EJA

Eur J Anaesthesiol 2026; **43:**1–10

OPEN

ORIGINAL ARTICLE

Post-operative delirium risk estimation and assessment with machine learning: development and validation of the Open-DREAM Model

A peri-operative dataset analysis

Armin Berger*, Mia Gisselbaek, Sarah Saxena, Joana Berger-Estilita* and Arnout Devos*

BACKGROUND Postoperative delirium (POD) is a common and serious complication in older surgical patients, associated with increased morbidity, prolonged hospitalisation and increased healthcare costs. Existing predictive models have limited clinical adoption due to implementation costs, suboptimal accuracy and poor generalisability.

OBJECTIVES To develop an open-access, interpretable machine learning (ML) model using only preoperative data to predict the risk of POD in patients aged at least 60 years undergoing surgery.

DESIGN Observational diagnostic study using prospectively collected routine care data.

SETTING Single secondary care hospital in Switzerland, January 2023 to January 2024.

PATIENTS A total of 1425 patients were screened; 748 met the inclusion criteria (≥60 years, noncardiac, nonintracranial surgery, no preoperative delirium).

INTERVENTIONS None.

MAIN OUTCOME MEASURE POD, defined as a Nursing Delirium Screening Scale (Nu-DESC) score at least 2 at any peri-operative assessment.

RESULTS Three ML algorithms were trained and compared: Support Vector Machines; Logistic Regression and XGBoost. The calibrated XGBoost model achieved an area under the receiver operating characteristic curve (AUC) of 0.80, with a sensitivity of 0.89 at the optimal probability threshold of 0.35. Feature selection using SHapley Additive exPlanations (SHAP)-derived rankings reduced the predictor set from 55 to 15 features without loss of AUC. Type of anaesthesia (spinal vs. general) was the strongest predictor.

CONCLUSIONS An open-access ML-based model using routinely collected preoperative variables can predict POD with high sensitivity and preserved interpretability. External validation is warranted. Future research should explore causal inference for modifiable risk factors, acknowledging the limitations of observational data. This open-access application is currently intended for research and educational use only until external validation confirms its performance in independent patient groups.

Published online xx month 2026

*AB, JBE and AD contributed equally.
From the ETH AI Center, Swiss Federal Institute of Technology Zurich (ETH Zurich), Zürich (AB, AD), Division of Anesthesiology, Department of Anesthesiology, Clinical Pharmacology, Intensive Care and Emergency Medicine, Geneva University Hospitals and Faculty of Medicine (MG), Unit of Development and Research in Medical Education (UDREM), Faculty of Medicine, University of Geneva, Geneva, Switzerland (MG), Department of Anesthesiology, Helora (SS), Department of Surgery, UMons, Research Institute for Health Sciences and Technology, University of Mons, Mons, Belgium (SS), Institute for Medical Education, University of Bern, Bern, Switzerland (JBE), RISE-Health, Centre for Health Technology and Services Research, Faculty of Medicine, University of Porto, Alameda Prof. Hernâni Monteiro, Porto, Portugal (JBE) and Institute for Anaesthesiology and Intensive Care, Salem Spital, Hirslanden Medical Group, Bern, Switzerland (JBE)

Correspondence to Joana Berger-Estilita, Institute for Medical Education, University of Bern, Mittelstrasse 43, 3012 Bern, Switzerland.
Tel: +41 31 684 62 01; e-mail: joanamberger@gmail.com

Received 13 August 2025 **Accepted** 18 January 2026

DOI:10.1097/EJA.0000000000002410

KEY POINTS

- Postoperative delirium (POD) is a common and serious complication in elderly surgical patients, with significant negative effect on outcomes and healthcare resources.
- We developed and validated an open-access machine learning model (XGBoost) using only preoperative data to predict POD risk with high accuracy (AUC = 0.80).
- The model performs comparably to existing tools (e.g. PIPRA) while using fewer and more accessible features, enhancing scalability.
- Type of anaesthesia (spinal vs. general) emerged as the most influential predictor, offering actionable insights for risk stratification.
- The model is freely available at https://delirium.streamlit.app, supporting real-world clinical implementation and equitable access to predictive analytics.

Introduction

Postoperative delirium (POD) remains one of the most common, underdiagnosed, and severe complications, particularly in older adults. It is an acute neurocognitive disorder characterised by a fluctuating disturbance in attention, awareness, and cognition that meets DSM-5 diagnostic criteria[1] and occurs during the postoperative period, up to 1 week after surgery or until hospital discharge, whichever occurs first.[2] Anaesthesia and surgery represent identifiable precipitating events. A lucid interval following emergence from anaesthesia is not required for the diagnosis, although its presence should be documented when observed. The reported incidence of POD varies significantly, from 15 to 50% among older adults and 2 to 30% in postanaesthesia care units (PACU).[3–5] Delirium identified in the PACU that meets diagnostic criteria represents early POD and has been associated with worse clinical outcomes, rather than a benign emergence phenomenon.[2,3,6–8]

The implications of POD extend far beyond the perioperative period. This condition is associated with prolonged hospital stays, long-term cognitive decline, increased mortality and a significant strain on healthcare systems.[4,9-11]

Preoperative prediction and mitigation of POD remain challenging due to its multifactorial nature and the variability in patient responses to surgical and anaesthetic interventions.[12–14] Nonetheless, early preoperative identification of patients at high risk for POD is critical for implementing personalised preventive strategies and tailoring postoperative follow-up. Advancements in machine learning (ML) have introduced new opportunities to enhance predictive accuracy and clinical decision-making.[15–18] Unlike traditional statistical models, ML algorithms can capture complex, nonlinear relationships within high-dimensional data,[19] making them suitable for multifactorial conditions like POD.[20]

Current clinical tools, such as the Postoperative Delirium Risk Assessment (PIPRA) model, offer some degree of predictive accuracy.[21] However, the standard implementation of this model is limited by the scope, interpretability, and per-patient cost of the included features. Furthermore, these tools usually lack generalisability across diverse clinical settings.[22,23]

We hypothesised that a ML model trained on preoperative clinical data could accurately predict POD risk in elderly patients undergoing surgery. As such, this study aimed to develop and validate an open-access ML model for predicting POD risk using preoperative data from a Swiss hospital.

Methods

Ethics statement

The Cantonal Ethics Committee of Bern (BASEC number: Req-2022-01502) reviewed the study protocol and granted a waiver for full ethics approval on 8 December 2022. The waiver was granted because the study used routinely collected peri-operative data, obtained as part of standard clinical care, with no additional interventions or deviation from standard practice. All data were de-identified before analysis, and no patient-identifiable information was stored in the research database.

According to the Committee's decision, written informed consent was not required; however, all included patients (or their legal representatives) provided verbal consent for their data to be used for quality improvement and research purposes at the time of preoperative assessment, following hospital policy.

The study adhered to the ethical principles outlined in the *Declaration of Helsinki*[24] and complied with the relevant Data Protection Acts applicable to participating academic institutions. Furthermore, it followed the *TRIPOD + AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods*[25] guidelines to ensure transparency and reproducibility.

Study design and setting

We used prospectively collected routine care data from a single secondary care hospital in Switzerland between January 2023 and January 2024. The hospital specialises in musculoskeletal medicine, spinal and general surgery, orthopaedics, urology, gynaecology and obstetrics. All peri-operative data were recorded in a predefined research database by a trained peri-operative team operating both day and night.

Several patient-reported symptoms (post-operative nausea and vomiting, shivering, thirst and stress) were collected at three peri-operative time points. For the purpose of model development, only preoperative measurements were used as predictors to ensure the model reflects purely preoperative risk estimation. Intraoperative and postoperative values were excluded from model training.

Data collection and predictors

For the model, we interrogated the records of the dataset. Patients meeting the following exclusion criteria were removed to align with comparable studies:[21] age below 60 years, undergoing intracranial or cardiac surgery, and preoperative symptoms of delirium. The consistent application of the Nu-DESC assessment across all patients aged 60 or older at standardised times ensured uniformity and reduced potential bias across sociodemographic groups.

We initially considered 55 routinely collected preoperative predictors, including personal variables (age, sex), ASA physical status, surgical speciality, anaesthesia type, airway management techniques and patient-reported outcomes (pain, anxiety, stress, thirst, nausea/vomiting, shivering, quality of recovery and satisfaction scores on a 0 to 10 numeric rating scale). The full list of variables is provided in Supplementary Digital File 1: Table S1, http://links.lww.com/EJA/B283.

Outcome definition

The primary outcome was POD, screened using the Nursing Delirium Screening Scale (Nu-DESC) at three peri-operative time points: before surgery, immediately after emergence from anaesthesia and at PACU discharge. The Nu-DESC evaluates five categories (disorientation, inappropriate behaviour, inappropriate communication, illusions/hallucinations and psychomotor retardation), each scored from 0 to 2. A total score of at least 2 was considered indicative of delirium. Nu-DESC is a validated screening tool but not a diagnostic instrument; positive screens were not confirmed by a diagnostic assessment such as the Confusion Assessment Method (CAM) or operationalised DSM-5/ICD-10 criteria. Therefore, the model predicts Nu-DESC-positive POD screens, which may differ from clinically diagnosed POD. Although Nu-DESC is not a diagnostic instrument, it has demonstrated high sensitivity (85 to 100%) and specificity (86 to 87%) in validation studies against DSM-based diagnoses.[26,27]

Missing data handling

Patients with missing outcome data (Nu-DESC score) or higher than 30% missing predictor values were excluded prior to model development (Supplemental Digital File 4: Figure S1, http://links.lww.com/EJA/B286). For the retained dataset, residual lack of data (3.8% overall; <8% per variable) was addressed using k-nearest neighbour (KNN) = 5 imputation, ensuring that partially missing values did not lead to exclusion.

Model development and training

The prediction task was framed as a binary classification problem. Three algorithms were compared:

(1) Support Vector Machine (SVM)
(2) Logistic Regression (L2 regularisation)
(3) Extreme Gradient Boosting (XGBoost)

Continuous variables were standardised (mean = 0, SD = 1) for SVM and logistic regression. Class imbalance was addressed using class weighting for SVM and logistic regression, and the scale-pos-weight function for XGBoost. No oversampling was applied. All models used regularisation factor $\lambda = 0.1$, selected to balance complexity and generalisability.

Feature selection

Feature selection was performed in two stages. First, a backward elimination process determined the approximate range of optimal feature count. Second, global SHapley Additive exPlanations (SHAP) values from the full 55-feature XGBoost model were used to rank feature importance. Model performance across increasing feature counts was evaluated using a *forward sequential addition* procedure based on the SHAP-derived global feature ranking. Each model was retrained, as features were cumulatively added from the most to least important, and AUC was calculated at each iteration to identify the minimal number of features required to maintain maximal performance. This ranking was then uniformly applied across all models (XGBoost, SVM and Logistic Regression) to enable cross-model comparability and ensure a consistent interpretability framework. Independent forward selection for each model was not performed to avoid overfitting and maintain reproducibility.

Feature importance was determined using SHAP values for XGBoost, absolute coefficients for logistic regression and absolute linear kernel coefficients for SVM. Feature elimination was performed in the training set, and model performance (AUC) was tracked to determine the optimal reduced set of predictors. Reducing predictors from 55 to the top 15 SHAP-ranked features preserved AUC and improved interpretability. To assess the marginal contribution of each predictor to model performance, we conducted a sequential feature addition experiment. Features were first ranked by importance within the XGBoost model using mean absolute SHAP values calculated from the training set. Starting with the single most important feature, predictors were added one at a time in order of decreasing importance. At each step, a new model was trained using only the included features, and performance was evaluated on the held-out test set. The AUC was recorded for each feature set size.

This process was repeated for Logistic Regression and SVM, using the same SHAP-derived ranking from the XGBoost model to ensure comparability across algorithms. Model training at each step followed the same hyperparameter settings as in the final tuned models. No information from the test set was used in the feature ranking or selection process to avoid data leakage.

Data preprocessing

Data were split into 80% training and 20% testing sets. Within the training set, five-fold cross-validation was employed to optimise hyperparameters. All preprocessing, including imputation and scaling, was performed within the training data during cross-validation to avoid data leakage. Missing values were imputed using KNN = 5 within the training set during cross-validation to avoid information leakage. The imputation was then applied to the test set using parameters derived solely from the training data. Categorical variables were one-hot encoded using panda's build in function (get dummies). Overall missingness was 3.8%, with no variable exceeding 8%. A KNN = 5 imputation approach was applied. A sensitivity analysis using complete-case data showed comparable model performance ($\Delta AUC \leq 0.01$), confirming robustness (Supplemental Digital File 2, http://links.lww.com/EJA/B284).

Model performance evaluation

The prediction task was framed as a binary classification problem, with POD defined as a Nu-DESC score at least 2. Three ML models – Support Vector Machines (SVM), Logistic Regression, and XGBoost – were trained. Continuous variables were standardised (mean = 0, SD = 1) before SVM and logistic regression training. Given the class imbalance (21.5% POD-positive cases), we applied class weighting in logistic regression and SVM, and used the scale_pos_weight feature in XGBoost to counteract bias toward the majority class. SMOTE or other synthetic oversampling techniques were not applied, as class weighting is often a more robust and stable approach for clinical datasets. Although XGBoost can handle imbalance reasonably well without weights, we explored scale_pos_weight as a tunable hyperparameter during cross-validation, with candidate values derived from the ratio of negative-to-positive class frequencies (range: 3 to 5 in our dataset). The optimal value was selected based on cross-validated AUC performance in the training set.

A regularisation parameter ($\lambda = 0.1$) was applied to all models to prevent overfitting while maintaining predictive performance. The role of λ varied across models. In XGBoost, λ controlled L2 regularisation (Ridge penalty) applied to the leaf weights of the decision trees, discouraging overly complex trees. In SVM, λ (also known as C in SVM terminology) managed the trade-off between margin size and classification error, ensuring a balanced decision boundary. In Logistic Regression, λ was an L2 penalty term that shrunk model coefficients to reduce multicollinearity and improve generalisation. A moderate $\lambda = 0.1$ was selected to balance model complexity and generalisability. Hyperparameter tuning using cross-validation was performed to confirm its suitability.

Model discrimination was primarily assessed using the area under the receiver operating characteristic curve (AUC), as this metric is robust with class imbalance and directly reflects the model's ability to distinguish between patients who develop POD and those who do not. Sensitivity (recall) was prioritised over specificity in our evaluation, given the clinical aim of identifying as many at-risk patients as possible for preventive interventions. We also report precision (positive-predictive value or PPV) to account for the clinical relevance of the predicted positive cases, particularly in a screening context.

To provide a comprehensive view of model behaviour, we included the F1-score (harmonic mean of precision and recall) and Brier score to capture overall probability calibration and accuracy. For model calibration assessment, calibration curves and the Hosmer–Lemeshow test were used, as these provide complementary graphical and statistical evaluations of probability alignment.

Software and packages

All analyses were conducted in Python (version 3.10.9) using:

(1) pandas (v1.5.3) for data handling
(2) numpy (v1.24.2) for numerical computations
(3) scikit-learn (v1.2.2) for model development, cross-validation, calibration, and evaluation
(4) xgboost (v1.7.5) for gradient boosting model training
(5) shap (v0.41.0) for feature importance and explainability
(6) matplotlib (v3.7.1) and seaborn (v0.12.2) for visualisation
(7) streamlit (v1.22.0) for the online tool interface
(8) The analytical code and processing pipeline are available from the corresponding author upon reasonable request.

Model calibration

To ensure that probability outputs were well calibrated, we applied posthoc calibration to the final XGBoost model.

Feature importance was assessed using SHapley Additive exPlanations (SHAP)[28] values for XGBoost, which provide individualised and global interpretability. Waterfall plots illustrated the varying impact of identical feature values (e.g. Wellbeing Score, NRS = 9) across predictions, providing granular insights beyond logistic

regression coefficients. For linear models, feature coefficients were inspected. These complementary approaches helped cross-validate insights and improve transparency. Further details on model performance, calibration outcomes, and feature selection are reported in the 'Results' section to improve clarity and avoid redundancy.

Feature importance and selection

Feature importance was assessed via normalised metrics

(1) Logistic Regression: absolute coefficients
(2) XGBoost: frequency of feature usage in decision splits
(3) SVM: absolute linear kernel coefficients

Model performance was averaged across seven random seeds to minimise variability.

Results

A total of 1425 patient records were examined. After accounting for missing data, 748 remained for analysis, yielding 161 events (Nu-DESC-positive POD), resulting in an events-per-variable ratio exceeding the standard minimum thresholds for multivariable prediction modelling. This was sufficient for feature selection and internal validation within the constraints of the dataset.

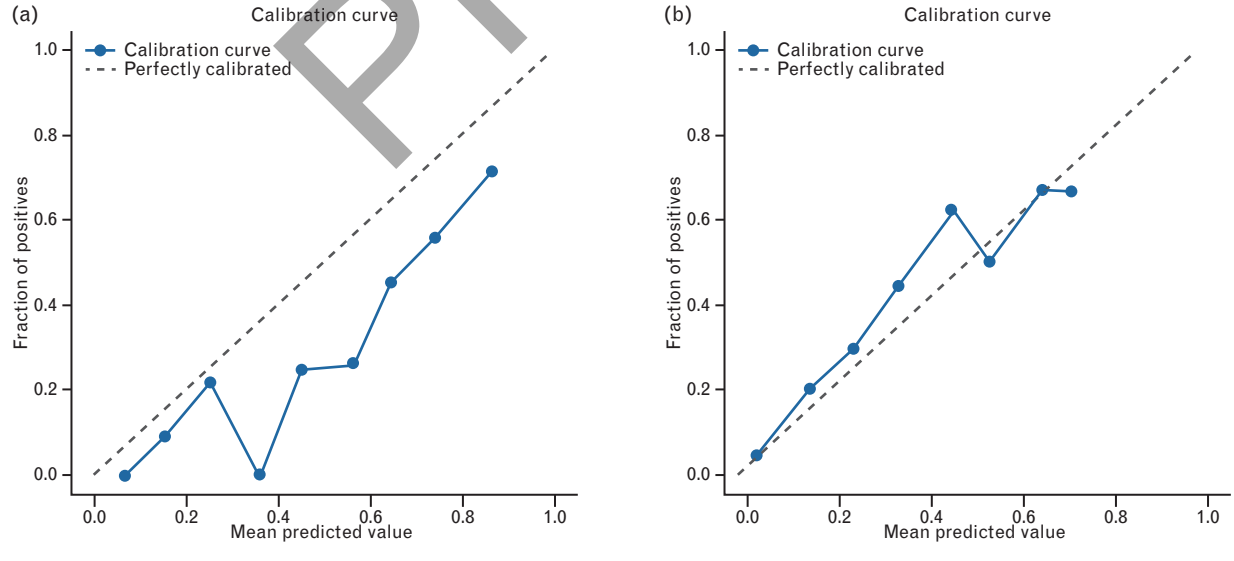
Feature selection

Figure 1 presents the calibration plots comparing the performance of the uncalibrated and calibrated models on the test set. The uncalibrated XGBoost model consistently overestimated risk across all risk groups. After calibration, the model demonstrated improved alignment with true (empirical) delirium probabilities, although it slightly underestimated risk, particularly for patients at higher risk (>0.5). In contrast, calibration was less effective for the SVM and Logistic Regression models. The SVM model systematically underestimated risk, limiting the effectiveness of calibration. On the other hand, the Logistic Regression model required minimal adjustment, as it was already relatively well calibrated from the outset. This is because Logistic Regression directly optimises log-likelihood, which naturally results in better probability estimates.[29] While calibration notably improved the probability interpretation of the XGBoost outputs, the differences in model behaviour highlighted the varying calibration needs across model architectures.

The importance scores of predictive features were aggregated and normalised across multiple models, allowing for a standardised comparison of their contributions to risk prediction. Based on their predictive power, features were categorised into high, medium and low importance.

In the high-importance category, two features stood out: spinal anaesthesia, which decreases predicted risk, and general anaesthesia, which increases predicted risk. The medium-importance category included features such as airway management with endotracheal intubation and EEG use, which were associated with an increased predicted risk. Finally, the low-importance group encompassed features such as target-controlled infusion and

Fig. 1 Calibration curves for the XGBoost model before and after calibration using Platt scaling with logistic regression applied to the validation set.



Calibration curve a (left) represents the uncalibrated model, while calibration curve b (right) shows the calibrated model. The *x*-axis represents the mean predicted probability, and the *y*-axis represents the observed fraction of positive cases (patients with postoperative delirium). The dashed diagonal line represents perfect calibration, where predicted probabilities match observed probabilities exactly. The calibrated model (b) aligns more closely with the diagonal, indicating improved probability estimates.

Table 1 Feature importance and predicted risk contribution across XGBoost, SVM and Logistic Regression models

Importance level	Feature name	Decreases predicted risk?
High	Spinal anaesthesia	Yes
	General anaesthesia	No
Medium	Airway management: endotracheal tube	No
	EEG usage	No
	Target controlled infusion (TCI)	Yes
	Gynaecology/obstetrics surgery	Yes
	Continuous temperature monitoring	No
	Airway management: Laryngeal mask (LMA)	No
	Eyeblinds/earplugs offered	Yes
Low	General surgery	Yes
	Patient uses dentures	No
	Vomiting score (NRS)	No
	Outpatient setting	Yes
	Stress score (NRS)	No
	Patient uses visual aids	No

Table presents the average importance of the top 15 features identified by XGBoost, SVM and Logistic Regression models. The impact on predicted risk (whether a feature increases or decreases risk) is inferred based on the coefficients from the Logistic Regression model, serving as a proxy for all three models. Features are categorised into high, medium, and low importance based on their contribution to model performance

gynaecology/obstetrics surgery, neither of which increased risk, also continuous temperature monitoring and airway management with supraglottic devices, both of which were linked to an increased predicted risk.

Notably, predictors in the *high* relevance cluster (e.g. type of anaesthesia) exhibit an order of magnitude that is more important than those in the *low* tier.

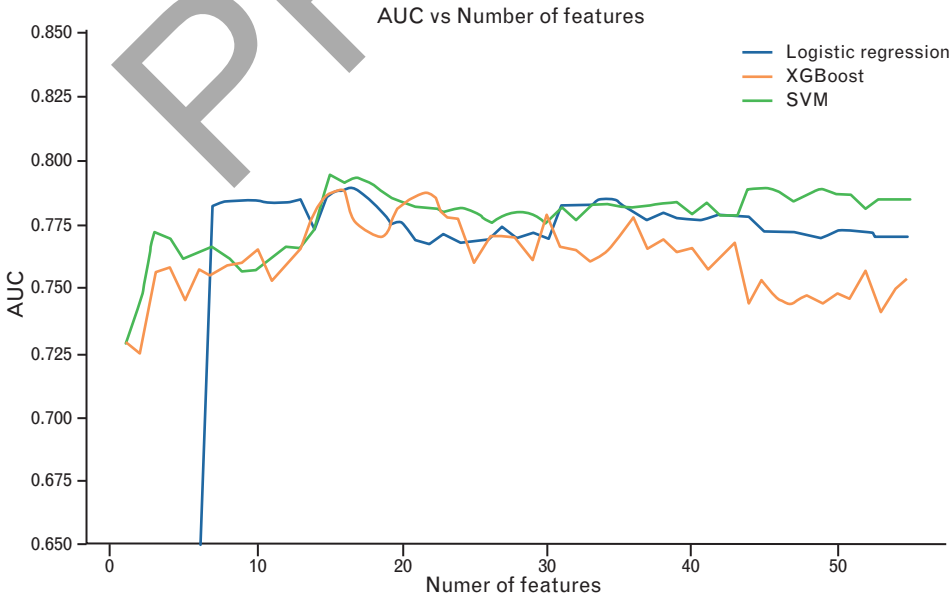
Table 1 presents the 15 most influential predictors of POD, ranked according to their contribution to model performance. The table also indicates whether each feature increases the predicted POD risk (positive coefficient in the logistic regression model).

As shown in Fig. 2, AUC increased with the number of included features up to approximately 15, after which performance plateaued. This finding guided the selection of the final 15-feature XGBoost model.

Model comparison and performance metrics

We performed pairwise DeLong's tests to assess the statistical significance of AUC differences between models. The results show no significant difference between XGBoost and Logistic Regression ($P = 0.42$), nor between XGBoost and SVM ($P = 0.37$). At the start, Logistic Regression experiences an initial dip in performance before stabilising as more features are added, ultimately maintaining a moderate and consistent AUC. In contrast, XGBoost showed a steady increase in AUC, reaching its peak early, but then exhibits fluctuations and a slight decline in performance, as the number of features increases. Meanwhile, SVM showed a relatively stable trend; an initial rise in AUC, followed by consistent performance across the feature range, with minor fluctuations.

Fig. 2 Sequential SHapley Additive exPlanations-based feature addition.



The plot illustrates AUC trends for XGBoost, SVM and Logistic Regression models as the number of top-ranked features (1 to 55) is progressively increased based on SHAP importance derived from the full XGBoost model. Lines represent mean area under the receiver operating characteristic curve (AUC) values across seven random seeds; shaded areas indicate ±1 standard deviation.

Overall, SVM emerged as the most stable and consistently high-performing model, particularly when the number of features increases. XGBoost performs well with a limited number of features but struggles to maintain consistency with additional ones. Logistic Regression delivers moderate performance but lacks the robustness seen in SVM. This suggests that while increasing the number of features may improve model performance up to a point, it can also introduce noise and reduce effectiveness, especially for specific algorithms like XGBoost. As feature importance was derived exclusively from the training set, no data leakage occurred during this stepwise reduction. Surprisingly, reducing the number of input features from 55 to 15 preserves model performance and slightly improves AUC. However, further reductions below 15 features results in a gradual decline in predictive accuracy. Despite this decline, the two most predictive features (type of anaesthesia: spinal or general) retain significant explanatory power, with the AUC stabilising around 0.73 for both Vector Machine (SVM) and XGBoost models when using only one or two top-ranked features. Full performance metrics (AUC, recall, precision, F1-score and Brier score) for all algorithms and feature-set sizes are reported in the Supplementary Digital File 3: Table S2, http://links.lww.com/EJA/B285. SHAP analysis confirmed that anaesthesia type, airway management and EEG monitoring were the dominant contributors to predicted POD risk in both the full and reduced models. The top 15 features preserved 98% of total SHAP explanatory variance from the 55-feature model, indicating minimal loss of information. Given its nearly identical discrimination to the full model ($\Delta AUC < 0.01$) and better calibration, the 15-feature XGBoost model was selected as the optimal balance of model performance and clinical feasibility.

Table 2 compares the predictive performance of our XGBoost-based model against the gold standard PIPRA method,[21] SVM, logistic regression and a baseline constant classifier. The ROC curve illustrates the trade-off between sensitivity and specificity, while the AUC quantifies the overall model discrimination. The F1 score, defined as the harmonic mean of precision and recall, was used to balance false positives and false negatives. Sensitivity analysis excluding imputed data yielded similar AUC and recall values across models, supporting the stability of results (Supplemental Digital File 2, http://links.lww.com/EJA/B284).

Given the clinical importance of minimising false negatives, we prioritised models with high sensitivity and strong F1 scores. The XGBoost model achieves an AUC of 0.80, demonstrating a favourable balance between recall (sensitivity) and precision (PPV). Importantly, our approach achieves equivalent AUC performance to existing models while requiring fewer input features, highlighting its efficiency and robustness in resource-constrained settings.

Discussion

Translating machine learning insights into clinical practice

To bridge the gap between ML advancements and clinical decision-making, we have developed a publicly accessible predictive tool for POD risk assessment, available at https://delirium.streamlit.app.

This platform, built on the XGBoost model, offers an intuitive interface for evaluating preoperative POD risk based on key patient data. Integrated SHAP value visualisations further enhance transparency by quantifying the contribution of each predictor to the final risk score.

By enabling clinicians to understand better and act upon individualised risk profiles, this tool supports targeted preventive strategies, ultimately improving peri-operative care outcomes. Openly available through 2025, it represents a step toward demoratising ML applications in healthcare and promoting evidence-based interventions in real-world settings.

This study highlights the transformative potential of open ML models in predicting POD risk using existing preoperative data. The calibrated XGBoost model achieved the highest overall AUC among the ML models tested and matched the AUC of the PIPRA reference model. However, at the predefined probability threshold of 0.35, it demonstrated higher sensitivity (0.89 vs. 0.77 for PIPRA) but lower precision (0.41 vs. 0.46), reflecting a trade-off between maximis detection of at-risk patients and reducing false positives.

Table 2 Performance metrics comparison across different models

Model	AUC[a]	Recall[b]	Precision[c]	F1	Brier	H-L[d]
XGBoost (our final model, $\lambda = 0.1$)	0.80	0.89	0.37	0.52	0.15	0.28
PIPRA method	0.80	0.68	0.77	–	–	
SVM ($\lambda = 0.1$)	0.77	0.91	0.36	0.51	0.16	0.22
Logistic Regression ($\lambda = 0.1$)	0.79	0.91	0.35	0.5	0.17	0.31
Constant Classifier (always delirium)	0.50	1.00	0.21			

The table compares the performance of our final XGBoost model with the existing PIPRA method,[21] SVM, Logistic Regression and a Constant Classifier that always predicts delirium. Key evaluation metrics include AUC (area under the curve), recall, and precision, which comprehensively assess each model's predictive capabilities. The regularisation parameter is specified in the model's name. [a] AUC (area under the curve): measures the overall ability of the model to distinguish between classes. [b] Recall (sensitivity): represents the proportion of actual positive cases correctly identified by the model. [c] Precision (positive-predictive value): indicates the proportion of predicted positive cases that are actually correct. [d] Hosmer–Lemeshow test.

Our findings establish a foundation for integrating predictive analytics into peri-operative care pathways, enabling more precise risk stratification. Our model is a freely accessible preoperative delirium prediction tool that achieves discrimination comparable to established approaches, such as the PIPRA model,[21] with an AUC of 0.80. Recent work by Jauk *et al.*[30] and Kocar *et al.*[31] has demonstrated models with higher discrimination; however, these often incorporate postoperative variables, advanced biomarkers or settings with different patient characteristics, which may limit generalisability and immediate integration into routine preoperative workflows. In our test set, using a probability cut-off of 0.35, the model achieved a sensitivity of 0.89, which reflects a prioritisation of identifying at-risk patients over maximising precision. Sensitivity can be adjusted by varying the classification threshold, and extremely high sensitivity (approaching 1) can be achieved at the expense of specificity, highlighting the need for context-specific threshold selection.

Compared with cohorts used in other delirium prediction studies, our cohort included a higher proportion of orthopaedic surgical patients and a lower representation of ASA class I patients, which may partly explain the differences in performance and feature importance rankings. Notably, prior comparative analyses have shown that linear models such as logistic regression can perform on par with nonlinear algorithms in delirium prediction tasks,[31–33] a finding we also observed in this work.

The study's XGBoost model achieved an AUC of 0.8, comparable to the existing PIPRA model,[21] while maintaining higher sensitivity (0.89). This performance is particularly noteworthy, given the model's reliance on a limited dataset of 748 patients, and highlights the efficacy of feature engineering and model calibration.

Feature importance analysis highlighted the predictive value of anaesthesia type (spinal vs. general) and monitoring techniques, enhancing model interpretability and guiding clinical decision-making to mitigate POD risk.

The impact of anaesthesia type on POD has been widely studied, with RCTs[34–36] identifying general anaesthesia as a major risk factor. In contrast, others report no significant difference between general anaesthesia and regional anaesthesia.[10,37,38] These discrepancies probably stem from differences in study design, patient selections, peri-operative practices and sedation strategies. However, our XGBoost model identified general anaesthesia as a high preoperative risk factor for POD compared with regional or spinal anaesthesia.

Notably, EEG monitoring did not show an association with reduced POD risk. While EEG is used to optimise anaesthetic depth, its lack of correlation with lower POD incidence suggests it may be used more frequently in high-risk patients.[39]

This study also highlights the scalability of ML models in reducing feature complexity without compromising predictive accuracy and reducing input features from 55 (from the PIPRA standard score[21]) to only 15 improved model performance, reflecting the importance of feature selection in mitigating the curse of dimensionality. This finding aligns with the literature indicating that simpler models often generalise better when trained on smaller datasets.[40] Integrating SHAP values enhances model interpretability, providing clinicians with a clear understanding of how predictions are made.[41] This commitment to interpretability and user-friendliness mitigates the risks associated with black-box AI systems in critical healthcare settings.[42]

Furthermore, making the XGBoost-based predictive tool publicly accessible exemplifies the importance of open access in promoting equitable healthcare innovation. By providing free access through an intuitive web platform, the tool empowers clinicians globally to leverage state-of-the-art predictive analytics, regardless of resource constraints. As such, AI and ML can address resource shortages in underserved regions and offer scalable solutions for early risk assessment, targeted prevention, and enhanced patient safety. Open-access, AI tools can democratise healthcare innovation, enabling global adoption regardless of economic constraints.

This study has several strengths. It is highly reproducible, with a robust methodology and clear documentation that enables validation across diverse settings. The application of ML (XGBoost) effectively captured complex relationships in POD risk, outperforming traditional statistical models. By using SHAP values, the study ensures feature interpretability, allowing clinicians to understand the contributions of individual predictors. The model's scalability is also highlighted by improved performance with fewer input features, making it suitable for broader implementation. Finally, its alignment with current clinical practices, mainly through comparison with the PIPRA model,[21] underscores its clinical relevance and potential for integration into existing workflows.

Although this model was developed using data from a single Swiss private hospital specialising in musculoskeletal and general surgery, the patients were heterogeneous in terms of age, comorbidities, and surgical types within these specialties. This environment provided standardised peri-operative workflows and high-quality, prospectively collected data, which are advantageous for initial model development. Furthermore, our feature set relied solely on routinely available preoperative variables, facilitating future external validation and adaptation in other clinical settings, enhancing the applicability and potential integration of the developed model into standard peri-operative workflows without incurring additional costs.

We characterise this tool as an 'open-access AI' model for three main reasons: its reliance solely on preoperative variables that are routinely available and do not require costly or invasive testing, its explainability through SHAP-based feature attribution, which allows clinicians to understand the drivers of each individual prediction, and its probability calibration, which helps ensure that predicted risks align more closely with observed event rates, supporting informed and transparent clinical decision-making.

The study has limitations. It was developed in a single, specialised Swiss centre, which restricts the generalisability of the findings. However, the use of widely collected preoperative features enhances the potential for external validation and adaptation in other surgical contexts. Delirium was assessed using the Nu-DESC at three standardised time points (preoperative, immediately after emergence and at PACU discharge). While this approach ensures data consistency and feasibility within peri-operative workflows, it probably underestimates the overall incidence of POD because cases that develop during the first postoperative days were not captured. The model, therefore, primarily reflects early-onset delirium risk and should be interpreted accordingly. We acknowledge that using the Nu-DESC as a screening rather than diagnostic tool may lead to false positives and underdiagnosis of hypo-active delirium subtypes. Nevertheless, its strong validation record, simplicity and widespread adoption in peri-operative research support its use in this large-scale screening context. Our outcome definition relied on the Nu-DESC screening tool without follow-up confirmation using a validated diagnostic method such as CAM or DSM-5 criteria. This pragmatic choice reflected real-world workflow constraints but may have led to misclassification bias. Future studies should integrate confirmatory diagnostic assessments (e.g. CAM or DSM-5) to strengthen diagnostic accuracy. Moreover, the limited dataset size may reduce the model's ability to capture rare clinically significant patterns, potentially affecting its robustness. In addition, the model relies solely on preoperative data, excluding potentially valuable intra-operative and postoperative variables, which might further enhance predictive accuracy. Although the present model was developed and internally validated using a single-centre dataset, external validation across multiple Swiss and European hospitals is ongoing to assess generalisability across populations and case-mix variations. We encourage independent replication and benchmarking using the openly available code and documentation to establish robustness and external transferability further. Despite imputation being restricted to the training data, some risk of data leakage remains. Future prospective studies with minimal missing data would help validate model robustness. On the other hand, while SHAP values improve model interpretability, it does not establish causality. For example, the observed association between EEG monitoring and increased POD risk requires further investigation to differentiate correlation from causation. Lastly, the study lacks external validation. An external validation study has been initiated at a partner institution. Patient enrolment and adherence to TRIPOD guidelines are underway, with results expected in 2026. Future research will focus on developing and validating a model based exclusively on preoperative variables and comparing its predictive performance against POD diagnosed through standardised multiday assessments such as 3D-CAM over at least 5 to 7 days. Such a design would further align predictive modelling with the current conceptual framework of POD and enhance clinical applicability.

Conclusion

The open-DREAM study allows for an open-access ML model to predict POD using preoperative data from a real-world peri-operative care program. The XGBoost model demonstrates strong predictive performance, matching established models while requiring fewer features and offering better interpretability. By identifying risk factors, the model supports early identification of high-risk patients, allowing targeted preventive strategies. However, the open-access application is currently intended for research and educational use only until external validation confirms its performance in independent patient groups. Making AI-driven predictive tools like this freely accessible ensures that physicians worldwide can implement data-driven approaches to optimise patient outcomes regardless of resources.

Acknowledgements relating to this article

Assistance with the article: Basile Morel.

Financial support and sponsorship: none.

Conflicts of interest: SS has received speaker's fees from Medtronic/Merck. JBE is a member of the Board of Directors of the European Society of Anaesthesiology and Intensive Care (ESAIC) and has received speaker's fees from Medtronic. The remaining authors declare no competing interests.

Presentation: preliminary data from this study were presented at the Euroanaesthesia Congress, Lisbon, Portugal, 25 to 27 May 2025, where it was awarded *Best Poster* in the Best Abstract Poster Competition.

Statement on the use of AI and AI-assisted technologies in scientific writing: while preparing this work, the authors used ChatGPT 4.0 and Grammarly (Grammarly, Inc., San Francisco, California, USA) to improve readability. After using these services, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

Regulatory disclaimer: the Open-DREAM predictive tool described in this study is intended for research and educational purposes only. It has not been reviewed, cleared, or approved by any regulatory authority (including but not limited to Swissmedic, the European Medicines Agency, or the U.S. Food and Drug Administration). The application should not be used as the sole basis for clinical decision-making, and clinicians must rely on their

own professional judgement and institutional protocols when assessing patient risk.

Patient and public involvement: there was no public or patient involvement in the design, conduct, or reporting of this study.

References

1 Diagnostic A. Statistical manual of mental disorders: DSM-5. Washington DC: American Psychiatric Association; 2013.
2 Evered L, Silbert B, Knopman DS, *et al.* Recommendations for the nomenclature of cognitive change associated with anaesthesia and surgery-2018. *Anesth Analg* 2018; **127**:1189–1195.
3 Bilotta F, Russo G, Verrengia M, *et al.* Systematic review of clinical evidence on postoperative delirium: literature search of original studies based on validated diagnostic scales. *J Anesth Analg Crit Care* 2021; **1**:18.
4 Gleason LJ, Schmitt EM, Kosar CM, *et al.* Effect of delirium and other major complications on outcomes after elective surgery in older adults. *JAMA Surg* 2015; **150**:1134–1140.
5 Igwe EO, Nealon J, O'Shaughnessy P, *et al.* Incidence of postoperative delirium in older adults undergoing surgical procedures: a systematic literature review and meta-analysis. *Worldviews Evid Based Nurs* 2023; **20**:220–237.
6 Neufeld KJ, Leoutsakos JM, Sieber FE, *et al.* Outcomes of early delirium diagnosis after general anesthesia in the elderly. *Anesth Analg* 2013; **117**: 471–478.
7 Berger-Estilita J, Marcolino I, Gisselbaek M, *et al.* Patient-reported outcomes as drivers of postoperative delirium in the postanaesthesia care unit: Data from a one-year prospective cohort study. *Eur J Anaesthesiol* 2025; **42**:1006–1016.
8 Berger-Estilita JCA, Marcolino I, Gisselbaek M, *et al.* The structured implementation of a perioperative brain healthcare bundle and its effectiveness on postoperative delirium incidence: a cohort study. *Eur J Anaesthesiol Intensive Care Med* 2025; **4**:e00001.
9 Bonhomme V, Putensen C, Böttiger BW, *et al.* Brain health: a concern for anaesthesiologists and intensivists. *Eur J Anaesthesiol Intensive Care* 2024; **3**:e0063.
10 Dolan MM, Hawkes WG, Zimmerman SI, *et al.* Delirium on hospital admission in aged hip fracture patients: prediction of mortality and 2-year functional outcomes. *J Gerontol A Biol Sci Med Sci* 2000; **55**: M527–M534.
11 Aranake-Chrisinger A, Avidan MS. Postoperative delirium portends descent to dementia. *Br J Anaesth* 2017; **119**:285–288.
12 Jellinger KA, Attems J. Neuropathological approaches to cerebral aging and neuroplasticity. *Dialogues Clin Neurosci* 2013; **15**:29–43.
13 Siddiqi N, Stockdale R, Britton AM, *et al.* Interventions for preventing delirium in hospitalised patients. *Cochrane Database Syst Rev* 2007; 2: CD005563.
14 Weiss Y, Zarour S, Neuman MD, *et al.* Shared decision-making for older adults in the peri-operative setting: a narrative review. *Eur J Anaesthesiol* 2025; **42**:767–773.
15 Adlung L, Cohen Y, Mor U, *et al.* Machine learning in clinical decision making. *Med* 2021; **2**:642–665.
16 Deo RC. Machine learning in medicine. *Circulation* 2015; **132**:1920–1930.
17 Uddin S, Khan A, Hossain ME, *et al.* Comparing different supervised machine learning algorithms for disease prediction. *BMC Med Inform Decis Making* 2019; **19**:281.
18 Saxena S, Barreto Chang OL, Suppan M, *et al.* A comparison of large language model-generated and published perioperative neurocognitive disorder recommendations: a cross-sectional web-based analysis. *Br J Anaesth* 2026; **136**:1275–1286.
19 Zhou W, Yan Z, Zhang L. A comparative study of 11 nonlinear regression models highlighting autoencoder, DBN, and SVR, enhanced by SHAP importance analysis in soybean branching prediction. *Sci Rep* 2024; **14**: 5905.
20 Liu Y, Shen W, Tian Z. Using machine learning algorithms to predict high-risk factors for postoperative delirium in elderly patients. *Clin Interv Aging* 2023; **18**:157–168.
21 Dodsworth BT, Reeve K, Falco L, *et al.* Development and validation of an international preoperative risk assessment model for postoperative delirium. *Age Ageing* 2023; **52**:afad086.
22 van Meenen LC, van Meenen DM, de Rooij SE, *et al.* Risk prediction models for postoperative delirium: a systematic review and meta-analysis. *J Am Geriatr Soc* 2014; **62**:2383–2390.
23 Schmutz N, Reeve K, Gaudet J, *et al.* Multicentre external validation of the machine learning-based PIPRA predictive model for postoperative delirium in elderly patients. *SSRN [Preprint]* 2025; doi:10.2139/ssrn.5435556.
24 WMA. WMA declaration of Helsinki: Ethical principles for medical research involving human subjects. 2013.
25 Collins GS, Moons KGM, Dhiman P, *et al.* TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ* 2024; **385**:e078378.
26 Gaudreau JD, Gagnon P, Harel F, *et al.* Fast, systematic, and continuous delirium assessment in hospitalized patients: the nursing delirium screening scale. *J Pain Symptom Manage* 2005; **29**:368–375.
27 Hargrave A, Bastiaens J, Bourgeois JA, *et al.* Validation of a nurse-based delirium-screening tool for hospitalized patients. *Psychosomatics* 2017; **58**:594–603.
28 Lundberg SM, Lee S-I. A unified approach to interpreting model predictions. In: Proceedings of the 31st Conference on Neural Information Processing Systems Long Beach, CA, USA. 2017.
29 Guo C, Pleiss G, Sun Y, *et al.* On calibration of modern neural networks. In: Proceedings of the 34th International Conference on Machine Learning. 2017.
30 Jauk S, Kramer D, Sumerauer S, *et al.* Machine learning-based delirium prediction in surgical in-patients: a prospective validation study. *JAMIA Open* 2024; **7**:ooae091.
31 Kocar TD, Wolf P, Leinert C, *et al.* SURGE-ahead postoperative delirium prediction: external validation and open-source library. *Eur Geriatr Med* 2025; **16**:851–859.
32 Benovic S, Ajlani AH, Leinert C, *et al.* Introducing a machine learning algorithm for delirium prediction-the Supporting SURgery with GEriatric Co-Management and AI project (SURGE-Ahead). *Age Ageing* 2024; **53**: afae101.
33 Kramer D, Veeranki S, Hayn D, *et al.* Development and validation of a multivariable prediction model for the occurrence of delirium in hospitalized gerontopsychiatry and internal medicine patients. *Stud Health Technol Inform* 2017; **236**:32–39.
34 Zhu J, Wang W, Shi H. The association between postoperative cognitive dysfunction and cerebral oximetry during geriatric orthopedic surgery: a randomized controlled study. *Biomed Res Int* 2021; **2021**:5733139.
35 Ko CC, Hung KC, Chang YP, *et al.* Association of general anesthesia exposure with risk of postoperative delirium in patients receiving transcatheter aortic valve replacement: a meta-analysis and systematic review. *Sci Rep* 2023; **13**:16241.
36 He M, Zhu Z, Jiang M, *et al.* Risk factors for postanesthetic emergence delirium in adults: a systematic review and meta-analysis. *J Neurosurg Anesthesiol* 2024; **36**:190–200.
37 Neuman MD, Feng R, Carson JL, *et al.* REGAIN Investigators. Spinal anesthesia or general anesthesia for hip surgery in older adults. *New Engl J Med* 2021; **385**:2025–2035.
38 Li T, Li J, Yuan L, *et al.* RAGA Study Investigators. Effect of regional vs general anesthesia on incidence of postoperative delirium in older patients undergoing hip fracture surgery: the RAGA Randomized Trial. *JAMA* 2022; **327**:50–58.
39 Sumner M, Deng C, Evered L, *et al.* Processed electroencephalography-guided general anaesthesia to reduce postoperative delirium: a systematic review and meta-analysis. *Br J Anaesth* 2023; **130**:e243–e253.
40 Hastie T, Tibshirani R, Friedman JH, *et al.* The elements of statistical learning: data mining, inference, and prediction. 2009; Springer,
41 Mienye ID, Obaido G, Jere N, *et al.* A survey of explainable artificial intelligence in healthcare: concepts, applications, and challenges. *Inform Med Unlocked* 2024; **51**:101587.
42 Xu H, Shuttleworth KMJ. Medical artificial intelligence and the black box problem: a view based on the ethical principle of "do no harm". *Intelligent Med* 2024; **4**:52–57.