

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

Disclaimer: © The authors. All rights reserved. This is a privileged document currently under peer-review/community review. Authors have provided JMIR Publications with an exclusive license to publish this preprint on its website for review purposes only. While the final peer-reviewed paper may be licensed under a CC BY license on publication, at this stage authors and publisher expressly prohibit redistribution of this draft paper other than for review purposes.

Table of Contents

Original Manuscript..... 5

Supplementary Files..... 24

 Figures 25

 Figure 1..... 26

 Figure 2..... 27

 Figure 3..... 28

 Figure 4..... 29

 Figure 5..... 30

 Figure 6..... 31

 Multimedia Appendixes 32

 Multimedia Appendix 0..... 33

 Multimedia Appendix 1..... 33

 Multimedia Appendix 2..... 33

 Multimedia Appendix 3..... 33

 Multimedia Appendix 4..... 33

 Multimedia Appendix 5..... 33

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 Lin¹ MMed; Jintao Duan⁵ BSc; Jing Huang⁶ BMed; Huixiu Hu¹ BMed; Ying Zhang¹ 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 visible to all users.

Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <https://preprints.jmir.org/preprint/67258>

No. Please do not make my accepted manuscript PDF available to anyone. I understand that if I later pay to participate in <https://preprints.jmir.org/preprint/67258>

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 Lin¹, MMed; Jintao Duan⁵, BSc; Jing Huang⁶, BMed; Huixiu Hu¹, BMed; Ying Zhang¹, BMed; Dengyun Xia^{7*}, MMed; Miao Zhou^{8*}, MD

¹Graduate School, Hebei North University, Zhangjiakou, Hebei, China.

²School of Anesthesiology, Naval Medical University, Shanghai, China.

³Information Center, The Second Affiliated Hospital of Naval Medical University, Shanghai, China.

⁴Department of Anesthesiology, Shanghai Ninth People's Hospital, Shanghai Jiao Tong University School of Medicine, Shanghai, China.

⁵School of Health Science and Engineering, University of Shanghai for Science and Technology, Shanghai, China.

⁶Graduate School, Wannan Medical College, Wuhu, Anhui, China.

⁷Department of Anesthesiology, The First Affiliated Hospital of Hebei North University, Zhangjiakou, Hebei, China.

⁸Department 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 state 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. Delirium predictions were 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, brier scores, decision curve analysis, and external validation.

Results

The MIMIC-IV and eICU-CRD datasets comprised 6129 and 709 patients, respectively, who were included in the analysis. A total of 54 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 0.848(95% CI: 0.826-0.869), 0.852(95% CI: 0.831-0.872), 0.851(95% CI: 0.831-0.871), and 0.844(95% CI: 0.823-0.863), respectively. The Area under the curve values for the external validation set were 0.777(95% CI: 0.726-0.825), 0.761(95% CI: 0.710-0.808), 0.753(95% CI: 0.704-0.798), and 0.737(95% CI: 0.695-0.777). Furthermore, the XGB model demonstrated the most accurate calibration across all prediction windows, with values of 0.129, 0.136, 0.144, and 0.148, 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 six most significant predictive features identified were the the first day's delirium assessment results, type of first care unit, minimum Glasgow Coma Scale score, Acute Physiology Score III, Acetaminophen and NSAIDs.

Conclusions

The high-performance XGB model for predicting postoperative delirium state in elderly ICU patients has been successfully developed and validated. The model predicts the delirium state 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 a common postoperative complication in elderly patients, characterized as an acute neuropsychiatric syndrome involving impairments in cognition, consciousness, attention, and psychomotor activity [1, 2]. This condition not only significantly prolongs hospital stays and increases healthcare costs, but is also closely associated with an increased incidence of postoperative complications and elevated mortality risk [3-7]. Of all hospital departments, intensive care unit (ICU) have the highest incidence of delirium, with rates of up to 80% in ICUs compared to 14-29% in non-ICU inpatients [8]. This disparity may be closely associated with the severe physiological disturbances and polypharmacy commonly present in ICU patients [9, 10]. Additionally, in elderly patients, the reduced ability to metabolize systemic anesthetics postoperatively leads to drug accumulation, which is significantly associated with an increased risk of ICU delirium [11, 12]. Fortunately, approximately 30% to 40% of cases can demonstrate significant improvement through preventive intervention [13-15]. It is therefore imperative that delirium is predicted at the earliest opportunity upon admission to the ICU, particularly in elderly patients.

Machine learning is an artificial intelligence technique that can process a substantial number of variables in a non-linear and highly interactive manner [16]. This technology autonomously uncovers inherent patterns within data to construct adaptive models, thereby enabling computer systems to achieve intelligent decision-making and accurate predictions without explicit programming instructions [17, 18]. Machine learning has been applied in the medical field for diagnosing diseases, recognizing medical images, providing treatment strategies and predicting outcomes [19]. To date, multiple studies have developed risk prediction models for postoperative delirium in elderly patients primarily based on preoperative and intraoperative parameters [20-23]. Compared with traditional assessment methods relying on clinical symptoms, these models demonstrate superior predictive performance. However, it is noteworthy that existing studies suffer from limitations such as restricted sample sizes or lack of external validation. Their clinical applicability and research quality still require further validation through multicenter, large-sample cohort studies.

In this study, we developed machine learning models to predict subsequent delirium status in elderly postoperative patients based on first-day characteristic data after ICU admission, comprehensively collected from the Medical Information Mart for Intensive Care IV (MIMIC-IV) database and the eICU Collaborative

Research Database (eICU-CRD). Unlike previous studies, this research innovatively established a prediction model based on real clinical scenarios. By retaining subjects with positive delirium assessment records during the baseline period and incorporating delirium evaluation results monitored within 24 hours of ICU admission as model features, the study effectively addresses the modeling limitations of traditional research regarding delirium state transitions (including complex clinical processes such as initial onset, recurrence, prolongation, and remission). In addition, the study adopted a multi-time window prediction framework to systematically evaluate the model's predictive performance for both short-term and long-term outcomes, as well as variations in delirium status. This study aims to provide evidence-based support to identify precise intervention windows and optimize personalized treatment plans for postoperative delirium, with the ultimate goal of improving patient outcomes.

Methods

Ethical review

The MIMIC-IV database and the eICU-CRD are publicly accessible and have been approved for use by the Institutional Review Boards of Beth Israel Deaconess Medical Centre and Massachusetts Institute of Technology, in accordance with the Declaration of Helsinki. Both databases received waivers for informed consent due to the deidentification of all protected health information. After passed the training course CITI Data or Specimens Only Research(certification number: 50199917), we had the access to extract data from both the MIMIC-IV and the eICU databases. Our study followed the Transparent Reporting of Multivariate Predictive Models for Individual Prognosis or Diagnosis statement [24].

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 [25]. 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 [26]. The study population comprised two cohorts: (1) patients aged ≥ 65 years with initial postoperative ICU admissions from the MIMIC-IV database, and (2) Patients aged ≥ 65 years with surgery-related diagnoses in the eICU Collaborative Research Database. To comprehensively and accurately include postoperative patients, we linked the surgical records from the `procedures_icd` table in the MIMIC database to each patient's first ICU admission record following their initial surgery. Since only the surgical date (without specific timestamps) was available and considering some surgeries might extend into the next day, we ultimately retained patients admitted to the ICU on both the day of the surgery and the following day to ensure clinical relevance between ICU data and surgical procedures. The inclusion criteria cover over 85% of patients admitted to the ICU for the first time postoperatively. In the eICU database, we included all the patient diagnosis records containing surgery-related entries, because a large number of postoperative patients did not have the corresponding postoperative diagnosis codes. Patients were excluded based on the following criteria: (1) ICU length of stay < 24 hours; (2) absence of valid delirium assessments in the corresponding observation and prediction windows; and (3) delirium assessment results in contradiction to diagnosis records and nursing note. For eICU data, after the initial screening, we further cleaned the non-surgical patients in the limited data. No such patients were found in the data of this study. In our study, we retained patients who experienced delirium on the first day of ICU admission to assess their prognosis or recurrence risk, with no restrictions imposed on prior history of delirium before ICU admission. In total, 6129 and 709 patients from the two databases were included in this study. The detailed flowchart illustrating the inclusion and exclusion criteria is presented in Figure 1.

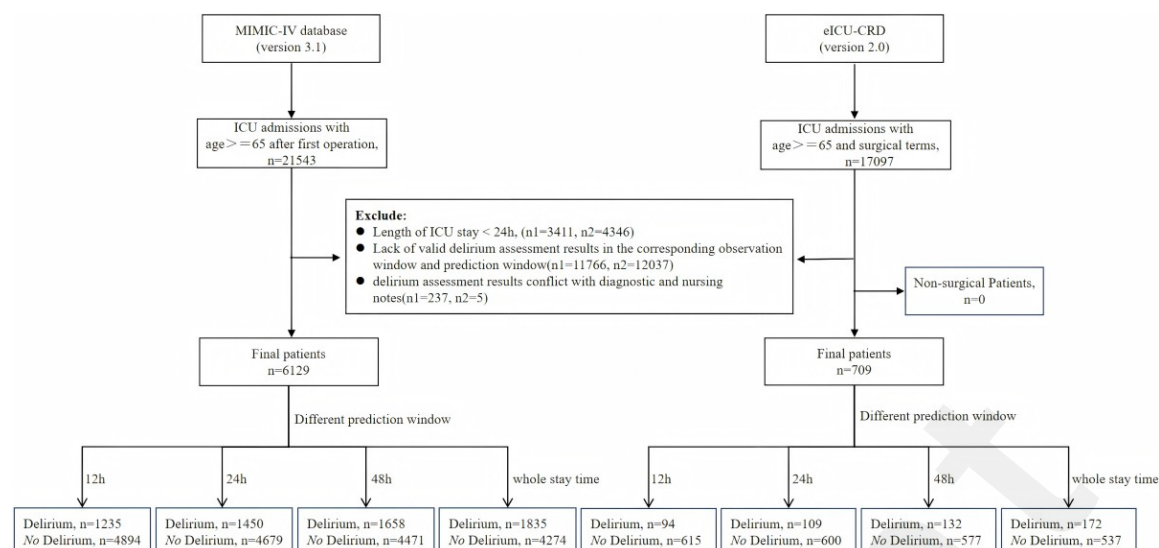


Figure 1. Cohort selection schema. MIMIC-IV: Medical Information Marketplace for Intensive Care IV; eICU-CRD: eICU Collaborative Research Database; ICU: intensive care unit

Delirium assessment

The primary outcome variable of this study was delirium, defined as a positive Confusion Assessment Method for the ICU screen or a score of 4 or more on the Intensive Care Delirium Screening Checklist, without any contradictions from diagnostic code information or nursing note information. 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. Specifically, we designated the first day following ICU admission as the observation window to predict delirium incidence during four distinct prediction windows: the subsequent 12, 24, 48 hours, and whole ICU stay time. Delirium was diagnosed if at least one delirium assessment 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 54 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 (DBP), mean blood pressure (MBP), temperature, respiratory rate, oxygen saturation

Laboratory results

- Hematocrit, hemoglobin, platelets, white blood cell, anion gap, calcium, blood urea nitrogen, chloride, creatinine, glucose, sodium, potassium, international normalized ratio (INR), prothrombin time (PT), partial thromboplastin time, urine output

Comorbidity

- Hypertension, diabetes, congestive heart failure, Chronic liver disease, chronic renal disease, chronic pulmonary disease, peptic ulcer disease, tumor, dementia.

Score

- Sequential Organ Failure Assessment (SOFA), Glasgow Coma Scale (GCS), Acute Physiology Score III (APSOIII)

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 (e.g., height was excluded due to its 34.64% missing rate in our MIMIC-IV dataset). Since patient ages above 90 years in the eICU database were masked to protect privacy, we imputed ages for these patients within the 90~100 range based on the age distribution trend of patients aged 80~89 in the database. Missing values for other features were addressed using multiple imputation[27]. Additionally, the data were shuffled to adjust the order of the samples.

Data Balancing

The extracted data exhibited a substantial class imbalance, with delirium-negative cases being significantly outnumbered by positive instances, leading the learning model to demonstrate a bias toward the negative class. To address this, we applied the Synthetic Minority Over-sampling Technique to generate synthetic positive samples, effectively balancing the class distribution and improving model generalizability [28].

Feature Selection

We initially performed multicollinearity analysis using Variance Inflation Factor, identifying multiple highly collinear feature groups (e.g., INR vs PT, MBP vs DBP, hematocrit vs hemoglobin). Subsequently, feature importance evaluation was conducted through a RFC tree-based model, which enabled the removal of less important features within each collinear group.

To further refine the feature set, we implemented the Boruta feature selection algorithm. 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 [29]. 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.

Brier Score is a performance metric primarily used to assess the accuracy and calibration of probabilistic models in binary classification tasks [30]. It measures the mean squared error between the predicted probabilities and the actual outcomes, with smaller values indicating more accurate predictions. A perfect score of 0 indicates ideal calibration where predictions exactly match reality, while higher scores (closer to the maximum value of 1) correspond to poorer performance. 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 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 χ^2 test. Normally distributed continuous variables are presented as mean $\pm$ SD, and group comparisons were analyzed with the Student's 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 6129 patients from the MIMIC-IV dataset, of whom 1855 (30.3%) were assessed as delirium during the remaining stay after the first day in the ICU. Additionally, 709 patients from the eICU-CRD database were included, of whom 172 (24.3%) 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=4274)	Delirium (n=1835)	P Value	No Delirium (n=537)	Delirium (n=172)	P Value
Demographic data						
Age, year, median (IQR)	75.0(69.0,82.0)	77.0(71.0,84.0)	<.001	74.0(69.0,80.0)	76.5(71.0,82.0)	.009
Gender, male, n (%)	2404.0(56.2%)	960.0(51.8%)	<.001	266.0(49.5%)	98.0(57.0%)	.09
Weight, kg, median (IQR)	78.6 (66.7, 91.6)	75.5(63.3, 89.1)	<.001	77.6 (65.6, 91.0)	75.6 (64.9, 91.3)	.79
Race, n (%)			<.001			.27
Black	296.0(6.9%)	179.0(9.6%)		59.0(11.0%)	27.0(15.7%)	
White	3073.0(71.9%)	1174.0(63.3%)		429.0(79.9%)	130.0(75.6%)	
Asian	106.0(2.5%)	29.0(1.6%)		2.0(0.4%)	1.0(0.6%)	
Hispanic	77.0(1.8%)	35.0(1.9%)		23.0(4.3%)	4.0(2.3%)	
Other/Unknown	722.0(16.9%)	438.0(23.6%)		24.0(4.5%)	10.0(5.8%)	
First care unit type, n (%)			<.001			.11
Cardiovascular ICU ^c	1912.0(44.7%)	400.0(21.6%)		109.0(20.3%)	41.0(23.8%)	
Neurological ICU	423.0(9.9%)	283.0(15.3%)		82.0(15.3%)	35.0(20.3%)	
Other ICU	1939.0(45.4%)	1172.0(63.2%)		346.0(64.4%)	96.0(55.8%)	
First 24h delirium assessment, n (%)			<.001			<.001
Negative	3640.0(85.2%)	677.0(36.5%)		500.0(93.1%)	100.0(58.1%)	
Positive	634.0(14.8%)	1178.0(63.5%)		37.0(6.9%)	71.0(41.3%)	
Vital signs, median (IQR)						
Heart rate, beats/min	79.4(71.1,89.0)	82.4(73.1,94.2)	<.001	83.0(74.4,91.4)	85.7(77.5,97.0)	.005
Systolic blood pressure, mmHg	115.5(106.8,126.6)	115.7(106.4,127.3)	.47	118.4(108.0,131.3)	119.2(106.9,129.1)	.51
Diastolic blood pressure, mmHg	58.7(53.0,65.6)	59.6(53.7,66.2)	.003	60.8(55.5,67.2)	60.6(55.3,66.8)	.93
Mean blood pressure, mmHg	75.2(69.9,82.1)	75.8(70.4,82.8)	.02	78.0(70.5,85.0)	76.9(69.8,85.2)	.46
Respiratory rate, beats/min	18.2(16.4,20.3)	18.9(16.9,21.4)	<.001	17.6(15.7,19.9)	18.0(15.7,21.1)	.20
Temperature, °C	36.8(36.6,37.0)	36.9(36.7, 37.2)	<.001	36.8(36.6,37.1)	36.8(36.6,37.1)	.46
Oxygen saturation, %	97.1(95.8,98.3)	97.5(96.0,98.8)	<.001	97.3(95.8,98.4)	97.4(96.1,98.6)	.23
Laboratory results, median (IQR)						
Hematocrit, %	31.5(28.0,35.5)	31.9(27.6,36.0)	.31	31.3(28.0,34.5)	30.0(26.2,33.7)	.03
Hemoglobin, g/dL	10.3 (9.1, 11.7)	10.4 (8.9, 11.7)	.21	10.4 (9.2, 11.5)	9.9(8.6,11.2)	.008
Platelet, 10 ⁹ /L	174.3(132.3,228.0)	180.0(133.3,243.0)	.02	187.6(142.0,234.0)	176.0(124.3,248.8)	.48
White blood cell, 10 ⁹ /L	11.1(8.4,14.8)	11.9(9.0,15.7)	<.001	11.7(9.2,15.2)	12.2(9.3,16.4)	.34
Anion gap, mmol/L	13.0(11.0,15.5)	14.3(12.3,17.0)	<.001	10.5(8.0,12.8)	11.0(8.4,14.2)	.06
Blood urea nitrogen, mg/dL	19.0(14.0,29.0)	24.0(16.5,39.3)	<.001	19.0(13.0,27.0)	22.0(14.8,35.8)	<.001
Calcium, mg/dL	8.3(8.0,8.7)	8.3(7.9,8.7)	.91	8.2(7.8,8.5)	8.2(7.7,8.7)	.77
Chloride, mmol/L	105.0(101.3,107.5)	105.0(101.0,108.0)	.36	104.5(102.0,108.0)	106.0(103.0,110.0)	<.001
Creatinine, mg/dL	1.0(0.8,1.4)	1.1(0.8, 1.7)	<.001	1.0(0.7,1.4)	1.2(0.9,1.7)	<.001
Glucose, mg/dL	127.0(108.7,150.0)	135.0(111.3,170.8)	<.001	139.8(118.0,162.0)	137.0(117.4,159.7)	.44
Sodium, mmol/L	138.3(136.0,140.3)	139.0(136.0,142.0)	<.001	138.3(136.0,140.5)	140.0(137.0,142.4)	<.001
Potassium, mmol/L	4.2(3.9,4.5)	4.2(3.9,4.5)	.56	4.2(3.9,4.6)	4.1(3.8,4.5)	.23
International normalized ratio	1.3(1.2,1.4)	1.3(1.2,1.5)	<.001	1.4(1.2,1.6)	1.4(1.2,1.7)	.15
Prothrombin time, s	14.0(12.5,15.4)	14.4(12.5,16.2)	<.001	16.1(14.3,18.3)	16.5(13.7,19.8)	.16
partial thromboplastin time, s	31.2(27.8,38.0)	31.8(27.7,39.3)	.20	35.5(35.4,35.6)	35.5(35.3,35.6)	.04
Urine output, ml	1485.0(993.0, 2155.0)	1225.0(745.0,1845.0)	<.001	1337.7(875.0,1775.0)	1096.5(650.0,1822.5)	.02

Comorbidity, n (%)						
Hypertension	3355.0(78.5%)	1482.0(79.9%)	.22	94.0(17.5%)	36.0(20.9%)	.31
Diabetes	1,390.0(32.5%)	686.0(37.0%)	<.001	70.0(13.0%)	26.0(15.1%)	.49
Congestive heart failure	1,434.0(33.6%)	750.0(40.4%)	<.001	43.0(8.0%)	13.0(7.6%)	.85
Chronic renal disease	1,043.0(24.4%)	575.0(31.0%)	<.001	39.0(7.3%)	19.0(11.0%)	.12
Chronic liver disease	343.0(8.0%)	169.0(9.1%)	.16	4.0(0.7%)	3.0(1.7%)	.37
Chronic pulmonary disease	1,097.0(25.7%)	571.0(30.8%)	<.001	45.0(8.4%)	15.0(8.7%)	.89
Peptic ulcer disease	123.0(2.9%)	73.0(3.9%)	.03	4.0(0.7%)	1.0(0.6%)	0.99
Tumor	714.0(16.7%)	272.0(14.7%)	.05	112.0(20.9%)	32.0(18.6%)	.52
Dementia	106.0(2.5%)	236.0(12.7%)	<.001	6.0(1.1%)	5.0(2.9%)	.12
Score, median (IQR)						
GCS ^d	15.0 (14.0, 15.0)	14.0(12.0, 15.0)	<.001	14.0(11.0, 15.0)	13.0 (8.0, 14.0)	<.001
SOFA ^e	4.0 (2.0, 6.0)	6.0 (3.0, 8.0)	<.001	5.0 (4.0, 7.0)	7.0 (5.0, 9.0)	<.001
APSI ^f	38.0 (30.0, 48.0)	49.0 (38.0, 63.0)	<.001	40.0 (30.0, 52.0)	49.5 (35.0, 69.5)	<.001
Treatment measures, n (%)						
Renal replacement therapy	130.0(3.0%)	100.0(5.4%)	<.001	18.0(3.4%)	6.0(3.5%)	.93
Invasive ventilation	1695.0(39.7%)	1190.0(64.2%)	<.001	199.0(37.1%)	82.0(47.7%)	.01
Acetaminophen	3338.0(78.1%)	1224.0(66.0%)	<.001	270.0(50.3%)	100.0(58.1%)	.07
Anticholinergics	1332.0(31.2%)	582.0(31.4%)	.87	33.0(6.1%)	22.0(12.8%)	.005
Anticoagulants	2759.0(64.6%)	1327.0(71.5%)	<.001	197.0(36.7%)	59.0(34.3%)	.57
Antihistamines	234.0(5.5%)	70.0(3.8%)	.005	34.0(6.3%)	8.0(4.7%)	.42
Antipsychotics	157.0(3.7%)	204.0(11.0%)	<.001	6.0(1.1%)	5.0(2.9%)	.15
Benzodiazepines	650.0(15.2%)	290.0(15.6%)	.67	87.0(16.2%)	36.0(20.9%)	.15
Diuretics	1739.0(40.7%)	713.0(38.4%)	.10	130.0(24.2%)	42.0(24.4%)	.96
General anesthetics	1576.0(36.9%)	1021.0(55.0%)	<.001	68.0(12.7%)	32.0(18.6%)	.05
NSAIDs ^g	2259.0(52.9%)	692.0(37.3%)	<.001	105.0(19.6%)	43.0(25.0%)	.13
Opioids	3672.0(85.9%)	1645.0(88.7%)	.003	276.0(51.4%)	99.0(57.6%)	.16
Vasopressors	2087.0(48.8%)	1004.0(54.1%)	<.001	87.0(16.2%)	48.0(27.9%)	<.001

^aMIMIC-IV: Medical Information Marketplace for Intensive Care IV.

^beICU-CRD: eICU Collaborative Research Database.

^cICU: intensive care unit.

^dGCS: Glasgow Coma Scale.

^eSOFA: Sequential Organ Failure Assessment.

^fAPSI: Acute Physiology Score III.

^gNSAIDs: 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 0.848(95% CI: 0.826-0.869), 0.852(95% CI: 0.831-0.872), 0.851(95% CI: 0.831-0.871), and 0.844(95% CI: 0.823-0.863), 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 0.846(95% CI: 0.824-0.869), 0.844(95% CI: 0.821-0.865), 0.844(95% CI: 0.822-0.863), and 0.839(95% CI: 0.820-0.858), respectively. Overall, both two models maintain a high performance in different prediction windows. 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 Multimedia Appendix 4.

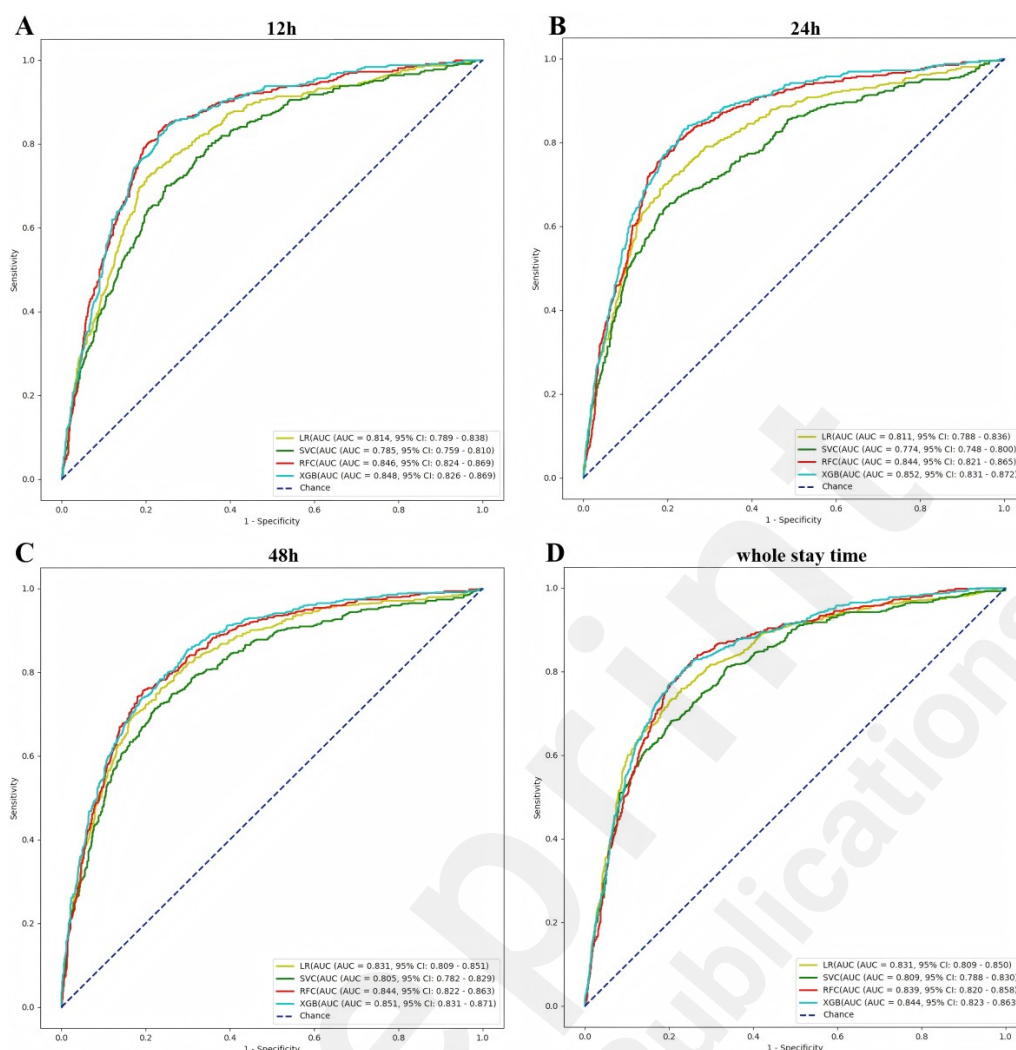


Figure 2. Receiver operating characteristic curves for all machine learning models in different prediction windows in the internal validation set. LR: logistic regression; SVC: support vector classifier; RFC: random forest classifier; XGB: extreme gradient boosting.

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	0.811 (0.793-0.829)	0.674 (0.626-0.722)	0.845 (0.827-0.864)	0.524 (0.479-0.569)	0.911 (0.896-0.926)	0.848 (0.826-0.869)
24h	0.809 (0.791-0.827)	0.703 (0.667-0.746)	0.842 (0.823-0.861)	0.580 (0.537-0.622)	0.902 (0.885-0.918)	0.852 (0.831-0.872)
48h	0.795 (0.777-0.813)	0.708 (0.668-0.748)	0.827 (0.807-0.847)	0.603 (0.563-0.642)	0.884 (0.867-0.902)	0.851 (0.831-0.871)
whole stay time	0.796 (0.778-0.814)	0.729 (0.692-0.766)	0.825 (0.804-0.846)	0.644 (0.607-0.682)	0.875 (0.856-0.894)	0.844 (0.823-0.863)

^aPPV: positive predictive value.

^bNPV: negative predictive value.

^cAUC: area under the curve.

We further evaluated the calibration performance of the model predictions by Brier score. As illustrated in Table 3, the XGB model demonstrates the best performance. The Brier scores for the XGB model in predicting delirium across different windows are 0.129, 0.136, 0.144, and 0.148, respectively, substantiating the reliability of our model. A decision curve is a tool used to evaluate the performance of a predictive model under different thresholds. As illustrated in Figure 3, 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.

Table 3. Brier scores for all machine learning models in different prediction windows in the internal validation

set.

Prediction window	LR ^a	SVC ^b	RFC ^c	XGB ^d
12h	0.135	0.159	0.134	0.129
24h	0.144	0.171	0.142	0.136
48h	0.146	0.169	0.148	0.144
whole stay time	0.151	0.175	0.156	0.148

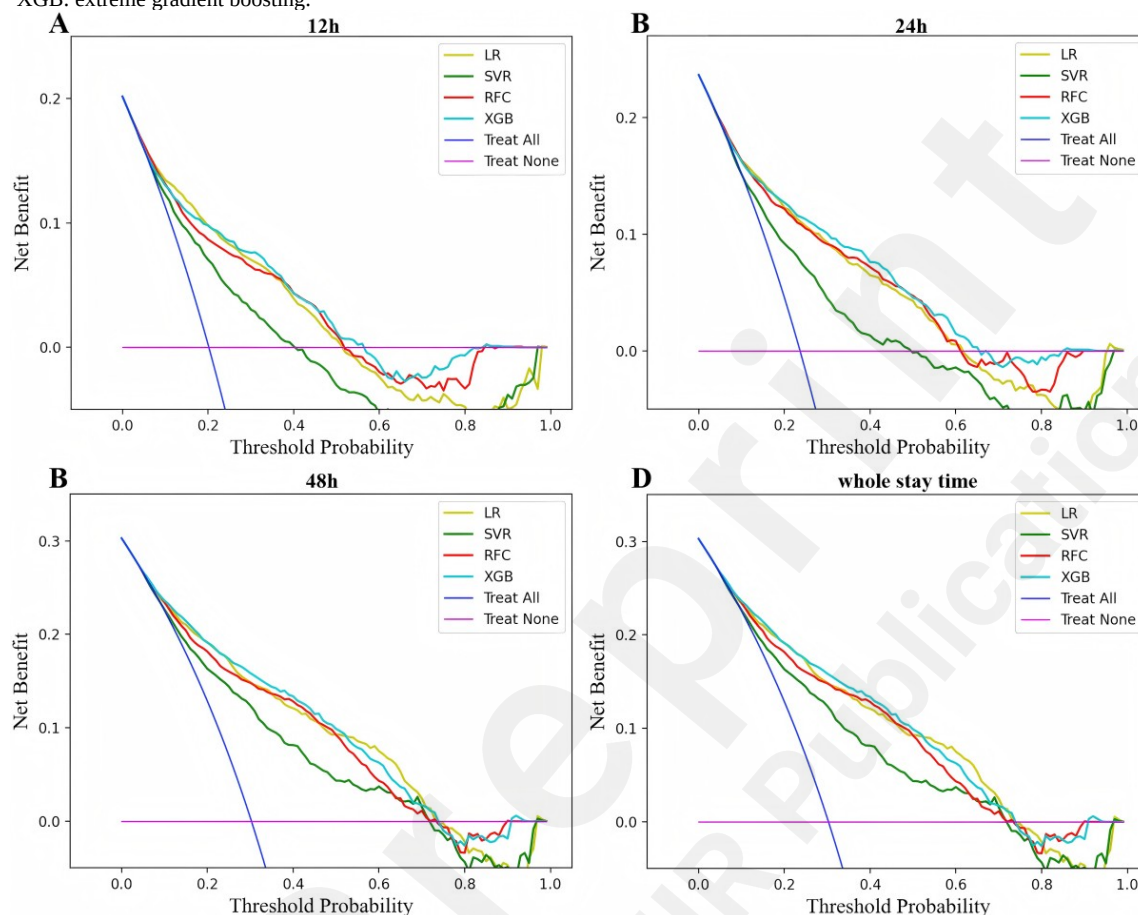
^aLR: logistic regression.^bSVC: support vector classifier.^cRFC: random forest classifier.^dXGB: extreme gradient boosting.

Figure 3.

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 4), the XGB model continues to demonstrate robust performance. The AUC values for the four prediction windows were 0.777(95% CI: 0.726-0.825), 0.761(95% CI: 0.710-0.808), 0.753(95% CI: 0.704-0.898), and 0.737(95% CI: 0.695-0.777), respectively. The comprehensive performance of the XGB model on the external validation set is presented in Table 4, and the associated confusion matrix plots are provided in Multimedia Appendix 5.

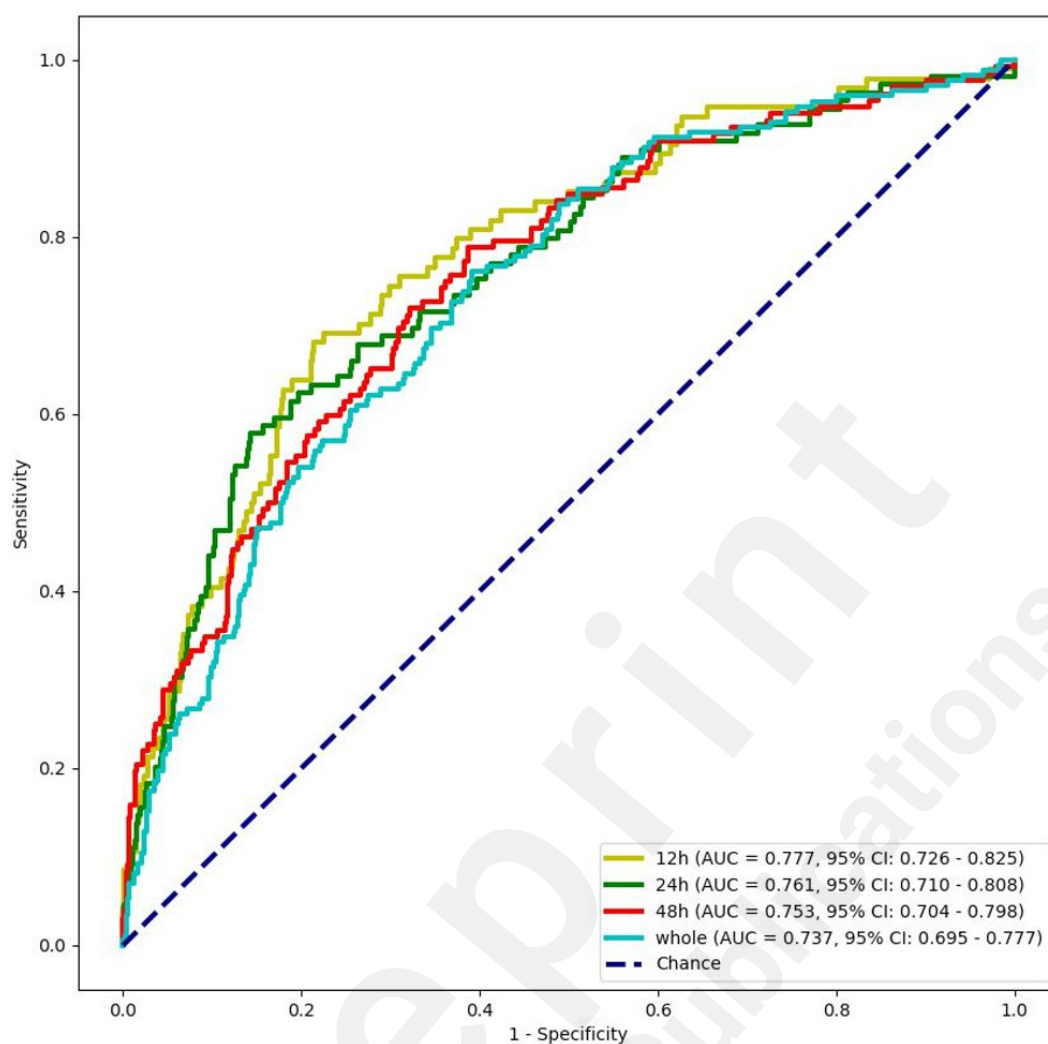


Figure 4. Receiver operating characteristic curves for extreme gradient boosting models in different prediction windows in the external validation set.

Table 4. 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	0.766 (0.735-0.797)	0.691 (0.598-0.785)	0.777 (0.744-0.810)	0.322 (0.257-0.386)	0.943 (0.923-0.963)	0.777 (0.726-0.825)
24h	0.745 (0.841-0.894)	0.670 (0.581-0.758)	0.758 (0.724-0.793)	0.335 (0.272-0.398)	0.927 (0.904-0.950)	0.761 (0.710-0.808)
48h	0.728 (0.695-0.761)	0.606 (0.523-0.689)	0.756 (0.721-0.791)	0.362 (0.299-0.425)	0.893 (0.866-0.921)	0.753 (0.704-0.898)
whole stay time	0.705 (0.672-0.739)	0.599 (0.526-0.672)	0.739 (0.702-0.776)	0.424 (0.362-0.486)	0.852 (0.820-0.884)	0.737 (0.695-0.777)

^aPPV: positive predictive value.

^bNPV: negative predictive value.

^cAUC: 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 5). 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, type of first care unit, minimum GCS score, APSIII, Acetaminophen, NSAIDs. Furthermore, invasive ventilation, SOFA score, mean body temperature, age, tumor and some laboratory metrics were also identified as relatively high-ranking features. The SHAP summary plot

(Figure 6) complements the above ranking by illustrating the impact of each feature on the model output.

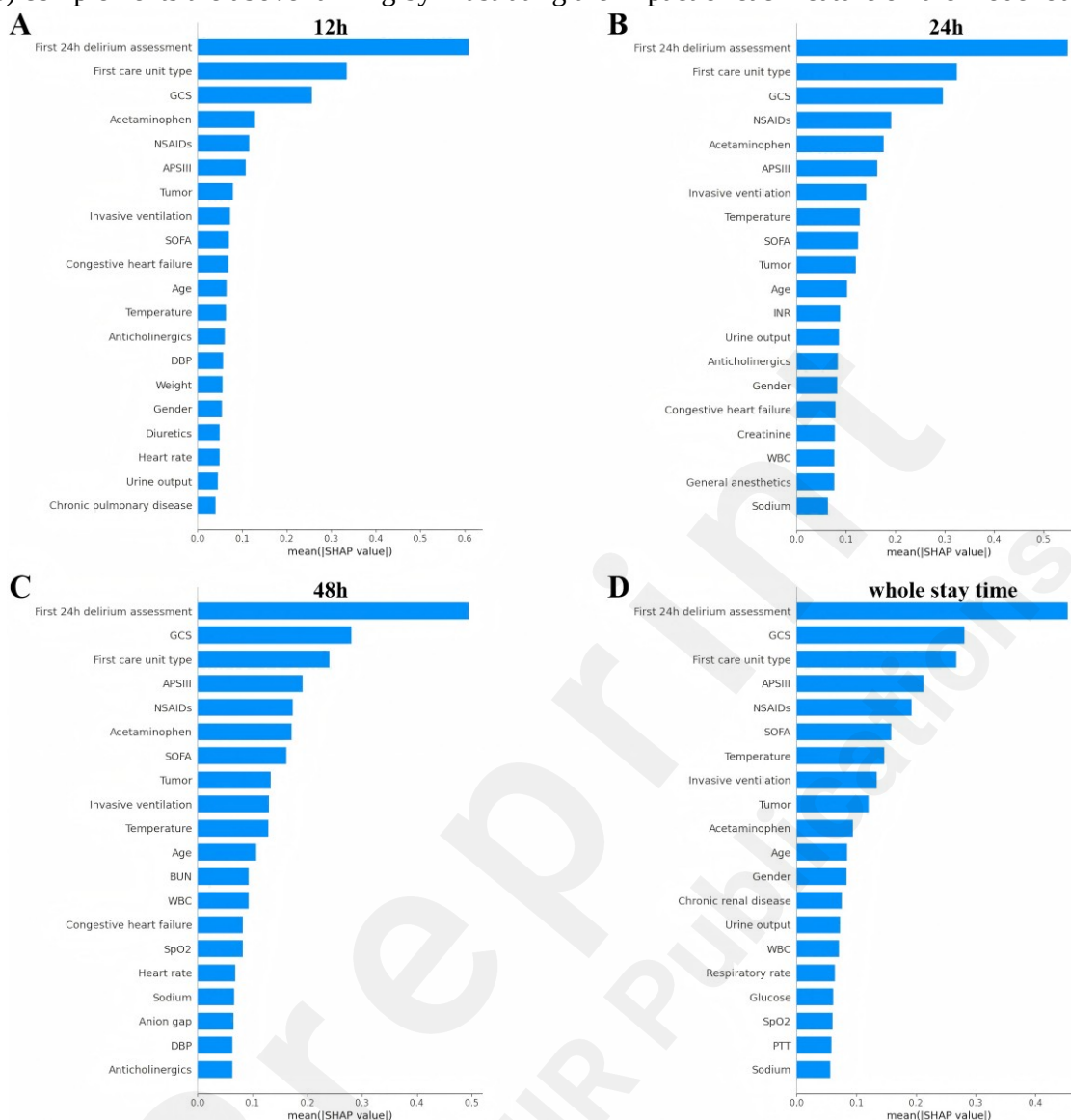


Figure 5. Feature importance ranking plot of the XGB machine learning models in different prediction windows (top 20 features). A, B, C, and D correspond to prediction windows of 12h, 24h, 48h, and whole stay time, respectively. GCS: Glasgow Coma Scale; SOFA: Sequential Organ Failure Assessment; APSI: Acute Physiology Score III; DBP: diastolic blood pressure; BUN: blood urea nitrogen; NSAIDs: Nonsteroidal Antiinflammatory Drugs; WBC: white blood cell; SpO2: oxygen saturation.

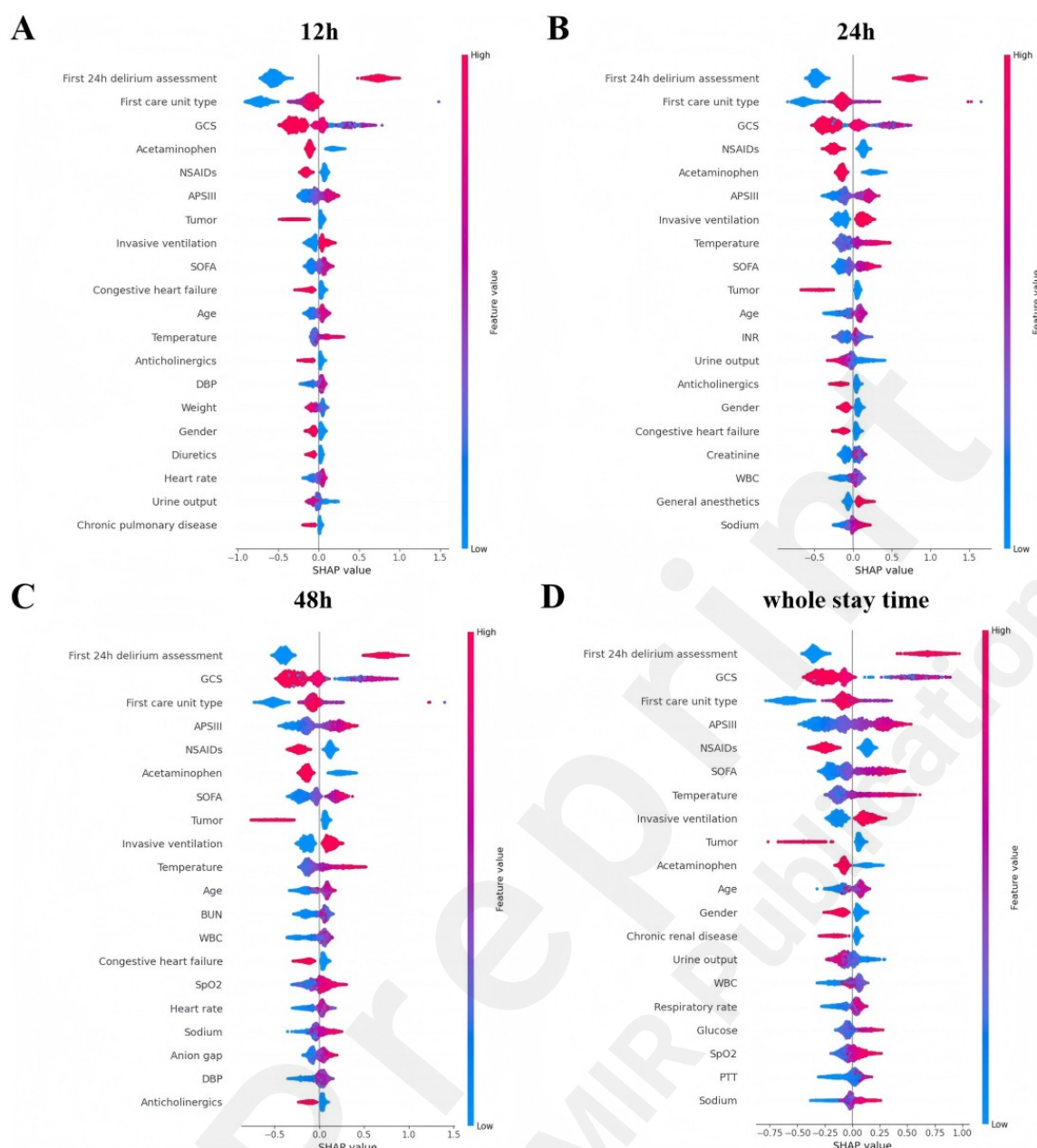


Figure 6. Shapley Additive Explanations (SHAP) summary plots of the XGB machine learning models in different prediction windows (top 20 features). A, B, C, and D correspond to prediction windows of 12h, 24h, 48h, and whole stay time, respectively. Each point in the plot in a given case corresponds to 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; APSIII: Acute Physiology Score III; DBP: diastolic 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. The study developed the early prediction models for postoperative delirium in elderly ICU patients by training four major machine learning algorithms on the MIMIC-IV dataset. The XGB model, which exhibited the best performance in identifying delirium status during internal validation, demonstrated acceptable recognition and generalization capabilities in the external validation on the eICU-CRD dataset. Additionally, we identified that the most significant features were the first day's delirium assessment results, type of first care unit, minimum GCS score, APSIII, Acetaminophen,

NSAIDs.

To the best of our knowledge, this is the inaugural predictive model to forecast postoperative delirium state in elderly patients in a windowed manner. The model leverages clinical data collected within the first 24 hours of ICU admission to predict delirium status at future intervals of 12, 24, and 48 hours, as well as throughout the entire ICU stay. By retaining subjects with positive delirium assessment records during the observation window and incorporating delirium evaluation outcomes as model features, our framework enables nuanced analysis of delirium (e.g., initial onset, recurrence, persistence, and remission) beyond binary classification. This tool assists clinicians in systematically evaluating patient status, optimizing clinical decision-making at critical junctures, and supplying preventive and personalized care to high-risk populations.

Comparison With Prior Work

At present, the CAM-ICU score is the most frequently employed method for the clinical diagnosis of delirium. However, it necessitates the administration of multiple assessments to ascertain a positive result [33]. 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 [34-37]. Although some predictive models for assessing the risk of postoperative delirium in elderly patients have been developed, existing studies still exhibit significant limitations in model construction and validation. Firstly, most current models utilize training cohorts with limited sample sizes ($n < 2000$) or lack rigorous external validation processes [20-23]. Some studies only employ temporal split-validation within single medical institutions [22], rendering their conclusions of questionable generalizability. Secondly, other models predominantly focus on specific surgical procedures (e.g., cardiac surgery, hip fracture repair), failing to meet the multidisciplinary demands of clinical practice. In contrast, this study developed a machine learning prediction model based on the MIMIC-IV database ($n = 6129$) and conducted external validation using the multicenter eICU-CRD cohort encompassing 208 medical institutions across North America. The results demonstrate that our model achieved acceptable predictive performance for delirium status in both internal and external validations, with cross-institutional and cross-procedural stability validation providing robust evidence for clinical implementation.

The findings of this study reveal that the prevalence of postoperative delirium among elderly patients in the MIMIC-IV database reached 40.6%, while in the eICU-CRD database it was 29.5%, significantly higher than previous similar studies [38-40]. This discrepancy may stem from our inclusion of delirium positive cases on the first day of admission, aiming to monitor the subsequent evolution of delirium status. Prior research has demonstrated that persistent or recurrent episodes of delirium are significantly associated with prolonged hospital stays and increased mortality rates [41]. Systematic clinical investigations and the establishment of effective monitoring mechanisms for this patient population hold substantial clinical value.

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. The present study identified six 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, type of first care unit, minimum GCS score, APACHE II, Acetaminophen, NSAIDs. Furthermore, invasive ventilation, SOFA score, mean body temperature, age, and tumor were also identified as relatively high-ranking features. This means that certain highly predictive features identified in previous studies (such as age, invasive ventilation, temperature, SOFA score, GCS score, and type of first care unit), were corroborated in the present study [38, 42, 43]. APACHE II As part of the APACHE score, suggesting a potential association with delirium [38]. In this study, the delirium status on the first day was consistently identified as the most critical predictive feature. Its prominent predictive value may be closely associated with the persistent characteristics of delirium pathophysiology. The onset of delirium on the first day often indicates that patients may have persistent underlying triggers that have not been effectively addressed (e.g., metabolic imbalances, occult infections, or cerebral hemodynamic abnormalities). The effects of these pathological factors may drive persistent delirium episodes through cascade effects, thereby influencing subsequent clinical assessments. Incorporating this feature into predictive models can indirectly capture the continuity of latent pathological states. It is noteworthy that in the MIMIC-IV dataset of this study, 1812 individuals were identified as having delirium on day one. But within the next 12 hours, 871 (48.1%) of these had moved to a non-delirious state. This finding suggests that while the first

day's delirium assessment holds significant predictive value, evaluation of delirium's dynamic trajectory requires integration with multidimensional indicators.

This study demonstrates a significant association between the use of acetaminophen or NSAIDs and the risk of delirium occurrence. As antipyretic-analgesic agents widely used in clinical practice, their pharmacological mechanisms involve the inhibition of cyclooxygenase enzymes, thereby blocking prostaglandin biosynthesis and downregulating the hypothalamic thermoregulatory set point to achieve antipyretic effects [44, 45]. Previous studies have revealed that prostaglandins not only participate in thermoregulation but also play a pivotal regulatory role in the pathophysiology of delirium through pathways affecting neurotransmitter release and neuroinflammatory responses[46]. These molecular findings corroborate clinical evidence from this study that identifies abnormal body temperature as an independent risk factor for delirium. This suggests that when developing body temperature management strategies for delirium patients, the neuroregulatory effects of such medications should be taken into consideration. At the same time, this study incorporated various chronic comorbidities to investigate the risk factors for delirium. The results revealed that the most significant association with delirium was observed in tumor patients. This correlation may stem from tumor-related metabolic disturbances leading to neurotransmitter imbalances, neurotoxic effects of anti-tumor treatments (such as chemotherapy drugs and radiotherapy), as well as the prevalent pain-induced stress and sleep disturbances among advanced-stage tumor patients [47, 48]. Other predictors such as blood sodium, oxygen saturation, blood urea nitrogen, blood pressure, urine output, and platelets have been validated by similar studies or predictive models [49-51].

Strengths and Limitations

It is important to note that our study has several notable contributions and strengths. First, 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. This approach not only guarantees the internal validity of model training but, more importantly, confirms the model's generalizability across diverse hospital populations through rigorous external validation procedures, significantly enhancing the clinical applicability of the research findings. Second, distinct from traditional single time-window prediction models, this study developed four progressive predictive models with 12h, 24h, 48h, and full-course monitoring windows. This innovative design enables both short-term and long-term prediction of delirium status, which is crucial for improving care continuity and optimizing resource allocation in resource-constrained clinical environments. Last but not least, all predictive variables were derived from routine monitoring data collected within the first 24 hours of ICU admission, requiring no additional testing for real-time implementation. Notably, this study incorporates delirium assessment features within the first 24 hours of ICU admission for the first time, considering the persistent clinical monitoring value of delirium recurrence, continuation, and improvement in real-world scenarios. The early and accurate prediction of delirium status achieved in this research empowers clinicians to adjust treatment strategies with enhanced temporal 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, during the cross-center external validation process, the research team was unable to fully reproduce the data screening criteria of the MIMIC-IV dataset within the eICU-CRD due to inherent structural differences between the databases. The particularly critical issue is that the eICU-CRD lacks explicit documentation of surgical timestamps, which hinders our ability to establish precise temporal correlations between ICU data and surgical events as effectively as we do in MIMIC-IV database. This limitation ultimately leads to performance degradation of the model during external validation. 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 delirium assessments data (65% 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 of the nonlinearities between features, which inevitably introduces correlation bias [52, 53].

Conclusions

In this study, we constructed and validated a high-performance prediction model for delirium in elderly ICU patients. This model can predict delirium state 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

The authors extend their gratitude to NM for the financial support and to all scholars who provided assistance during the research. Special acknowledgment is given to the Institutional Review Boards of Beth Israel Deaconess Medical Centre and Massachusetts Institute of Technology for granting access to the MIMIC-IV database and the eICU-CRD.

Funding Statement

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

Data Availability

The datasets used and analyzed during the current study are available from the corresponding author on reasonable request.

Authors' Contributions

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

APSI : Acute Physiology Score III

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*. 2020 Oct 6;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*. 2017 Sep 26;318(12):1161-74. 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*. 2021 Jun 8;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*. 2018 Jul 13;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*. 2015 Dec;150(12):1134-40. 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*. 2021 Aug;12(4):787-91. 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*. 2010 Jul 28;304(4):443-51. PMID: 20664045. doi: 10.1001/jama.2010.1013.
8. 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*. 2020 Jul 15;202(2):193-201. PMID: 32078780. doi: 10.1164/rccm.201904-0816CI.
9. 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*. 2021 Sep 1;133(3):713-22. PMID: 33433117. doi: 10.1213/ANE.0000000000005365.
10. 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.
11. 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*. 2022 Jan 4;327(1):50-8. PMID: 34928310. doi: 10.1001/jama.2021.22647.
12. 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*. 2020 Mar;16(3):572-80. PMID: 31761478. doi: 10.1016/j.jalz.2019.09.080.
13. 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*. 2021 Feb;42(1):112-26. PMID: 32746469. doi: 10.1055/s-0040-1710572.
14. 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*. 2020 Feb;11(1):71-81. PMID: 32297241. doi: 10.1007/s41999-019-00260-7.
15. Hshieh TT, Yue J, Oh E, Puella M, Dowal S, Travison T, et al. Effectiveness of multicomponent nonpharmacological delirium interventions: a meta-analysis. *JAMA Intern Med*. 2015 Apr;175(4):512-20. PMID: 25643002. doi: 10.1001/jamainternmed.2014.7779.
16. Mullainathan S, Spiess J. Machine learning: an applied econometric approach. *J Econ Perspect*. May 01, 2017;31(2):87-106. doi: 10.1257/jep.31.2.87.17. Cuocolo R, Caruso M, Perillo T, Uggla L, Petretta M. Machine Learning in oncology: A clinical appraisal. *Cancer Lett*. 2020 Jul 1;481:55-62. PMID: 32251707. doi: 10.1016/j.canlet.2020.03.032.
18. 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*. 2020 Dec;478(12):2751-64. PMID: 32740477. doi: 10.1097/CORR.0000000000001360.

19. 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*. 2017 Dec;2(4):230-43. PMID: 29507784. doi: 10.1136/svn-2017-000101.

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. 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*. 2023 Jan;29(1):158-67. PMID: 36217732. doi: 10.1111/cns.13991.

22. 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;18:157-68. PMID: 36789284. doi: 10.2147/CIA.S398314.

23. Sheng W, Tang X, Hu X, Liu P, Liu L, Miao H, et al. Random forest algorithm for predicting postoperative delirium in older patients. *Front Neurol*. 2023;14:1325941. PMID: 38274882. doi: 10.3389/fneur.2023.1325941.

24. 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*. 2015 Jan 7;350:g7594. PMID: 25569120. doi: 10.1136/bmj.g7594.

25. 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*. 2023 Jan 3;10(1):1. PMID: 36596836. doi: 10.1038/s41597-022-01899-x.

26. Pollard TJ, Johnson AEW, Raffa JD, Celi LA, Mark RG, Badawi O. The eICU Collaborative Research Database, a freely available multi-center database for critical care research. *Sci Data*. 2018 Sep 11;5:180178. PMID: 30204154. doi: 10.1038/sdata.2018.178.

27. Zhang Z. Multiple imputation with multivariate imputation by chained equation (MICE) package. *Ann Transl Med*. 2016 Jan;4(2):30. PMID: 26889483. doi: 10.3978/j.issn.2305-5839.2015.12.63.

28. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WPJJoAIR. SMOTE: Synthetic Minority Over-sampling Technique. 2002;16(1):321-57. doi: 10.1613/jair.953.

29. 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.

30. Rufibach K. Use of Brier score to assess binary predictions. *J Clin Epidemiol*. 2010 Aug;63(8):938-9; author reply 9. 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*. 2006 Nov-Dec;26(6):565-74. 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*. 2024 Jun 28;100(1185):512-5. PMID: 38453146. doi: 10.1093/postmj/qgae027.

33. 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*. 2021 Mar;9(3):239-50. PMID: 33428871. doi: 10.1016/S2213-2600(20)30552-X.

34. 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*. 2022 Mar 3;26(1):53. PMID: 35241132. doi: 10.1186/s13054-022-03931-3.

35. 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*. 2012 Feb 9;344:e420. PMID: 22323509. doi: 10.1136/bmj.e420.

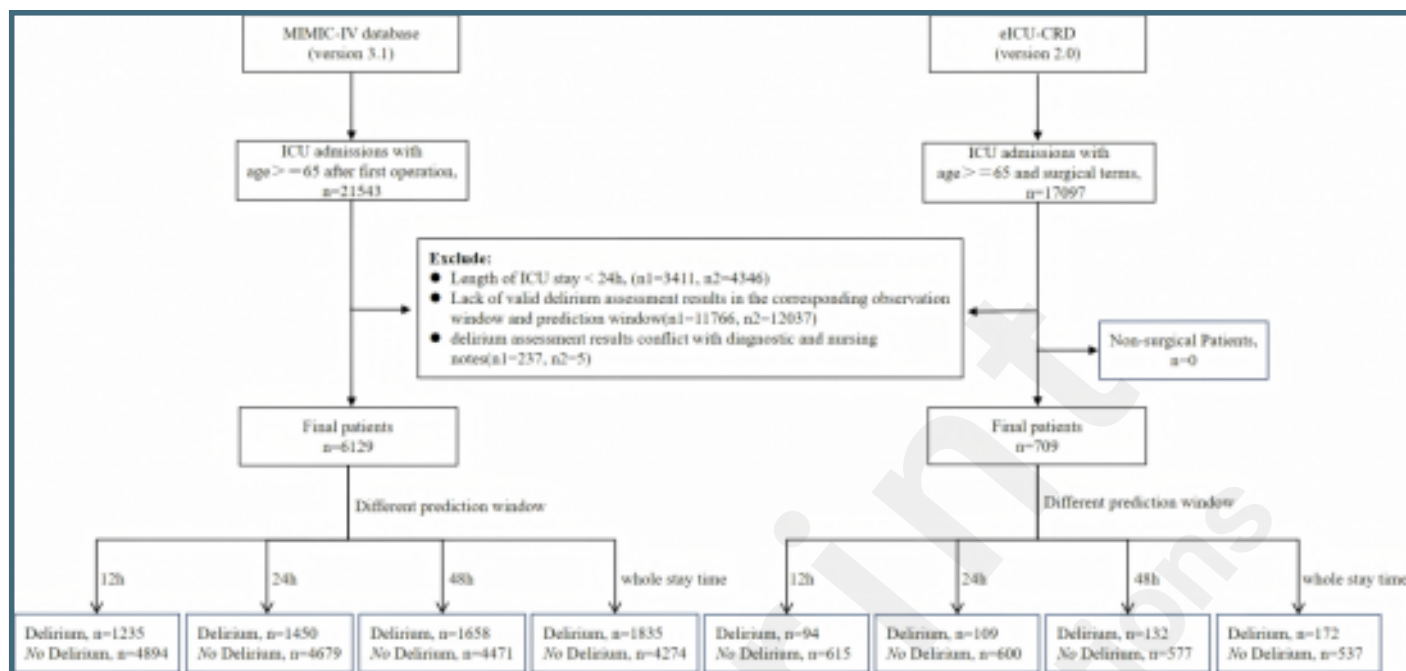
36. Dos Santos FCM, Rego AS, Montenegro WS, de Carvalho S, Cutrim RC, Junior AAM, et al. Delirium in the intensive care unit: identifying difficulties in applying the Confusion Assessment Method for the Intensive Care Unit (CAM-ICU). *BMC Nurs*. 2022 Nov 23;21(1):323. PMID: 36419158. doi: 10.1186/s12912-022-01103-w.

37. 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*. 2017 Sep 2;17(1):120. PMID: 28865447. doi: 10.1186/s12871-017-0415-2.
38. 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*. 2023 Mar 1;138(3):299-311. PMID: 36538354. doi: 10.1097/ALN.0000000000004478.
39. 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.
40. 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*. 2016 Aug 21;20:264. PMID: 27544077. doi: 10.1186/s13054-016-1445-8.
41. 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*. 2022 May;48(5):559-69. PMID: 35322288. doi: 10.1007/s00134-022-06650-z.
42. 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*. 2020 Feb 24;24(1):66. PMID: 32093710. doi: 10.1186/s13054-020-2796-8.
43. 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*. 2021 Jul 26;9(7):e23401. PMID: 34309567. doi: 10.2196/23401.
44. Graham GG, Scott KF. Mechanism of action of paracetamol. *Am J Ther*. 2005 Jan-Feb;12(1):46-55. PMID: 15662292. doi: 10.1097/00045391-200501000-00008.
45. Murphy PJ, Myers BL, Badia P. Nonsteroidal anti-inflammatory drugs alter body temperature and suppress melatonin in humans. *Physiol Behav*. 1996 Jan;59(1):133-9. PMID: 8848472. doi: 10.1016/0031-9384(95)02036-5.
46. Cunningham C, Skelly DT. Non-steroidal anti-inflammatory drugs and cognitive function: are prostaglandins at the heart of cognitive impairment in dementia and delirium? *J Neuroimmune Pharmacol*. 2012 Mar;7(1):60-73. PMID: 21932048. doi: 10.1007/s11481-011-9312-5.
47. Yang Y, Chen J, Wen Q, Jin G, Liu F, Yu L, et al. Effects of preoperative neoadjuvant chemotherapy on postoperative delirium in patients with gynecological tumor surgery: an observational study. *J Cancer Res Clin Oncol*. 2024 Nov 14;150(11):497. PMID: 39542945. doi: 10.1007/s00432-024-06006-8.
48. Wang P, Xiao W. Development and Validation of Subsyndromal Delirium Prediction Model in Patients With Advanced Malignant Tumor: A Case-Control Study. *Cancer Nurs*. 2023 Nov 13. PMID: 37962137. doi: 10.1097/NCC.0000000000001290.
49. 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*. 2023 Aug 4;13(1):12697. PMID: 37542106. doi: 10.1038/s41598-023-38650-4.
50. Lucini FR, Stelfox HT, Lee J. Deep Learning-Based Recurrent Delirium Prediction in Critically Ill Patients. *Crit Care Med*. 2023 Apr 1;51(4):492-502. PMID: 36790184. doi: 10.1097/CCM.0000000000005789.
51. 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*. 2024 Jan 25;14(1):57. PMID: 38267405. doi: 10.1038/s41398-024-02762-w.
52. 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.
53. Lundberg S, Lee SI. A unified approach to interpreting model predictions. *NIPS*. 2017;30 (0):4768-4777. doi: 10.5555/3295222.3295230.

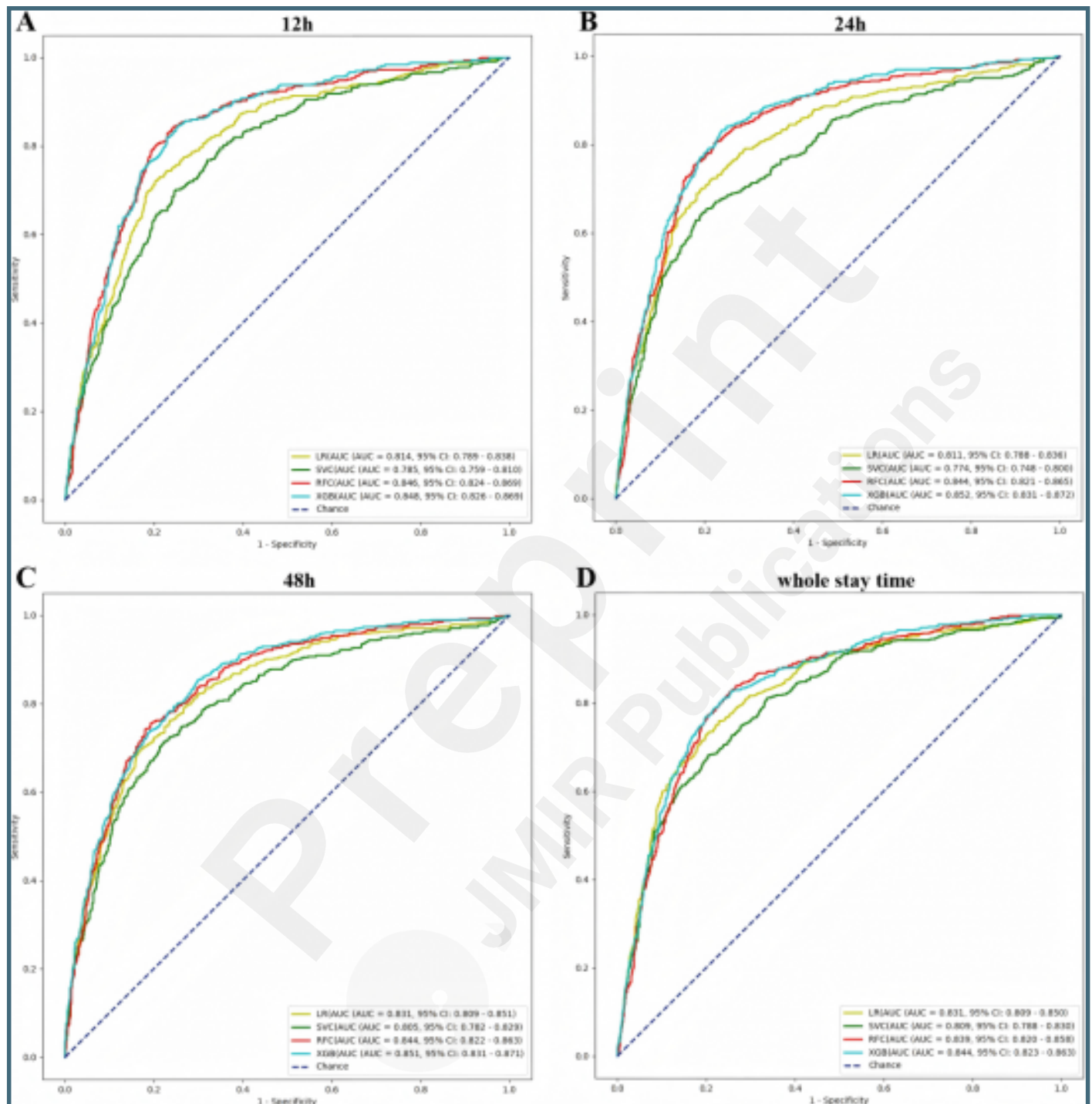
Supplementary Files

Figures

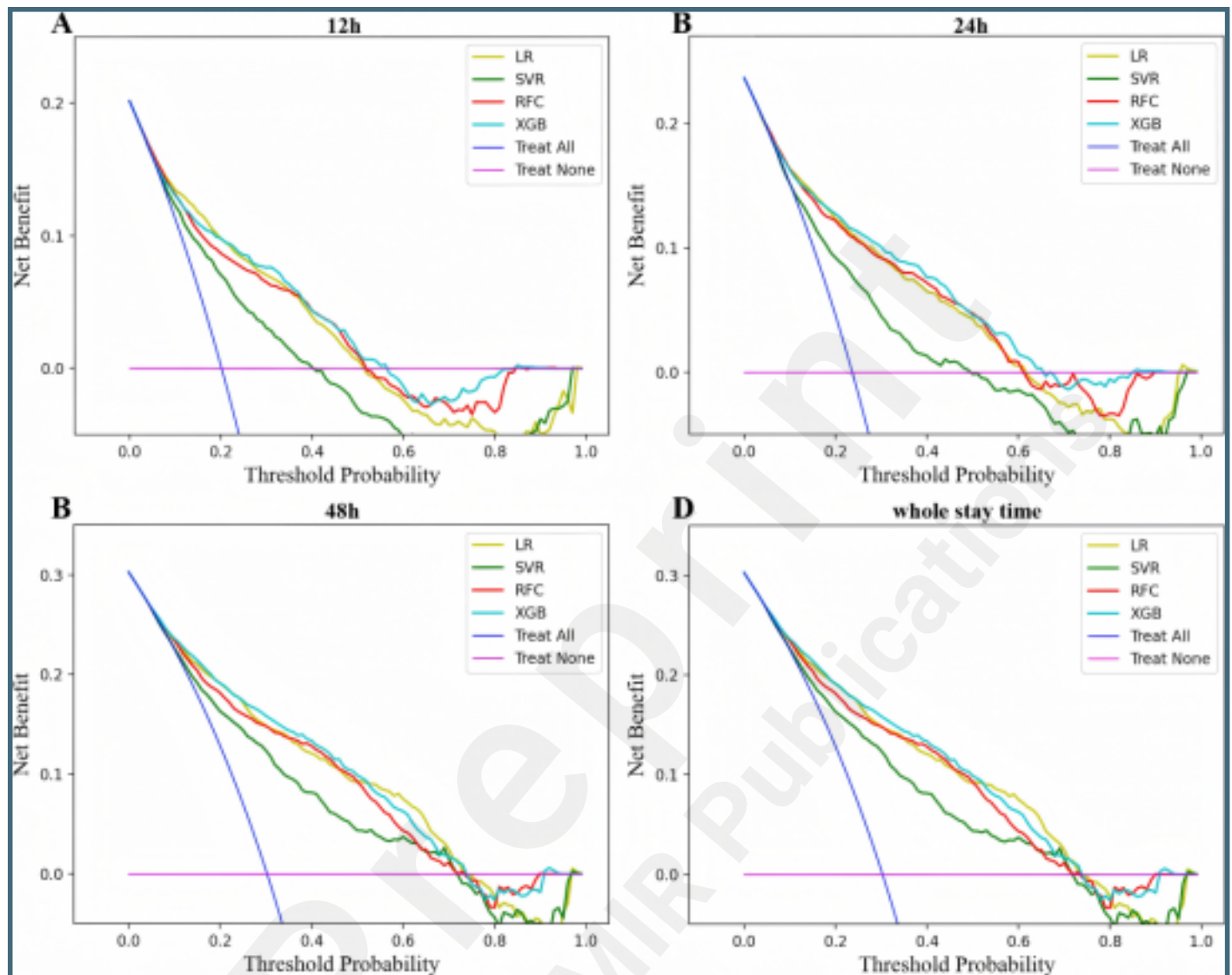
Cohort selection schema.



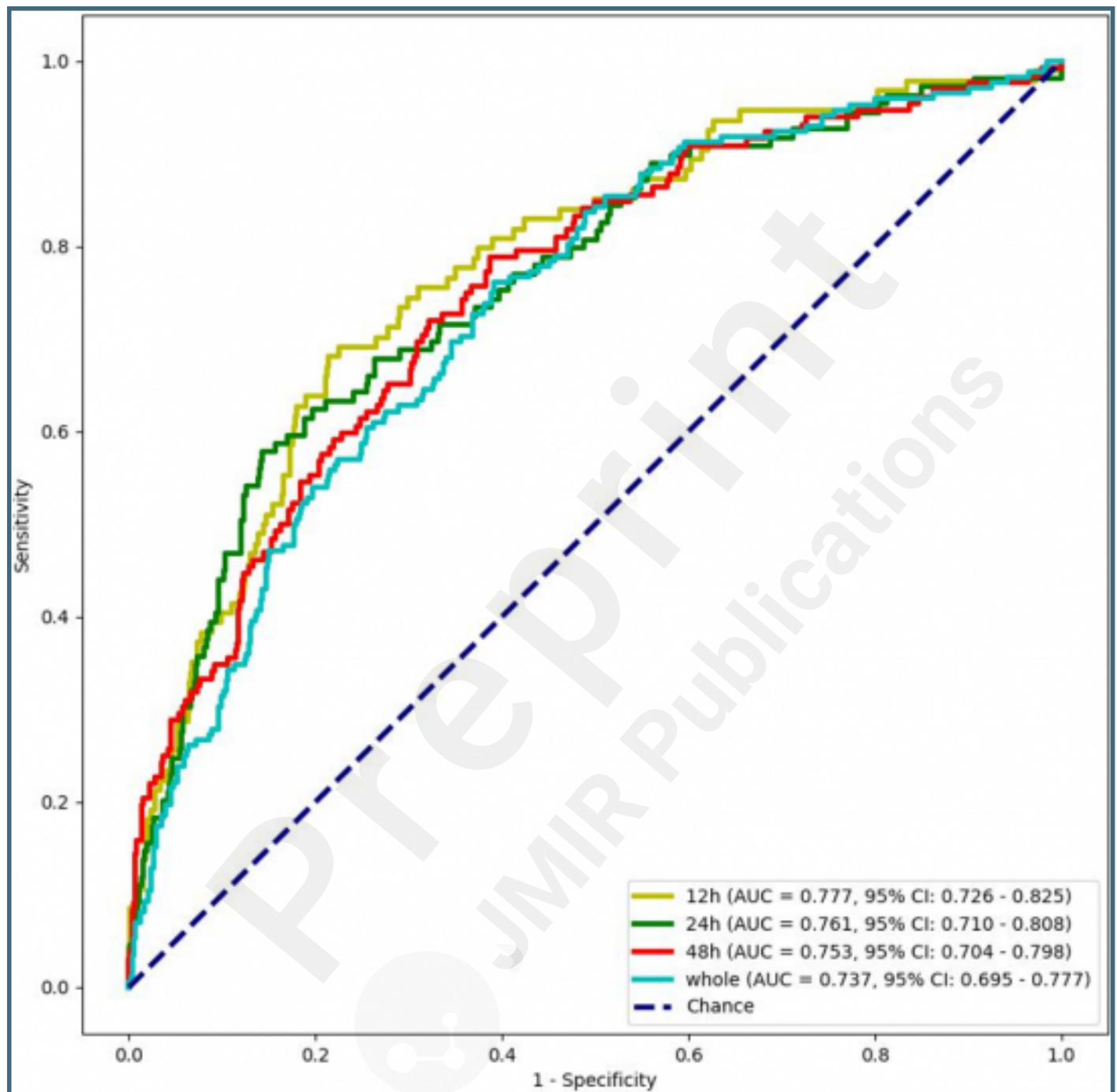
Receiver operating characteristic curves for all machine learning models in different prediction windows in the internal validation set.



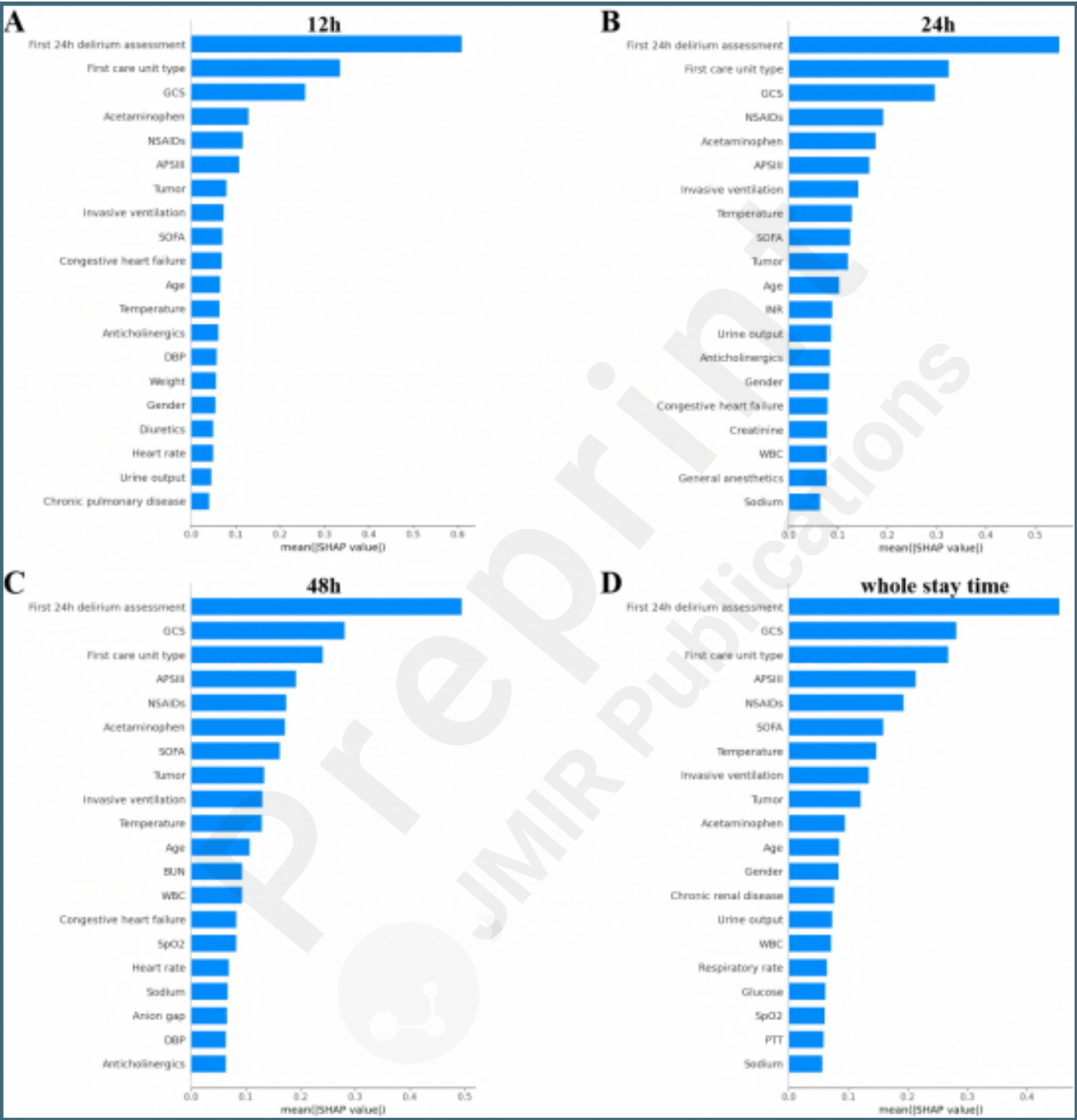
Decision curves for all machine learning models in different prediction windows in the internal validation set.



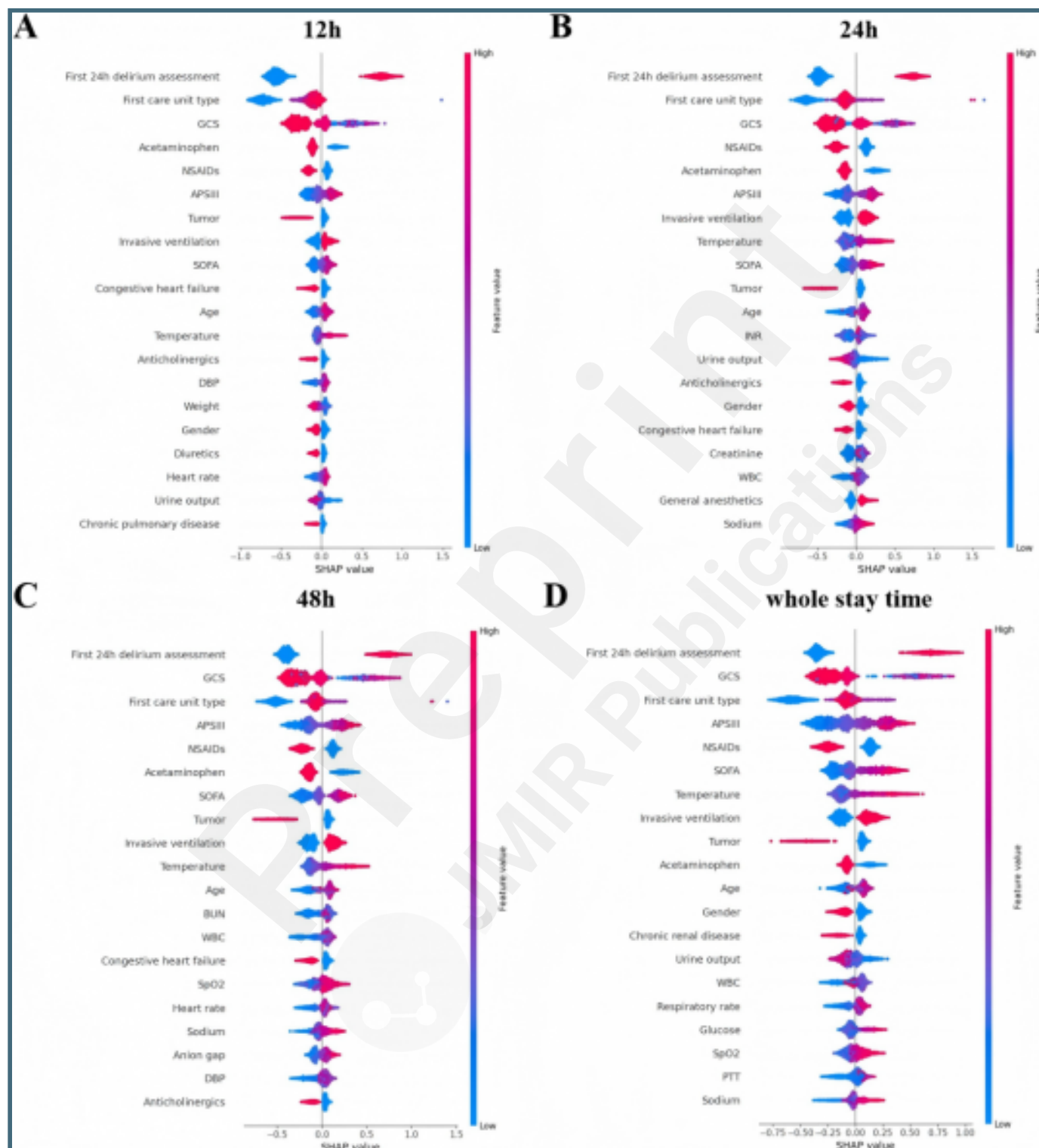
Receiver operating characteristic curves for extreme gradient boosting models in different prediction windows in the external validation set.



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

track changes in Revised Ms.

URL: <http://asset.jmir.pub/assets/8558b232f6edda9b1e06b7fc920b2e39.docx>

Baseline Characteristics of delirium and Non-delirium Patients in the 12h prediction window.

URL: <http://asset.jmir.pub/assets/bea5ff78976037d909d1a328210d65d0.docx>

Baseline Characteristics of delirium and Non-delirium Patients in the 24h prediction window.

URL: <http://asset.jmir.pub/assets/a625fc977f7f246e493cf06da68933e.docx>

Baseline Characteristics of delirium and Non-delirium Patients in the 48h prediction window.

URL: <http://asset.jmir.pub/assets/b4e75ecdda2cefd2aff9e02b93bf1a37.docx>

Confusion Matrix of XGB Models for Different Prediction windows in the internal validation set.

URL: <http://asset.jmir.pub/assets/27e1650bda1380fc1ea3d701497925d.docx>

Confusion Matrix of XGB Models for Different Prediction windows in the external validation set.

URL: <http://asset.jmir.pub/assets/fa6c9a59b74638cf3e3990244ef58c08.docx>