

Development and temporal validation of a prediction model for emergence agitation in older patients undergoing orthopedic surgery with general anesthesia: a single-center retrospective cohort study

Lizhen WU

Hainan General Hospital (Hainan Affiliated Hospital of Hainan Medical University)

Lei Chen

Hainan General Hospital (Hainan Affiliated Hospital of Hainan Medical University)

Xiaoman Huang

Hainan General Hospital (Hainan Affiliated Hospital of Hainan Medical University)

Dianxiu Zhang

Hainan General Hospital (Hainan Affiliated Hospital of Hainan Medical University)

Manli Chen

Hainan General Hospital (Hainan Affiliated Hospital of Hainan Medical University)

Zhihua Wang

Hainan General Hospital (Hainan Affiliated Hospital of Hainan Medical University)

Cejun Fu

fcj-mengxiang@163.com

Hainan General Hospital (Hainan Affiliated Hospital of Hainan Medical University)

Yingliang Wu

Hainan General Hospital (Hainan Affiliated Hospital of Hainan Medical University)

Research Article

Keywords: Machine learning, Older patients, Orthopedic surgery, Emergence agitation, Prediction model, SHAP analysis

Posted Date: October 18th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10801002/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.
[Read Full License](#)

Additional Declarations: No competing interests reported.

Abstract

Background

To develop and validate a risk prediction model for emergence agitation (EA) in older patients undergoing orthopedic surgery under general anesthesia, thereby providing an adjunctive tool for early risk stratification in the post-anesthesia care unit (PACU).

Methods

This single-center retrospective cohort study included 1,339 patients treated between January and December 2025 in the model development cohort and 670 patients treated between January and June 2026 in a same-center temporal validation cohort. Predictors were selected using the least absolute shrinkage and selection operator (LASSO). Five candidate models were developed using logistic regression (LR), classification and regression tree (CART), random forest (RF), extreme gradient boosting (XGBoost), and support vector machine (SVM). Model performance was evaluated using nested stratified fivefold cross-validation. The final model was selected by jointly considering discrimination, calibration, the Brier score, and decision curve analysis, and was interpreted using SHapley Additive exPlanations (SHAP).

Results

The incidence of EA was 29.6% in the development cohort and 22.2% in the temporal validation cohort. LASSO selected seven predictors: drinking history, history of hypertension, history of stroke, sevoflurane use, patient-controlled analgesia pump setup, indwelling urinary catheter, and PaCO₂. Considering overall performance and parsimony, LR was selected as the final model. Its area under the receiver operating characteristic curve (AUC) was 0.929 (95% confidence interval [CI] 0.913–0.945) in internal validation and 0.904 (95% CI 0.876–0.932) in temporal validation. The calibration intercept was – 0.797 and the calibration slope was 0.729, indicating some overestimation of risk.

Conclusions

The LR model showed good discrimination and may support risk stratification in the PACU. Multicenter external validation and evaluation of calibration in new target populations are required before broader implementation.

Introduction

Emergence agitation (EA) is an acute behavioral disturbance commonly observed during early recovery from general anesthesia. It is characterized by psychomotor agitation, excessive motor activity, and

perceptual disturbance, often accompanied by disorientation, involuntary struggling, or attempts to remove medical devices [1, 2]. The reported incidence of EA in adults undergoing general anesthesia ranges from 2.5% to 25.5% [3, 4]. With population aging and the increasing volume of orthopedic procedures in older adults, EA has become an important concern in this population because of reduced physiological reserve, multiple comorbidities, and substantial surgical stress. Severe EA may result in device removal, falls, wound bleeding, and other adverse events that threaten perioperative safety [5]. An incidence as high as 37.6% has been reported among older patients undergoing hip or knee arthroplasty [6]. Therefore, risk stratification at the time of tracheal extubation using available perioperative information may help identify patients who require enhanced monitoring and prompt intervention in the post-anesthesia care unit (PACU).

Evidence on the prediction of EA in older patients undergoing orthopedic surgery under general anesthesia remains limited. Previous studies have primarily examined associated factors or developed models using conventional regression methods [7]. Some have been limited by an unclear prediction time point, reliance on univariable screening, inadequate internal validation, or the absence of independent temporal validation. Moreover, perioperative variables may have complex nonlinear relationships and interactions that are not fully captured by a single modeling approach. Machine learning methods can accommodate different forms of predictor–outcome relationships and have increasingly been applied to clinical risk prediction [8–11]. In this study, we used five algorithms to develop candidate models for predicting EA in older patients undergoing orthopedic surgery under general anesthesia. We compared their performance using nested cross-validation and temporally validated the final model in patients treated at the same center during a subsequent period. SHAP was used to explain the contributions of individual predictors to model output. Our aim was to develop an accurate and interpretable EA risk prediction model to support early risk stratification and targeted monitoring in the PACU.

Methods

Study design

This single-center retrospective cohort prediction model study comprised model development and same-center temporal validation. The study was approved by the Ethics Committee of Hainan General Hospital (approval No. EC-YLY-75-01). Because anonymized retrospective clinical data were used, the requirement for informed consent was waived by the ethics committee.

Study population

The development cohort comprised 1,339 older patients who underwent orthopedic surgery with endotracheal general anesthesia at Hainan General Hospital between January and December 2025 and

were subsequently admitted to the PACU. An additional 670 patients who underwent surgery at the same center between January and June 2026 constituted the temporal validation cohort.

The inclusion criteria were: (1) age ≥ 60 years; (2) elective or emergency orthopedic surgery; (3) general anesthesia with endotracheal intubation; and (4) postoperative admission to the PACU for observation. The exclusion criteria were: (1) preoperative documentation of delirium, severe cognitive impairment, or another condition that prevented the patient from following basic instructions; (2) missing time of endotracheal extubation, precluding determination of the prediction time point (T0); and (3) absence of a Richmond Agitation–Sedation Scale (RASS) assessment during the PACU stay, precluding ascertainment of EA.

Sample size calculation

Thirty-one candidate predictors were prespecified. Surgical type had four categories and was represented by three dummy variables, whereas the American Society of Anesthesiologists (ASA) physical status classification had four levels and was represented by three dummy variables. The initial sample size estimate was based on the conventional rule of at least 10 outcome events per candidate parameter. Accordingly, at least 350 EA events were considered necessary for model development. Assuming an EA incidence of 37.6% based on a previous study [6] and allowing for approximately 10% missing data, at least 1,035 patients were required. The final development cohort included 1,339 patients, of whom 397 experienced EA, thereby meeting the prespecified sample size requirement.

Data sources and variable collection

Data were obtained from the electronic medical record system, anesthesia information management system, and PACU clinical records of Hainan General Hospital. Clinical data were extracted according to a prespecified data dictionary with standardized variable definitions, units, and coding rules. After extraction, the data were anonymized and checked for duplicate records, missing values, outliers, and temporal inconsistencies. Questionable entries were verified against the original clinical records.

Prediction time point

Endotracheal extubation was defined as the model prediction time point (time zero, T0), based on the extubation time documented in the electronic anesthesia record. The recording time of every candidate predictor was verified, and only variables that had been administered, measured, and documented before extubation were eligible. Variables that could not be confirmed as preceding T0 were excluded from the primary model.

Candidate predictors

Thirty-one candidate predictors were prespecified through a literature review, expert consultation, clinical experience, and data availability. These comprised: (1) demographic and general clinical characteristics: sex, age, body mass index (BMI), smoking history, drinking history, and preoperative histories of hypertension, diabetes mellitus, and stroke; (2) surgical characteristics: surgical urgency, type of surgery, ASA physical status, duration of surgery, and intraoperative blood loss; (3) anesthetic medications and perioperative management: use of sevoflurane, propofol, etomidate, or benzodiazepines before T0, and administration of reversal or emergence-promoting agents, patient-controlled analgesia pump setup, nerve block, indwelling urinary catheter, and surgical drain; and (4) laboratory measurements obtained before T0: pH, arterial partial pressure of oxygen (PaO₂), arterial partial pressure of carbon dioxide (PaCO₂), bicarbonate (HCO₃⁻), potassium, glucose, sodium, lactate, and preoperative albumin.

Surgical urgency was classified as elective or emergency. According to the primary procedure, surgery was categorized as joint arthroplasty, spinal surgery, internal fixation of a fracture, or another orthopedic procedure. When multiple laboratory measurements were available before T0, the value closest to T0 and confirmed to have been obtained before T0 was used. Pain scores were recorded after extubation and close to the first EA assessment; because they could not be confirmed to precede EA onset, they were excluded from the primary prediction model.

Outcome definition

The primary outcome was EA. After endotracheal extubation, PACU nurses who had received standardized training assessed each patient's level of sedation and agitation using the RASS. Scores range from - 5 to + 4; negative scores indicate increasing levels of sedation, whereas higher positive scores indicate more severe agitation. The outcome assessment window extended from extubation until discharge from the PACU. The highest RASS score documented during this window was used to determine the outcome. EA was defined as a maximum RASS score $\geq + 2$, whereas a maximum score $\leq + 1$ indicated no EA [12].

Data preprocessing and handling of missing data

Data exported from the clinical information systems were first subjected to logical and range checks. Apparent data-entry errors, unit inconsistencies, and clinically implausible values were checked against the original medical records, anesthesia records, or laboratory reports. Values that could be verified were corrected accordingly; otherwise, they were treated as missing. Extreme values that were verified and clinically plausible were retained.

Baseline characteristics were summarized using the verified data before model-based imputation, and the number and proportion of missing values were reported for each variable. In each outer iteration of nested cross-validation, missing continuous variables were imputed using the median of the outer training set, and missing categorical variables were imputed using its mode. Categorical variables were

dummy-coded using prespecified reference categories. Imputation parameters, coding rules, and standardization parameters were estimated exclusively from the outer training set and then applied unchanged to the corresponding outer validation set. The outer validation set was not used to determine any preprocessing parameter.

During nested cross-validation, continuous variables were standardized as z scores using the mean and standard deviation of the corresponding outer training set before LASSO selection and before fitting models that required scaling. The resulting parameters were applied to the corresponding outer validation set. After LR had been selected as the final model, it was refitted in the complete 2025 development cohort using predictors on their original clinical scales, thereby producing a formula that could be applied directly without standardizing continuous predictors. Missing-data imputation and categorical-variable coding rules were estimated from the complete development cohort and then applied unchanged to the 2026 temporal validation cohort.

Descriptive statistical analysis

Demographic and perioperative characteristics were summarized descriptively. Because most continuous variables were non-normally distributed, they are presented as medians and interquartile ranges [M (P25, P75)] and were compared using the Mann–Whitney U test. Categorical variables are presented as counts and percentages [n (%)] and were compared using Pearson's χ^2 test or Fisher's exact test, as appropriate according to expected cell counts. Comparisons between patients with and without EA were performed solely to describe baseline characteristics and were not used for predictor screening. All prespecified candidate variables meeting the data-quality requirements were entered into the subsequent LASSO logistic regression.

Predictor selection

Predictors were selected using the least absolute shrinkage and selection operator (LASSO) [13]. LASSO applies an L1 regularization penalty to shrink regression coefficients, with some coefficients reduced to zero, thereby enabling variable selection and dimensionality reduction. Continuous variables were standardized before LASSO analysis. P values from baseline group comparisons or univariable analyses were not used as inclusion criteria. During internal validation, the penalty parameter was selected within each outer training set using stratified fivefold cross-validation. The one-standard-error rule was used to select λ_{1se} , defined as the largest λ for which the cross-validation error was within one standard error of the minimum. Predictors with nonzero coefficients at λ_{1se} were used as common input features for all five candidate models in that outer iteration. If no predictors were retained at λ_{1se} , predictors with nonzero coefficients at λ_{min} were used. The frequency with which each predictor was selected across the outer training sets was recorded to assess selection stability.

After internal validation and model comparison, λ_{1se} was re-estimated in the complete 2025 development cohort using stratified tenfold cross-validation to obtain a fixed set of predictors for the

final model. No predictor was subsequently removed on the basis of its P value in the multivariable logistic regression.

Model development and nested cross-validation

We developed LR, CART, RF, XGBoost, and SVM models to predict the risk of EA. Internal validation was performed using nested stratified fivefold cross-validation. The outer cross-validation loop was used to estimate generalization performance, whereas the inner loop was used to select the LASSO penalty parameter and tune model hyperparameters.

The development cohort was randomly partitioned into five mutually exclusive folds. In each outer iteration, four folds were used for training and the remaining fold for validation, such that each patient served once as an outer validation observation. Within each iteration, imputation parameters, categorical-variable coding rules, and standardization parameters were estimated using only the outer training set and were then applied to the outer validation set. The outer validation set did not contribute to estimation of any preprocessing parameter.

Within each outer training set, stratified fivefold cross-validation was used to select the LASSO penalty parameter using binomial deviance as the performance measure, and λ_{1se} was selected according to the one-standard-error rule. Predictors with nonzero coefficients at λ_{1se} were used as common inputs for the five candidate models in that outer iteration. If no predictor was retained at λ_{1se} , variables with nonzero coefficients at λ_{min} were used. Predictor inclusion across the outer training sets was recorded to assess selection stability.

Using the LASSO-selected predictors, a separate stratified fivefold cross-validation procedure was then performed within the outer training set to conduct grid searches for CART, RF, SVM, and XGBoost. The hyperparameter combination with the highest mean AUC across the inner validation folds was selected. No additional hyperparameter tuning was performed for LR. Each candidate model was fitted to the complete outer training set using the selected predictors and optimal hyperparameters and then applied to the corresponding outer validation set to generate out-of-fold predicted probabilities. The outer validation set was not involved in predictor selection, hyperparameter optimization, or model fitting.

Out-of-fold predicted probabilities from the five outer validation sets were pooled to calculate internal validation performance. Performance estimates were not obtained by simply averaging fold-specific metrics. Fixed random seeds were used for all stratified sampling and model fitting to ensure reproducibility.

Model performance assessment and final model selection

Model performance was evaluated using the pooled out-of-fold predictions from the five outer validation folds. Discrimination was assessed using the AUC and its 95% CI. Overall prediction error and calibration were assessed using the Brier score, calibration plots, calibration intercept, and calibration slope. Decision curve analysis was used to estimate clinical net benefit across a range of threshold

probabilities. A classification threshold of 0.50 was prespecified to calculate accuracy, positive predictive value, sensitivity, specificity, and the F1 score; this threshold was used only to compare the classification performance of candidate models and was not intended as a recommended clinical intervention threshold.

The final model was selected by jointly considering discrimination, calibration, the Brier score, clinical net benefit, model complexity, and interpretability. Once selected, the model was refitted in the complete development cohort using the final predictor set to obtain the definitive intercept and regression coefficients. The model structure, preprocessing rules, regression coefficients, and classification threshold were then locked before the model was applied directly to the temporal validation cohort.

Temporal validation

The locked final model was applied directly to the 2026 same-center temporal validation cohort without repeating predictor selection, model fitting, parameter adjustment, or threshold optimization. Missing-data handling and categorical-variable coding followed the rules established in the development cohort. PaCO₂ was entered into the final LR equation on its original scale in mmHg, and no preprocessing parameter was re-estimated using temporal validation data. Performance was evaluated using the AUC with its 95% CI, the Brier score, calibration intercept, calibration slope, and calibration plot. Accuracy, positive predictive value, negative predictive value, sensitivity, specificity, and the F1 score were calculated at the prespecified threshold of 0.50. Decision curve analysis was used to evaluate clinical net benefit across threshold probabilities.

Interpretability analysis of the final model

SHAP was used to interpret the final LR model fitted in the complete 2025 development cohort [14]. Exact SHAP values on the log-odds scale were calculated from the LR coefficients and the mean value of each predictor in the development cohort to quantify the magnitude and direction of each variable's contribution to individual predictions. Mean absolute SHAP values were used to rank global feature importance, and a SHAP direction plot was used to show how predictor values contributed positively or negatively to EA predictions. Positive SHAP values indicated an increase in the model-predicted risk of EA, whereas negative values indicated a decrease. SHAP values were used to explain contributions to model output and not to infer causal relationships with EA.

Statistical software

Statistical analyses and model development were performed using SPSS version 26.0 and R version 4.5.3. LASSO, CART, RF, SVM, and XGBoost were implemented using the glmnet, rpart, randomForest, e1071, and xgboost packages, respectively; LR was fitted using the stats package in base R. ROC curves, AUCs, and 95% CIs were calculated using pROC. Nested cross-validation, grid search, calibration analysis, and decision curve analysis were implemented using custom R scripts, and SHAP results were visualized using shapviz. Fixed random seeds were used for all data partitioning and resampling. Regression coefficients, odds ratios (ORs), 95% CIs, and P values are reported for the final LR model.

After LASSO selection, variables were not removed on the basis of P values from the multivariable LR model; P values are reported only as supplementary descriptions of model parameters and were not used for variable selection, model selection, or assessment of predictor importance. All baseline group comparisons used two-sided tests, with $P < 0.05$ indicating statistical significance; these comparisons were descriptive only and were not used for predictor selection.

Results

Patient characteristics and baseline comparisons

The development cohort included 1,339 older patients who underwent orthopedic surgery under general anesthesia and were admitted to the PACU at Hainan General Hospital between January and December 2025. EA occurred in 397 patients (29.6%). Most candidate predictors were complete, with missing data limited to several laboratory variables: pH, 3 patients (0.2%); HCO_3^- , 11 (0.8%); PaCO_2 , 3 (0.2%); PaO_2 , 3 (0.2%); potassium, 5 (0.4%); sodium, 7 (0.5%); glucose, 5 (0.4%); lactate, 6 (0.4%); and preoperative albumin, 41 (3.1%). The proportion of missing data was below 5% for every candidate predictor. After prespecified preprocessing, all eligible patients ($n = 1,339$) were included in model development and internal validation. Demographic and perioperative characteristics are presented in Table 1.

Table 1
Baseline and perioperative characteristics of the development cohort

Variable	Category	NoEA(n = 942)	EA(n = 397)	Pvalue	Missing,n (%)
Sex	Female	552 (58.6)	217 (54.7)	0.183	0 (0.0)
	Male	390 (41.4)	180 (45.3)		
Age (years)		69 [64, 72]	68 [63, 72]	0.609	0 (0.0)
BMI (kg/m ²)		23.44 [21.33, 25.30]	23.44 [21.26, 25.09]	0.928	0 (0.0)
Smoking history	No	577 (61.3)	229 (57.7)	0.223	0 (0.0)
	Yes	365 (38.7)	168 (42.3)		
Drinking history	No	899 (95.4)	294 (74.1)	< 0.001	0 (0.0)
	Yes	43 (4.6)	103 (25.9)		
Surgical urgency	Elective	911 (96.7)	374 (94.2)	0.034	0 (0.0)
Surgical urgency	Emergency	31 (3.3)	23 (5.8)		
History of hypertension	No	672 (71.3)	212 (53.4)	< 0.001	0 (0.0)
History of hypertension	Yes	270 (28.7)	185 (46.6)		
History of diabetes mellitus	No	739 (78.5)	319 (80.4)	0.435	0 (0.0)
History of diabetes mellitus	Yes	203 (21.5)	78 (19.6)		
History of stroke	No	894 (94.9)	214 (53.9)	< 0.001	0 (0.0)
History of stroke	Yes	48 (5.1)	183 (46.1)		
Type of surgery	Joint arthroplasty	186 (19.7)	54 (13.6)	< 0.001	0 (0.0)
Type of surgery	Spinal surgery	386 (41.0)	209 (52.6)		
Type of surgery	Internal fixation of a fracture	175 (18.6)	62 (15.6)		
Type of surgery	Other orthopedic surgery	195 (20.7)	72 (18.1)		

Variable	Category	NoEA(n = 942)	EA(n = 397)	Pvalue	Missing,n (%)
Sevoflurane	No	108 (11.5)	22 (5.5)	< 0.001	0 (0.0)
Sevoflurane	Yes	834 (88.5)	375 (94.5)		
Propofol	No	606 (64.3)	220 (55.4)	0.002	0 (0.0)
Propofol	Yes	336 (35.7)	177 (44.6)		
Etomidate	No	494 (52.4)	183 (46.1)	0.034	0 (0.0)
Etomidate	Yes	448 (47.6)	214 (53.9)		
Benzodiazepines	No	908 (96.4)	375 (94.5)	0.107	0 (0.0)
Benzodiazepines	Yes	34 (3.6)	22 (5.5)		
Reversal or emergence-promoting agent	No	581 (61.7)	220 (55.4)	0.033	0 (0.0)
Reversal or emergence-promoting agent	Yes	361 (38.3)	177 (44.6)		
Patient-controlled analgesia pump setup	No	155 (16.5)	313 (78.8)	< 0.001	0 (0.0)
Patient-controlled analgesia pump setup	Yes	787 (83.5)	84 (21.2)		
Nerve block	No	816 (86.6)	359 (90.4)	0.052	0 (0.0)
Nerve block	Yes	126 (13.4)	38 (9.6)		
ASA physical status	Ⅰ	3 (0.3)	0 (0.0)	0.043	0 (0.0)
ASA physical status	Ⅱ	551 (58.5)	207 (52.1)		
ASA physical status	Ⅲ	387 (41.1)	188 (47.4)		
ASA physical status	Ⅳ	1 (0.1)	2 (0.5)		
Indwelling urinary catheter	No	495 (52.5)	142 (35.8)	< 0.001	0 (0.0)
Indwelling urinary catheter	Yes	447 (47.5)	255 (64.2)		
Indwelling surgical drain	No	440 (46.7)	223 (56.2)	0.002	0 (0.0)
Indwelling surgical drain	Yes	502 (53.3)	174 (43.8)		
Intraoperative blood loss (mL)		100.0 [30.0, 200.0]	150.0 [50.0, 300.0]	< 0.001	0 (0.0)

Variable	Category	NoEA(n = 942)	EA(n = 397)	Pvalue	Missing,n (%)
Duration of surgery (min)		173.0 [135.0, 235.0]	200.0 [161.0, 255.0]	< 0.001	0 (0.0)
pH		7.37 [7.34, 7.42]	7.37 [7.33, 7.40]	< 0.001	3 (0.2)
HCO ₃ ⁻ (mmol/L)		23.70 [22.10, 25.40]	23.70 [22.10, 25.40]	0.936	11 (0.8)
PaCO ₂ (mmHg)		40.00 [36.00, 44.00]	42.00 [38.00, 46.00]	< 0.001	3 (0.2)
PaO ₂ (mmHg)		193.00 [113.00, 248.00]	198.00 [104.00, 271.00]	0.723	3 (0.2)
Potassium (mmol/L)		3.70 [3.40, 4.10]	3.80 [3.50, 4.10]	0.061	5 (0.4)
Sodium (mmol/L)		136.00 [135.00, 138.00]	137.00 [134.00, 138.00]	0.273	7 (0.5)
Glucose (mmol/L)		6.40 [5.50, 8.10]	6.40 [5.50, 7.70]	0.868	5 (0.4)
Lactate (mmol/L)		1.20 [0.80, 1.50]	1.10 [0.80, 1.60]	0.748	6 (0.4)
Preoperative albumin (g/L)		34.50 [32.20, 37.10]	35.40 [31.60, 37.10]	0.721	41 (3.1)

Data are presented as median [interquartile range] or n (%). EA, emergence agitation; BMI, body mass index; ASA, American Society of Anesthesiologists physical status; PaCO₂, arterial partial pressure of carbon dioxide; PaO₂, arterial partial pressure of oxygen.

Predictor selection using LASSO regression

Candidate predictors were screened using LASSO logistic regression, with the penalty parameter determined by stratified tenfold cross-validation. The cross-validation procedure yielded $\lambda_{\min} = 0.005483$ and $\lambda_{1se} = 0.024295$. Under the one-standard-error rule, seven predictors were retained at λ_{1se} : drinking history, history of hypertension, history of stroke, sevoflurane use, patient-controlled analgesia pump setup, indwelling urinary catheter, and PaCO₂. Predictors selected at λ_{1se} were used in subsequent modeling to obtain a more parsimonious and stable feature set. The LASSO coefficient paths and tenfold cross-validation curve are shown in Fig. 1.

Model performance based on out-of-fold predictions from nested stratified fivefold cross-validation

In nested stratified fivefold cross-validation, missing-data handling, variable coding and standardization, LASSO selection, hyperparameter optimization, and model fitting were repeated within each outer training fold, and predictions were generated in the corresponding outer validation fold. The five outer training folds retained 8, 7, 8, 7, and 7 parameters, respectively. Drinking history, history of hypertension, history of stroke, sevoflurane use, patient-controlled analgesia pump setup, indwelling urinary catheter, and PaCO₂ were selected in all five outer training folds, whereas duration of surgery was selected in two folds.

Based on the pooled out-of-fold predicted probabilities, XGBoost had the numerically highest AUC at 0.929 (95% CI 0.914–0.945). The AUCs for LR and RF were 0.929 (95% CI 0.913–0.945) and 0.928 (95% CI 0.912–0.943), respectively, and were similar to that of XGBoost. SVM and CART had lower AUCs. Overall, XGBoost, LR, and RF showed good discrimination, but their AUCs and 95% CIs were very similar, with no clear discriminatory advantage for any single model (Table 2 and Fig. 2A).

At the prespecified classification threshold of 0.50, XGBoost achieved an accuracy of 0.875, precision of 0.819, sensitivity of 0.743, specificity of 0.931, and F1 score of 0.779; its precision and specificity were the highest among the five models. SVM had the highest accuracy (0.877) and F1 score (0.784), whereas CART had the highest sensitivity (0.763) but the lowest AUC. The model favored by each performance measure differed, indicating that a single classification metric was insufficient for selecting the optimal model.

Calibration analysis showed that LR had the lowest Brier score (0.090), with a calibration intercept of –0.002 and slope of 0.966, which were closest to the ideal values of 0 and 1, respectively. This indicated good agreement between predicted probabilities and observed EA outcomes. XGBoost also showed good calibration, with a Brier score of 0.091, calibration intercept of 0.038, and calibration slope of 1.095. By comparison, RF showed weaker calibration, with an intercept of 0.308 and slope of 0.568. The calibration plots likewise showed that LR and XGBoost were generally closer to the ideal calibration line, whereas the other models deviated to varying degrees in some ranges of predicted risk (Fig. 2B).

Decision curve analysis showed that all five models provided greater net benefit than the “treat-all” and “treat-none” strategies over most of the examined threshold range. Positive clinical net benefit was observed at threshold probabilities of approximately 0.02–0.80 for LR, 0.05–0.80 for CART, and 0.03–0.80 for RF, SVM, and XGBoost. The decision curves overlapped substantially, and no model showed a clear clinical utility advantage (Fig. 2C).

Considering discrimination, calibration, the Brier score, and model complexity, both LR and XGBoost performed well. Although XGBoost had the numerically highest AUC, its improvement over LR was

limited. LR had the lowest Brier score, calibration intercept and slope closest to their ideal values, and a more parsimonious structure.

Table 2
Internal validation performance of the candidate models

Model	AUC_95CI	Accuracy	Precision	Sensitivity	F1	Specificity	Brier
LR	0.929 (0.913–0.945)	0.8723	0.8054	0.7506	0.7771	0.9236	0.0897
CART	0.912 (0.894–0.931)	0.8715	0.7953	0.7632	0.7789	0.9172	0.0966
RF	0.928 (0.912–0.943)	0.8701	0.8072	0.7380	0.7711	0.9257	0.0933
XGBoost	0.929 (0.914–0.945)	0.8753	0.8194	0.7431	0.7794	0.9310	0.0912
SVM	0.915 (0.897–0.932)	0.8768	0.8169	0.7531	0.7837	0.9289	0.0985

Performance estimates were calculated from pooled out-of-fold predictions generated by nested stratified fivefold cross-validation. AUC, area under the receiver operating characteristic curve; CI, confidence interval; LR, logistic regression; CART, classification and regression tree; RF, random forest; XGBoost, extreme gradient boosting; SVM, support vector machine.

Selection of the final model

On the basis of the overall results of nested stratified fivefold cross-validation, LR was selected as the final prediction model. The model was refitted in the complete 2025 development cohort using the seven prespecified predictors to estimate the final intercept and regression coefficients. Predictors were entered on their original clinical scales. PaCO₂ was entered in mmHg; its regression coefficient and OR therefore represent the change in the log odds and odds of EA, respectively, for each 1-mmHg increase. Drinking history, history of hypertension, history of stroke, sevoflurane use, indwelling urinary catheter, and higher PaCO₂ were associated with a higher predicted probability of EA, whereas patient-controlled analgesia pump setup was associated with a lower predicted probability. Final model parameters are presented in Table 3.

Table 3
Parameters of the logistic regression prediction model

Predictor	β	Standard error	OR	95% CI	P value
Intercept	-4.379	0.725		—	<0.001
Drinking history	2.214	0.293	9.148	5.148–16.258	<0.001
History of hypertension	0.590	0.190	1.804	1.244–2.618	0.002
History of stroke	2.809	0.246	16.593	10.251–26.858	<0.001
Sevoflurane use	1.414	0.358	4.113	2.038–8.300	<0.001
Patient-controlled analgesia pump setup	-3.419	0.211	0.033	0.022–0.050	<0.001
Indwelling urinary catheter	1.199	0.195	3.317	2.262–4.865	<0.001
PaCO ₂	0.054	0.015	1.056	1.025–1.087	<0.001

β , regression coefficient; OR, odds ratio; CI, confidence interval. Binary predictors were coded as no = 0 and yes = 1. PaCO₂ was entered in mmHg.

Final model equation: $\text{logit}(P) = -4.379 + 2.214X_1 + 0.590X_2 + 2.809X_3 + 1.414X_4 - 3.419X_5 + 1.199X_6 + 0.054X_7$, and $P = \exp[\text{logit}(P)] / \{1 + \exp[\text{logit}(P)]\}$. X_1 , drinking history; X_2 , history of hypertension; X_3 , history of stroke; X_4 , sevoflurane use; X_5 , patient-controlled analgesia pump setup; X_6 , indwelling urinary catheter; X_7 , PaCO₂ (mmHg).

Temporal validation results

The 2026 temporal validation cohort included 670 patients, of whom 149 developed EA (22.2%). Predictor selection, model fitting, parameter adjustment, and classification-threshold optimization were not repeated in the temporal validation cohort.

The AUC in the temporal validation cohort was 0.904 (95% CI 0.876–0.932), which was close to the internal validation AUC of 0.929 and indicated that the model retained good discrimination in subsequent patients. The Brier score was 0.105, the calibration intercept was –0.797, and the calibration slope was 0.729. Although the calibration plot showed a broadly similar trend between predicted probabilities and observed EA rates, observed rates were lower than predicted rates across most risk ranges, indicating overall risk overestimation and overly dispersed predictions.

At the prespecified threshold of 0.50, the model achieved an accuracy of 0.843, sensitivity of 0.725, specificity of 0.877, precision of 0.628, negative predictive value of 0.918, and F1 score of 0.673. It correctly identified 108 patients with EA and 457 without EA, with 64 false-positive and 41 false-negative classifications.

Decision curve analysis showed that the LR model provided greater net benefit than the “treat-all” and “treat-none” strategies at threshold probabilities of approximately 0.03–0.79, suggesting potential clinical utility within this range. Overall, the LR model retained good discrimination and net benefit across a broad range of thresholds in the same-center temporal validation cohort, although it overestimated absolute risk to some extent. The ROC, calibration, and decision curves for temporal validation are shown in Fig. 3.

SHAP-based interpretation of the prediction model

SHAP was used to interpret the LR model. Global feature importance analysis showed that patient-controlled analgesia pump setup had the highest mean absolute SHAP value, followed by history of stroke, indwelling urinary catheter, drinking history, history of hypertension, sevoflurane use, and PaCO₂ (Fig. 4A).

The SHAP direction analysis showed positive SHAP values for history of stroke, indwelling urinary catheter, drinking history, history of hypertension, and sevoflurane use, indicating that the presence of these features increased the model-predicted risk of EA. Patient-controlled analgesia pump setup had negative SHAP values, indicating a reduction in predicted risk. SHAP values generally increased with PaCO₂, suggesting that higher PaCO₂ was associated with a higher model-predicted risk of EA (Fig. 4B).

Discussion

Using perioperative clinical data from older patients undergoing orthopedic surgery under general anesthesia, we developed and validated a model to predict EA in the PACU. LASSO logistic regression selected seven predictors: drinking history, history of hypertension, history of stroke, sevoflurane use, patient-controlled analgesia pump setup, indwelling urinary catheter, and PaCO₂. All five candidate models based on these predictors showed good discrimination, with similar AUCs for LR, RF, and XGBoost. After jointly considering discrimination, calibration, the Brier score, and parsimony, LR was selected as the final model. It retained good discrimination in same-center temporal validation, although absolute risk was overestimated to some extent.

The incidence of EA was 29.6% in the development cohort and 22.2% in the temporal validation cohort. Reported rates vary substantially across studies, potentially because of differences in study populations, surgical procedures, anesthetic techniques, EA definitions, assessment instruments, and observation windows [15, 16]. The present study focused on older orthopedic patients under general anesthesia and defined EA using the highest RASS score recorded between extubation and PACU discharge. Differences in population and outcome assessment likely contributed to the discrepancy between our incidence estimates and those of previous studies. The difference between the development and temporal validation cohorts also suggests that case mix, clinical practice, or baseline outcome risk may have changed over time.

Older patients may be particularly vulnerable to pain, medications, metabolic disturbances, and environmental stimuli because of reduced neurocognitive reserve, a greater burden of comorbidity, and altered sensitivity to anesthetic agents. Substantial postoperative pain after orthopedic surgery further complicates recovery management [17–19]. Risk stratification using clinically available information before the EA assessment window may therefore help guide enhanced PACU monitoring and allocation of nursing resources.

We did not select the final model solely on the basis of AUC; discrimination, calibration, the Brier score, and clinical net benefit were considered together. Nested cross-validation showed nearly identical AUCs and overlapping 95% CIs for LR, RF, and XGBoost, with no clear discrimination advantage for the more complex models. The model favored by individual classification metrics also varied: XGBoost had higher positive predictive value and specificity, SVM had higher accuracy and F1 score, and CART had higher sensitivity. Thus, judgments about model performance depend on the metric considered and should not be based on a single measure [20, 21]. Furthermore, with limited sample size and a modest number of candidate predictors, complex machine learning methods may not outperform LR, and tree-based methods may require more data to achieve stable calibration [22, 23]. In this study, LR had the lowest Brier score, calibration parameters closest to their ideal values, and the most parsimonious structure, supporting its selection as the final model.

To enhance interpretability, we applied SHAP to the final LR model, thereby improving the clinical readability of its predictions [24, 25]. Patient-controlled analgesia pump setup made the largest relative contribution to model output. Drinking history, history of hypertension, history of stroke, sevoflurane use, and indwelling urinary catheter generally increased predicted EA risk, whereas patient-controlled analgesia pump setup generally decreased it; model-predicted risk also tended to increase with PaCO₂. For an LR model, SHAP complements regression coefficients and ORs by visually quantifying each variable's contribution to individual predictions, but it does not establish causality.

Long-term alcohol consumption may be accompanied by central nervous system adaptation, alcohol tolerance, potential withdrawal, and altered anesthetic requirements and metabolism, all of which may influence behavior during emergence [26]. Histories of hypertension and stroke may reflect cardiovascular and cerebrovascular disease burden and neurological vulnerability. Previous stroke may be associated with structural brain injury, reduced cognitive reserve, and impaired neurological compensation, potentially increasing sensitivity to anesthetic agents, hemodynamic fluctuations, and environmental stimuli [27]. Direct evidence linking hypertension to EA in adults remains inconsistent. Its predictive contribution may instead reflect chronic vascular injury, altered blood pressure regulation, or associated comorbidities and therefore requires validation in other centers and populations [28].

Previous adult studies have associated sevoflurane use and indwelling urinary catheters with EA [29, 30]. Rapid changes in the concentration of inhaled anesthetics and their emergence characteristics may affect behavior during recovery, whereas catheter-related irritation and discomfort may aggravate agitation. Patient-controlled analgesia pump setup made a substantial contribution in the SHAP

analysis. Adequate analgesia may reduce nociceptive stimulation and the stress response, promoting a smoother emergence. Studies of older patients undergoing joint replacement and adults undergoing other procedures have associated postoperative pain with EA, suggesting that adequate analgesia may improve the emergence period [31–33]. However, sevoflurane use, analgesia pump setup, and urinary catheterization are influenced by procedure type, patient condition, and clinician decisions and may therefore be subject to confounding by indication. These findings should not be interpreted as causal treatment effects.

Higher PaCO₂ was associated with higher model-predicted risk. PaCO₂ reflects ventilatory status before the prediction time point; an elevated value may accompany hypoventilation, residual respiratory depression, or acid–base disturbance and may affect emergence through changes in cerebral blood flow. Previous evidence suggests that carbon dioxide levels may influence emergence and recovery after general anesthesia [34].

The LR model achieved an AUC of 0.904 in the temporal validation cohort, slightly lower than the internal validation AUC of 0.929 but still indicative of good discrimination. The calibration intercept of – 0.797 and slope of 0.729 indicated overall overestimation of absolute EA risk and predictions that were too extreme. This may reflect the lower incidence of EA in the temporal validation cohort and temporal changes in case mix or clinical practice. The model therefore distinguished relatively high- and low-risk patients well, but its absolute risk estimates should be interpreted cautiously.

At the fixed threshold of 0.50, the model identified 72.5% of patients with EA and correctly classified 87.7% of those without EA, but 41 EA cases were classified as low risk. Decision curve analysis suggested potential net benefit at threshold probabilities of approximately 0.03–0.79; however, this does not demonstrate that use of the model improves patient outcomes. At present, the model should be considered an adjunct to enhanced PACU observation and risk management rather than a replacement for routine clinical monitoring or standardized EA assessment.

Limitations

This study has several limitations. First, it was a single-center retrospective study. Although same-center temporal validation was performed, transportability across regions, patient populations, and clinical workflows requires multicenter external validation. Second, potentially relevant predictors—including preoperative cognitive function, anxiety, frailty, depth of anesthesia, cumulative opioid dose, and dynamic physiological measurements—were unavailable. In addition, retrospective RASS documentation may have varied in timing and frequency, resulting in unmeasured predictive information and some outcome misclassification. Third, sevoflurane use, patient-controlled analgesia pump setup, and urinary catheterization may have been influenced by indications and clinical decisions; the model identifies predictive associations rather than causal effects. Simple imputation and use of a common LASSO-derived feature set may also have limited the ability of complex models to capture nonlinear relationships and interactions. Finally, the model overestimated risk in the temporal validation cohort and

has not undergone prospective clinical impact evaluation. Independent external validation, recalibration where necessary, and clinical implementation studies are therefore required before widespread use.

Conclusions

Using perioperative clinical information available before endotracheal extubation, we developed a model to predict EA in older patients undergoing orthopedic surgery under general anesthesia. The LR model, comprising drinking history, history of hypertension, history of stroke, sevoflurane use, patient-controlled analgesia pump setup, indwelling urinary catheter, and PaCO₂, showed good discrimination in both nested cross-validation and same-center temporal validation, although it overestimated risk in the temporal validation cohort. The model may serve as an adjunctive tool for EA risk stratification and enhanced monitoring in the PACU. Multicenter external validation, recalibration where necessary, and prospective clinical impact evaluation are required before broader implementation.

Declarations

Ethics approval and consent to participate

This study was approved by the Ethics Committee of Hainan General Hospital (approval No. EC-YLY-75-01) and was conducted in accordance with the Declaration of Helsinki and applicable ethical guidelines. Because this retrospective study used anonymized existing clinical data, the ethics committee waived the requirement for informed consent.

Clinical trial number

Clinical trial number: not applicable.

Consent for publication

The manuscript contains no identifiable personal information, images, or videos; therefore, consent for publication was not applicable.

Competing interests

The authors declare that they have no competing interests.

Funding

Availability of data and materials

The datasets used in this study are not publicly available because of patient privacy considerations and institutional data-management regulations.

Authors' contributions

L.W. contributed to the study design and supervision. L.C. contributed to data collection, data analysis, model development, and drafting of the manuscript. X.H., D.Z., and M.C. contributed to data collection and verification. Z.W. provided clinical guidance and critically revised the manuscript. Y.W. and C.F. supervised the study and approved the final version. All authors read and approved the final manuscript. L.W. and L.C. contributed equally as co-first authors. Y.W. and C.F. are co-corresponding authors.

Acknowledgements

The authors thank Shantian Hu, the liaison for the hospital clinical information system, for technical assistance in developing and standardizing the variable extraction list and configuring the data export process, which improved the efficiency of data extraction and subsequent verification against the original clinical records.

References

1. Wei B, Feng Y, Chen W, Ren D, Xiao D, Chen B. Risk factors for emergence agitation in adults after general anesthesia: a systematic review and meta-analysis. *Acta Anaesthesiol Scand*. 2021;65(6):719–29. 10.1111/aas.13774.
2. Zhang C, Feng J, Wang S, Gao P, Xu L, Zhu J, et al. Incidence of and trends in hip fracture among adults in urban China: a nationwide retrospective cohort study. *PLoS Med*. 2020;17(8). 10.1371/journal.pmed.1003180.
3. Moon SH, Hwang SY. Risk factors for emergence agitation in the postanesthesia care unit: a propensity score-matched analysis. *J Perianesth Nurs*. 2026;41(1):191–8. 10.1016/j.jopan.2025.06.006.
4. Song SS, Lin L, Li L, Han XD. Influencing factors and risk prediction model for emergence agitation after general anesthesia for primary liver cancer. *World J Gastrointest Surg*. 2024;16(7):2194–201. 10.4240/wjgs.v16.i7.2194.
5. Li J, Zhu H, Wang Y, Chen J, He K, Wang S. Remifentanyl is superior to propofol for treating emergence agitation in adults after general anesthesia. *Drug Des Devel Ther*. 2024;18:341–50.

- 10.2147/DDDT.S433155.
6. Wang N, Hao J, Zhang J, Du J, Luo Z. Risk factors for emergence agitation during the awakening period in elderly patients after total joint arthroplasty: a retrospective cohort study. *BMJ Open*. 2023;13(5). 10.1136/bmjopen-2022-068284.
 7. Xu L, Zhang Y, Zhang J, Xiao W, Liu Y, Li Q, et al. A Bayesian network-based predictive model for postoperative delirium following coronary artery bypass grafting. *BMC Psychiatry*. 2025;25(1):822. 10.1186/s12888-025-07299-w.
 8. Hashimoto DA, Witkowski E, Gao L, Meireles O, Rosman G. Artificial intelligence in anesthesiology: current techniques, clinical applications, and limitations. *Anesthesiology*. 2020;132(2):379–94. 10.1097/ALN.0000000000002960.
 9. Bihorac A, Ozrazgat-Baslanti T, Ebadi A, Motaei A, Madkour M, Pardalos PM, et al. MySurgeryRisk: development and validation of a machine-learning risk algorithm for major complications and death after surgery. *Ann Surg*. 2019;269(4):652–62. 10.1097/SLA.0000000000002706.
 10. Lu Y, Li Y, Chi S, Feng Y, Li G, Lin X, et al. Comparison of machine learning and logistic regression models for predicting emergence delirium in elderly patients: a prospective study. *Int J Med Inf*. 2025;199:105888. 10.1016/j.ijmedinf.2025.105888.
 11. 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;14(1):57. 10.1038/s41398-024-02762-w.
 12. Sessler CN, Gosnell MS, Grap MJ, Brophy GM, O'Neal PV, Keane KA, et al. The Richmond Agitation-Sedation Scale: validity and reliability in adult intensive care unit patients. *Am J Respir Crit Care Med*. 2002;166(10):1338–44. 10.1164/rccm.2107138.
 13. Bodinier B, Filippi S, Nøst TH, Chiquet J, Chadeau-Hyam M. Automated calibration for stability selection in penalised regression and graphical models. *J R Stat Soc Ser C Appl Stat*. 2023;72(5):1375–93. 10.1093/jrssc/qlad058.
 14. Ponce-Bobadilla AV, Schmitt V, Maier CS, Mensing S, Stodtman S. Practical guide to SHAP analysis: explaining supervised machine learning model predictions in drug development. *Clin Transl Sci*. 2024;17(11). 10.1111/cts.70056.
 15. Heily M, Gerditz M, Jarden R, Yap Y, Bellomo R. Anaesthetic emergence agitation in adults following general surgery: a scoping review. *Int J Nurs Stud Adv*. 2025;8:100320. 10.1016/j.ijnlsa.2025.100320.
 16. Lee SJ, Sung TY. Emergence agitation: current knowledge and unresolved questions. *Korean J Anesthesiol*. 2020;73(6):471–85. 10.4097/kja.20097.
 17. Aldecoa C, Bettelli G, Bilotta F, Sanders RD, Aceto P, Audisio R, et al. Update of the European Society of Anaesthesiology and Intensive Care Medicine evidence-based and consensus-based guideline on postoperative delirium in adult patients. *Eur J Anaesthesiol*. 2024;41(2):81–108. 10.1097/EJA.0000000000001876.

18. Pinho C, Cruz S, Santos A, Abelha FJ. Postoperative delirium: age and low functional reserve as independent risk factors. *J Clin Anesth.* 2016;33:507–13. 10.1016/j.jclinane.2015.09.002.
19. Duan J, Ju X, Wang X, Liu N, Xu S, Wang S. Effects of remimazolam and propofol on emergence agitation in elderly patients undergoing hip replacement: a clinical, randomized, controlled study. *Drug Des Devel Ther.* 2023;17:2669–78. 10.2147/DDDT.S419146.
20. Efthimiou O, Seo M, Chalkou K, Debray T, Egger M, Salanti G. Developing clinical prediction models: a step-by-step guide. *BMJ.* 2024;386. 10.1136/bmj-2023-078276.
21. Ojeda FM, Jansen ML, Thiéry A, Blankenberg S, Weimar C, Schmid M, et al. Calibrating machine learning approaches for probability estimation: a comprehensive comparison. *Stat Med.* 2023;42(29):5451–78. 10.1002/sim.9921.
22. Michelsen C, Jørgensen CC, Heltberg M, Jensen MH, Lucchetti A, Petersen PB, et al. Machine-learning vs. logistic regression for preoperative prediction of medical morbidity after fast-track hip and knee arthroplasty: a comparative study. *BMC Anesthesiol.* 2023;23(1):391. 10.1186/s12871-023-02354-z.
23. Austin PC, Lee DS, Wang B. The relative data hungriness of unpenalized and penalized logistic regression and ensemble-based machine learning methods: the case of calibration. *Diagn Progn Res.* 2024;8(1):15. 10.1186/s41512-024-00179-z.
24. Ladbury C, Zarinshenas R, Semwal H, Tam A, Vaidehi N, Rodin AS, et al. Utilization of model-agnostic explainable artificial intelligence frameworks in oncology: a narrative review. *Transl Cancer Res.* 2022;11(10):3853–68. 10.21037/tcr-22-1626.
25. Ocampo Osorio F, Alzate-Ricaurte S, Mejia Vallecilla TE, Cruz-Suarez GA. The anesthesiologist's guide to critically assessing machine learning research: a narrative review. *BMC Anesthesiol.* 2024;24(1):452. 10.1186/s12871-024-02840-y.
26. Fields A, Huang J, Schroeder D, Sprung J, Weingarten T. Agitation in adults in the post-anaesthesia care unit after general anaesthesia. *Br J Anaesth.* 2018;121(5):1052–8. 10.1016/j.bja.2018.07.017.
27. Hu Y, Yang M. A predictive scoring system for postoperative delirium in elderly patients with intertrochanteric fracture. *BMC Surg.* 2023;23(1):154. 10.1186/s12893-023-02065-9.
28. He M, Zhu Z, Jiang M, Liu X, Wu R, Zhou J, et al. Risk factors for postanesthetic emergence delirium in adults: a systematic review and meta-analysis. *J Neurosurg Anesthesiol.* 2024;36(3):190–200. 10.1097/ANA.0000000000000942.
29. Kim HJ, Kim DK, Kim HY, Kim JK, Choi SW. Risk factors of emergence agitation in adults undergoing general anesthesia for nasal surgery. *Clin Exp Otorhinolaryngol.* 2015;8(1):46–51. 10.3342/ceo.2015.8.1.46.
30. Ramroop R, Hariharan S, Chen D. Emergence delirium following sevoflurane anesthesia in adults: prospective observational study. *Braz J Anesthesiol.* 2019;69(3):233–41. 10.1016/j.bjan.2018.12.003.
31. Jung JY, You HJ, Kwon W, Sung TY. Comparison of emergence agitation between sevoflurane and desflurane in elderly orthopedic patients: a retrospective cohort study. *Minerva Anesthesiol.*

2026;92(5):399–408. 10.23736/S0375-9393.25.19364-4.

32. Wan Y, Ye H, Liu K. Analysis of influencing factors and clinical application of a predictive model for emergence agitation from general anesthesia after abdominal surgery. *Am J Transl Res*. 2025;17(3):1742–55. 10.62347/SSEP8225.
33. Isazadehfard K, Entezariasl M, Abitorabi A, Yousefian M. The effect of dexmedetomidine on emergence agitation and pain in post-anesthesia recovery of patients undergoing upper limb elective orthopedic surgery: a double-blind randomized clinical trial. *BMC Anesthesiol*. 2026;26(1):378. 10.1186/s12871-026-03850-8.
34. Saghaei M, Matin G, Golparvar M. Effects of intra-operative end-tidal carbon dioxide levels on the rates of post-operative complications in adults undergoing general anesthesia for percutaneous nephrolithotomy: a clinical trial. *Adv Biomed Res*. 2014;3:84. 10.4103/2277-9175.127997.

Figures

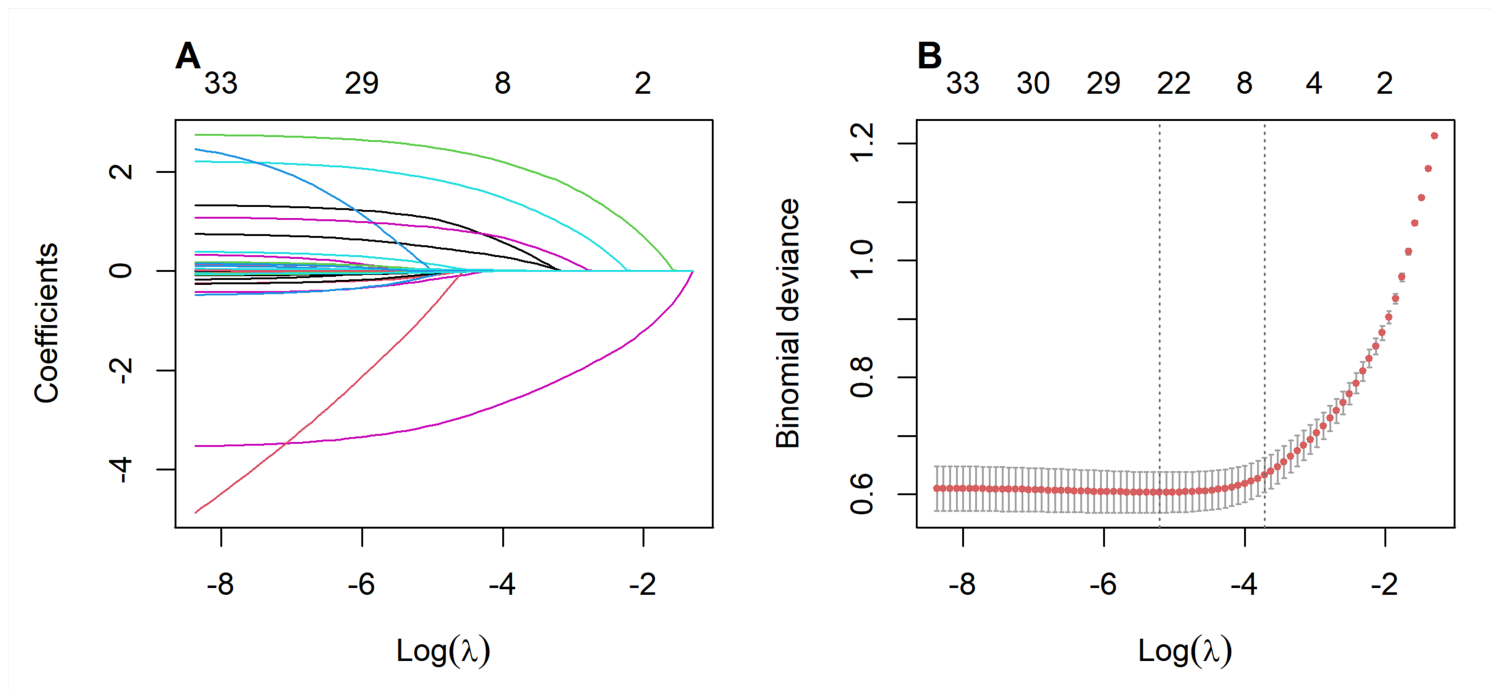


Figure 1

Predictor selection using LASSO regression.

A, LASSO coefficient paths; B, tenfold cross-validation error curve. The two vertical dashed lines in panel B indicate $\lambda_{\min}$ and λ_{1se} , respectively.

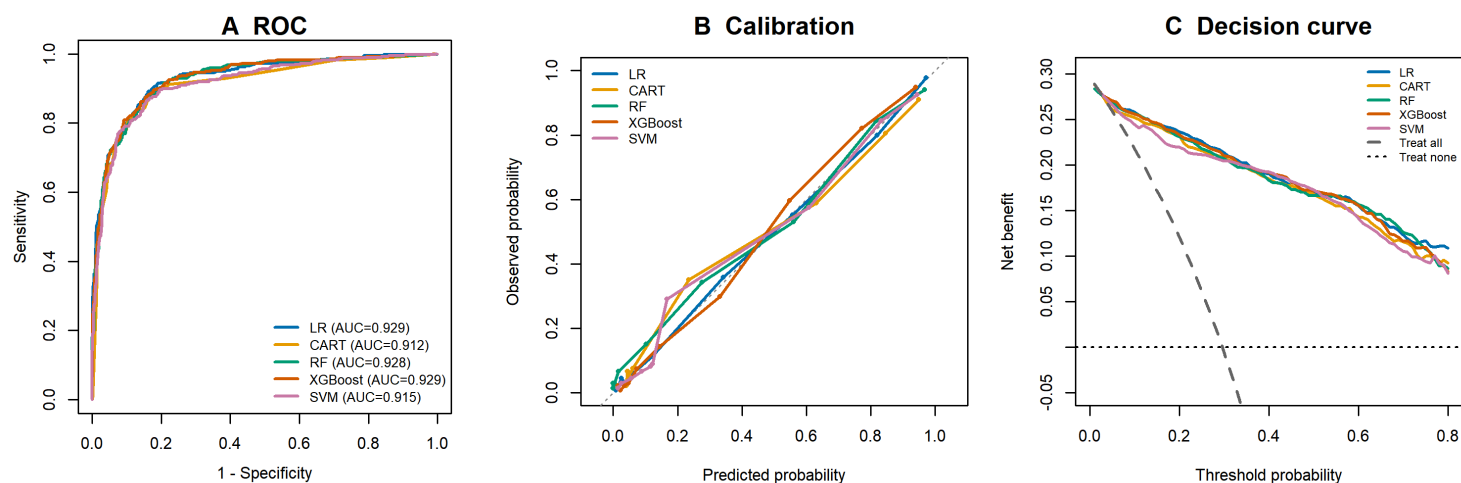


Figure 2

Internal validation performance of the candidate models.

A, receiver operating characteristic curves based on pooled out-of-fold predicted probabilities from nested stratified fivefold cross-validation; B, calibration plots, with the dashed line indicating ideal calibration; C, decision curves, with the gray dashed and black dotted lines indicating the 'treat-all' and 'treat-none' strategies, respectively.

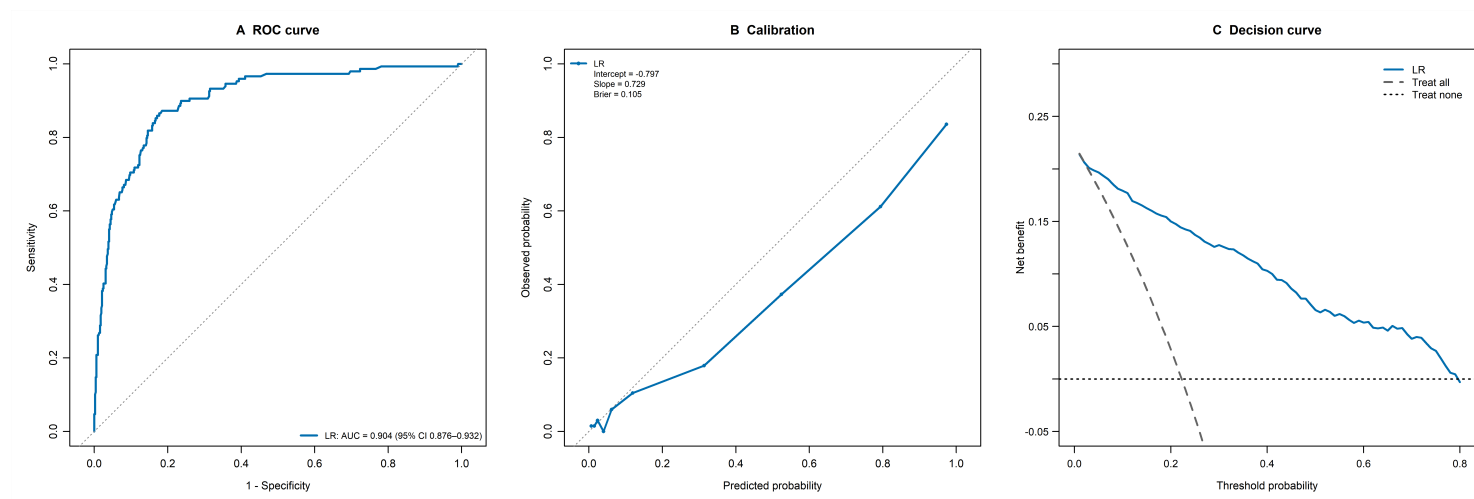


Figure 3

Performance of the logistic regression model in the temporal validation cohort.

A, receiver operating characteristic curve; B, calibration plot, with the gray dotted line indicating ideal calibration; C, decision curve, with the gray dashed and black dotted lines indicating the 'treat-all' and 'treat-none' strategies, respectively.

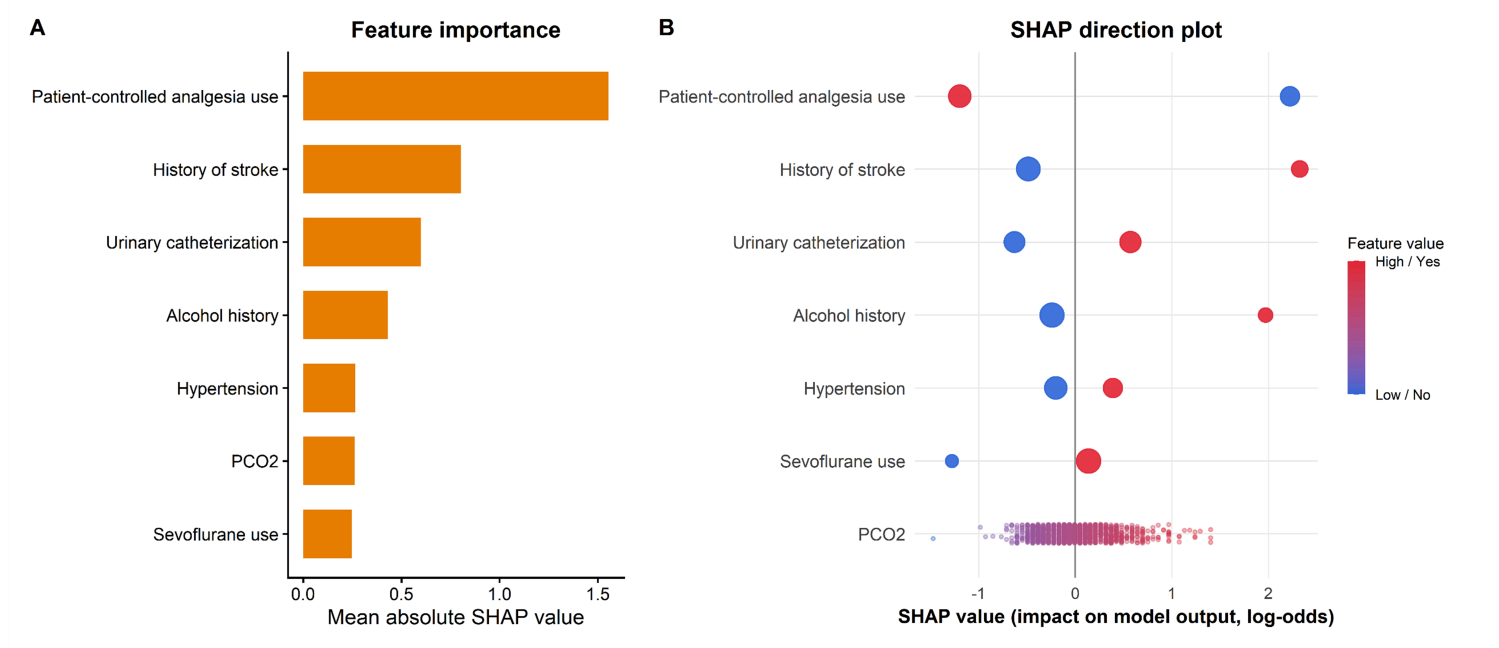


Figure 4

SHAP interpretation of the logistic regression model.

A, global feature importance ranked by mean absolute SHAP value; B, direction of predictor contributions to model output. Red indicates a value of 1 for binary predictors or a higher value for continuous predictors, whereas blue indicates a value of 0 for binary predictors or a lower value for continuous predictors. Positive SHAP values increase the model-predicted risk of emergence agitation, whereas negative SHAP values decrease it.