scientific reports

Check for updates

OPEN

Development and validation comparison of multiple models for perioperative neurocognitive disorders during hip arthroplasty

Gang Wang[1,4], Yi Xie[2,4], XiaRui Bai[1], Yuming Zhang[1] & Jiao Guo [1,3]✉

This study aims to develop optimal predictive models for perioperative neurocognitive disorders (PND) in hip arthroplasty patients, thereby advancing clinical practice. Data from all hip arthroplasty patients in the MIMIC-IV database were utilized to predict PND. With 62 variables, we applied multiple logistic regression, artificial neural network (ANN), Naive Bayes, support vector machine, and decision tree (XgBoost) algorithms to forecast PND. Feature analysis, receiver operating characteristic curve (ROC) and calibration curve plotting, and sensitivity, specificity, and F-measure β = 1 (F1-score) assessments were conducted on both training and validation sets for classifying models' effectiveness. Brier score and Index of prediction accuracy (IPA) were employed to compare prediction capabilities in both sets. Among 3,292 hip arthroplasty patients in the MIMIC database, 331 developed PND. Five models using different algorithms were constructed. After thorough comparison and validation, the ANN model emerged as the most effective model. Performance metrics on the training set for the ANN model were: ROC: 0.954, Accuracy: 0.938, Precision: 0.758, F1-score: 0.657, Brier Score: 0.048, IPA: 90.8%. On the validation set, the ANN model performed as follows: ROC: 0.857, Accuracy: 0.903, Precision: 0.539, F1-score: 0.432, Brier Score: 0.071, IPA: 71.4%. An online visualization tool was developed (https://xyyy.pythonanywhere.com/).

Keywords Perioperative neurocognitive disorders (PND), Hip arthroplasty, Model development, Artificial neural network (ANN)

Abbreviations

PND	Perioperative neurocognitive disorders
MLR	Multiple logistic regression
ANN	Artificial neural network
NB	Naive Bayes
SVM	Support vector machine
XgBoost	Decision tree
ROC	Receiver operating characteristic
IPA	Prediction accuracy
DCA	Decision curve analysis
SOFA	Sequential Organ Failure Assessment
ASA	American Society of Anesthesiologists
LASSO	The Least Absolute Shrinkage and Selection Operator
AUC	The area under the curve

In 2018, the International Perioperative Cognitive Expert Group formally introduced the term "perioperative neurocognitive disorders (PND)" to replace cognitive change associated with surgery and anesthesia. This inclusive nomenclature encompasses a spectrum of cognitive alterations, including preoperative cognitive decline, early postoperative delirium, delayed neurocognitive recovery, postoperative cognitive dysfunction, and various degrees of cognitive deficits ranging from mild to severe[1]. Perioperative neurocognitive disorders have

[1]Department of Anesthesiology, Department of Emergency Surgery, Shaanxi Provincial People's Hospital, Youyi West Road No. 256, Beilin District, Xi'an 710000, China. [2]Faculty of Business, The Hong Kong Polytechnic University, Kowloon, China. [3]Xi'an Medical University, HanGuang North Road No. 74, Beilin District, Xi'an 710000, China. [4]Gang Wang and Yi Xie share first authorship, they contributed equally to this work. ✉email: guojane1989@gmail.com

lasting effects on patients and families beyond surgery. Many with mild cognitive issues develop Alzheimer's within three years, raising dementia risk 12 times and increasing postoperative mortality significantly[2,3]. An editorial revealed that 65% of Australian patients fear lasting cognitive issues after surgery[4]. This highlights the need for more research into perioperative neurocognitive disorders due to heightened patient concern.

Perioperative neurocognitive disorders vary in prevalence among different surgeries. A major study found a notable 19.6% rate of cognitive impairment after total hip arthroplasty, ranking it highest outside cardiac procedures[5]. Another study found a significant 54% detection of neurocognitive issues in 99 patients after total joint replacement[6]. Evidently, cognitive impairment is prevalent throughout the entire joint replacement process.

Existing research have primarily focused on risk factors, with recent findings highlighting hronic illnesses, alcohol intake, auditory deficits, infections, high Apache II scores at admission, sedative and analgesic use, catheter insertion, and sleep issues as independent risks[7,8]. Furthermore, certain studies have delved into some laboratory values, revealing connections between perioperative cognitive dysfunction and serum lipid, uric acid, perioperative lactate, and inflammation-related factors. These links suggest potential predictive factors[9–11]. Systematic reviews also analyze clinical parameters associated with cognitive impairment after hip arthroplasty. Findings suggest elderliness, lower BMI, smoking, visual impairment, preoperative delirium, dementia, hypertension, and COPD as significant risk factors[12]. In addition, previous studies mainly focus on predictive factors in the general population, lacking tailored predictive analytics. There's a crucial need for precise research to deliver personalized predictions for tailored interventions. Anticipating cognitive issues during surgery not only benefits patients but also helps healthcare professionals improve monitoring and care processes[13].

Built upon the MIMIC-IV database, this study involves data from over 3000 hip arthroplasty patients, including 30+ clinical variables like demographics, lab tests, vital signs, complications, medications, and treatments. Its main goal is to create a predictive model for neurocognitive issues during hip arthroplasty using machine learning. After thorough validation and comparison, we'll determine the most effective model for precise predictions. Also, we'll develop an online tool to identify patients at risk of neurocognitive disorder during hip arthroplasty.

Methods

Data source

This study utilized patient data from the extensive MIMIC-IV database to develop and validate a predictive model for perioperative neurocognitive disorders following hip arthroplasty. Covering clinical data from over 190,000 patients and 450,000 hospital admissions at the Beth Israel Deaconess Medical Center between 2008 and 2019, the MIMIC-IV database is a valuable and publicly accessible clinical resource. Researchers can share data after completing a training program. Importantly, patient identifiers in MIMIC-IV have been meticulously removed to the fullest extent possible, ensuring the utmost protection of patient privacy.

Throughout this research, a team member, duly authorized to access the MIMIC database, employed the MIMIC-IV repository to compile a dataset that comprehensively encompasses inpatient data concerning perioperative neurocognitive disorders during individuals undergoing hip joint surgeries. This dataset incorporates an array of detailed information, encompassing demographics, laboratory findings, perioperative medication history, vital sign recordings, surgical interventions, treatment, disease diagnoses, medication regimens, and subsequent survival outcomes.

Participants and outcomes

We gathered information from all patients admitted for hip joint replacement or revision in the MIMIC-IV database, which covered surgeries conducted through various approaches, including both partial and total hip joint replacement procedures. The identification of these surgeries was based on codes from the 9th edition revision (ICD-9) and the 10th edition revision of the MIMIC-IV database (see Supplementary Appendix S1).

The main outcome of our study was on perioperative cognitive disorders occurring during hospitalization. In accordance with the guidelines outlined in the 5th edition of the Diagnostic and Statistical Manual of Mental Disorders (DSM-5), we classified delirium, mild cognitive impairment (MCI) and dementia as neurocognitive disorders in our study[14]. Data related to cognitive disorders were extracted using both ICD-9 and ICD-10 codes (see Supplementary Appendix S2).

Predictive variables

Following a comprehensive review of previously identified risk factors associated with cognitive impairment, we extracted the relevant risk factor variables from the MIMIC database, along with other pertinent variables. The included variables encompassed various domains: (1) Demographic data, including age, gender, admission type, marital status, ICU stay time, and in-hospital mortality rate; (2) Vital signs, including heart rate, systolic blood pressure, diastolic blood pressure, respiratory rate, temperature, and pulse oximetry (SpO$_2$); (3) Laboratory variables, including creatinine, urea nitrogen, hemoglobin, platelet count, international normalized ratio (INR), prothrombin time (PT), white blood cell count, lymphocytes, neutrophils, uric acid, total cholesterol, triglycerides, low-density lipoprotein (LDL), high-density lipoprotein (HDL), sodium, potassium, pH, glucose, and lactate; (4) Identification of complications based on ICD-9 and ICD-10 codes, including hypertension, diabetes, cardiovascular diseases, chronic bronchopulmonary disease, liver disease, kidney disease, craniocerebral disease, and sleep disorders(see Supplementary Appendix S3-S10); (5) Medications: including analgesics and sedatives (morphine, fentanyl, midazolam, propofol, etomidate, dexmedetomidine) and cardiovascular drugs (norepinephrine, epinephrine, nitroglycerin, nifedipine, captopril, hydrocortisone, dobutamine, and dopamine); (6) Scoring systems, including Simplified Acute Physiology Score II (SAPS II), Sequential Organ Failure Assessment (SOFA) score, and Glasgow Coma Scale (GCS); (7) Treatment strategies, including mechanical ventilation, renal replacement therapy, blood transfusion (see Supplementary Appendix

S11). For laboratory data, vital signs, and related information, we opted for the earliest available results or the first postoperative results.

Data acquisition and missing management
In this study, we extracted various patient data from the MIMIC-IV database. Demographic information was obtained from the admissions and patients tables, while laboratory results came from the lab events table. We collected complications data from the "diagnoses_icd" table and medication records from the prescriptions table. Scores for the Sequential Organ Failure Assessment (SOFA), Simplified Acute Physiology Score II (SAPS II), and Glasgow Coma Scale (GCS) were retrieved from the corresponding derived tables: sofa, SAPS II, and GCS. To identify patients with perioperative cognitive issues, we queried the "d_icd_diagnoses" and "diagnoses_icd" tables. Each record in the MIMIC-IV database is uniquely identified by the "hadm_id" (hospital admission ID), which we used to link the different data sources for further analysis.

During data mining, we obtained 62 variables from the MIMIC database. Two variables lacked data and were excluded. Variables with over 30% missing data were also omitted. Missing values in continuous variables were filled with median values, and categorical variables were encoded using ordinal techniques. Consequently, 27 variables were removed during model building, including parameters like heart rate, blood pressure, respiratory rate, oxygen saturation, temperature, and various medications and treatments such as captopril, dobutamine, dopamine, dexmedetomidine, epinephrine, etomidate, fentanyl, hydrocortisone, midazolam, morphine, nifedipine, nitroglycerin, propofol, mechanical ventilation, renal replacement therapy, and transfusion.

Sample size
In the MIMIC-IV database, 3,292 patients underwent hip arthroplasty, with 331 developing perioperative neurocognitive disorders. This subset formed our study cohort, which was then randomly split into a development group and a validation group in a 7:3 ratio. To improve the stability and robustness of model performance, we incorporated fivefold cross-validation in the model development and evaluation process, addressing potential variability due to random seed selection. Adhering to the guideline advocating for a minimum of 10 occurrences per predictive variable in machine learning algorithms, the inclusion of 3,292 patients in our study cohort met the recommended sample size criteria for rigorous statistical analysis[15].

Statistical analyses
The normality of the data was examined through the Shapiro–Wilk test. Continuous variables are presented as mean ± standard deviation for normally distributed data, or as median [P25, P75] for non-normally distributed data, while categorical variables are expressed as frequencies (percentages). To assess disparities between the PND and non-PND groups within both the development and validation sets, the Pearson chi-square test was employed to analyze distinctions among categorical variable groups. For continuous variables exhibiting non-normal distribution or heterogeneity, the Mann–Whitney U test or Kruskal–Wallis test was utilized accordingly.

Model development and validation
In this study, we developed five models employing various methodologies, encompassing multiple logistic regression (MLR), artificial neural network (ANN), Naive Bayes (NB), support vector machine (SVM), and decision tree (XgBoost) algorithms. For the MLR and SVM models, we applied the Least Absolute Shrinkage and Selection Operator (LASSO) regression for variable selection, with λ set at 1. This procedure led to the reduction of variables from 35 to 22 (see Supplementary Appendix S12). Subsequently, these selected predictive variables were utilized as inputs for the MLR and SVM to facilitate the machine learning process aimed at identifying pivotal features. Conversely, in the case of the remaining models developed using different algorithms, feature selection involved inputting all predictive variables, allowing the algorithms to autonomously discern significant features.

We conducted feature analysis plots and generated Receiver Operating Characteristic (ROC) curves for the five predictive models. The Area Under the Curve (AUC) from the ROC curves was evaluated to compare their performance. Additionally, sensitivity, specificity, and F1 score were assessed for each model on both the development and validation sets to ascertain their classification efficacy. Calibration curves were constructed to examine the alignment between predicted probabilities and actual outcomes in both datasets. Furthermore, the Brier score and the index of prediction accuracy (IPA) were computed to compare the calibration and prediction capabilities of the models. Finally, Decision Curve Analysis (DCA) plots were generated to estimate and compare the potential clinical utility of the models.

Results
Details of included cases and correlation among variables
The MIMIC-IV database encompasses records of 3,272 patients who underwent either hip arthroplasty or revision surgery. Among these, 330 cases experienced perioperative neurocognitive disorders (PND), while 2,942 cases did not. All cases were incorporated into the final analysis, comprising 264 PND cases and 2,353 non-PND cases in the training set, and 66 PND cases and 589 non-PND cases in the validation set. Table 1 provides a comprehensive overview of the characteristics exhibited by PND and non-PND patients across distinct datasets.

Correlation analysis was undertaken on all variables included in the study. The interrelations among these variables are depicted graphically as a heatmap in Fig. 1. It is evident from the heatmap that there exists a correlation between high density lipoprotein (HDL), low density lipoprotein (LDL), and total cholesterol. Furthermore, a statistically significant correlation is observed between the percentages of lymphocytes and neutrophils.

Variables	Train set			Test set		
	PND group 264	Non-PND group 2353	P1	PND group 66	Non-PND group 589	P2
Age	81.0 [73.0,87.0]	65.0[55.0, 75.0]	< 0.001	79.0 [70.0,88.0]	64.0 [53.0, 74.0]	< 0.001
Gender, n (%)			0.398			0.079
Male	125 (42.09%)	1187 (44.84%)		13 (39.39%)	128 (43.39%)	
Female	172 (57.91%)	1460 (55.16%)		20 (60.61%)	167 (56.61%)	
Marital Status, n (%)			< 0.001			< 0.001
Married	125(43.55%)	1182 (45.39%)		13 (41.94%)	120 (41.24%)	
Widowed	84(29.27%)	390 (14.98%)		11 (35.48%)	42 (14.43%)	
Single	56(19.51%)	798 (30.65%)		6 (19.35%)	101 (34.71%)	
Divorced	22(7.67%)	234 (8.99%)		1 (3.23%)	28 (9.62%)	
Admission type, n (%)			< 0.001			< 0.001
Elective	49 (16.5%)	1583 (59.8%)		12 (36.4%)	176 (59.7%)	
Urgent	248 (83.5%)	1064 (40.2%)		21 (63.6%)	119 (40.3%)	
ICU Stay, Time (h)	8.55 ± 5.90	8.10 ± 5.27	0.029	7.48 ± 5.11	8.60 ± 5.48	0.149
Death During Hospitalization, n (%)	12 (4.04%)	29 (1.10%)	< 0.001	2 (6.06%)	3 (1.02%)	1.000
Vital signs, M [P15, P45]						
Heart Rate (bpm)	85.5 [72.25,99.0]	91.0 [78.0, 105.0]	0.132	89 [81,94]	93.5 [75.0, 106.75]	0.102
Respiratory Rate (bpm)	19 [16.25, 24.0]	18 [14, 21]	0.079	16 [14.25, 17]	17 [14, 20.25]	0.617
Systolic Blood Pressure (mmHg)	106.5 [97.0, 122.5]	110.0 [94.0, 128.0]	0.912	112 [90.5, 126.0]	128.5 [107.25, 150.0]	0.741
Diastolic Blood Pressure (mmHg)	57.0 [50.0, 72.0]	57.0 [49.0, 71.5]	0.924	52 [46, 55.75]	62.0 [53.25, 78.75]	0.741
SpO_2 (%)	98 [97,100]	98 [95, 100]	0.448	97 [94.25, 99.75]	100 [96.5, 100]	0.269
Temperature (F)	98.1 [97.75,98.88]	98.2 [97.6,98.7]	0.943	98.0 [97.8, 98.58]	98.3 [97.35,99.1]	0.406
Laboratory Parameters, M [P15, P45]						
Lactate (mmol/L)	1.5 [1.2, 1.9]	1.5 [1.1, 2.1]	0.775	1.7 [1.35,2.13]	1.65 [1.1, 2.28]	0.368
PH	7.38[7.33,7.43]	7.38[7.34, 7.42]	0.714	7.40 [7.37,7.44]	7.39[7.35,7.43]	0.830
Cholesterol, HDL (mmol/L)	0.72[0.65,0.94]	0.87 [0.60, 1.14]	0.591	0.83 [0.80, 0.85]	0.83 [0.80, 0.85]	0.177
Cholesterol, LDL (mmol/L)	1.89 [1.68, 2.06]	1.50 [1.24, 1.94]	0.614	1.53 [1.40, 1.66]	1.53 [1.40, 1.66]	0.178
Cholesterol, Total (mmol/L)	3.19 [2.92, 3.41]	2.96 [2.38, 3.63]	0.611	2.92 [2.74, 3.10]	2.92 [2.74, 3.10]	1.000
Triglycerides(mg/dL)	1.08 [0.81, 1.41]	1.28 [0.98, 2.02]	0.033	1.59 [1.21, 2.02]	1.40 [1.16, 1.57]	0.077
Glucose (mg/dL)	120 [106.75,140.0]	121.0 [105.0,143.0]	0.624	113.0 [103.0, 134.0]	126.0 [106.0,146.0]	0.859
Potassium (mmol/L)	4.1 [3.8, 4.5]	4.1 [3.9, 4.5]	0.666	4.0 [3.6, 4.3]	4.15 [3.9, 4.57]	0.218
Sodium (mmol/L)	138.0 [136.0,141.0]	138.0 [135.0, 140.0]	< 0.001	138.0 [137.0, 141.0]	137.0 [136.0, 140.0]	0.098
Urea Nitrogen (mg/dL)	20.5 [14.0,26.0]	15.0 [12.0, 21.0]	< 0.001	18.0 [13.0, 24.0]	15.0 [12.0, 20.0]	< 0.001
Uric Acid (mg/dL)	3.9 [3.2, 7.6]	5.1 [3.98, 6.98]	0.050	4.32 [3.89, 5.43]	4.35 [3.7, 6.4]	0.177
Creatinine (mg/dL)	0.9 [0.7, 1.3]	0.8 [0.7,1.1]	< 0.001	0.9 [0.7, 1.1]	0.9 [0.7, 1.1]	0.146
Hemoglobin (g/dL)	10.95 [9.88,12.2]	10.6 [9.5, 11.9]	< 0.001	11.2 [10.5, 11.9]	10.5 [9.4, 11.63]	0.429
INR	1.2 [1.1, 1.3]	1.2 [1.1, 1.3]	0.029	1.1 [1.0, 1.3]	1.2 [1.1, 1.3]	0.963
Lymphocytes (%)	10.0[7.85,15.5]	12.7 [8.17, 18,5]	< 0.001	11.2 [5.90, 13.1]	12.8 [7.7, 19.8]	
Neutrophils (%)	80.2[73.95, 85.4]	77.8 [69.7, 83.6]	0.056	82.15[77.88,85.07]	76.3 [68.75, 84.0]	
Platelet Count (×10⁹/L)	196 [156, 250]	207.0 [166.0, 263.0]	0.011	177 [146, 221]	208 [164, 259]	0.405
White Blood Cells (×10⁹/L)	10.35 [7.7, 13.17]	9.8 [7.6, 12.6]	0.216	9.1 [8.2, 11.8]	10.0 [7.7, 12.6]	0.296
PT (s)	13.1[12.1,14.8]	13.2 [12.0,14.7]	0.170	12.3 [11.38,13.88]	12.9[11.9,14.02]	0.955
Comorbidity, n (%)						
Hypertension	97 (32.66%)	704 (26.61%)	0.031	11 (33.33%)	83 (28.14%)	0.385
Diabetes	66 (22.22%)	484 (18.29%)	0.117	6 (18.18%)	55 (18.64%)	0.023
Cardiovascular Disease	84 (28.28%)	428 (16.18%)	< 0.001	8 (24.24%)	50 (16.95%)	< 0.001
Chronic Bronchopulmonary Disease	43 (14.48%)	482 (18.22%)	0.129	7 (21.21%)	43 (14.58%)	0.735
Craniocerebral Diseases	238 (80.13%)	858 (32.43%)	< 0.001	28 (84.85%)	93 (31.53%)	< 0.001
Kidney Disease	106 (35.69%)	474 (17.91%)	< 0.001	10 (30.30%)	45 (15.25%)	0.065
Liver Disease	21(7.07%)	138 (5.22%)	0.228	1 (3.03%)	17 (5.76%)	0.049
Sleep Disorder	46 (15.49%)	401 (15.15%)	0.947	2 (6.06%)	51 (17.29%)	0.336
Continued						

Variables	Train set			Test set		
	PND group 264	Non-PND group 2353	P1	PND group 66	Non-PND group 589	P2
Score System, M [IQR]						
GCS	14.0[14.0,15.0]	15.0[15.0, 15.0]	< 0.001	14.5 [13.0, 15.0]	15.0[15.0, 15.0]	< 0.001
SAPSII	41.0[33.5,49.0]	35.0[27.5, 43.5]	0.006	37.5 [31.25, 45.25]	33.5[29.5,41.75]	0.063
SOFA_24hours	3.0 [1.5,4.0]	2.0 [1.0,3.0]	0.003	2.5 [2.0, 3.0]	2.0 [1.0,3.0]	0.011
Treatment, n (%)						
Mechanical Ventilation	7 (2.59%)	33 (1.44%)	0.038	1 (3.33%)	4 (1.54%)	0.887
Renal Replacement Therapy	3 (1.11%)	18 (0.78%)	0.591	0 (0%)	2 (0.77%)	1.000
Blood Transfusion	0 (0%)	0 (0%)	1.000	0 (0%)	1 (0.38%)	1.000
Medicines, n (%)						
Analgesic and sedative drugs						
Dexmedetomidine	6 (2.27%)	13 (0.55%)	0006	2 (3.03%)	3 (0.51%)	0.137
Etomidate	1 (0.38%)	4 (0.17%)	1.000	0 (0%)	0 (0%)	1.000
Midazolam	9 (3.41%)	36 (1.53%)	0.048	1 (1.52%)	12 (2.04%)	1.000
Morphine	81 (30.68%)	811 (34.47%)	0.245	22 (33.33%)	206 (34.97%)	0.897
Propofol	2 (0.76%)	4 (0.17%)	0.225	0 (0%)	2 (0.34%)	1.000
Fentanyl	20 (7.58%)	107 (4.55%)	0.043	2 (3.03%)	28 (4.75%)	0.745
Cardiovascular system drugs						
Captopril	1 (0.38%)	14 (0.59%)	0.991	1 (1.52%)	3 (0.51%)	0.872
Dobutamine	1 (0.38%)	2 (0.08%)	0.705	1 (1.52%)	1 (0.17%)	0.483
Dopamine	2 (0.76%)	1 (0.04%)	0.022	1 (1.52%)	0 (0%)	0.184
Epinephrine	2 (0.76%)	13 (0.55%)	1.000	0 (0%)	5 (0.85%)	0.995
Nifedipine	2 (0.76%)	7 (0.30%)	0.512	0 (0%)	6 (1.02%)	0.887
Norepinephrine	6 (2.27%)	49 (2.08%)	1.000	4 (6.06%)	8 (1.36%)	0.027
Nitroglycerin	0 (0%)	8 (0.34%)	0.718	0 (0%)	1 (0.17%)	1.000
Hydrocortisone	3 (1.14%)	42 (1.78%)	0.604	0 (0%)	17 (2.89%)	0.322

Table 1. The characteristics of PND and non-PND patients from train set and test set. P1 represents the comparison between the PND and non-PND groups in the training set. P2 represents the comparison between the PND and non-PND groups in the validation set.

The key features of models established by different algorithms

Figure 2 illustrates the essential features and their weight rankings in models established by each algorithm. Notably, models developed through different algorithms prioritize distinct key features. In the multivariate logistic regression model (Fig. 2a), the top 10 crucial variables include age, craniocerebral disease, serum creatinine, triglycerides, hemoglobin, international normalized ratio (INR), prothrombin time (PT), SOFA score, lymphocytes, and neutrophils. For the ANN predictive model (Fig. 2b), the algorithm assigns the highest weights to the following 10 variables: age, gender, serum creatinine, GCS score, and whether the patient has diabetes, hypertension, chronic disease, kidney disease, craniocerebral disease, or sleep disorder. Support vector machines (SVM), acclaimed for their classification prowess, stand out in delineating data into distinct categories by identifying the optimal hyperplane. In our SVM-based predictive model (Fig. 2c), the algorithm identifies the top 10 most crucial features, such as age, hemoglobin, SOFA score, sodium, potassium, PT, and indicators of craniocerebral disease, hypertension, liver disease, or kidney disease.

In the Naive Bayes model, the conditional probabilities of each feature under different categories are derived from the training data. By analyzing these logarithmic probabilities, the algorithm can assess the contribution of each feature to the classification task. Figure 2d illustrates the top 10 most crucial features for the occurrence of PND and non-PND cases in the Naive Bayes model. For PND cases, the most important features include hemoglobin, sodium, craniocerebral disease, age, GCS score, pH, neutrophils, total cholesterol, HDL, and LDL. Conversely, for non-PND cases, the top 10 features comprise potassium, age, sodium, pH, neutrophils, GCS score, total cholesterol, HDL, LDL, and the patient's admission type. In practice, we obtain logarithmic probabilities for each feature under each category. A higher logarithmic probability for a specific category indicates a higher likelihood of the feature under that given category. In the decision tree (Xgboost) model, each node represents a feature, with branches signifying different values of that feature, and leaf nodes indicating the model's prediction for input samples (Fig. 2e). Various splitting criteria, such as the Gini coefficient, are employed to determine the optimal feature for reducing data uncertainty through splitting. This process occurs at each node of the tree. Key nodes in the decision tree include age, the presence of craniocerebral disease, neutrophils (%), lactate, GCS score, admission type, the presence of chronic bronchopulmonary disease, hemoglobin, white blood cell count, and serum sodium. These nodes play pivotal roles in guiding the model's predictions by effectively partitioning the data based on their respective features.
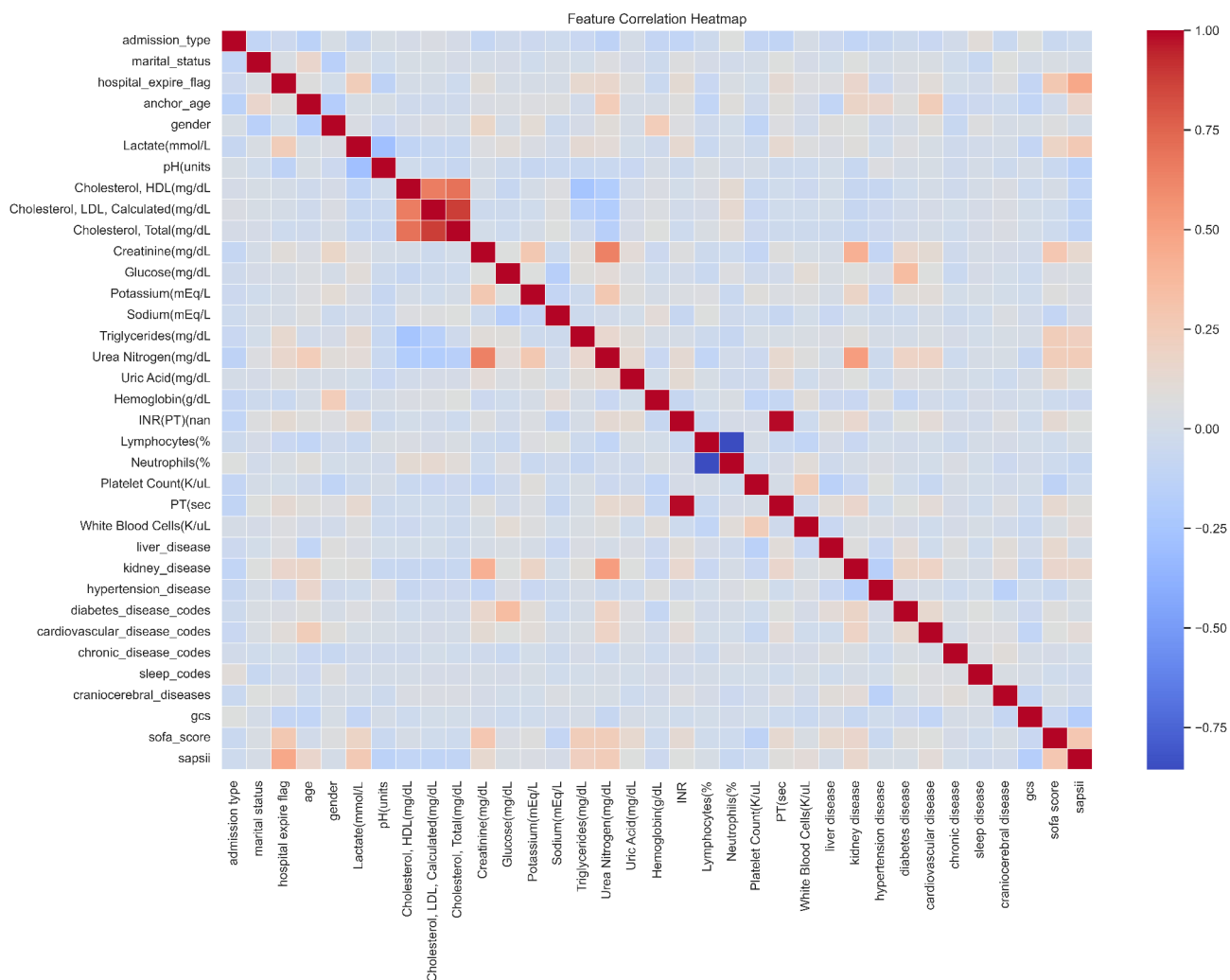
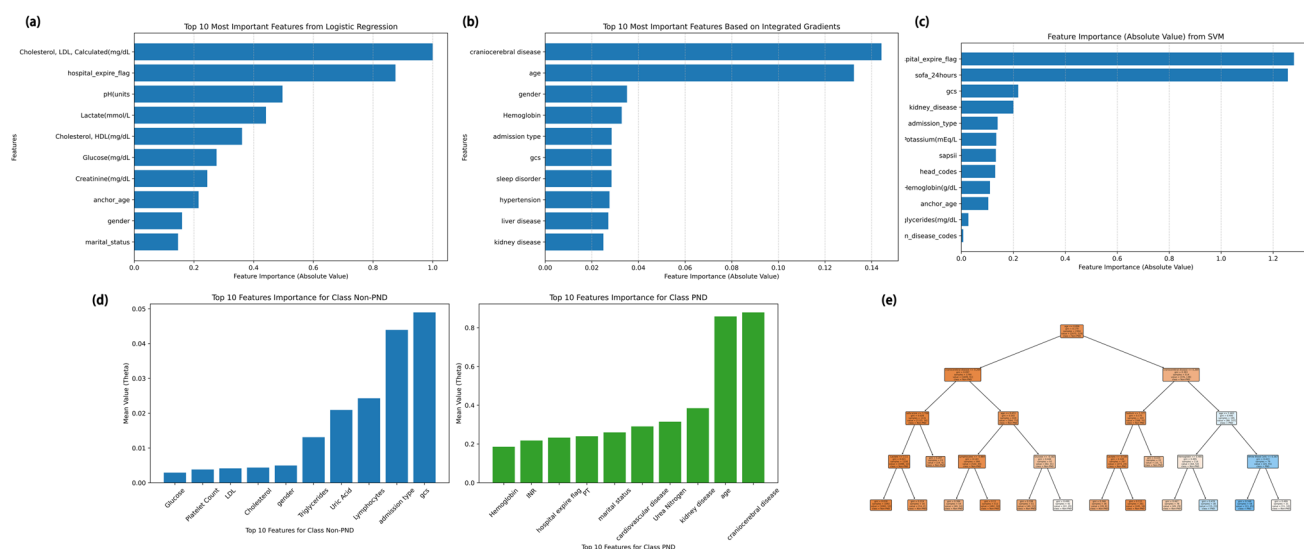**Fig. 1**. The relationships among the included variables.

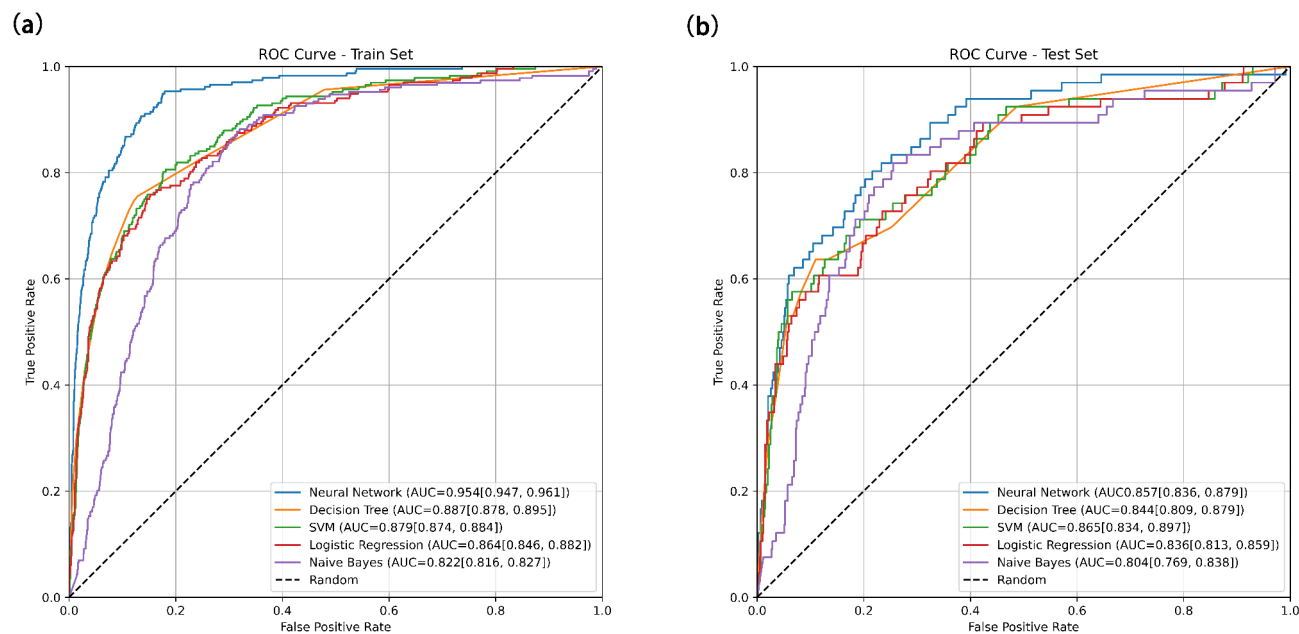**Fig. 2**. The key features of models established by different algorithms.

(a)

(b)

Fig. 3. ROC curves based on different models in the training and validation sets.

Algorithm	Dataset	Accuracy [95%CI]	Precision [95%CI]	Recall [95%CI]	F1-score [95%CI]	RocAUC [95%CI]	Brier Score	IPA (%)
MLR	Train	0.910 [0.904, 0.915]	0.715 [0.515, 0.915]	0.218 [0.000, 0.442]	0.290 [0.063, 0.518]	0.864 [0.846, 0.882]	0.069 [0.065, 0.074]	72.8 [69, 76]
	Test	0.902 [0.892, 0.912]	0.510 [0.223, 0.797]	0.179 [0.000, 0.413]	0.229 [0.000, 0.488]	0.836 [0.813, 0.859]	0.076 [0.066, 0.086]	67.2 [63, 72]
Decision tree	Train	0.920 [0.916, 0.924]	0.675 [0.622, 0.729]	0.397 [0.255, 0.539]	0.489 [0.396, 0.582]	0.887 [0.878, 0.895]	0.060 [0.058, 0.063]	77.4 [75,79]
	Test	0.903 [0.889, 0.917]	0.552 [0.431, 0.672]	0.3356 [0.171, 0.500]	0.400 [0.300, 0.499]	0.844 [0.809, 0.879]	0.070 [0.062, 0.078]	68.9 [62, 76]
ANN	Train	0.938 [0.932, 0.945]	0.758 [0.749, 0.767]	0.583 [0.492, 0.673]	0.657 [0.599, 0.714]	0.954 [0.947, 0.961]	0.048 [0.043, 0.052]	90.8 [89, 92]
	Test	0.903 [0.891, 0.914]	0.539 [0.440, 0.638]	0.372 [0.288, 0.455]	0.432 [0.381, 0.483]	0.857 [0.836, 0.879]	0.071 [0.063, 0.079]	71.4 [67, 76]
NB	Train	0.726 [0.505, 0.946]	0.258 [0.170, 0.346]	0.682 [0.432, 0.932]	0.356 [0.281, 0.432]	0.822 [0.816, 0.827]	0.224 [0.049, 0.400]	64.3 [63, 65]
	Test	0.726 [0.496, 0.956]	0.263 [0.167, 0.359]	0.674 [0.403, 0.945]	0.357 [0.272, 0.443]	0.804 [0.769, 0.838]	0.228 [0.041, 0.416]	60.7 [54, 68]
SVM	Train	0.916 [0.914, 0.918]	0.650 [0.639, 0.660]	0.335 [0.289, 0.381]	0.441 [0.401, 0.481]	0.879 [0.874, 0.884]	0.064 [0.063, 0.065]	75.7 [75, 76]
	Test	0.911 [0.905, 0.917]	0.612 [0.556, 0.668]	0.320 [0.238, 0.403]	0.417 [0.345, 0.489]	0.865 [0.834, 0.897]	0.067 [0.063, 0.071]	73.0 [66, 79]

Table 2. Models performance by different algorithms.

Model performance and comparation

As illustrated in Fig. 3 and summarized in Table 2, the artificial neural network (ANN)-based model exhibits superior performance compared to others on the training dataset, demonstrating a larger area under the Receiver Operating Characteristic (ROC) curve (AUC: 0.954; 95% CI 0.947–0.961). With an accuracy of 93.8%, precision of 75.8%, and an F1-score of 65.7%, the ANN model showcases notable predictive capabilities, surpassing models founded on multiple logistic regression (MLR), decision tree, support vector machine (SVM), and Naive Bayes (NB). However, it is noteworthy that the SVM-based model also presents commendable performance, notably excelling in accuracy and precision metrics (AUC: 0.879; 95% CI: 0.874–0.884; accuracy: 91.6%, precision: 65.0%, F1-score: 44.1%). These findings indicate that both artificial neural network (ANN) and support vector machine (SVM) models exhibit robust potential for predictive accuracy, with the ANN model demonstrating slightly superior overall performance. Moreover, upon assessing model performance on the validation dataset, the SVM-derived model emerges as noteworthy (AUC: 0.865; 95% CI 0.834–0.897), surpassing alternative algorithms with an accuracy of 91.1%, precision of 61.2%, and an F1-score of 41.7%.
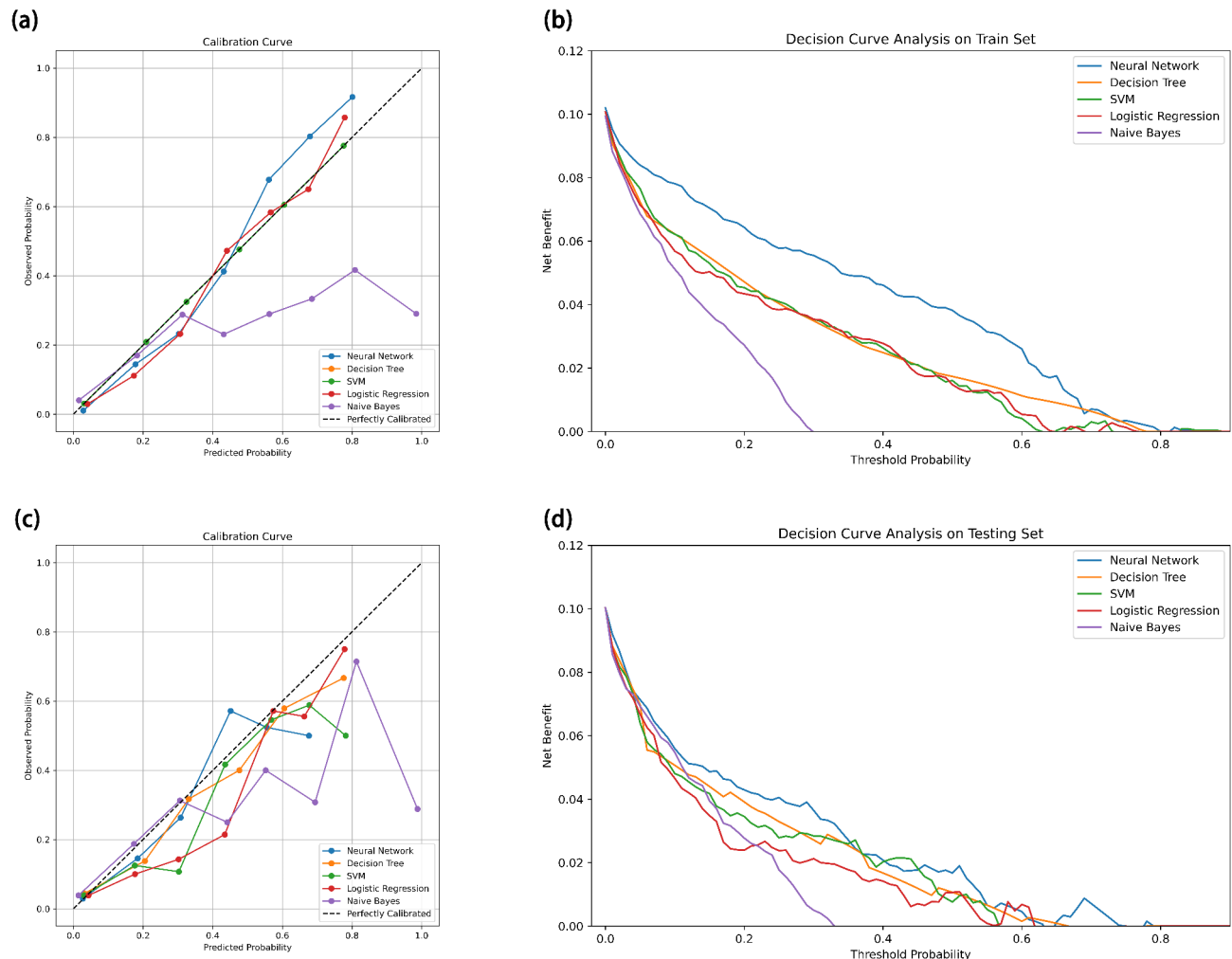
Fig. 4. Calibration plots and Decision Curve Analysis based on different models in the training and validation sets.

In terms of calibration, we generated calibration plots and computed Brier Scores along with the index of prediction accuracy (IPA). Table 2 illustrates that the Brier Scores for all five predictive models consistently remain below 0.1, indicating outstanding accuracy. Notably, on the training dataset, models based on the decision tree, ANN and SVM algorithms show lower Brier Scores and higher IPA values compared to others (decision tree: 0.060; SVM: 0.064; ANN: 0.048; IPA, decision tree: 77.4%; SVM: 75.7%; ANN: 90.8%). Similarly, on the validation dataset, models derived from the decision tree and SVM algorithms demonstrate relatively superior accuracy (decision tree: 0.070; SVM: 0.067; ANN: 0.071; IPA, decision tree: 68.9%; SVM: 73.0%; ANN: 71.4%). These results underscore the remarkable accuracy and calibration performance of the ANN and SVM models across both training and validation datasets (Fig. 4a, c).

Based on the training dataset, we constructed Decision Curve Analysis (DCA) plots for the five models and performed a comparative assessment. The DCA plots reveal that within the predictive risk threshold range of 0% to 80%, models based on artificial neural networks (ANN) outperform others. This indicates that, in this risk range, patients would benefit more from the use of ANN-based predictive models than from alternative approaches.

The decreasing trend observed in the DCA lines suggests that, as the threshold probability increases beyond a certain point, the net benefit of using the predictive models diminishes. This implies that higher threshold probabilities may lead to more false positive predictions, thereby reducing the model's clinical utility. The points where the DCA curves cross the x- and y-axes are critical: when a model's curve crosses the y-axis, it shows the net benefit of treating all patients (i.e., the model has no predictive value). When the curve crosses the x-axis, it suggests the threshold probability beyond which the model's predictions are no longer useful, as the net benefit becomes negative. Similar findings are observed in the validation dataset, where for predictive risk thresholds between 0 and 80%, ANN-based models again show superior performance compared to other models (Fig. 4b, d).

Constructure online tool

In summary, considering the model's performance, DCA, and predictive accuracy, the ANN-based predictive model stands out as superior, showing greater potential for clinical use. Using this ANN model, we've created an online tool (https://xyyy.pythonanywhere.com/) to identify high-risk patients for PND during hip arthroplasty.

Discussion

In this investigation, we developed a rapid and convenient online tool based on a dataset of 3272 patients undergoing hip arthroplasty. The primary objective of this tool is to effectively identify high-risk patients susceptible to perioperative cognitive disorders (PND), thereby aiding in their perioperative management. Our findings underscore the significance of various clinical factors, including craniocerebral diseases, age, gender, hemoglobin, liver disease, kidney disease, admission type, GCS, sleep disorder and hypertension as pivotal features for predicting the onset of PND in patients undergoing hip joint replacement surgery.

Age and gender are widely acknowledged as a crucial risk factor for the occurrence of PND in patients during the perioperative period[16–18]. Our study's advanced model confirms that both factors significantly influence PND incidence, with older patients facing a higher risk compared to younger ones, and female patients exhibiting a greater likelihood of developing PND than their male counterparts. Additionally, our analysis identifies craniocerebral diseases as the primary predictors of PND, supporting clinical findings that vascular mechanisms are integral to PND-associated cognitive issues[19]. The Glasgow Coma Scale (GCS), essential for assessing cognitive functions and mental states, is also highlighted as a pivotal predictor of PND according to our data analysis[20]. Furthermore, our findings indicate that renal and hepatic diseases, as well as sleep disorders, substantially impact PND risk through mechanisms like disrupted drug metabolism and toxin accumulation. Meta-analyses of existing cohort studies establish a strong correlation between chronic kidney disease and functional brain disorders, with individuals having elevated urine albumin-to-creatinine ratios facing increased risks of Alzheimer's disease and vascular dementia[21]. Additionally, previous research underscores a significant rise in sleep disorders among individuals experiencing cognitive decline, reinforcing the profound link between sleep disturbances and cognitive deterioration[22].

Observational studies have shown that individuals with hypertension are more susceptible to cognitive dysfunctions[23]. Extensive research, including retrospective analyses and meta-analyses, has identified hemoglobin levels as a predictive factor for postoperative delirium in orthopedic patients, with lower hemoglobin levels associated with an increased likelihood of this condition[24–26]. Consistent with prior studies, our results affirm that hypertension and low hemoglobin are key predictors of PND.

In the past, several studies have proposed similar models. However, these studies were based on a limited amount of data for model development. These studies had small sample sizes and incorporated numerous variables, which could potentially introduce statistical biases[27–29]. Our study is built upon an extensive dataset comprising 3272 cases, from which we meticulously curated a comprehensive set of variables, totaling over 30. These variables were carefully selected to facilitate both screening and deep learning processes, aiming to develop a robust model targeting perioperative cognitive disorders in patients undergoing hip replacement or revision surgery. Furthermore, we employed a diverse array of five different methods for variable selection, including filter, wrapper, and embedded techniques, in contrast to the prevalent reliance on LASSO regression seen in most studies[27,30]. Lastly, our study employed a variety of algorithms for model prediction and construction. Through thorough comparisons, we identified and developed the optimal model (ANN-based predictive model). Notably, the resulting model is not only highly effective but also straightforward, convenient, and user-friendly.

Nonetheless, our research does have some limitations. Firstly, being a retrospective study, it is contingent upon the completeness of the collected data. Initially, we selected over 60 variables; however, due to a high rate of missing data for certain variables, deletions were necessary. Furthermore, additional validation using external datasets is warranted for this study.

Conclusion

In summary, we have developed a simple and convenient big data-based visualization online tool for perioperative screening of high-risk patients for postoperative neurocognitive disorders (PND) in hip replacement or revision surgeries. This tool aims to assist healthcare professionals in making decisions regarding perioperative management strategies.

Data availability

All data included in this study are available upon request by contact with the corresponding author.

References

1. Evered, L. et al. Recommendations for the nomenclature of cognitive change associated with anaesthesia and surgery-2018. *Br. J. Anaesth.* **121**, 1005–1012. https://doi.org/10.1016/j.bja.2017.11.087 (2018).
2. Goldstein, D. P. et al. Cognitive impairment and delirium in older patients undergoing major head and neck surgery. *Otolaryngol. Head Neck Surg.* **167**, 97–99. https://doi.org/10.1177/01945998211045293 (2022).
3. Cao, E., Ma, D., Nayak, S. & Duong, T. Q. Deep learning combining FDG-PET and neurocognitive data accurately predicts MCI conversion to Alzheimer's dementia 3-year post MCI diagnosis. *Neurobiol Dis* **187**, 106310. https://doi.org/10.1016/j.nbd.2023.106310 (2023).
4. Rowley, P. et al. What do people expect of general anaesthesia?. *Br. J. Anaesth.* **118**, 486–488. https://doi.org/10.1093/bja/aex040 (2017).

5. Liu, F. et al. Subjective cognitive decline may mediate the occurrence of postoperative delirium by P-tau undergoing total hip replacement: The PNDABLE study. *Front Aging Neurosci.* **14**, 978297. https://doi.org/10.3389/fnagi.2022.978297 (2022).

6. Labaste, F. et al. Postoperative delirium is a risk factor of institutionalization after hip fracture: an observational cohort study. *Front Med. (Lausanne)* **10**, 1165734. https://doi.org/10.3389/fmed.2023.1165734 (2023).

7. Wang, Y. et al. Identification of risk factors for postoperative delirium in elderly patients with hip fractures by a risk stratification index model: A retrospective study. *Brain Behav.* **11**, e32420. https://doi.org/10.1002/brb3.2420 (2021).

8. Fan, H. et al. Development and validation of a dynamic delirium prediction rule in patients admitted to the Intensive Care Units (DYNAMIC-ICU): A prospective cohort study. *Int. J. Nurs. Stud.* **93**, 64–73. https://doi.org/10.1016/j.ijnurstu.2018.10.008 (2019).

9. Wang, F. et al. Potential value of serum uric acid in the identification of postoperative delirium in geriatric patients undergoing knee replacement. *Front Aging Neurosci.* **14**, 909738. https://doi.org/10.3389/fnagi.2022.909738 (2022).

10. Lin, Y. et al. Potential value of serum lipid in the identication of postoperative delirium undergoing knee/hip arthroplasty: The perioperative neurocognitive disorder and biomarker lifestyle study. *Front Psychiatry* **13**, 870317. https://doi.org/10.3389/fpsyt.2022.870317 (2022).

11. Wang, R. et al. Perioperative platelet count in peripheral blood is associated with the early stage of PND after major orthopedic surgery: A prospective observational study. *BMC Geriatr* **22**, 200. https://doi.org/10.1186/s12877-022-02899-7 (2022).

12. Wu, J., Yin, Y., Jin, M. & Li, B. The risk factors for postoperative delirium in adult patients after hip fracture surgery: A systematic review and meta-analysis. *Int. J. Geriatr. Psychiatry* **36**, 3–14. https://doi.org/10.1002/gps.5408 (2021).

13. Dilmen, O. K., Meco, B. C., Evered, L. A. & Radtke, F. M. Postoperative neurocognitive disorders: A clinical guide. *J. Clin. Anesth.* **92**, 111320. https://doi.org/10.1016/j.jclinane.2023.111320 (2024).

14. Sachdev, P. S. et al. Classifying neurocognitive disorders: the DSM-5 approach. *Nat. Rev. Neurol.* **10**, 634–642. https://doi.org/10.1038/nrneurol.2014.181 (2014).

15. Collins, G. S., Reitsma, J. B., Altman, D. G. & Moons, K. G. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): The TRIPOD statement. *BMJ* **350**, g7594. https://doi.org/10.1136/bmj.g7594 (2015).

16. Chaiwat, O. et al. Postoperative delirium in critically ill surgical patients: incidence, risk factors, and predictive scores. *BMC Anesthesiol.* **19**, 39. https://doi.org/10.1186/s12871-019-0694-x (2019).

17. Kim, E. M., Li, G. & Kim, M. Development of a risk score to predict postoperative delirium in patients with hip fracture. *Anesth. Analg.* **130**, 79–86. https://doi.org/10.1213/ane.0000000000004386 (2020).

18. Haight, T. N. & Marsh, E. B. Identifying delirium early after stroke: A new prediction tool for the intensive care unit. *J. Stroke Cerebrovasc. Dis.* **29**, 105219. https://doi.org/10.1016/j.jstrokecerebrovasdis.2020.105219 (2020).

19. Mendyk-Bordet, A. M. et al. Endothelial dysfunction and pre-existing cognitive disorders in stroke patients. *Biomolecules* **14**, 721. https://doi.org/10.3390/biom14060721 (2024).

20. Lelis, R. G. et al. Apolipoprotein E4 genotype increases the risk of postoperative cognitive dysfunction in patients undergoing coronary artery bypass graft surgery. *J. Cardiovasc. Surg. (Torino)* **47**, 451–456 (2006).

21. Tang, X. et al. Association of kidney function and brain health: A systematic review and meta-analysis of cohort studies. *Ageing Res. Rev.* **82**, 101762. https://doi.org/10.1016/j.arr.2022.101762 (2022).

22. Pistacchi, M., Gioulis, M., Contin, F., Sanson, F. & Marsala, S. Z. Sleep disturbance and cognitive disorder: Epidemiological analysis in a cohort of 263 patients. *Neurol. Sci.* **35**, 1955–1962. https://doi.org/10.1007/s10072-014-1870-x (2014).

23. Butters, A. F. et al. Brain health clinics—An evolving clinical pathway?. *J. Prev. Alzheimers Dis.* **12**, 100051. https://doi.org/10.1016/j.tjpad.2024.100051 (2025).

24. Kim, N. et al. Analysis of the incidence and risk factors of postoperative delirium in patients with degenerative cervical myelopathy. *Neurospine* **19**, 323–333. https://doi.org/10.14245/ns.2142778.389 (2022).

25. Mariz, J. et al. Risk and clinical-outcome indicators of delirium in an emergency department intermediate care unit (EDIMCU): an observational prospective study. *BMC Emerg. Med.* **13**, 2. https://doi.org/10.1186/1471-227x-13-2 (2013).

26. Zhang, H. J., Ma, X. H., Ye, J. B., Liu, C. Z. & Zhou, Z. Y. Systematic review and meta-analysis of risk factor for postoperative delirium following spinal surgery. *J. Orthop. Surg. Res.* **15**, 509. https://doi.org/10.1186/s13018-020-02035-4 (2020).

27. Chen, D. et al. Predicting postoperative delirium after hip arthroplasty for elderly patients using machine learning. *Aging Clin. Exp. Res.* **35**, 1241–1251. https://doi.org/10.1007/s40520-023-02399-7 (2023).

28. Choi, J. Y. et al. Development and validation of a prognostic classification model predicting postoperative adverse outcomes in older surgical patients using a machine learning algorithm: Retrospective observational network study. *J. Med. Internet Res.* **25**, e42259. https://doi.org/10.2196/42259 (2023).

29. Song, A. L. et al. Dynamic nomogram for predicting the risk of perioperative neurocognitive disorders in adults. *Anesth. Analg.* **137**, 1257–1269. https://doi.org/10.1213/ane.0000000000006746 (2023).

30. Li, Y. et al. Development and validation of prediction models for neurocognitive disorders in adult patients admitted to the ICU with sleep disturbance. *CNS Neurosci. Ther.* **28**, 554–565. https://doi.org/10.1111/cns.13772 (2022).

Acknowledgements

Author contributions

Xie Yi contributed to data extraction and analysis, while Guo Jiao contributed to paper writing and data analysis. Wang Gang oversaw the paper revision, and Bai Xiarui contributed to the research design. Zhang Yuming provided financial support in this study. Each author contributed important intellectual content during manuscript drafting or revision and accepts personal accountability for the overall work, and agrees to ensure that questions pertaining to the accuracy or integrity of any portion of the work are appropriately investigated and resolved.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-025-93324-7.

Correspondence and requests for materials should be addressed to J.G.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.