


ORIGINAL ARTICLE

Automated machine learning-based model predicts postoperative delirium using readily extractable perioperative collected electronic data

Xiao-Yi Hu^{1,2} | He Liu³ | Xue Zhao¹ | Xun Sun^{1,2} | Jian Zhou^{1,2} | Xing Gao^{2,4} | Hui-Lian Guan^{2,5} | Yang Zhou² | Qiu Zhao^{1,2} | Yuan Han⁶ | Jun-Li Cao^{1,2} 

¹Department of Anesthesiology, The Affiliated Hospital of Xuzhou Medical University, Jiangsu Province, Xuzhou City, China

²Jiangsu Province Key Laboratory of Anesthesiology & NMPA Key Laboratory for Research and Evaluation of Narcotic and Psychotropic Drugs, Xuzhou Medical University, Jiangsu Province, Xuzhou City, China

³Department of Anesthesiology, The Affiliated Huzhou Hospital, Zhejiang University School of Medicine, Huzhou Central Hospital, Zhejiang Province, Huzhou City, China

⁴Department of Anesthesiology, Changzhou First People's Hospital, Changzhou, Jiangsu, China

⁵Department of Anesthesiology, The First Affiliated Hospital of Bengbu Medical College, Bengbu, Anhui, China

⁶Department of Anesthesiology, Eye & ENT Hospital of Fudan University, Shanghai, China

Correspondence

Yuan Han, M.D., Ph.D. Associate Professor of Anesthesiology, Department of Anesthesiology, Eye & ENT Hospital of Fudan University, Shanghai, 200031, China.

E-mail: yuan.han@fdeent.org

Prof. Jun-Li Cao, Department of Anesthesiology, The Affiliated Hospital of Xuzhou Medical University, NO. 99 Huaihai Road, Quanshan District, Xuzhou City 221002; and Jiangsu Province Key Laboratory of Anesthesiology & NMPA Key Laboratory for Research and Evaluation of Narcotic and Psychotropic Drugs, Xuzhou Medical University, NO. 209 Tongshan Road, Yunlong District, Xuzhou City 221004, Jiangsu Province, China.

Email: caojl0310@aliyun.com

Funding information

This study was supported in part by grants from the National Natural Science Foundation of China (NSFC81720108013, NSFC31771161 and NSFC81230025 to J.L. Cao; NSFC81300957 to H. Liu), Jiangsu Provincial Special Program of Medical Science (BL2014029 to J.L.

Abstract

Objective: Postoperative delirium (POD) is a common postoperative complication that is relevant to poor outcomes. Therefore, it is critical to find effective methods to identify patients with high risk of POD rapidly. Creating a fully automated score based on an automated machine-learning algorithm may be a method to predict the incidence of POD quickly.

Materials and methods: This is the secondary analysis of an observational study, including 531 surgical patients who underwent general anesthesia. The least absolute shrinkage and selection operator (LASSO) was used to screen essential features associated with POD. Finally, eight features (age, intraoperative blood loss, anesthesia duration, extubation time, intensive care unit [ICU] admission, mini-mental state examination score [MMSE], Charlson comorbidity index [CCI], postoperative neutrophil-to-lymphocyte ratio [NLR]) were used to established models. Four models, logistic regression, random forest, extreme gradient boosted trees, and support vector machines, were built in a training set (70% of participants) and evaluated in the remaining testing sample (30% of participants). Multivariate logistic regression analysis was used to explore independent risk factors for POD further.

Results: Model 1 (logistic regression model) was found to outperform other classifier models in testing data (area under the curve [AUC] of 80.44%, 95% confidence

Xiao-Yi Hu and He Liu contributed equally to this work.

Yuan Han is the co-corresponding author.

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2021 The Authors. *CNS Neuroscience & Therapeutics* published by John Wiley & Sons Ltd.

Cao), Jiangsu Provincial Natural Science Foundation (BK20181145 to H. Liu), the Clinical Technical Research and Study Plan Project (2018211006 to H. Liu)

interval [CI] 72.24%–88.64%) and achieve the lowest Brier Score as well. These variables including age (OR = 1.054, 95%CI: 1.017–1.093), extubation time (OR = 1.027, 95%CI: 1.012–1.044), ICU admission (OR = 2.238, 95%CI: 1.313–3.793), MMSE (OR = 0.929, 95%CI: 0.876–0.984), CCI (OR = 1.197, 95%CI: 1.038–1.384), and postoperative NLR (OR = 1.029, 95%CI: 1.002–1.057) were independent risk factors for POD in this study.

Conclusions: We have built and validated a high-performing algorithm to demonstrate the extent to which patient risk changes of POD during the perioperative period, thus leading to a rational therapeutic choice.

KEYWORDS

delirium, machine learning, model prediction, nomogram, postoperative

1 | INTRODUCTION

Postoperative delirium (POD) is an acute fluctuating neurocognitive syndrome caused by reversible neuronal disruption due to an underlying systemic perturbation, which usually occurs a few hours to a few days after surgery and mainly manifests as a decline in consciousness, attention disorders, and thinking disorders.¹ It has been reported that the incidence of POD in elderly surgical patients ranges from 10% to 70%.^{2,3}

Previous studies have demonstrated that early interventions can help reduce or even prevent POD,⁴ while many patients with POD can't be identified efficiently. In clinical settings, the diagnosis of POD is still mainly based on clinical observation.⁵ However, the type of hypoactive POD is about 71% and very hard to notice. Therefore, it is critical to find methods to identify patients with a high risk of POD rapidly.

In recent years, basic and clinical studies have found that many risk factors or biomarkers may affect the occurrence of POD.^{6,7} For instance, many inflammatory markers investigated in scientific and clinical studies, such as CRP, were believed to be associated with POD.^{8–10} Therefore, disease prediction models conveniently screen high-risk patients, and the nomogram could be easily used in clinical settings. However, some prediction models for POD were based on a single statistical method, which may be limited in predictive performance.^{11,12} Recently, it has been reported that using machine-learning techniques to establish various disease prediction models could improve the predictive performance of these models.^{13,14}

Thus, in the current study, we used machine-learning technology to extract the clinical data of 531 surgical patients who underwent general anesthesia before and on the first day after surgery and established four predictive models of POD using different methods. Finally, we compared these models and created a model with optimal predictive performance, which can assist in diagnosing and identifying patients with a high risk of POD. Furthermore, to increase the availability of the optimal model, the optimal model was transformed into the form of a nomogram.

2 | MATERIALS AND METHODS

2.1 | Data source and extraction

The secondary analysis was based on an observational study (the Ethical Committee of the Affiliated Hospital of Xuzhou Medical University approved it, Certification No. XYFY2018-KL091). The written informed consent was obtained from all subjects participating, a legal surrogate, or the parents in this trial. Inclusion criteria were as follows: non-history of clear neurological disease; patients who underwent major noncardiac or non-neurological surgery with general anesthesia; expected a hospital stay of ≥ 3 days; Exclusion criteria were as follows¹⁵: significant impairments of vision; hearing or motor skills; history of neurological disease; liver or kidney dysfunction (such as severe hepatitis, pyelonephritis); severe trauma or surgical history within one year; history of severe physical illness and alcoholism; mini-mental state examination (MMSE) score < 17 ; refuse to sign informed consent.

2.2 | Model endpoint definition

We built classification models to predict the in-hospital incidence of POD as a binary outcome.

2.3 | Delirium assessment

Delirium was assessed using rigorous methodologies. In this trial, CAM¹⁶ was applied to patients who could be communicated with. The CAM-ICU¹⁷ was applied to patients admitted to the intensive care unit (ICU) and cannot be communicated with due to endotracheal intubation. We assessed for delirium 2 h after the surgery and then repeated the assessment twice a day for three days after the surgery in the morning, afternoon, or evening. There was at least 6 h interval between these two assessments.¹⁸ Additionally, evidence of delirium, including confusion, agitation, sedation, hallucinations,

and delusions, was obtained from the nurses, families, and medical records. The evaluation of delirium was carried out by trained researchers who neither knew the patient's perioperative characteristics nor data entry and statistical analysis.

2.4 | Model input features

Forty-nine potential useful features including basic information such as age, sex, BMI, education degree; American society of anesthesiologists (ASA) degree; laboratory data obtained before surgery, such as serum sodium, potassium, creatinine, and blood cell counts; and surgery-specific information such as the surgery type were collected.

Least absolute shrinkage and selection operator (LASSO) was used to select important features associated with POD. Finally, eight features (age, intraoperative blood loss, anesthesia duration, extubation time, ICU admission, mini-mental state examination score [MMSE], Charlson comorbidity index [CCI], postoperative neutrophil-to-lymphocyte ratio [NLR]) were included to established models.

To achieve the highest predictive performance, four models were established, including logistic regression model (LR), random forest (RF), extreme gradient boosted trees (XGB) classifier, and support vector machine (SVM) classifier. Furthermore, in order to further explore the relationship between the above eight features and POD, multivariate logistic regression was used to confirm the independent risk factors for POD in this study.

2.5 | Sample size and statistical analysis

For the two-class prediction model, one of the sample size calculation methods proposed in the article is¹⁹ $n = \exp\left(\frac{-0.508 + 0.259\ln(\varphi) + 0.504\ln(P) - \ln(MAPE)}{0.544}\right)$, where φ is the proportion of ending events ($\varphi = 0.23$), P is the number of predictors ($P = 8$), $MAPE$ is the average absolute error between the observed and true outcome probability ($MAPE = 0.05$). According to the above formula, the sample size of training set is calculated as at least 330. To achieve a sample size of at least 330 in the training set, we randomly split the total dataset ($n = 531$) into a training set ($n = 400$) and a testing set ($n = 131$) at a ratio of 7:3 in this study. Any patients who appeared in the testing set would be removed from the training set in case of information leakage.

All analyses were performed with R version 3.6.1. (R Development Core Team). The normal distribution of numeric variables was tested by the Shapiro-Wilk test. Continuous variables with a normal distribution were expressed as the mean \pm standard deviation (SD) and were compared using the independent-sample *t*-test. The Mann-Whitney *U* test presented continuous variables with a non-normal distribution. Categorical data were presented as a number (%) and were analyzed using the chi-square test or Fisher's exact probability

test. The importance of each variable in the training datasets was assessed by LASSO regression analysis.

The selection of model hyperparameters used 10-fold cross-validation on training datasets. In 10-fold cross-validation, the datasets were divided into ten partitions, where nine-tenths of the data were used to build the models, and the remaining one-tenths were used as the testing datasets. This process was repeated such that each partition was used as testing datasets only once and training datasets nine times. Cross-validation made ensures a better assessment of model performance by averaging metrics over multiple trials.

The role of missing data imputation is described as follows. If the missing value percentage is more significant than 20%, it will be excluded from the final completed dataset. If the rate of missing value is smaller than 20%, the random forest regression method would be used for imputation.

Discrimination and calibration were used to verify the predictive ability of the model. The AUROC expressed measurement of discrimination, and the Youden index (sensitivity + specificity - 1) was used to find the best critical value (cutoff value). The performance of models was evaluated by accuracy, sensitivity, specificity, recall, and precision. Model calibration was measured by Brier score and calibration curve. Brier score was the average squared distance between the predicted probability of the outcome and the true label, and the lower Brier score indicated the better performance of the model.

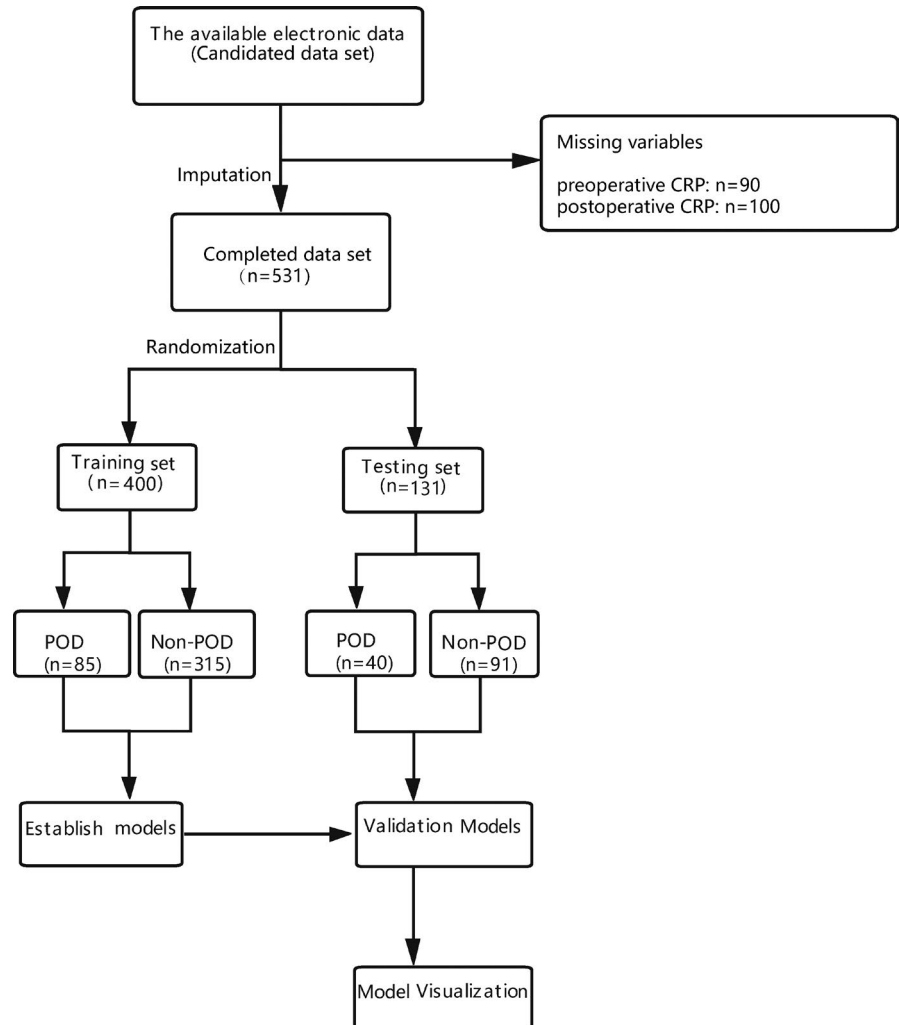
The LR and RF classifiers were implemented with glmnet package and randomForest package in R version 3.6.1, and the XGB classifier was implemented with the XGBoost package in R version 3.6.1. All performance metrics were calculated on the held-out testing datasets. We generated confidence intervals (CIs) for performance metrics with epiR package of R software in training and testing datasets.

3 | RESULTS

3.1 | Patient characteristics

A total of 531 patients were included in this study. Among those screened, the incidence of POD was approximately 23.54%. The variables, including preoperative C-reactive protein (CRP) and postoperative CRP, have missing values. Missing parts of these variables accounted for 16.9% and 18.8% of the total data, respectively. The missing data were imputed by random forest regression. The dataset ($n = 531$) is randomly divided into the training set and testing set at the ratio of 7:3. Four hundred patients formed a training dataset. One hundred thirty-one patients formed testing datasets. The data collected from training datasets were used to assess important variables associated with POD and to establish the predictive models. Patients were divided into POD group ($n = 125$) and Non-POD group ($n = 406$) according to whether or not delirium occurred within the first three days after surgery. The data collected from the testing dataset were aimed to validate predictive models. The patients' recruitment flowchart is shown in Figure 1. Detailed information on

FIGURE 1 Patient recruitment flowchart



patient characteristics can be found in Table 1. There was no significant statistical difference between the features of patients in the training datasets and the testing datasets. The selection of the best parameter (λ) in the LASSO model uses 10-fold cross-validation. Dotted vertical lines were drawn at the optimal values by using the minimum criteria and the 1 SE of the minimum criteria (the 1 - SE criteria).

A vertical line was drawn at the value selected using 10-fold cross-validation, where optimal λ resulted in eight features with non-zero coefficients (Figure 2A,B). We selected eight non-zero characteristic variables in the LASSO regression results, including age, intraoperative blood loss, anesthesia duration, extubation time, ICU admission, MMSE score, CCI score, and postoperative NLR (Table 2).

3.2 | Model performance

We used four algorithms to build predictive models of POD, and the following values in the training datasets were found: LR classifier (AUC value = 73.99% (95%CI: 67.63%-80.35%), accuracy = 0.708 (95%CI: 0.660-0.752), precision = 0.701 (95%CI:

0.652-0.753), recall = 0.913 (95%CI: 0.872-0.942)); RF classifier algorithm (AUC value = 99.06% (95%CI: 97.74%-100%), accuracy = 0.993 (95%CI: 0.978-0.999), precision = 0.991 (95%CI: 0.982-1.000), recall = 1.0000 (95%CI: 0.982-1.000)); XGB classifier (AUC value = 89.77% (95%CI: 86.21%-93.32%), accuracy = 0.868 (95%CI: 0.8303-0.899), precision = 0.931 (95%CI: 0.892-0.953), recall = 0.911 (95%CI: 0.870-0.941)); SVM classifier (AUC value = 87.39% (95%CI: 82.84%-91.94%), accuracy = 0.913 (95%CI: 0.880-0.938), precision = 0.941 (95%CI: 0.910-0.964), recall = 0.951 (95%CI: 0.922-0.974)) (Table 3).

For the testing dataset, the following values in the test group were found: LR classifier (AUC value = 80.44% (95%CI: 72.24%-88.64%), accuracy = 0.687 (95%CI: 0.600-0.765), precision = 0.661 (95%CI: 0.563-0.754), recall = 0.891 (95%CI: 0.791-0.950); RF classifier (AUC value = 70.36% (95%CI: 61.35%-79.37%), accuracy = 0.801 (95%CI: 0.723-0.866), precision = 0.912 (95%CI: 0.832-0.962), recall = 0.842 (95%CI: 0.751-0.904); XGB classifier (AUC value = 76.83% (95%CI: 66.77%-86.89%), accuracy = 0.779 (95%CI: 0.698-0.847), precision = 0.881 (95%CI: 0.792-0.930), recall = 0.832 (95%CI: 0.753-0.901); SVM classifier (AUC value = 68.44% (95%CI: 59.13%-77.74%), accuracy = 0.702 (95%CI: 0.616-0.779), precision = 0.723 (95%CI: 0.622-0.812),

TABLE 1 Comparison of demographic characteristics and perioperative and postoperative variables between the training dataset and the testing dataset

Property	Training dataset	Testing dataset	<i>p</i>
Patients, <i>n</i>	400	131	
Sex			0.324
Female, <i>n</i> (%)	165 (41.2%)	47 (35.9%)	
Male, <i>n</i> (%)	235 (58.8%)	84 (64.1%)	
Age (median ± IQR)	68.00 (65.00, 73.25)	68.00 (64.00, 72.0)	0.443
Height (median ± IQR)	165.00 (159.80, 170.00)	162.00 (158.00, 170.00)	0.247
Weight (median ± IQR)	62.50 (56.00, 70.00)	61.00 (55.00, 70.00)	0.483
BMI (median ± IQR)	23.80 (21.25, 25.90)	23.40 (21.47, 25.95)	0.867
Education degree, <i>n</i> (%)			0.945
Illiteracy	114 (28.5%)	40 (30.5%)	
Primary education	106 (26.5%)	34 (26.0%)	
Junior high school education	109 (27.3%)	31 (23.7%)	
High school education	53 (13.3%)	21 (16.0%)	
University degree	14 (3.5%)	4 (3.1%)	
University degree above	4 (1.0%)	1 (0.8%)	
ASA degree, <i>n</i> (%)			0.390
I	12 (3.0%)	5 (3.8%)	
II	324 (81.0%)	106 (80.9%)	
III	64 (16.0%)	19 (14.5%)	
IV	0 (0.0%)	1 (0.8%)	
Smoking, <i>n</i> (%)			0.836
None	250 (62.5%)	80 (61.1%)	
Yes	150 (37.5%)	51 (38.9%)	
Alcohol, <i>n</i> (%)			0.026*
None	296 (74.0%)	83 (63.4%)	
Yes	104 (26.0%)	48 (36.6%)	
Hypertension, <i>n</i> (%)			0.672
None	265 (66.2%)	84 (64.1%)	
Yes	135 (33.8%)	47 (35.9%)	
Diabetes, <i>n</i> (%)			0.993
None	349 (87.3%)	115 (87.8%)	
Yes	51 (12.8%)	16 (12.2%)	
Hemoglobin (median ± IQR)	130.00 (118.00, 141.00)	131.00 (118.5, 144.0)	0.726
Albumin (median ± IQR)	42.15 (38.70, 45.00)	41.70 (38.0, 44.80)	0.408
ALT (median ± IQR)	16 (12, 24)	16 (11, 22.5)	0.240
AST (median ± IQR)	19 (15, 23.0)	18 (15, 21)	0.093
BUN (median ± IQR)	5.10 (4.19, 6.10)	5.20 (4.25, 6.65)	0.248
Cr (median ± IQR)	62.00 (54.00, 71.00)	64.00 (53.00, 71.5.00)	0.834
Blood volume (median ± IQR)	100 (100, 200)	100 (100, 250)	0.261
Urine volume (median ± IQR)	400 (300, 400)	400 (300, 500)	0.170
Crystalloid solution (median ± IQR)	1250 (1000, 1500)	1350 (1000, 1550)	0.298
Ethoxyl volume (median ± IQR)	500 (500, 500)	500 (500, 500)	0.668
Gelatin volume (median ± IQR)	0 (0, 0)	0 (0, 0)	0.517
Blood transfusion (median ± IQR)	0 (0, 0)	0 (0, 0)	0.087
Surgery time (median ± IQR)	161.00 (110.00, 225.00)	165.00 (120.00, 220.00)	0.875

(Continues)

TABLE 1 (Continued)

Property	Training dataset	Testing dataset	p
Anesthesia duration (median ± IQR)	200.00 (150.00, 260.00)	190.00 (150.00, 252.50)	0.914
Extubation time (median ± IQR)	17.00 (10.00, 23.00)	12.00 (12.00, 24.50)	0.809
ICU admission, n (%)			0.493
None	333 (83.3%)	113 (86.5%)	
Yes	67 (16.8%)	18 (13.7%)	
MMSE (median ± IQR)	25.50 (23.00, 28.00)	26.00 (23.50, 28.00)	0.737
CCI (median ± IQR)	3.00 (2.00, 4.00)	3.00 (2.00, 4.00)	0.088
PSMS (median ± IQR)	6.00 (6.00, 6.00)	6.00 (6.00, 6.00)	0.334
IADL (median ± IQR)	8.00 (8.00, 8.00)	8.00 (8.00, 8.00)	0.648
QoR40 preoperative (median ± IQR)	195.00 (190.00, 198.00)	196.00 (190.00, 198.00)	0.770
PCA pump, n (%)			0.677
None	149 (37.2%)	46 (35.1%)	
Yes	251 (62.7%)	85 (64.9%)	
Nerve block, n (%)			1
None	286 (71.5%)	94 (71.8%)	
Yes	114 (28.5%)	37 (28.2%)	
Surgery type, n (%)			0.229
Thoracic operation	130 (32.5%)	44 (33.6%)	
Abdominal operation	183 (45.8%)	57 (43.5%)	
Urinary operation	68 (17%)	24 (18.3%)	
Orthopedic operation	19 (4.8%)	6 (4.6%)	
K+ (median ± IQR)	4.01 (3.75, 4.27)	4.01 (3.75, 4.27)	0.832
Glu (median ± IQR)	5.26 (4.81, 5.96)	5.26 (4.83, 5.91)	0.992
CRP preoperative (median ± IQR)	2.70 (1.10, 11.32)	3.60 (1.20, 10.70)	0.591
CRP postoperative (median ± IQR)	73.45 (42.90, 103.0)	76.80 (37.90, 124.00)	0.251
Cholesterol (median ± IQR)	4.59 (3.87, 5.13)	4.37 (3.79, 5.06)	0.392
Preoperative White blood cell count (median ± IQR)	5.90 (4.80, 7.20)	5.50 (4.80, 6.70)	0.394
Preoperative neutrophil count (median ± IQR)	3.63 (2.74, 4.72)	3.37 (2.81, 4.24)	0.538
Preoperative lymphocyte count (median ± IQR)	1.60 (1.20, 2.00)	1.50 (1.20, 1.90)	0.706
Postoperative White blood cell count (median ± IQR)	10.20 (8.40, 12.55)	10.10 (8.25, 12.20)	0.581
Postoperative neutrophil count (median ± IQR)	8.80 (7.02, 10.91)	8.34 (6.42, 10.86)	0.294
Postoperative lymphocyte count (median ± IQR)	0.90 (0.60, 1.20)	0.90 (0.70, 1.15)	0.701
Postoperative NLR (median ± IQR)	9.98 (6.54, 14.90)	9.66 (5.67, 14.12)	0.349
Preoperative NLR (median ± IQR)	2.27 (1.58, 3.26)	2.19 (1.72, 3.00)	0.934
Postoperative delirium, n (%)			0.040*
None	315 (78.7%)	91 (69.5%)	
Yes	85 (21.3%)	40 (30.5%)	

Abbreviations: ALT, alanine transaminase; ASA, American society of anesthesiologists; AST, glutamic oxalacetic transaminase; BMI, body mass index (kg/m²); BUN, blood urea nitrogen; CCI, Charlson comorbidity index; Cr, serum creatinine; CRP, C-reactive protein; IADL, instrumental activities of daily living; MMSE, mini-mental state examination score; NLR, neutrophil-to-lymphocyte ratio; PCA, postoperative analgesia pump; PSMS, physical self-maintenance scale; QoR40, recovery quality rating scale.

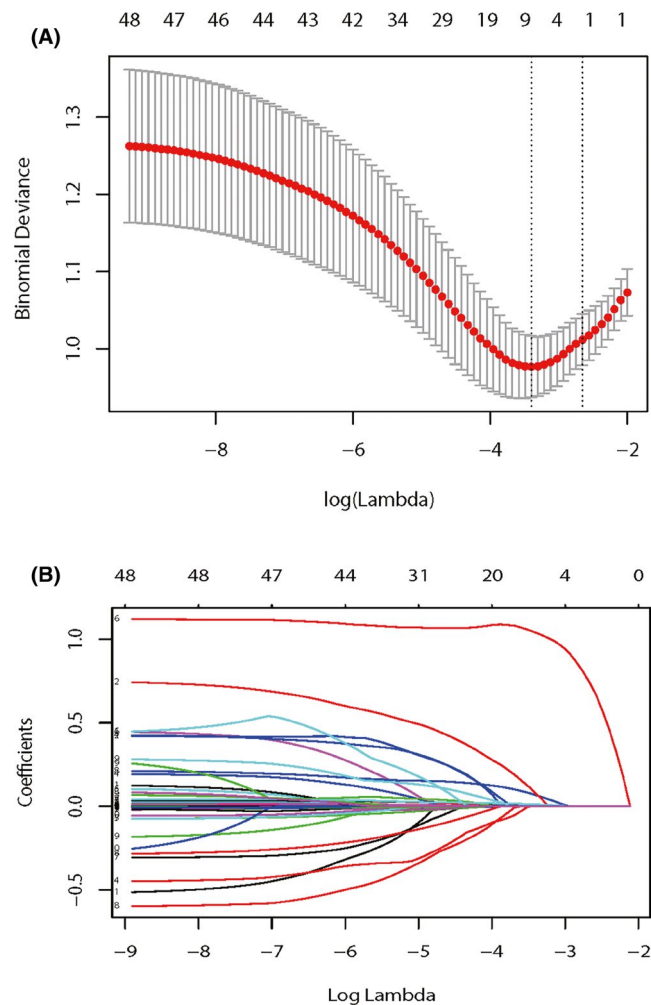


FIGURE 2 Demographic and clinical feature selection using the LASSO regression

recall = 0.852 (95%CI: 0.763–0.924)) (Table 4). The ROC of models in testing dataset and training dataset is shown in Figure 3A,B, and the AUROC for each model is shown in Table 5.

TABLE 2 LASSO regression results of important variables related to POD (training dataset)

Variables	Coefficient	Lambda.min
Age	0.011537909	0.0345332
Intraoperative blood loss	0.0002223647	
Anesthesia duration	0.0017088601	
Extubation time	0.004272257	
ICU admission	0.951368637	
MMSE score	0.0066777804	
CCI	0.088073881	
Postoperative NLR	0.010530093	

Abbreviations: CCI, Charlson comorbidity index; MMSE, mini-mental state examination score; NLR, neutrophil-to-lymphocyte ratio.

The LR achieved much lower (better) Brier scores compared with the other models. Calibration plots of four models in the training dataset and testing dataset are shown in Figure 3C,D. The curve at 45° between the X-axis and the Y-axis indicates good consistency of the model.

Finally, the LR model is transformed into a nomogram to understand and use the model (Figure 4). The two-class prediction outcome is generated based on the optimal cutoff value of the optimal model. Comparing the prediction outcome with the actual occurrence of delirium, the optimal model has shown that the prediction outcome has good performance. The optimal cutoff value of the LR model corresponds to the optimal score of the nomogram. The optimal score of the nomogram was determined to be 109 according to the optimal cutoff value of the LR model. If the sum of the scores corresponding to each entry in the nomogram is greater than 109 points, then the patients who underwent surgery have a higher risk of developing POD. At this moment, nursing staff and doctors should pay attention to the situation of patients. Table 6 presents that these variables including age (OR = 1.054, 95%CI: 1.017–1.093), extubation time (OR = 1.027, 95%CI: 1.012–1.044), ICU admission (OR = 2.238, 95%CI: 1.313–3.793), MMSE (OR = 0.929, 95%CI: 0.876–0.984), CCI (OR = 1.197, 95%CI: 1.038–1.384), and postoperative NLR (OR = 1.029, 95%CI: 1.002–1.057) were independent risk factors for POD in this study (Table 6).

4 | DISCUSSION

The accumulation of multiple risk factors is critical for the occurrence of POD, and there is currently no single treatment to prevent the occurrence of POD. The combination of non-drug therapy and drug therapy is one of the best methods to treat POD. POD has been reported to occur in 10% to 70% of all elderly patients,^{1–3} causing increased mortality, prolonged hospital stays, reduced functional abilities,^{20,21} long-term cognitive dysfunction,²² and even dementia.^{15,23} Therefore, the prevention and treatment of POD is a clinical problem that needs to be solved. In many clinical studies on POD, researchers have tried to find powerful biomarkers that can accurately predict POD, such as S100 β protein,²⁴ neuron-specific enolase (NSE),²⁵ tau protein,²⁶ and inflammatory mediators.²⁷ Researchers are also trying to find better ways to reduce the occurrence of POD. Although these biomarkers have a relatively high ability to predict POD, they cannot be popularized clinically because of the complexity and high cost of sampling. They are always used to explore scientific questions in clinical trials. Therefore, the emergence of disease prediction models may provide a solution for the prevention of POD. Neuroinflammation and the oxidative stress response may be involved in the pathophysiological process of POD.^{5,6} Inflammatory markers investigated in scientific studies have been associated with delirium.^{7–9} To further increase the general applicability of the model, this study included easily available laboratory test items, including some inflammatory mediators, such as CRP and the NLR variables.

TABLE 3 Performance metrics for four models in training dataset

Model	Accuracy	F1 score	Precision	Recall	Specificity
LR	0.708 (0.660, 0.752)	0.494	0.701 (0.652, 0.753)	0.913 (0.872, 0.942)	0.382 (0.302, 0.464)
RF	0.993 (0.978, 0.999)	0.981	0.991 (0.982, 1.000)	1.000 (0.982, 1.000)	0.971 (0.912, 1.000)
XGB	0.868 (0.8303, 0.899)	0.654	0.931 (0.892, 0.953)	0.911 (0.870, 0.941)	0.682 (0.564, 0.782)
SVM	0.913 (0.880, 0.938)	0.785	0.941 (0.910, 0.964)	0.951 (0.922, 0.974)	0.763 (0.664, 0.851)

Accuracy = (TP + TN)/(TP + TN + FP + FN). Precision = TP/(TP + FP). Recall = TP/(TP + FN). Specificity = TN/(TN + FP). F1 score = 2/([1/Recall] + [1/Precision]). FN, false negatives; FP, false positives; TN, true negatives; TP, true positives.

TABLE 4 Performance metrics for four models in testing dataset

Model	Accuracy	F1 score	Precision	Recall	Specificity
LR	0.687 (0.600, 0.765)	0.559	0.661 (0.563, 0.754)	0.891 (0.791, 0.950)	0.442 (0.311, 0.583)
RF	0.801 (0.723, 0.866)	0.567	0.912 (0.832, 0.962)	0.842 (0.751, 0.904)	0.651 (0.442, 0.833)
XGB	0.779 (0.698, 0.847)	0.539	0.881 (0.792, 0.930)	0.832 (0.753, 0.901)	0.592 (0.390, 0.761)
SVM	0.702 (0.616, 0.779)	0.530	0.723 (0.622, 0.812)	0.852 (0.763, 0.924)	0.452 (0.311, 0.602)

Accuracy = (TP + TN)/(TP + TN + FP + FN). Precision = TP/(TP + FP). Recall = TP/(TP + FN). Specificity = TN/(TN + FP). F1 score = 2/([1/Recall] + [1/Precision]). FN, false negatives; FP, false positives; TN, true negatives; TP, true positives.

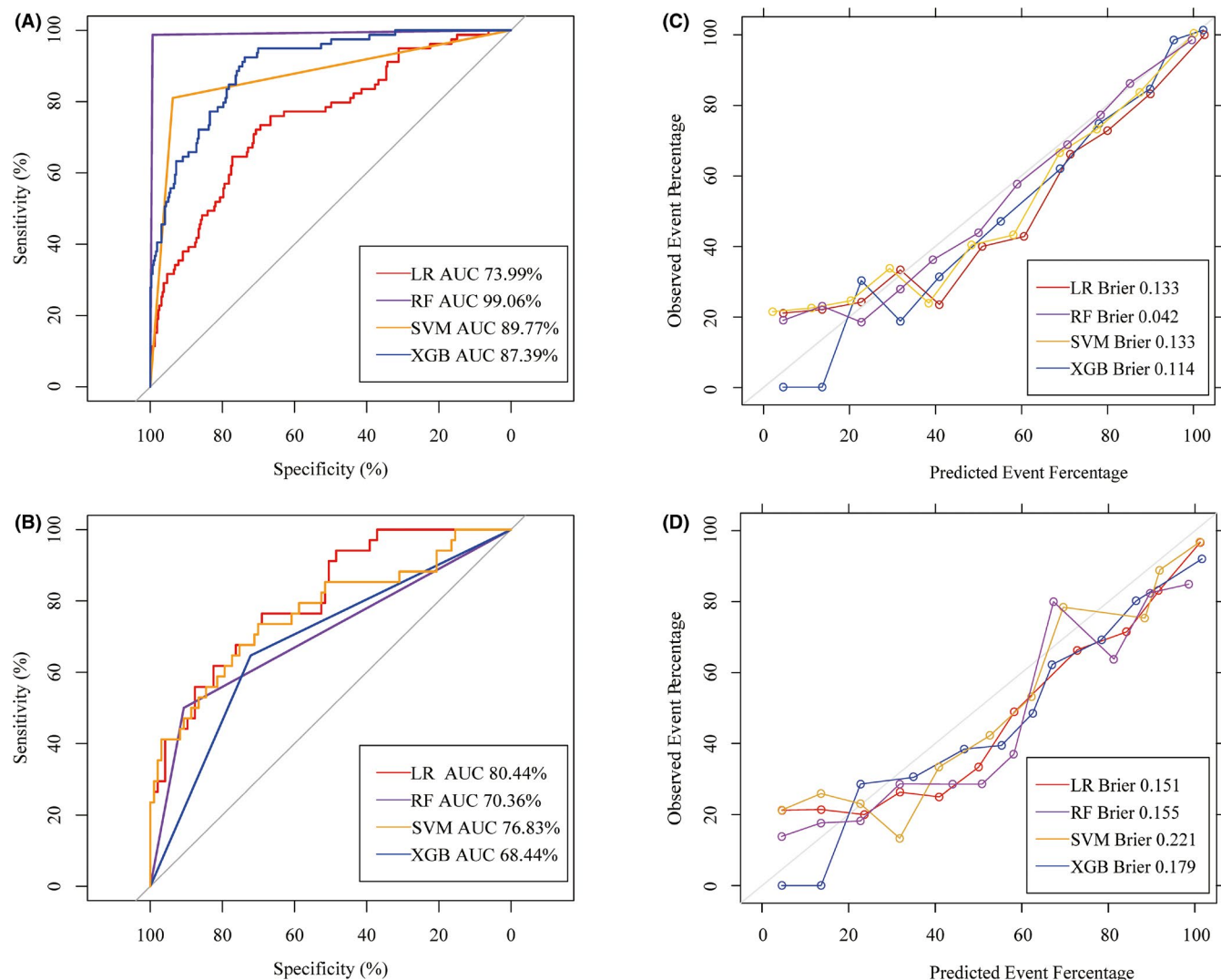


FIGURE 3 ROC of models and calibration plot in training dataset and testing dataset (A and C represented training dataset. B and D represented testing dataset)

TABLE 5 Delirium prediction performance using AUROC

Model	Training sets AUC (95%CI)	Testing sets AUC (95%CI)
LR	73.99% (67.63%–80.35%)	80.44% (72.24%–88.64%)
RF	99.06% (97.74%–100%)	70.36% (61.35%–79.37%)
XGB	89.77% (86.21%–93.32%)	76.83% (66.77%–86.89%)
SVM	87.39% (82.84%–91.94%)	68.44% (59.13%–77.74%)

Abbreviations: LR, logistic regression; RF, random forest; SVM, support vector machine; XGB, extreme gradient boosting.

There are many ways to build a POD prediction model, but many mathematical terms are always involved.^{12,28} This is not conducive to the understanding and use of a model by medical staff. At the same time, many disease prediction models are transformed into certain formulas, limiting the availability of prediction models.^{11,12} Therefore, the model established in this study was transformed into a nomogram to increase the availability of the model further.

In this study, we established a predictive model and incorporated the following eight variables into its construction: age, intraoperative blood loss, anesthesia duration, extubation time, ICU admission, MMSE score, CCI score, and postoperative NLR. The optimal predictive model was represented by a nomogram. It is a new concept to use a nomogram to estimate the risk of POD. The LR model performed well, with AUCs of 73.99% and 80.44% in the training and testing datasets, respectively. The calibrations of the models were compared quantitatively using Brier scores. The calibration of the LR model showed good agreement between the prediction outcome and the actual observed outcome. For the application of this model, the sum of the scores corresponding to each entry in the nomogram was more significant than 109 points, and patients who underwent surgery had a higher risk of developing

POD. Based on this predictive model, the nomogram can be used as a tool to screen out patients with a high risk of POD. Thus, targeted interventions can be made for high-risk patients.

Finally, eight variables were included in the multivariate logistic regression analysis. We found that age, extubation time, ICU admission, MMSE score, CCI score, and postoperative NLR were independent risk factors for POD. Advanced age is known to be the most relevant risk factor for POD, and some basic systemic diseases before surgery may also increase the incidence of POD.^{29,30} Entering the ICU after surgery may also increase the incidence of POD, which may be related to the ICU environment, long-term mechanical ventilation, and the severity of the patient's disease.³¹ The MMSE assesses cognitive function in patients and is associated with POD.³² These findings are consistent with our study. Extubation time is related to residual anesthetic drugs at the end of the anesthesia maintenance period and the patient's disease state before surgery. This study also confirmed that extubation time is a risk factor for POD. The postoperative NLR is also related to POD, but CRP variable was excluded when we screened for important features in this study.

On the one hand, we infer that the NLR, a parameter derived from different white blood cell counts, is a synthesized marker of both inflammation and oxidative stress and a stronger inflammatory factor than CRP variable.^{7,8} On the other hand, CRP variable has a certain amount of missing data. Although we imputed missing data, this could still affect the screening of important features. Considering the above two aspects, the missing data could lead to the exclusion of the CRP variable and the NLR inclusion. However, two variables, intraoperative blood loss and anesthesia duration, were excluded by multivariate logistic regression. These were inconsistent with some previous research findings.^{33,34} Considering that these two variables may be potential risk factors for POD and the principle of the minimum Akaike information criterion (AIC) and the maximum AUROC of the prediction model, we finally included these two variables in the

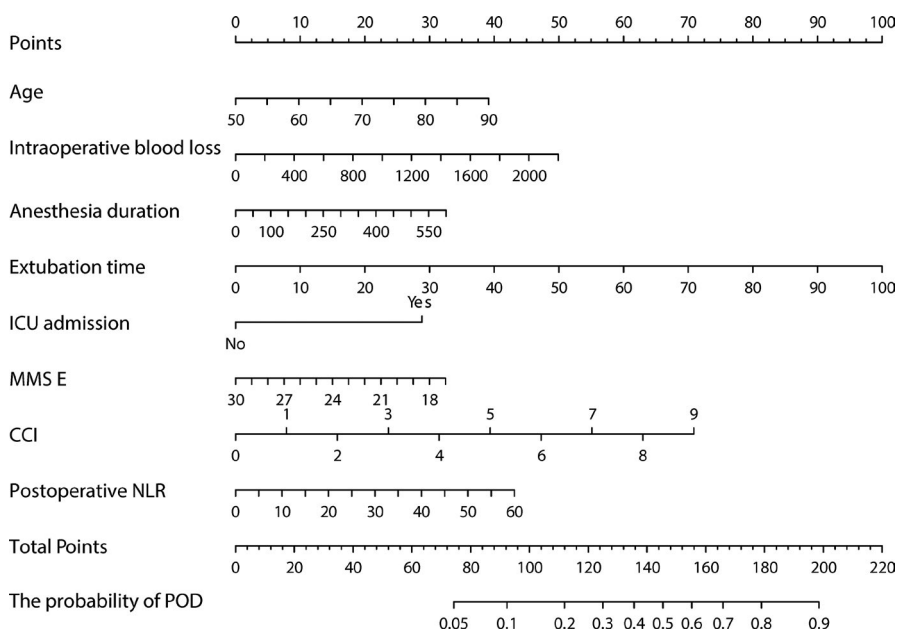


FIGURE 4 Nomogram for estimation of POD

TABLE 6 Multivariate logistic regression analysis results in training set

Variables	β coefficient	OR (95%CI)	p-value
Age	0.053	1.054 (1.017-1.093)	0.003*
Intraoperative blood loss	0.001	1.000 (0.999-1.001)	0.149
Anesthesia duration	0.002	1.002 (0.999-1.005)	0.076
Extubation time	0.027	1.027 (1.012-1.044)	0.006*
ICU admission	0.806	2.238 (1.313-3.793)	0.002*
MMSE	-0.074	0.929 (0.876-0.984)	0.012*
CCI	0.179	1.197 (1.038-1.384)	0.014*
Postoperative NLR	0.028	1.029 (1.002-1.057)	0.034*

*p < 0.05

Abbreviations: CCI, Charlson comorbidity index; MMSE, mini-mental state examination score; NLR, neutrophil-to-lymphocyte ratio.

established models. Predictive models built using independent risk factors could ignore this principle and fail to achieve the best model.

However, there are still several limitations of this study. First, this is a small sample study, and the predictive model requires a larger sample for verification. Second, the interpolation of missing data is a complex problem because data were considered to be randomly missing. In fact, there is a large field of research that builds optimal imputation algorithms, and suboptimal imputation algorithms will decrease the performance of the predictive model. This may be a possible reason why the performance of our model is lower than that of the previous models.³⁵ However, our choice to use imputation algorithms,^{36,37} while not optimal, was better than using mean imputation. Third, the data in this study came from a single large academic medical center. Thus, this model may not have similar effects when used in other medical institutions. Most likely, the model will need to be recalibrated when used by another institution. The exact weights of the features may change through such recalibration. Finally, this model requires an independent dataset to test the extrapolation and generalization of the model. We hope to collect enough external validation datasets to improve this model in the future further.

The benefits of machine-learning technology are large, especially in the medical industry. For example, using machine-learning technology to establish disease prediction and risk assessment models can help clinicians better identify the factors that truly drive the occurrence and development of diseases.

5 | CONCLUSIONS

We developed four different POD prediction models and calibrated them with Brier Score to select the model with the best performance. We believe that the model is an important tool that should be utilized to screen out the high-risk group of POD.

ACKNOWLEDGMENTS

We would like to thank the study participants, data collectors and obstetricians and nurses for their unreserved help. Finally, we are grateful to those who directly or indirectly supported us.

DISCLOSURE STATEMENT

The authors declare that they have no competing interest.

AUTHORS' CONTRIBUTIONS

Jun-Li Cao designed the study, critically reviewed the manuscript, approves the final version, and is accountable for the work. Xiao-Yi Hu, He Liu, and Yuan Han designed the study, conducted the study, collected the data, prepared the manuscript, critically reviewed the manuscript, approved the final version, and are accountable for the work. Xing Gao, Yang Zhou, and Jian Zhou helped conduct the study and collected the data. Hui-Lian Guan and Xun Sun analyzed and interpreted the data. Xue Zhao and Qiu Zhao helped prepare the manuscript and critically reviewed the manuscript.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from the corresponding author upon reasonable request.

ORCID

Jun-Li Cao  <https://orcid.org/0000-0003-1868-9603>

REFERENCES

- Schenning KJ, Deiner SG. Postoperative delirium in the geriatric patient. *Anesthesiol Clin*. 2015;33(3):505-516.
- Deiner S, Luo X, Lin H-M, et al. Intraoperative infusion of dexmedetomidine for prevention of postoperative delirium and cognitive dysfunction in elderly patients undergoing major elective noncardiac surgery: a randomized clinical trial. *JAMA Surg*. 2017;152(8):e171505.
- Avidan MS, Fritz BA, Maybrier HR, et al. The Prevention of Delirium and Complications Associated with Surgical Treatments (PODCAST) study: protocol for an international multicenter randomized controlled trial. *BMJ Open*. 2014;4(9):e005651.
- Janssen TL, Alberts AR, Hooft L, et al. Prevention of postoperative delirium in elderly patients planned for elective surgery: systematic review and meta-analysis. *Clin Interv Aging*. 2019;14:1095-1117. Published 2019 June 19.
- Inouye SK, Westendorp RGJ, Saczynski JS. Delirium in elderly people. *Lancet*. 2014;383(9920):911-922.
- Zaal IJ, Slooter AJ. Delirium in critically ill patients: epidemiology, pathophysiology, diagnosis and management. *Drugs*. 2012;72(11):1457-1471.
- Maldonado JR. Neuropathogenesis of delirium: review of current etiologic theories and common pathways. *Am J Geriatr Psychiatry*. 2013;21(12):1190-1222.
- Egberts A, Mattace-Raso FU. Increased neutrophil-lymphocyte ratio in delirium: a pilot study. *Clin Interv Aging*. 2017;12:1115-1121. Published 2017 July 14.
- Vasunilashorn SM, Ngo LH, Jones RN, et al. The association between C-reactive protein and postoperative delirium differs by Catechol-O-Methyltransferase genotype. *Am J Geriatr Psychiatry*. 2019;27(1):1-8.

10. Kotfis K, Bott-Olejnik M, Szylińska A, et al. Could Neutrophil-to-Lymphocyte Ratio (NLR) serve as a potential marker for delirium prediction in patients with acute ischemic stroke? A prospective observational study. *J Clin Med*. 2019;8(7):1075. Published 2019 July 22.
11. Huang H-W, Zhang G-B, Li H-Y, et al. Development of an early prediction model for postoperative delirium in neurosurgical patients admitted to the ICU after elective craniotomy (E-PREPOD-NS): a secondary analysis of a prospective cohort study. *J Clin Neurosci*. 2021;90:217-224.
12. Wang G, Zhang L, Qi Y, et al. Development and validation of a postoperative delirium prediction model for elderly orthopedic patients in the intensive care unit. *J Healthc Eng*. 2021;2021:9959077.
13. Rajkomar A, Oren E, Chen K, et al. Scalable and accurate deep learning for electronic health records. *NPJ Digit Med*. 2018;1:18.
14. Fogel AL, Kvedar JC. Artificial intelligence powers digital medicine. *NPJ Digit Med*. 2018;1:5.
15. Rudolph JL, Marcantonio ER, Culley DJ, et al. Delirium is associated with early postoperative cognitive dysfunction. *Anesthesia*. 2008;63(9):941-947.
16. Inouye SK, van Dyck CH, Alessi CA, et al. Clarifying confusion: the confusion assessment method. A new method for detection of delirium. *Ann Intern Med*. 1990;113(12):941-948.
17. Ely EW, Margolin R, Francis J, et al. Evaluation of delirium in critically ill patients: validation of the Confusion Assessment Method for the Intensive Care Unit (CAM-ICU). *Crit Care Med*. 2001;29(7):1370-1379.
18. Avidan MS, Maybrier HR, Abdallah AB, et al. Intraoperative ketamine for prevention of postoperative delirium or pain after major surgery in older adults: an international, multicentre, double-blind, randomised clinical trial. *Lancet*. 2017;390(10091):267-275.
19. Riley RD, Ensor J, Snell KIE, et al. Calculating the sample size required for developing a clinical prediction model. *BMJ*. 2020;368:m441.
20. van Eijk MM, Slooter AJ. Duration of ICU delirium, severity of the underlying disease, and mortality. *Am J Respir Crit Care Med*. 2010;181(4):419-420.
21. Bellelli G, Mazzola P, Morandi A, et al. Duration of postoperative delirium is an independent predictor of 6-month mortality in older adults after hip fracture. *J Am Geriatr Soc*. 2014;62(7):1335-1340.
22. Hudetz JA, Byrne AJ, Patterson KM, et al. Postoperative delirium is associated with postoperative cognitive dysfunction at one week after cardiac surgery with cardiopulmonary bypass. *Psychol Rep*. 2009;105(3 Pt 1):921-932.
23. van den Boogaard M, Schoonhoven L, Evers AWM, et al. Delirium in critically ill patients: impact on long-term health-related quality of life and cognitive functioning. *Crit Care Med*. 2012;40(1):112-118.
24. Erikson K, Ala-Kokko TI, Koskenkari J, et al. Elevated serum S-100 β in patients with septic shock is associated with delirium. *Acta Anaesthesiol Scand*. 2019;63(1):69-73.
25. Gailiūšas M, Andrejaitienė J, Širvinskis E, et al. Association between serum biomarkers and postoperative delirium after cardiac surgery. *Acta Med Litu*. 2019;26(1):8-10.
26. Ballweg T, White M, Parker M, et al. Association between plasma tau and postoperative delirium incidence and severity: a prospective observational study. *Br J Anaesth*. 2021;126(2):458-466.
27. Kazmierski J, Banys A, Latek J, et al. Mild cognitive impairment with associated inflammatory and cortisol alterations as independent risk factor for postoperative delirium. *Dement Geriatr Cogn Disord*. 2014;38(1-2):65-78.
28. Xing H, Zhou W, Fan Y, Wen T, Wang X, Chang G. Development and validation of a postoperative delirium prediction model for patients admitted to an intensive care unit in China: a prospective study. *BMJ Open*. 2019;9(11):e030733.
29. Stoffels JMJ, van Munster BC, Muller M. Delirium in the elderly; article for education and training purposes. *Ned Tijdschr Geneesk*. 2020;164:D4953.
30. Es O, Tg F, Tt H, et al. Delirium in older persons: advances in diagnosis and treatment. *JAMA*. 2017;318:1161-1174.
31. Luetz A, Grunow JJ, Mörgeli R, et al. Innovative ICU solutions to prevent and reduce delirium and post-intensive care unit syndrome. *Semin Respir Crit Care Med*. 2019;40(5):673-686.
32. Price CC, Garvan C, Hizel LP, et al. Delayed recall and working memory MMSE domains predict delirium following cardiac surgery. *J Alzheimers Dis*. 2017;59(3):1027-1035.
33. Ravi B, Pincus D, Choi S, et al. Association of duration of surgery with postoperative delirium among patients receiving hip fracture repair. *JAMA Netw Open*. 2019;2(2):e190111. Published 2019 February 1.
34. Hasegawa T, Saito I, Takeda D, et al. Risk factors associated with postoperative delirium after surgery for oral cancer. *J Craniomaxillofac Surg*. 2015;43(7):1094-1098.
35. Wang Y, Lei L, Ji M, et al. Predicting postoperative delirium after microvascular decompression surgery with machine learning. *J Clin Anesth*. 2020;66:109896.
36. Zhang Z. Multiple imputation with multivariate imputation by chained equation (MICE) package. *Ann Transl Med*. 2016;4(2):30.
37. De Silva AP, Moreno-Betancur M, De Livera AM, et al. Multiple imputation methods for handling missing values in a longitudinal categorical variable with restrictions on transitions over time: a simulation study. *BMC Med Res Methodol*. 2019;19(1):14.

How to cite this article: Hu X-Y, Liu H, Zhao X, et al. Automated machine learning-based model predicts postoperative delirium using readily extractable perioperative collected electronic data. *CNS Neurosci Ther*. 2022;28:608-618. <https://doi.org/10.1111/cns.13758>