP: Diastolic blood pressure
eICU-CRD: eICU Collaborative Research Database
GCS: Glasgow Coma Scale
ICU: intensive care unit
LR: logistic regression
MBP: Mean blood pressure
MIMIC-IV: Medical Information Marketplace for Intensive Care IV
NPV: negative predictive value
NSAIDs: Nonsteroidal Antiinflammatory Drugs
PPV: positive predictive value
RFC: random forest classifier
ROC: receiver operating characteristic
SBP: systolic blood pressure
SHAP: Shapley Additive Explanations
SOFA: Sequential Organ Failure Assessment
SpO2: oxygen saturation
SVC: support vector classifier
WBC: white blood cell
XGB: extreme gradient boosting
Multimedia Appendix 1
Baseline Characteristics of delirium and Non-delirium Patients in the 12h prediction window.
Multimedia Appendix 2
Baseline Characteristics of delirium and Non-delirium Patients in the 24h prediction window.
Multimedia Appendix 3
Baseline Characteristics of delirium and Non-delirium Patients in the 48h prediction window.
Multimedia Appendix 4
Confusion Matrix of XGB Models for Different Prediction windows in the internal validation set.
Multimedia Appendix 5
Confusion Matrix of XGB Models for Different Prediction windows in the external validation set.
References
1. Mattison MLP. Delirium. Ann Intern Med. Oct 06, 2020;173(7):ITC49-ITC64. PMID: 33017552. doi: 10.7326/AITC202010060.
2. Oh ES, Fong TG, Hshieh TT, Inouye SK. Delirium in Older Persons: Advances in Diagnosis and Treatment. JAMA. Sep 26, 2017;318(12):1161-1174. PMID: 28973626. doi: 10.1001/jama.2017.12067.
3. Wilcox ME, Girard TD, Hough CL. Delirium and long term cognition in critically ill patients. BMJ. Jun 8, 2021;373:n1007. PMID: 34103334. doi: 10.1136/bmj.n1007.
4. Schubert M, Schurch R, Boettger S, Garcia Nunez D, Schwarz U, Bettex D, et al. A hospital-wide evaluation of delirium prevalence and outcomes in acute care patients - a cohort study. BMC Health Serv Res. Jul 13, 2018;18(1):550. PMID: 30005646. doi: 10.1186/s12913-018-3345-x.
5. Gleason LJ, Schmitt EM, Kosar CM, Tabloski P, Saczynski JS, Robinson T, et al. Effect of Delirium and Other Major Complications on Outcomes After Elective Surgery in Older Adults. JAMA Surg. Dec 2015;150(12):1134-1140. PMID: 26352694. doi: 10.1001/jamasurg.2015.2606.
6. Chalmers LA, Searle SD, Whitby J, Tsui A, Davis D. Do specific delirium aetiologies have different associations with death? A longitudinal cohort of hospitalised patients. Eur Geriatr Med. Aug 2021;12(4):787-791. PMID: 33725336. doi: 10.1007/s41999-021-00474-8.
7. Witlox J, Eurelings LS, de Jonghe JF, Kalisvaart KJ, Eikelenboom P, van Gool WA. Delirium in elderly patients and the risk of postdischarge mortality, institutionalization, and dementia: a meta-analysis. JAMA. Jul

28, 2010;304(4):443-451. PMID: 20664045. doi: 10.1001/jama.2010.1013.
8. Bai J, Liang Y, Zhang P, Liang X, He J, Wang J, et al. Association between postoperative delirium and mortality in elderly patients undergoing hip fractures surgery: a meta-analysis. Osteoporos Int. Feb 2020;31(2):317-326. PMID: 31741024. doi: 10.1007/s00198-019-05172-7.
9. Goldberg TE, Chen C, Wang Y, Jung E, Swanson A, Ing C, et al. Association of Delirium With Long-term Cognitive Decline: A Meta-analysis. JAMA Neurol. Nov 01, 2020;77(11):1373-1381. PMID: 32658246. doi: 10.1001/jamaneurol.2020.2273.
10. Liu Y, Shen W, Tian Z. Using Machine Learning Algorithms to Predict High-Risk Factors for Postoperative Delirium in Elderly Patients. Clin Interv Aging. 2023;18:157-168. PMID: 36789284. doi: 10.2147/CIA.S398314.
11. Mart MF, Williams Roberson S, Salas B, Pandharipande PP, Ely EW. Prevention and Management of Delirium in the Intensive Care Unit. Semin Respir Crit Care Med. Feb 2021;42(1):112-126. PMID: 32746469. doi: 10.1055/s-0040-1710572.
12. Salvi F, Young J, Lucarelli M, Aquilano A, Luzi R, Dell'Aquila G, et al. Non-pharmacological approaches in the prevention of delirium. Eur Geriatr Med. Feb 2020;11(1):71-81. PMID: 32297241. doi: 10.1007/s41999-019-00260-7.
13. Hshieh TT, Yue J, Oh E, Puelle M, Dowal S, Travison T, et al. Effectiveness of multicomponent nonpharmacological delirium interventions: a meta-analysis. JAMA Intern Med. Apr 2015;175(4):512-520. PMID: 25643002. doi: 10.1001/jamainternmed.2014.7779.
14. Mullainathan S, Spiess J. Machine learning: an applied econometric approach. J Econ Perspect. May 01, 2017;31(2):87-106. doi: 10.1257/jep.31.2.87.
15. Cuocolo R, Caruso M, Perillo T, Ugga L, Petretta M. Machine Learning in oncology: A clinical appraisal. Cancer Lett. Jul 01, 2020;481:55-62. PMID: 32251707. doi: 10.1016/j.canlet.2020.03.032.
16. Groot OQ, Bongers MER, Ogink PT, Senders JT, Karhade AV, Bramer JAM, et al. Does Artificial Intelligence Outperform Natural Intelligence in Interpreting Musculoskeletal Radiological Studies? A Systematic Review. Clin Orthop Relat Res. Dec 2020;478(12):2751-2764. PMID: 32740477. doi: 10.1097/CORR.0000000000001360.
17. Jiang F, Jiang Y, Zhi H, Dong Y, Li H, Ma S, et al. Artificial intelligence in healthcare: past, present and future. Stroke Vasc Neurol. Dec 2017;2(4):230-243. PMID: 29507784. doi: 10.1136/svn-2017-000101.
18. Bishara A, Chiu C, Whitlock EL, Douglas VC, Lee S, Butte AJ, et al. Postoperative delirium prediction using machine learning models and preoperative electronic health record data. BMC Anesthesiol. Jan 3, 2022;22(1):8. PMID: 34979919. doi: 10.1186/s12871-021-01543-y.
19. Lindroth H, Bratzke L, Purvis S, Brown R, Coburn M, Mrkobrada M, et al. Systematic review of prediction models for delirium in the older adult inpatient. BMJ Open. Apr 28, 2018;8(4):e019223. PMID: 29705752. doi: 10.1136/bmjopen-2017-019223.
20. Rohr V, Blankertz B, Radtke FM, Spies C, Koch S. Machine-learning model predicting postoperative delirium in older patients using intraoperative frontal electroencephalographic signatures. Front Aging Neurosci. 2022;14:911088. PMID: 36313029. doi: 10.3389/fnagi.2022.911088.
21. Gong KD, Lu R, Bergamaschi TS, Sanyal A, Guo J, Kim HB, et al. Predicting Intensive Care Delirium with Machine Learning: Model Development and External Validation. Anesthesiology. Mar 01, 2023;138(3):299-311. PMID: 36538354. doi: 10.1097/ALN.0000000000004478.
22. Zhang Y, Hu J, Hua T, Zhang J, Zhang Z, Yang M. Development of a machine learning-based prediction model for sepsis-associated delirium in the intensive care unit. Sci Rep. Aug 04, 2023;13(1):12697. PMID: 37542106. doi: 10.1038/s41598-023-38650-4.
23. Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. BMJ. Jan 07, 2015;350:g7594. PMID: 25569120. doi: 10.1136/bmj.g7594.
24. Johnson AEW, Bulgarelli L, Shen L, Gayles A, Shammout A, Horng S, et al. MIMIC-IV, a freely accessible electronic health record dataset. Sci Data. Jan 03, 2023;10(1):1. PMID: 36596836. doi: 10.1038/s41597-022-01899-x.
25. Pollard TJ, Johnson AEW, Raffa JD, Celi LA, Mark RG, Badawi O. The eICU Collaborative Research Database, a freely available multi-center database for critical care research. Sci Data. Sep 11, 2018;5:180178. PMID: 30204154. doi: 10.1038/sdata.2018.178.
26. Zhang Z. Multiple imputation with multivariate imputation by chained equation (MICE) package. Ann

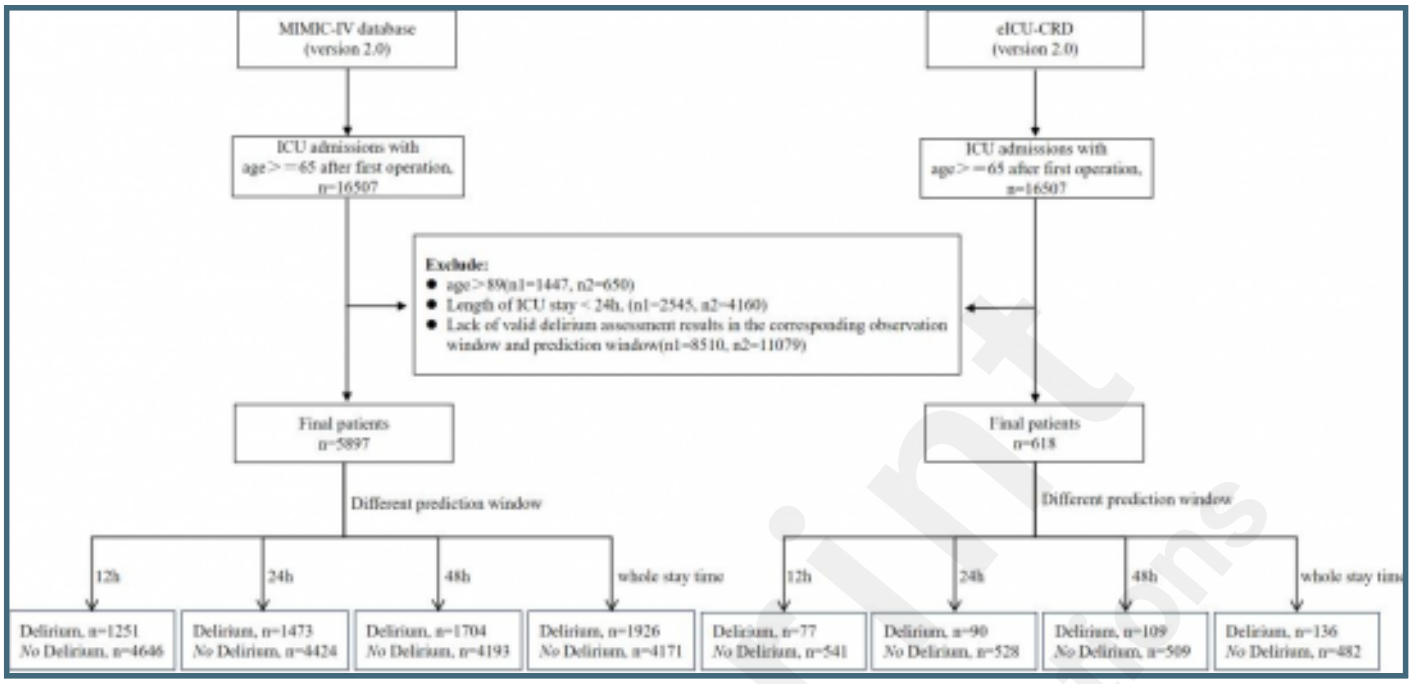
Transl Med. Jan 2016;4(2):30. PMID: 26889483. doi: 10.3978/j.issn.2305-5839.2015.12.63.
27. Lei J, Sun T, Jiang Y, Wu P, Fu J, Zhang T, et al. Risk Identification of Bronchopulmonary Dysplasia in Premature Infants Based on Machine Learning. Front Pediatr. 2021;9:719352. PMID: 34485204. doi: 10.3389/fped.2021.719352.
28. Cook NR. Statistical evaluation of prognostic versus diagnostic models: beyond the ROC curve. Clin Chem. Jan 2008;54(1):17-23. PMID: 18024533. doi: 10.1373/clinchem.2007.096529.
29. Cearns M, Hahn T, Clark S, Baune BT. Machine learning probability calibration for high-risk clinical decision-making. Aust N Z J Psychiatry. Feb 2020;54(2):123-126. PMID: 31707786. doi: 10.1177/0004867419885448.
30. Rufibach K. Use of Brier score to assess binary predictions. J Clin Epidemiol. Aug 2010;63(8):938-939; author reply 939. PMID: 20189763. doi: 10.1016/j.jclinepi.2009.11.009.
31. Vickers AJ, Elkin EB. Decision curve analysis: a novel method for evaluating prediction models. Med Decis Making. Nov-Dec 2006;26(6):565-574. PMID: 17099194. doi: 10.1177/0272989X06295361.
32. Zhao L, Leng Y, Hu Y, Xiao J, Li Q, Liu C, et al. Understanding decision curve analysis in clinical prediction model research. Postgrad Med J. Jun 28, 2024;100(1185):512-515. PMID: 38453146. doi: 10.1093/postmj/qgae027.
33. Honarmand K, Lalli RS, Priestap F, Chen JL, McIntyre CW, Owen AM, et al. Natural History of Cognitive Impairment in Critical Illness Survivors. A Systematic Review. Am J Respir Crit Care Med. Jul 15, 2020;202(2):193-201. PMID: 32078780. doi: 10.1164/rccm.201904-0816CI.
34. Boncyk CS, Farrin E, Stollings JL, Rumbaugh K, Wilson JE, Marshall M, et al. Pharmacologic Management of Intensive Care Unit Delirium: Clinical Prescribing Practices and Outcomes in More Than 8500 Patient Encounters. Anesth Analg. Sep 01, 2021;133(3):713-722. PMID: 33433117. doi: 10.1213/ANE.0000000000005365.
35. Ghasemiyeh P, Vazin A, Zand F, Haem E, Karimzadeh I, Azadi A, et al. Pharmacokinetic assessment of vancomycin in critically ill patients and nephrotoxicity prediction using individualized pharmacokinetic parameters. Front Pharmacol. 2022;13:912202. PMID: 36091788. doi: 10.3389/fphar.2022.912202.
36. Li T, Li J, Yuan L, Wu J, Jiang C, Daniels J, et al. Effect of Regional vs General Anesthesia on Incidence of Postoperative Delirium in Older Patients Undergoing Hip Fracture Surgery: The RAGA Randomized Trial. JAMA. Jan 04, 2022;327(1):50-58. PMID: 34928310. doi: 10.1001/jama.2021.22647.
37. Vasunilashorn SM, Ngo LH, Inouye SK, Fong TG, Jones RN, Dillon ST, et al. Apolipoprotein E genotype and the association between C-reactive protein and postoperative delirium: Importance of gene-protein interactions. Alzheimers Dement. Mar 2020;16(3):572-580. PMID: 31761478. doi: 10.1016/j.jalz.2019.09.080.
38. Pun BT, Badenes R, Heras La Calle G, Orun OM, Chen W, Raman R, et al. Prevalence and risk factors for delirium in critically ill patients with COVID-19 (COVID-D): a multicentre cohort study. Lancet Respir Med. Mar 2021;9(3):239-250. PMID: 33428871. doi: 10.1016/S2213-2600(20)30552-X.
39. la Cour KN, Andersen-Ranberg NC, Weihe S, Poulsen LM, Mortensen CB, Kjer CKW, et al. Distribution of delirium motor subtypes in the intensive care unit: a systematic scoping review. Crit Care. Mar 03, 2022;26(1):53. PMID: 35241132. doi: 10.1186/s13054-022-03931-3.
40. van den Boogaard M, Pickkers P, Slooter AJ, Kuiper MA, Spronk PE, van der Voort PH, et al. Development and validation of PRE-DELIRIC (PREdiction of DELIRium in ICu patients) delirium prediction model for intensive care patients: observational multicentre study. BMJ. Feb 09, 2012;344:e420. PMID: 22323509. doi: 10.1136/bmj.e420.
41. Dos Santos FCM, Rego AS, Montenegro WS, de Carvalho S, Cutrim RC, Junior AAM, et al. Delirium in the intensive care unit: identifying difficulties in applying the Confusion Assessment Method for the Intensive Care Unit (CAM-ICU). BMC Nurs. Nov 23, 2022;21(1):323. PMID: 36419158. doi: 10.1186/s12912-022-01103-w.
42. Kotfis K, Zegan-Baranska M, Zukowski M, Kusza K, Kaczmarczyk M, Ely EW. Multicenter assessment of sedation and delirium practices in the intensive care units in Poland - is this common practice in Eastern Europe? BMC Anesthesiol. Sep 02, 2017;17(1):120. PMID: 28865447. doi: 10.1186/s12871-017-0415-2.
43. Green C, Bonavia W, Toh C, Tiruvoipati R. Prediction of ICU Delirium: Validation of Current Delirium Predictive Models in Routine Clinical Practice. Crit Care Med. Mar 2019;47(3):428-435. PMID: 30507844. doi: 10.1097/CCM.0000000000003577.
44. Song YX, Yang XD, Luo YG, Ouyang CL, Yu Y, Ma YL, et al. Comparison of logistic regression and

machine learning methods for predicting postoperative delirium in elderly patients: A retrospective study. CNS Neurosci Ther. Jan 2023;29(1):158-167. PMID: 36217732. doi: 10.1111/cns.13991.
45. Wassenaar A, Schoonhoven L, Devlin JW, van Haren FMP, Slooter AJC, Jorens PG, et al. Delirium prediction in the intensive care unit: comparison of two delirium prediction models. Crit Care. May 05, 2018;22(1):114. PMID: 29728150. doi: 10.1186/s13054-018-2037-6.
46. Tang D, Ma C, Xu Y. Interpretable machine learning model for early prediction of delirium in elderly patients following intensive care unit admission: a derivation and validation study. Front Med (Lausanne). 2024;11:1399848. PMID: 38828233. doi: 10.3389/fmed.2024.1399848.
47. Rudiger A, Begdeda H, Babic D, Kruger B, Seifert B, Schubert M, et al. Intra-operative events during cardiac surgery are risk factors for the development of delirium in the ICU. Crit Care. Aug 21, 2016;20:264. PMID: 27544077. doi: 10.1186/s13054-016-1445-8.
48. Young M, Holmes N, Kishore K, Marhoon N, Amjad S, Serpa-Neto A, et al. Natural language processing diagnosed behavioral disturbance vs confusion assessment method for the intensive care unit: prevalence, patient characteristics, overlap, and association with treatment and outcome. Intensive Care Med. May 2022;48(5):559-569. PMID: 35322288. doi: 10.1007/s00134-022-06650-z.
49. Favre E, Bernini A, Morelli P, Pasquier J, Miroz JP, Abed-Maillard S, et al. Neuromonitoring of delirium with quantitative pupillometry in sedated mechanically ventilated critically ill patients. Crit Care. Feb 24, 2020;24(1):66. PMID: 32093710. doi: 10.1186/s13054-020-2796-8.
50. Hur S, Ko RE, Yoo J, Ha J, Cha WC, Chung CR. A Machine Learning-Based Algorithm for the Prediction of Intensive Care Unit Delirium (PRIDE): Retrospective Study. JMIR Med Inform. Jul 26, 2021;9(7):e23401. PMID: 34309567. doi: 10.2196/23401.
51. Liu Z, Yang B. CTRP6(C1q/Tumor Necrosis Factor (TNF)-related protein-6) alleviated the sevoflurane induced injury of mice central nervous system by promoting the expression of p-Akt (phosphorylated Akt). Bioengineered. Dec 2021;12(1):5716-5726. PMID: 34516328. doi: 10.1080/21655979.2021.1967838.
52. Sadean MR, Glass PS. Pharmacokinetics in the elderly. Best Pract Res Clin Anaesthesiol. Jun 2003;17(2):191-205. PMID: 12817914. doi: 10.1016/s1521-6896(03)00002-8.
53. Smit L, Dijkstra-Kersten SMA, Zaal IJ, van der Jagt M, Slooter AJC. Haloperidol, clonidine and resolution of delirium in critically ill patients: a prospective cohort study. Intensive Care Med. Mar 2021;47(3):316-324. PMID: 33591422. doi: 10.1007/s00134-021-06355-9.
54. Lucini FR, Stelfox HT, Lee J. Deep Learning-Based Recurrent Delirium Prediction in Critically Ill Patients. Crit Care Med. Apr 01, 2023;51(4):492-502. PMID: 36790184. doi: 10.1097/CCM.0000000000005789.
55. Song Y, Zhang D, Wang Q, Liu Y, Chen K, Sun J, et al. Prediction models for postoperative delirium in elderly patients with machine-learning algorithms and SHapley Additive exPlanations. Transl Psychiatry. Jan 25, 2024;14(1):57. PMID: 38267405. doi: 10.1038/s41398-024-02762-w.
56. Aas K, Jullum M, Lland A. Explaining individual predictions when features are dependent: More accurate approximations to Shapley values. Artif Intell. Sep 01, 2021;298:103502. doi: 10.1016/j.artint.2021.103502.
57. Lundberg S, Lee SI. A unified approach to interpreting model predictions. NIPS. 2017;30 (0):4768-4777. doi: 10.5555/3295222.3295230.

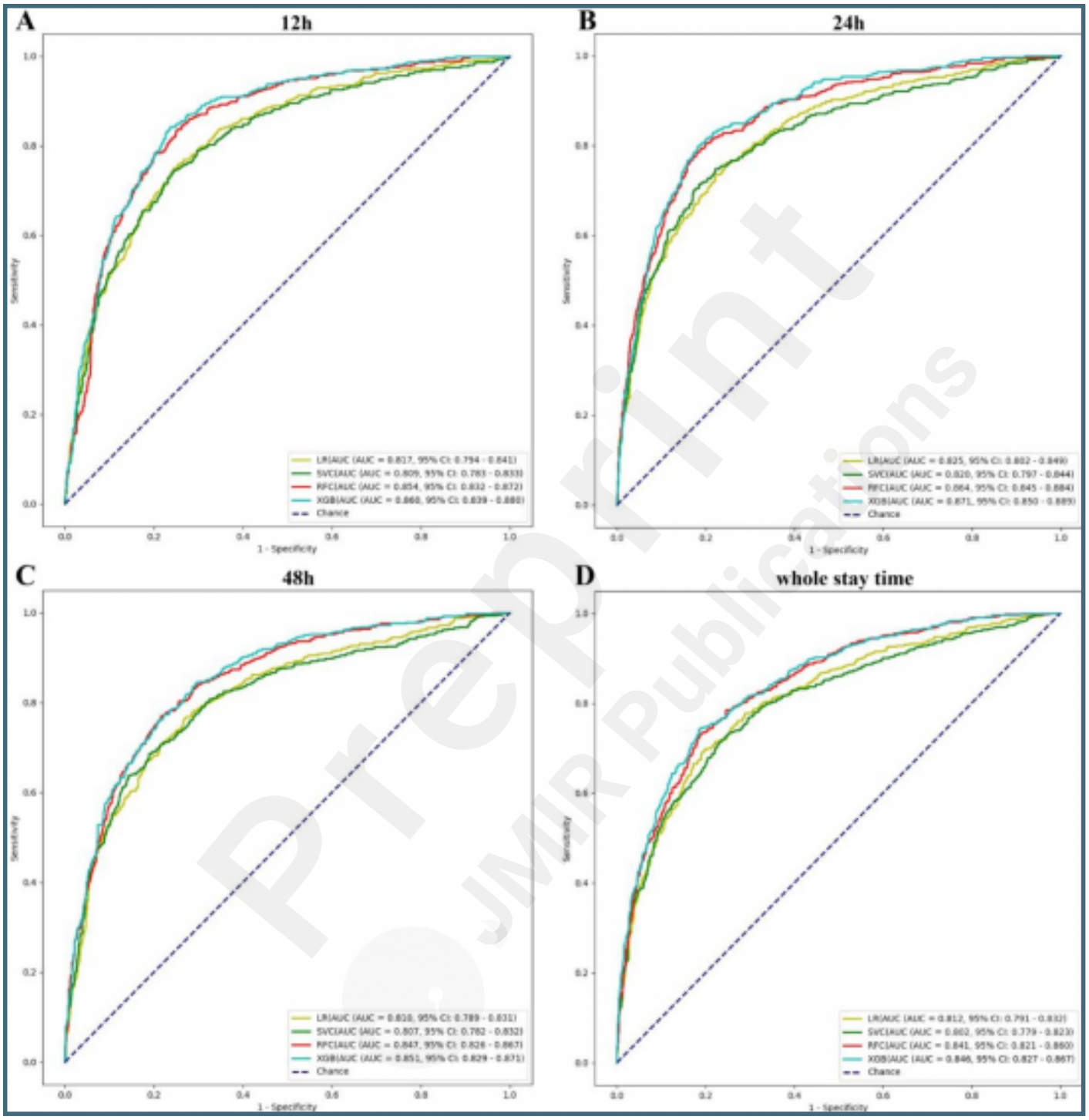
Supplementary Files

Figures

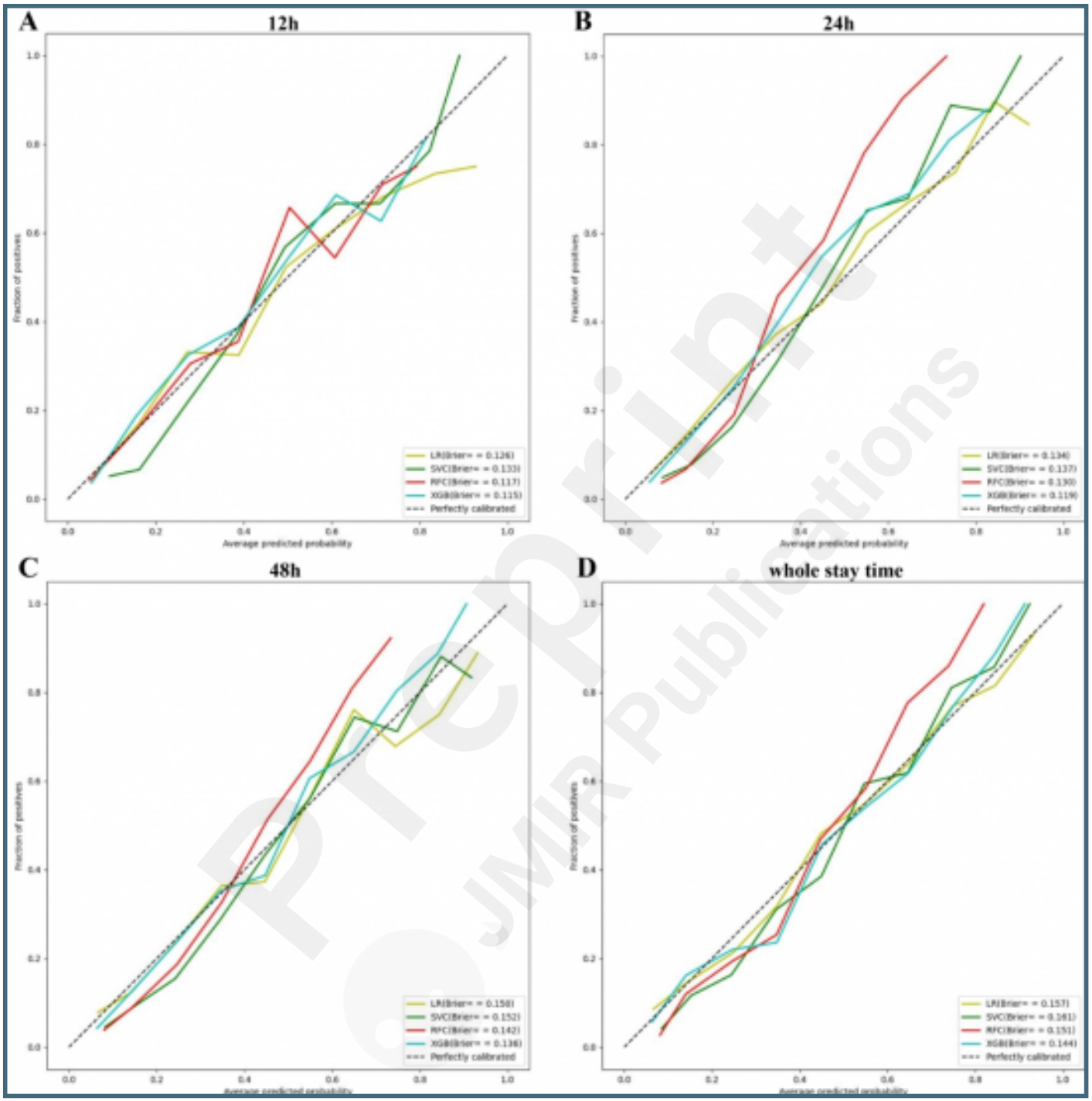
Cohort selection schema.



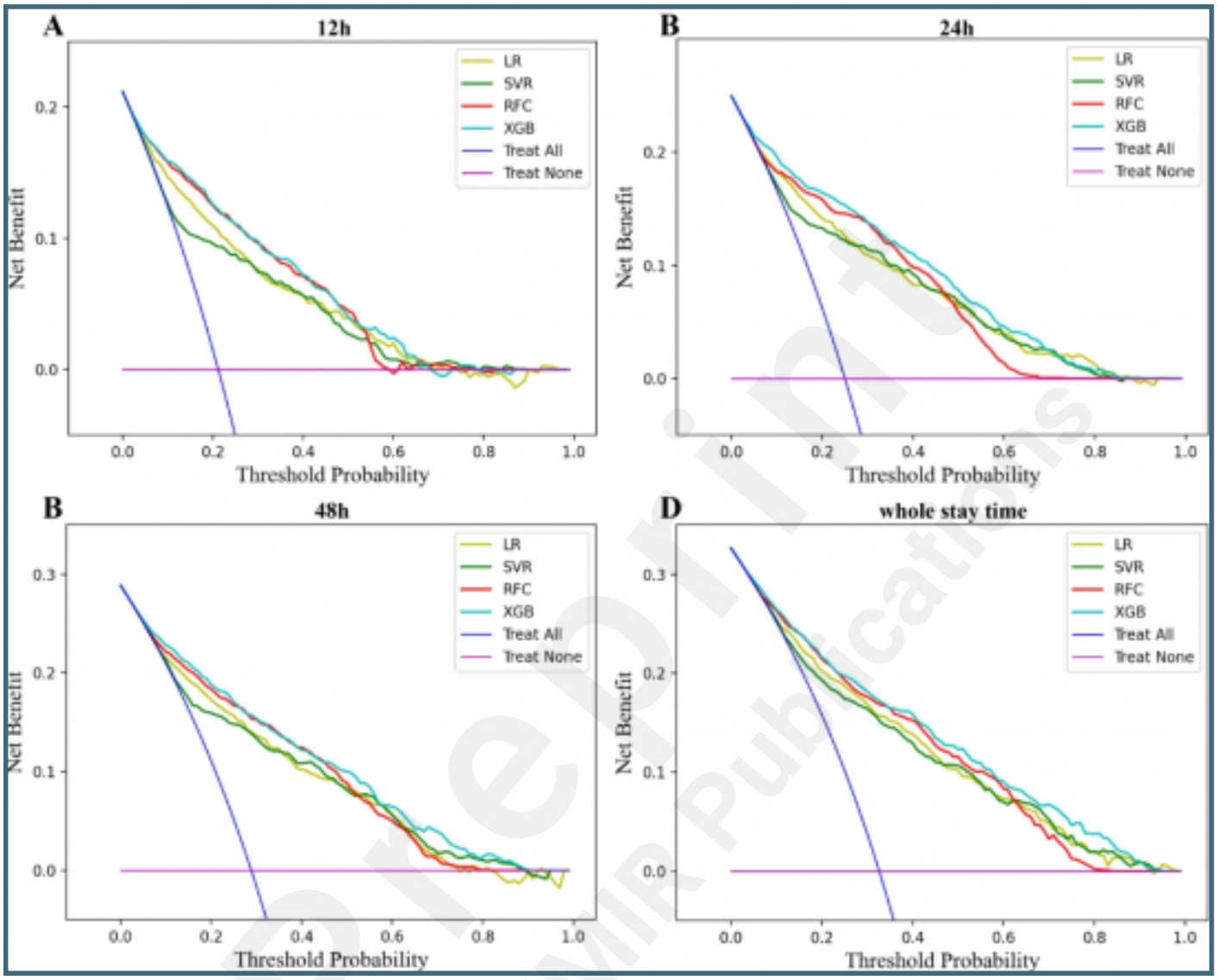
Receiver operating characteristic curves for all machine learning models in different prediction windows in the internal validation set.



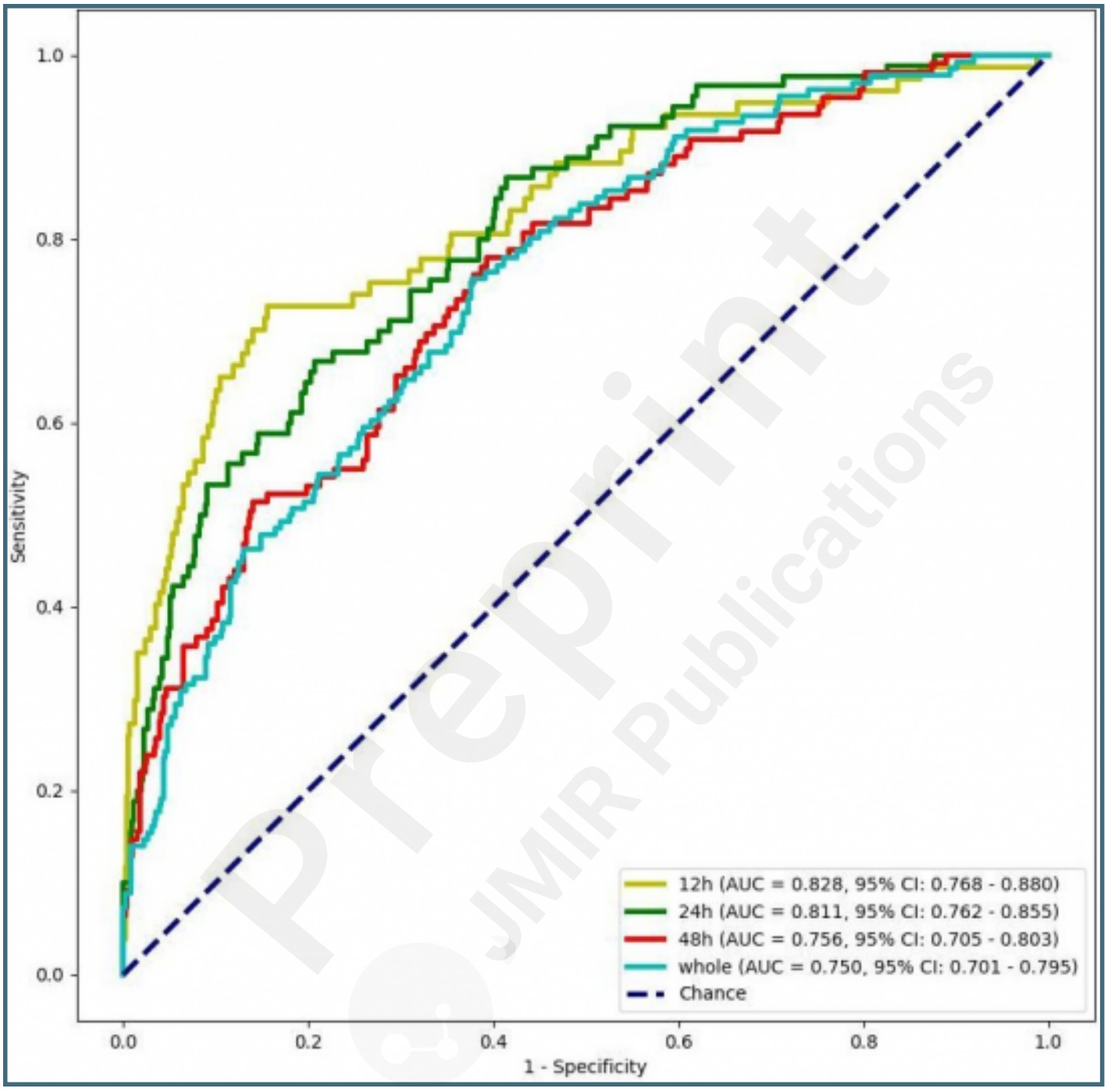
Calibration curves for all machine learning models in different prediction windows in the internal validation set.



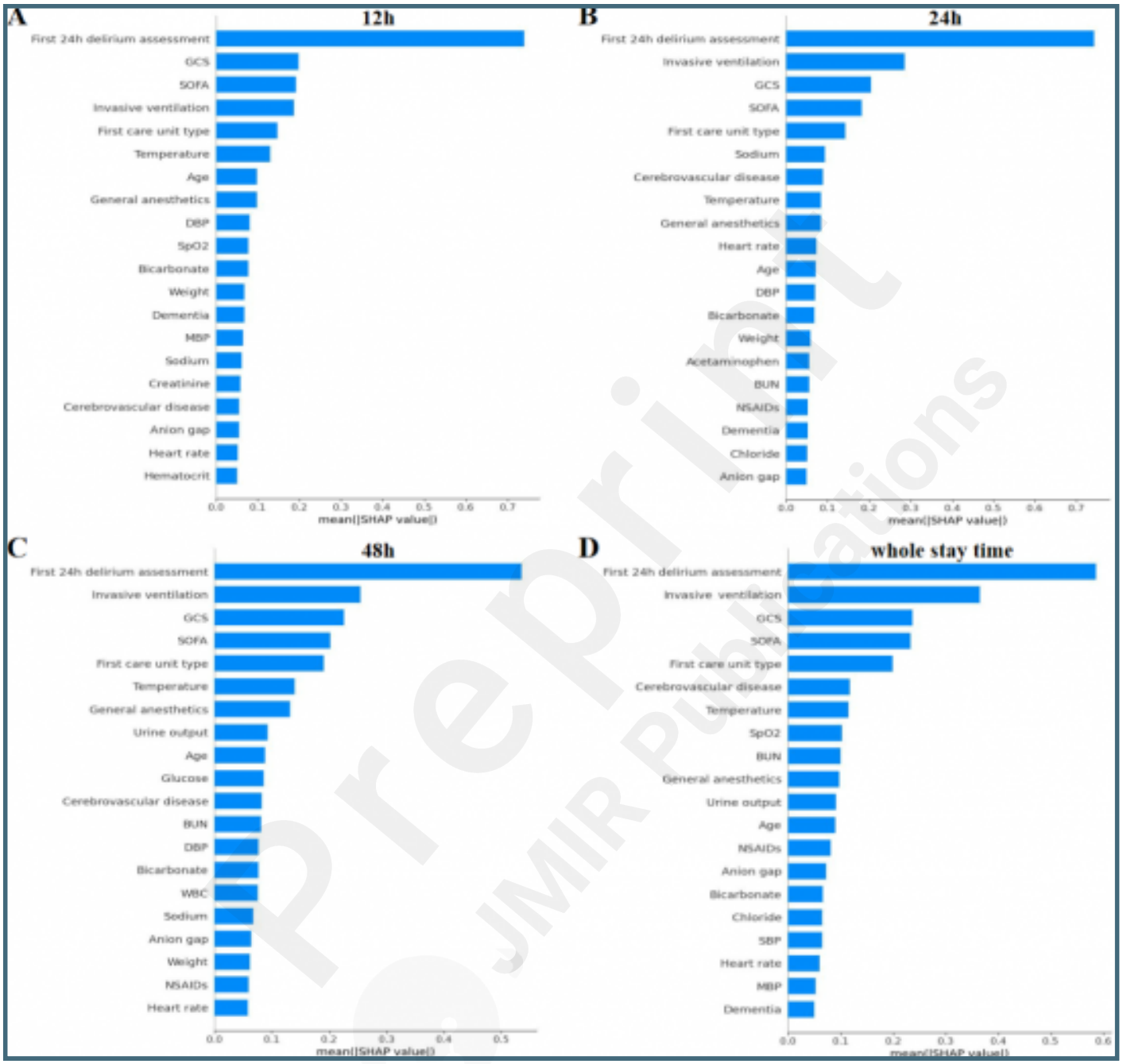
Decision curves for all machine learning models in different prediction windows in the internal validation set.



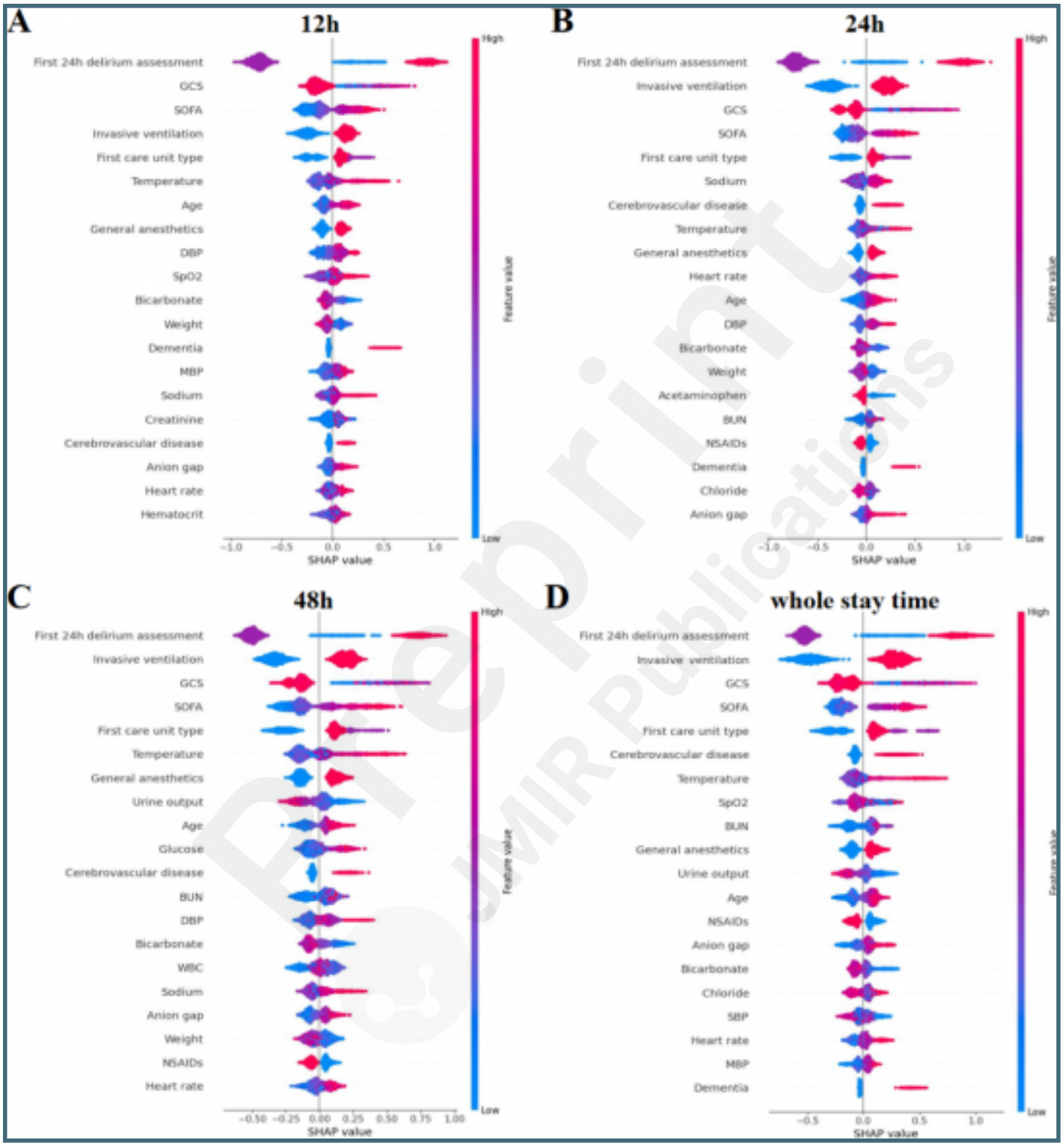
Receiver operating characteristic curves for extreme gradient boosting models in different prediction windows in the external validation set.



Feature importance ranking plot of the XGB machine learning models in different prediction windows (top 20 features).



Shapley Additive Explanations (SHAP) summary plots of the XGB machine learning models in different prediction windows (top 20 features).



Multimedia Appendixes

Baseline Characteristics of delirium and Non-delirium Patients in the 12h prediction window.
URL: http://asset.jmir.pub/assets/25718c2c35a5aaba36687369248aecde.doc

Baseline Characteristics of delirium and Non-delirium Patients in the 24h prediction window.
URL: http://asset.jmir.pub/assets/f9c0b0efa99658a1a37aebaa0b630f06.doc

Baseline Characteristics of delirium and Non-delirium Patients in the 48h prediction window.
URL: http://asset.jmir.pub/assets/c9ab98bf813b7cc1d4a45b10ece69508.doc

Confusion Matrix of XGB Models for Different Prediction windows in the internal validation set.
URL: http://asset.jmir.pub/assets/20ca513eec1ce3bada8d6ebee39eeb93.doc

Confusion Matrix of XGB Models for Different Prediction windows in the external validation set.
URL: http://asset.jmir.pub/assets/e5ce003bdf7faaa26592039af623720c.doc