

Development and internal validation of explainable machine learning models for predicting failure of first ambulation within 12 h after cesarean delivery: A retrospective cohort study

LIN XIONG¹, YOUMING JIN¹, WEI LI¹, JUNJUN XIONG² and HONGMEI DING¹

¹Department of Obstetrics, The Affiliated Hospital, Southwest Medical University, Luzhou, Sichuan 646000, P.R. China;

²Department of Health Management Center, Luzhou Maternal & Child Health Hospital, Luzhou, Sichuan 646000, P.R. China

Received June 17, 2026; Accepted September 16, 2026

DOI: 10.3892/etm.2026.13320

Abstract. Early ambulation is a notable component of post-operative recovery after cesarean delivery, but some women fail to complete first ambulation within the early postoperative period because of pain, dizziness, orthostatic discomfort, residual lower limb numbness, nausea or vomiting, hypotension or poor activity tolerance. Early identification of women at high risk of ambulation failure may support timely and individualized nursing interventions. The present study aimed to develop and internally validate explainable machine learning models for predicting failure of first ambulation within 12 h after cesarean delivery using routinely collected obstetric peri-operative nursing assessment data. The present single-center retrospective cohort study included relatively stable women who underwent cesarean delivery between January 2022 and December 2025 and were considered clinically eligible for early mobilization. The primary outcome was failure of first ambulation within 12 h after surgery, defined as no documented successful ambulation within 12 h or discontinuation of an ambulation attempt because of nursing-documented intolerance. Candidate predictors included demographic, obstetric, pregnancy-related, laboratory, anesthetic, surgical and early postoperative nursing assessment variables. Postoperative symptoms and vital signs were assessed within predefined early postoperative intervals, whereas first turning time was treated as a dynamically updated indicator that became available when the first spontaneous or assisted turning event was documented. The model was therefore interpreted as a dynamic early postoperative risk-assessment

approach rather than a prediction model applied exclusively at a fixed 4-h time point. The final analytic dataset contained no item-level missing values; therefore, no imputation was performed. The cohort was randomly divided into training and internal validation cohorts at a ratio of 7:3. Logistic regression, random forest, support vector machine, Extreme Gradient Boosting (XGBoost) and LightGBM models were developed and compared. Model performance was assessed using discrimination, classification metrics, calibration, Brier score and decision curve analysis. SHapley Additive exPlanations (SHAP) analysis was used to interpret the final model. A total of 1,486 women were included, and 326 women experienced failure of first ambulation within 12 h, with an incidence of 21.9%. The training and validation cohorts included 1,040 and 446 women, respectively. In the validation cohort, LightGBM achieved an area under the receiver operating characteristic curve (AUC) of 0.809 and a Brier score of 0.132, whereas XGBoost achieved an AUC of 0.805 and a Brier score of 0.134. The paired DeLong test showed no statistically significant difference between the validation AUCs of the two models ($P=0.647$). XGBoost achieved higher sensitivity, negative predictive value and F1 score than LightGBM (0.714 vs. 0.653, 0.906 vs. 0.892 and 0.567 vs. 0.561, respectively). Because these metrics were particularly relevant to identifying high-risk women and minimizing missed high-risk cases, XGBoost was selected as the final model. Calibration curves showed acceptable agreement between predicted and observed risks, and decision curve analysis demonstrated net benefit across relevant threshold probabilities. The 10 highest-ranked predictors according to mean absolute SHAP value were postoperative pain NRS score at 2 h, emergency cesarean delivery, time to first documented turning, lower limb numbness within 0–4 h, systolic blood pressure at 2 h, body mass index, preoperative hemoglobin, age, intraoperative fluid volume and dizziness or orthostatic discomfort within 0–4 h. In conclusion, in the present single-center retrospective cohort of relatively stable women considered clinically eligible for early mobilization, the explainable XGBoost model showed promising internal performance for predicting failure of first ambulation within 12 h after cesarean delivery. The model should not be extrapolated to women transferred to intensive care, women with

Correspondence to: Dr Hongmei Ding, Department of Obstetrics, The Affiliated Hospital, Southwest Medical University, 8 Kangcheng Road Section 2, Jiangyang, Luzhou, Sichuan 646000, P.R. China
E-mail: 15283014743@163.com

Key words: cesarean delivery, early ambulation, nursing assessment, machine learning, Extreme Gradient Boosting, SHapley Additive exPlanations, prediction model

severe postpartum hemorrhage, women with explicit medical restrictions on ambulation or women with severe postoperative complications. These findings support further evaluation but do not establish clinical utility, transportability or readiness for implementation. Temporal and external validation, followed by prospective assessment of workflow feasibility and clinical impact, is required before clinical use can be considered.

Introduction

Cesarean delivery is one of the most commonly performed surgical procedures worldwide and represents a substantial component of modern obstetric care. Although cesarean delivery is often lifesaving for mothers and neonates, it is also associated with postoperative pain, delayed functional recovery, impaired maternal-infant interaction, prolonged hospital stay and increased nursing workload when early postoperative recovery is suboptimal (1). In recent years, enhanced recovery after surgery and enhanced recovery after cesarean delivery pathways have emphasized multimodal analgesia, early oral intake, early urinary catheter removal and early mobilization as key elements of standardized postoperative care (2). Among these components, early ambulation is particularly important because it reflects not only physical recovery but also the patient's ability to tolerate activity after anesthesia and surgery.

Early postoperative ambulation after cesarean delivery has several potential benefits, including promoting gastrointestinal recovery, reducing venous stasis, improving maternal comfort, facilitating breastfeeding and newborn care and supporting earlier return to daily activity (3). Conversely, failure to ambulate during the early postoperative period may indicate inadequate pain control, hemodynamic intolerance, residual lower limb numbness, nausea or vomiting, dizziness, orthostatic discomfort or poor overall activity tolerance. These symptoms are commonly encountered by obstetric nurses during postoperative assessment and may delay progressive mobilization despite the absence of major surgical complications (4). Therefore, identifying women who are unlikely to complete first ambulation within a clinically meaningful early time window may allow nurses to intensify assessment, optimize analgesia, provide assisted mobilization and implement individualized recovery planning.

Current postoperative nursing practice after cesarean delivery mainly relies on routine vital sign monitoring, pain assessment, symptom inquiry and clinical judgment to determine readiness for mobilization. However, failure of early ambulation may be influenced by multiple interacting factors, including body mass index, emergency cesarean delivery, anesthesia and surgical variables, blood loss, intraoperative hypotension, postoperative pain, systolic blood pressure, nausea or vomiting, lower limb numbness, dizziness or orthostatic discomfort and early turning status (5). These factors may not be adequately captured by conventional single-factor assessment, highlighting the need for individualized risk stratification.

Clinical prediction models offer a structured approach to estimating individualized risk and may support more timely and objective decision-making in postoperative nursing care (6). Traditional regression-based models are widely used

because of their interpretability, but they may have limited ability to capture complex non-linear relationships and interactions among perioperative variables. Machine learning methods, including random forest, support vector machine, extreme gradient boosting (XGBoost) and light gradient boosting machine algorithms, have increasingly been applied to perioperative and obstetric prediction tasks because they can integrate multidimensional clinical data and may improve predictive performance (7). In the obstetric setting, machine learning has been used to predict outcomes such as cesarean delivery, postpartum hemorrhage, postoperative pain and other maternal complications, suggesting its potential value for risk assessment in maternal care (8).

Despite these advances, few studies have specifically focused on failure of first ambulation after cesarean delivery as a nursing-sensitive postoperative outcome. Most existing enhanced recovery studies evaluate broader endpoints, such as postoperative pain, length of stay, opioid consumption, breastfeeding or overall recovery quality, while early ambulation failure is often treated as a process indicator rather than a prediction target (9). However, first ambulation within the early postoperative period is a concrete and clinically observable milestone that can be directly assessed and acted upon by nurses. A model designed around this outcome may be particularly useful in obstetric wards because it translates routinely collected nursing assessment data into individualized risk estimates that can guide early bedside interventions.

Another notable consideration is model interpretability. Although machine learning models may achieve strong predictive performance, their clinical adoption may be limited when the basis of individual predictions is unclear. Explainable artificial intelligence methods, particularly SHapley Additive exPlanations (SHAP), can quantify the contribution of each predictor to model output and provide both global and individualized explanations (10). Such interpretability is important in nursing applications, where model outputs should be clinically understandable and aligned with bedside assessment. For example, identifying pain, time to first documented turning, lower limb numbness, blood pressure or dizziness as influential predictive signals may help direct additional assessment. However, SHAP values describe contributions to model predictions and do not establish that these variables causally determine ambulation failure.

Gradient boosting algorithms are especially suitable for tabular clinical data. XGBoost is a scalable tree-boosting method that has shown strong performance across prediction tasks and incorporates regularization strategies that may reduce overfitting (11). LightGBM is another efficient gradient boosting framework that can handle high-dimensional structured data and has demonstrated competitive predictive accuracy with improved computational efficiency (12). Comparing these algorithms with conventional logistic regression and other machine learning approaches may help identify a model that balances discrimination, calibration, sensitivity and clinical usability. In postoperative nursing risk prediction, sensitivity and negative predictive value are particularly important because missed high-risk patients may lose the opportunity for timely pain management, assisted mobilization and fall-risk prevention.

To ensure transparent reporting and methodological rigor, prediction model studies using regression or machine learning methods should clearly describe study design, participant selection, predictor definitions, outcome ascertainment, model development, internal validation, performance evaluation, calibration and clinical utility (13). In addition, decision curve analysis can help determine whether a model provides net clinical benefit across plausible threshold probabilities, which is important when considering whether risk-based nursing interventions are justified (14).

Therefore, the present study aimed to develop and internally validate explainable machine learning models for predicting failure of first ambulation within 12 h after cesarean delivery using routinely collected perioperative nursing assessment data. By comparing logistic regression, random forest, support vector machine, XGBoost and LightGBM models, and by applying SHAP-based interpretation to the final model, this study sought to provide an interpretable risk stratification tool that may support individualized early mobilization planning and postoperative nursing decision-making after cesarean delivery.

Materials and methods

Study design. The present study was a single-center retrospective cohort study designed to develop and internally validate explainable machine learning models for predicting failure of first ambulation within 12 h after cesarean delivery. The present study used routinely collected obstetric perioperative nursing assessment data, together with demographic, obstetric, laboratory, anesthetic and surgical information. No prospective intervention was performed. The full cohort was randomly divided into a training cohort and a validation cohort at a ratio of 7:3. The training cohort was used for model development, threshold selection and model interpretation, whereas the validation cohort was used for internal validation and final performance evaluation.

The outcome of interest was failure of first ambulation within 12 h after cesarean delivery. The intended prediction task was dynamic early postoperative risk assessment before the 12-h outcome was determined, rather than risk estimation at one fixed postoperative time point. Demographic, obstetric, laboratory, anesthetic and surgical variables were available at or shortly after completion of surgery. Pain and systolic blood pressure were assessed at 2 h, and dizziness or orthostatic discomfort, nausea or vomiting, lower limb numbness and sedation or sleepiness were recorded during the prespecified 0–4-h postoperative assessment interval. These fixed-time observations were eligible as predictors only when they were documented before the outcome was determined. First turning time was treated as a time-updated functional recovery indicator that entered the available predictor set when the first spontaneous or assisted turning event was documented; it was not treated as a fixed 4-h measurement. Variables documented only during or after a failed ambulation attempt, or after the 12-h outcome had been determined, were not used as pre-outcome predictor observations. Accordingly, the model should be interpreted as a dynamic early postoperative assessment model whose available information may be updated during the early recovery period.

Setting. The present study was conducted in the Department of Obstetrics at The Affiliated Hospital, Southwest Medical University (Luzhou, China). Consecutive women who underwent cesarean delivery between January 2022 and December 2025 were retrospectively screened. Clinical data were obtained from the electronic medical record system, obstetric perioperative records, anesthesia records, laboratory information system and structured nursing assessment records. In routine clinical care at the study center, women undergoing cesarean delivery were encouraged to begin progressive postoperative mobilization when clinically stable, including assisted turning, sitting, standing and first ambulation under nursing supervision.

Participants. Women were eligible if they underwent cesarean delivery during the study period and had perioperative nursing assessment records that allowed determination of first ambulation status within 12 h after surgery. The exclusion criteria were transfer to the intensive care unit after surgery, severe postpartum hemorrhage requiring intensive monitoring or emergency intervention, explicit postoperative medical orders restricting ambulation, severe postoperative complications that made early mobilization clinically inappropriate, inability to determine whether first ambulation was attempted or completed within 12 h or insufficient records to establish the required analytic variables. Of 1,612 records initially screened, 126 were excluded: 18 because of postoperative intensive-care transfer, 14 because of severe postpartum hemorrhage, 21 because of explicit ambulation restrictions, 17 because of severe postoperative complications, 24 because the ambulation outcome was indeterminate and 32 because the required predictor information was insufficient. The remaining 1,486 women constituted the final analytic cohort. All 1,486 participants were women who underwent cesarean delivery; male participants were not part of the eligible obstetric population. The median age was 31 years, with a range of 18–45 years. The final analytic dataset contained complete outcome and predictor information for all included women, with 0% item-level missingness among the retained records. Of these, 1,040 were assigned to the training cohort and 446 to the validation cohort. The participant screening and eligibility process is summarized in Fig. S1.

Variables. The primary outcome was failure of first ambulation within 12 h after cesarean delivery. The 12-h window was calculated from the documented completion time of surgery. Successful first ambulation was defined as standing and walking at least several steps away from the bedside, with or without nursing assistance, as documented in the structured nursing mobilization record. Sitting at the bedside, dangling, standing without walking or turning in bed alone was not considered successful ambulation. Failure of first ambulation was defined as either: i) No documented successful ambulation within 12 h when the nursing record was sufficient to determine the patient's ambulation status during the entire assessment window; or ii) an ambulation attempt that was discontinued because of nursing-documented intolerance, including pain, dizziness, orthostatic discomfort, hypotension, nausea or vomiting, lower limb numbness, incision discomfort, weakness or another documented safety concern. Symptoms

recorded during routine pre-ambulation assessment before an attempt were treated as pre-outcome predictor observations. When a symptom was first documented only during the ambulation attempt, it was used to describe the reason for discontinuation and was not treated as a separate pre-outcome predictor observation. If the available records were insufficient to determine whether ambulation was completed or attempted within 12 h, the outcome was classified as indeterminate and the record was excluded rather than treated as a successful or failed outcome. Nursing documentation followed the structured postoperative assessment and mobilization-recording process used at the study center.

Candidate predictors were selected according to clinical relevance, availability in routine records, temporal relationship with the outcome and nursing interpretability. Variables included demographic characteristics, obstetric information, pregnancy-related conditions, preoperative laboratory indicators, anesthesia and surgical factors, and early postoperative nursing assessment indicators.

Demographic and obstetric variables included age, body mass index, gestational age, primipara status, previous cesarean delivery, emergency cesarean delivery and multiple pregnancy. Pregnancy-related conditions included hypertensive disorders of pregnancy and gestational diabetes mellitus. Preoperative laboratory variables included hemoglobin, preoperative anemia and albumin. Anesthesia and surgical variables included anesthesia type, surgery duration, estimated blood loss, intraoperative hypotension and intraoperative fluid volume. Early postoperative nursing assessment variables measured within predefined intervals included systolic blood pressure and postoperative pain numeric rating scale score at 2 h, as well as dizziness or orthostatic discomfort, nausea or vomiting, lower limb numbness and sedation or sleepiness within 0–4 h. First turning time was analyzed separately as a dynamically updated functional recovery indicator, defined by the earliest documented spontaneous or assisted turning event after surgery. Completion of early mobilization education and preoperative anxiety or fear were also included.

Data sources. All data were extracted from routinely collected hospital records. Demographic and obstetric variables were obtained from admission records and obstetric documentation. Laboratory indicators were obtained from the most recent preoperative laboratory tests before cesarean delivery. Anesthesia type, intraoperative hypotension, intraoperative fluid volume, surgery duration and estimated blood loss were obtained from anesthesia and operative records. Early postoperative variables were obtained from structured nursing assessment records.

Pain intensity was assessed using the numeric rating scale, ranging from 0 to 10, with higher scores indicating more severe pain. Dizziness or orthostatic discomfort, nausea or vomiting, lower limb numbness, sedation or sleepiness and completion of mobilization education were recorded as binary variables based on nursing documentation. First turning time was defined as the elapsed time from completion of surgery to the earliest documented spontaneous or assisted turning in bed and was recorded in hours. It was a dynamically updated observation that became available when the turning event was documented, rather than a baseline predictor measured

at a fixed postoperative time point. Because first turning may reflect pain, residual anesthetic effects, lower limb numbness, fatigue, activity tolerance and the need for nursing assistance, this variable was interpreted as an early warning marker of functional recovery status and not as a causal determinant of ambulation failure. Intraoperative hypotension was defined according to anesthesia records as a documented episode of clinically significant blood pressure reduction requiring monitoring or intervention.

Missing data. Data completeness was assessed for the primary outcome and all candidate predictors before model development. Among the 1,486 women retained in the final analytic cohort, no item-level missing values were identified for any demographic, obstetric, pregnancy-related, laboratory, anesthetic, surgical or postoperative nursing assessment variable; the variable-specific missingness rate was therefore 0%, and no imputation was performed. This complete-case result applies to the retained analytic cohort and does not imply that all initially screened records were complete. Records with insufficient information to determine the outcome or establish the required predictors were excluded during eligibility assessment. Because no missing observations remained in the analytic dataset, an imputation-based sensitivity analysis was not applicable. Nevertheless, exclusion of records with insufficient information may have introduced selection bias and may limit generalizability if documentation completeness differed across patient groups or time periods.

Bias. Several strategies were used to reduce potential bias. First, consecutive eligible patients during the study period were screened to minimize selection bias. Second, predictor timing and availability were explicitly considered. Fixed-time postoperative assessments were distinguished from first turning time, which was treated as a dynamically updated recovery indicator rather than a baseline predictor. No variables documented after determination of the 12-h ambulation outcome were included. Third, women with explicit medical restrictions to ambulation were excluded because their failure to ambulate would reflect medical contraindication rather than ambulation intolerance. Fourth, model development and evaluation were separated using training and internal validation cohorts derived from the same source population. Fifth, model thresholds were determined in the training cohort and then applied to the validation cohort to avoid optimistic estimation of classification performance. Finally, SHAP analysis was used to improve model transparency; however, SHAP contributions were interpreted as predictive associations rather than causal effects.

Study size. The study size was determined by the number of consecutive eligible cesarean-delivery cases available during the prespecified study period after application of the inclusion and exclusion criteria; no separate hypothesis-testing power calculation was performed because the primary objective was prediction-model development and internal validation. The final cohort included 1,486 women, of whom 326 experienced failure of first ambulation within 12 h after cesarean delivery, corresponding to an event rate of 21.9%. The training cohort included 1,040 women with 228 outcome events, and the

validation cohort included 446 women with 98 outcome events. The 26 candidate predictors were counted before dummy encoding of categorical variables, yielding 8.8 outcome events per prespecified candidate predictor in the training cohort. This ratio is provided for transparency but should not be interpreted as a formal adequacy criterion for the machine-learning models, because the effective number of parameters depends on categorical encoding, model architecture, regularization and tuning procedures. For the planned development strategy, the available sample was considered suitable for initial single-center internal development because the models used a prespecified predictor set, stratified five-fold cross-validation within the training cohort, no class rebalancing or resampling and constrained boosting configurations. However, the sample size cannot establish model stability or transportability, and the results require confirmation in larger independent temporal and geographic cohorts.

Quantitative variables. Continuous variables were retained in their original continuous form for model development whenever possible, including age, body mass index, gestational age, preoperative hemoglobin, preoperative albumin, surgery duration, estimated blood loss, intraoperative fluid volume, systolic blood pressure at 2 h and postoperative pain score at 2 h. Binary variables were coded as 0 or 1. Anesthesia type was treated as a categorical variable and converted into dummy variables for machine learning modeling. No clinically meaningful continuous variable was arbitrarily categorized for the primary model, in order to preserve information and avoid loss of predictive power.

Statistical methods. Baseline characteristics are presented as mean $\pm$ standard deviation for continuous variables and frequency with percentage for categorical variables. Comparisons between the training and validation cohorts were performed to assess baseline balance. Comparisons between women with successful and failed first ambulation in the training cohort were also conducted to describe outcome-related differences. Welch's t-test was used for continuous variables, and Pearson's χ^2 test or Fisher's exact test was used for categorical variables, as appropriate. Before model development, data completeness was summarized as the number and percentage of missing observations for each variable. Because no missing values were identified in the final analytic dataset, all 1,486 included women were retained in the analysis and no imputation was performed.

A total of five prediction models were developed using the same prespecified predictor set. Categorical variables were dummy coded, and continuous variables were standardized when required by the algorithm. Model tuning was performed within the training cohort using stratified five-fold cross-validation with a fixed random seed (20260822), without class rebalancing or resampling. For the two leading boosting models, the final XGBoost configuration used a maximum tree depth of 3, a learning rate of 0.03 and 100 boosting iterations, whereas the final LightGBM configuration used 7 leaves, a learning rate of 0.03, and 100 boosting iterations.

The sample-size considerations were interpreted in relation to the prespecified model-development procedure. The 228 training-cohort events provided information for

estimating risk patterns across the 26 prespecified candidate predictor variables before categorical encoding. Because tree-based machine-learning models do not have a one-to-one correspondence between candidate predictors and regression coefficients, adequacy was not judged solely by a conventional events-per-variable threshold. Instead, model complexity was controlled through prespecification of the predictor set, stratified five-fold cross-validation, and constrained boosting hyperparameters. The resulting sample size was therefore considered appropriate for exploratory single-center internal development, while recognizing that additional data are required to establish reliable calibration, model stability, and transportability.

Model performance was evaluated in both the training and validation cohorts. Discrimination was assessed using the area under the receiver operating characteristic curve (AUC), with 95% confidence intervals calculated using DeLong's method. Because XGBoost and LightGBM were evaluated in the same validation participants, their validation AUCs were compared using a two-sided paired DeLong test. Classification performance was assessed using accuracy, sensitivity, specificity, positive predictive value, negative predictive value and F1 score. Overall prediction error was assessed using the Brier score. Calibration of the final model was evaluated using calibration curves based on grouped predicted risks. Clinical utility was assessed using decision curve analysis by comparing the net benefit of the final model with the treat-all and treat-none strategies across a range of threshold probabilities.

The final model was selected according to overall validation performance and the intended clinical purpose. When the leading models showed comparable discrimination, sensitivity, negative predictive value and F1 score were given particular consideration because the primary purpose of the model was to identify women at high risk of early ambulation failure and minimize missed high-risk cases. Accordingly, XGBoost was selected as the final model for subsequent calibration, decision curve analysis and SHAP-based interpretation.

To interpret the final XGBoost model, SHAP analysis was performed in the validation cohort. Predictors were ranked according to their mean absolute SHAP values. The direction of association between each predictor and predicted risk was interpreted from the SHAP value distributions. SHAP summary plots and dependence plots were used to visualize the contribution of key predictors to individualized model predictions. All statistical analyses and model development were performed using R software (version 4.3.2; RStudio, Inc.). A two-sided $P < 0.05$ was considered to indicate a statistically significant difference.

Results

Baseline characteristics of the training and validation cohorts.

A total of 1,486 women who underwent cesarean delivery were included in the present study, comprising 1,040 patients in the training cohort and 446 patients in the internal validation cohort. The overall incidence of failure of first ambulation within 12 h after cesarean delivery was 21.9% (326/1,486), with similar event rates between the training and internal validation cohorts (21.9 vs. 22.0%). Baseline characteristics were generally well balanced between the training and internal

Table I. Baseline characteristics of the overall, training and internal validation cohorts.

Characteristic	Overall (n=1,486)	Training cohort (n=1,040)	Validation cohort (n=446)	P-value
Age, years	30.7±4.6; 31 (18-45)	30.8±4.6	30.6±4.6	0.554
Body mass index, kg/m ²	29.0±3.7	29.0±3.7	28.9±3.6	0.804
Gestational age, weeks	38.6±1.3	38.6±1.3	38.6±1.3	0.893
Primipara	846 (56.9)	599 (57.6)	247 (55.4)	0.463
Previous cesarean delivery	323 (21.7)	221 (21.3)	102 (22.9)	0.532
Emergency cesarean delivery	426 (28.7)	299 (28.8)	127 (28.5)	0.964
Multiple pregnancy	51 (3.4)	38 (3.7)	13 (2.9)	0.574
Hypertensive disorders of pregnancy	137 (9.2)	99 (9.5)	38 (8.5)	0.608
Gestational diabetes mellitus	207 (13.9)	155 (14.9)	52 (11.7)	0.116
Preoperative hemoglobin, g/l	117.7±10.9	117.7±11.0	117.5±10.8	0.788
Preoperative anemia	354 (23.8)	248 (23.8)	106 (23.8)	>0.999
Preoperative albumin, g/l	36.3±3.2	36.3±3.2	36.3±3.2	0.805
Anesthesia type				0.699
Combined spinal-epidural anesthesia	364 (24.5)	249 (23.9)	115 (25.8)	
General anesthesia	41 (2.8)	30 (2.9)	11 (2.5)	
Spinal anesthesia	1,081 (72.7)	761 (73.2)	320 (71.7)	
Surgery duration, min	56.6±14.4	56.8±14.1	56.1±15.1	0.395
Estimated blood loss, ml	408.5±143.5	407.5±140.1	410.7±151.3	0.698
Intraoperative hypotension	201 (13.5)	132 (12.7)	69 (15.5)	0.176
Intraoperative fluid volume, ml	1,918.8±406.7	1,911.3±403.7	1,936.2±413.6	0.284
Systolic blood pressure at 2 h, mmHg	110.4±10.3	110.7±10.3	109.9±10.2	0.211
Postoperative pain NRS score at 2 h	4.3±1.5	4.3±1.5	4.3±1.5	0.738
Dizziness or orthostatic discomfort within 0-4 h	234 (15.7)	166 (16.0)	68 (15.2)	0.788
Nausea or vomiting within 0-4 h	307 (20.7)	216 (20.8)	91 (20.4)	0.929
Lower limb numbness within 0-4 h	248 (16.7)	173 (16.6)	75 (16.8)	0.992
Urinary catheter in place at 6 h	466 (31.4)	340 (32.7)	126 (28.3)	0.103
Time to first documented turning, h	5.1±1.4	5.1±1.4	4.9±1.4	0.049
Sedation or sleepiness within 0-4 h	172 (11.6)	131 (12.6)	41 (9.2)	0.073
Early mobilization education completed	611 (41.1)	424 (40.8)	187 (41.9)	0.720
Preoperative anxiety or fear	307 (20.7)	207 (19.9)	100 (22.4)	0.304
Failure of first ambulation within 12 h	326 (21.9)	228 (21.9)	98 (22.0)	>0.999

Data are presented as mean ± standard deviation or n (%), unless otherwise indicated. For the overall cohort, age is additionally presented as median (range): 31 (18-45) years. All participants were women who underwent cesarean delivery. P-values compare the training and internal validation cohorts. NRS, numeric rating scale.

validation cohorts. No statistically significant differences were observed for most demographic, obstetric, perioperative and early postoperative nursing assessment variables. The only variable showing a slight between-cohort difference was first turning time (5.1±1.4 vs. 4.9±1.4 h; $P=0.049$), whereas the event rate was similar between cohorts (21.9 vs. 22.0%; $P>0.999$) (Table I).

Characteristics according to first ambulation outcome in the training cohort. In the training cohort, 812 women completed first ambulation within 12 h after cesarean delivery, whereas 228 failed to do so. Compared with women who successfully completed first ambulation, those with ambulation failure had a higher body mass index (29.6±3.8 vs. 28.8±3.7 kg/m²;

$P=0.008$), a substantially higher proportion of emergency cesarean delivery (52.2 vs. 22.2%; $P<0.001$), lower preoperative hemoglobin levels (116.1±11.3 vs. 118.1±10.8 g/l; $P=0.017$) and a higher prevalence of preoperative anemia (32.5 vs. 21.4%; $P<0.001$). Significant differences were also observed in anesthesia type, surgery duration, estimated blood loss and intraoperative hypotension. Specifically, the failed ambulation group had longer surgery duration (60.3±15.1 vs. 55.9±13.7 min; $P<0.001$), greater estimated blood loss (447.2±158.2 vs. 396.4±132.6 ml; $P<0.001$) and a higher incidence of intraoperative hypotension (19.3 vs. 10.8%; $P=0.001$). Regarding early postoperative nursing assessment indicators, women with ambulation failure had lower systolic blood pressure at 2 h after surgery (108.6±10.0

Table II. Characteristics according to first ambulation outcome in the training cohort.

Characteristic	Training cohort overall (n=1,040)	Successful first ambulation (n=812)	Failed first ambulation (n=228)	P-value
Age, years	30.8±4.6	30.8±4.7	30.7±4.4	0.773
Body mass index, kg/m ²	29.0±3.7	28.8±3.7	29.6±3.8	0.008
Gestational age, weeks	38.6±1.3	38.6±1.3	38.6±1.2	0.611
Primipara	599 (57.6)	460 (56.7)	139 (61.0)	0.276
Previous cesarean delivery	221 (21.3)	175 (21.6)	46 (20.2)	0.721
Emergency cesarean delivery	299 (28.8)	180 (22.2)	119 (52.2)	<0.001
Multiple pregnancy	38 (3.7)	35 (4.3)	3 (1.3)	0.054
Hypertensive disorders of pregnancy	99 (9.5)	82 (10.1)	17 (7.5)	0.283
Gestational diabetes mellitus	155 (14.9)	120 (14.8)	35 (15.4)	0.913
Preoperative hemoglobin, g/l	117.7±11.0	118.1±10.8	116.1±11.3	0.017
Preoperative anemia	248 (23.8)	174 (21.4)	74 (32.5)	<0.001
Preoperative albumin, g/l	36.3±3.2	36.3±3.2	36.2±3.2	0.512
Anesthesia type				0.016
Spinal anesthesia	761 (73.2)	607 (74.8)	154 (67.5)	
Combined spinal-epidural anesthesia	249 (23.9)	187 (23.0)	62 (27.2)	
General anesthesia	30 (2.9)	18 (2.2)	12 (5.3)	
Surgery duration, min	56.8±14.1	55.9±13.7	60.3±15.1	<0.001
Estimated blood loss, ml	407.5±140.1	396.4±132.6	447.2±158.2	<0.001
Intraoperative hypotension	132 (12.7)	88 (10.8)	44 (19.3)	0.001
Intraoperative fluid volume, ml	1,911.3±403.7	1,908.1±403.2	1,922.9±405.9	0.625
Systolic blood pressure at 2 h, mmHg	110.7±10.3	111.2±10.3	108.6±10.0	<0.001
Postoperative pain NRS score at 2 h	4.3±1.5	4.0±1.4	5.2±1.4	<0.001
Dizziness or orthostatic discomfort within 0-4 h	166 (16.0)	102 (12.6)	64 (28.1)	<0.001
Nausea or vomiting within 0-4 h	216 (20.8)	154 (19.0)	62 (27.2)	0.009
Lower limb numbness within 0-4 h	173 (16.6)	102 (12.6)	71 (31.1)	<0.001
Urinary catheter in place at 6 h	340 (32.7)	271 (33.4)	69 (30.3)	0.421
Time to first documented turning, h	5.1±1.4	4.9±1.4	5.9±1.4	<0.001
Sedation or sleepiness within 0-4 h	131 (12.6)	94 (11.6)	37 (16.2)	0.079
Early mobilization education completed	424 (40.8)	336 (41.4)	88 (38.6)	0.497
Preoperative anxiety or fear	207 (19.9)	168 (20.7)	39 (17.1)	0.270

Data are presented as mean ± standard deviation or n (%). P-values were calculated using Welch's t-test for continuous variables and Pearson's χ^2 test or Fisher's exact test for categorical variables, as appropriate. The training cohort included 1,040 patients, of whom 812 completed first ambulation within 12 h after cesarean delivery and 228 failed to do so. Time to first documented turning was defined as the elapsed time from completion of surgery to the earliest documented spontaneous or assisted turning in bed. It was treated as a dynamically updated marker of early functional recovery and not as a causal determinant of ambulation failure. NRS, numeric rating scale.

vs. 111.2±10.3 mmHg; $P<0.001$), higher postoperative pain NRS scores at 2 h (5.2±1.4 vs. 4.0±1.4; $P<0.001$) and longer first turning time (5.9±1.4 vs. 4.9±1.4 h; $P<0.001$). They also had higher rates of dizziness or orthostatic discomfort within 0-4 h (28.1 vs. 12.6%; $P<0.001$), nausea or vomiting (27.2 vs. 19.0%, $P=0.009$) and lower limb numbness (31.1 vs. 12.6%; $P<0.001$). No significant differences were observed in age, gestational age, parity status, previous cesarean delivery, hypertensive disorders of pregnancy, gestational diabetes mellitus, intraoperative fluid volume, early mobilization education completion or preoperative anxiety or fear (Table II).

Model performance and selection of the final model. The predictive performance of the five models, all developed using the same prespecified set of 26 candidate predictors described in the Variables subsection, is summarized in Table III and illustrated by the ROC curves in Fig. 1. In the training cohort, random forest showed the highest apparent discrimination, with an AUC of 0.996, followed by support vector machine, XGBoost, LightGBM and logistic regression, with AUC values of 0.946, 0.924, 0.919 and 0.841, respectively. However, random forest performance declined substantially in the validation cohort, with an AUC of 0.800, suggesting potential overfitting. In the internal validation cohort, LightGBM achieved an AUC

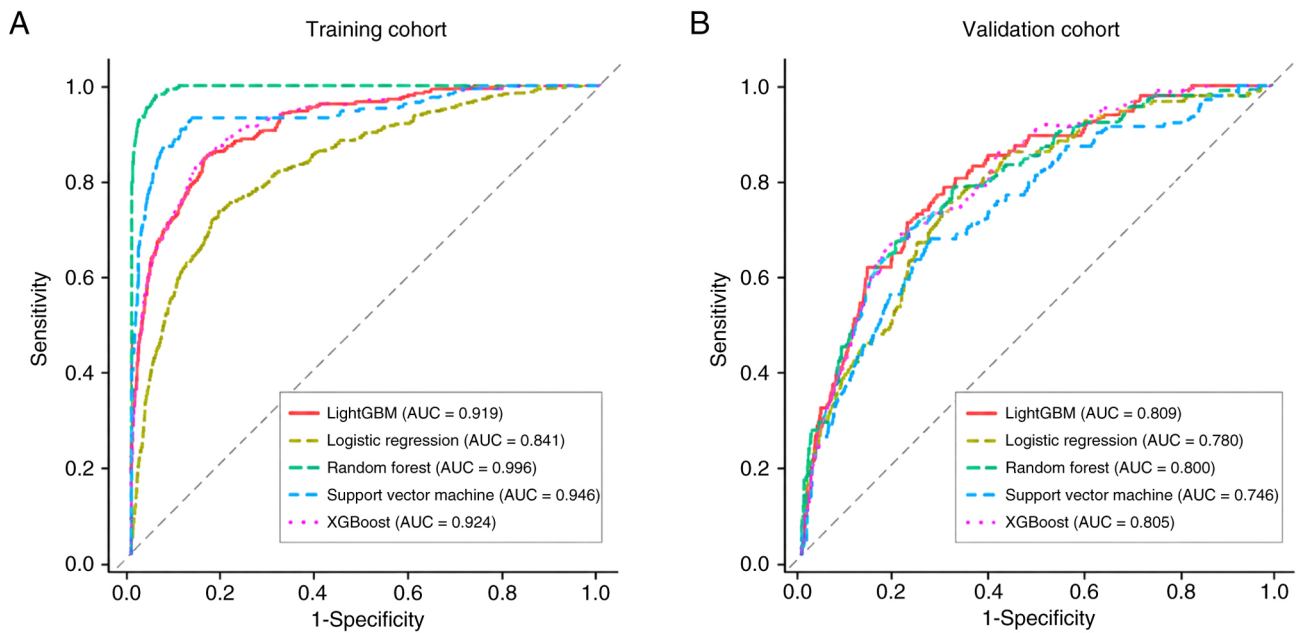


Figure 1. Receiver operating characteristic curves of the five prediction models in the training and validation cohorts. (A) Receiver operating characteristic curves of logistic regression, random forest, support vector machine, XGBoost and LightGBM models in the training cohort. (B) Receiver operating characteristic curves of the five models in the internal validation cohort. The AUC is shown for each model. Although random forest showed the highest apparent discrimination in the training cohort, its performance decreased in the validation cohort, suggesting potential overfitting. In the validation cohort, LightGBM achieved an AUC of 0.809 and XGBoost achieved an AUC of 0.805; the paired DeLong test showed no statistically significant difference between the two models ($P=0.647$). XGBoost was selected as the final model because it showed higher sensitivity, negative predictive value and F1 score, which were more closely aligned with the clinical objective of identifying high-risk women and minimizing missed high-risk cases. AUC, area under the receiver operating characteristic curve; XGBoost, Extreme Gradient Boosting.

of 0.809 (95% CI, 0.762-0.857), accuracy of 0.776, sensitivity of 0.653, specificity of 0.810, positive predictive value of 0.492, negative predictive value of 0.892, F1 score of 0.561 and Brier score of 0.132. XGBoost achieved an AUC of 0.805 (95% CI, 0.758-0.852), accuracy of 0.760, sensitivity of 0.714, specificity of 0.773, positive predictive value of 0.470, negative predictive value of 0.906, F1 score of 0.567 and Brier score of 0.134. The paired DeLong test showed no statistically significant difference between the validation AUCs of XGBoost and LightGBM ($P=0.647$), indicating that the two models had comparable discrimination. Although LightGBM achieved a marginally higher AUC and a lower Brier score, XGBoost demonstrated higher sensitivity, negative predictive value, and F1 score in the validation cohort. Because the primary clinical objective was to identify women at high risk of failure of first ambulation and minimize missed high-risk cases, XGBoost was selected as the final model for subsequent calibration, decision curve analysis, and SHAP-based interpretability analysis (Table III and Fig. 1).

Calibration and decision curve analysis of the XGBoost model.

The calibration curve showed general agreement between predicted and observed risks in the internal validation cohort, although some deviation from ideal calibration was observed at higher predicted probabilities. The validation calibration intercept was 1.290 and the calibration slope was 2.161; the corresponding optimism-corrected estimates were 1.268 and 2.137. These findings suggest that calibration was not uniform across the full predicted-risk range and should be reassessed in future temporal and external validation. Decision curve analysis showed that the XGBoost model provided higher

estimated net benefit than the treat-all and treat-none strategies across much of the threshold range displayed in Fig. 2. Because the analysis was based on single-center internal validation, it should be interpreted as preliminary decision-analytic evidence rather than proof of clinical utility.

SHAP-based interpretation of the final XGBoost model.

SHAP analysis was performed in the validation cohort to interpret the final XGBoost model. The 10 highest-ranked predictors according to mean absolute SHAP value, in descending order, were postoperative pain NRS score at 2 h, emergency cesarean delivery, time to first documented turning, lower limb numbness within 0-4 h, systolic blood pressure at 2 h, body mass index, preoperative hemoglobin, age, intraoperative fluid volume and dizziness or orthostatic discomfort within 0-4 h. Postoperative pain NRS score at 2 h had the highest mean absolute SHAP value (0.6876). The SHAP summary plot indicated that higher pain scores, emergency cesarean delivery, later time to first documented turning, lower limb numbness, higher body mass index, lower preoperative hemoglobin, lower systolic blood pressure and dizziness or orthostatic discomfort generally shifted predictions toward a higher estimated risk of ambulation failure. Although age and intraoperative fluid volume did not show statistically significant between-group differences in the univariable comparisons presented in Table II, they ranked among the 10 highest predictors by mean absolute SHAP value in the multivariable XGBoost model. This is not contradictory, because SHAP importance reflects a variable's contribution to model predictions, including possible non-linear and interaction effects, rather than statistical significance in univariable hypothesis testing. No simple directional interpretation was

Table III. Predictive performance of different models in the training and validation cohorts.

Model	Cohort	AUC	AUC 95% CI	Accuracy	Sensitivity	Specificity	PPV	NPV	F1 score	Brier score
Logistic regression	Training cohort	0.841	0.811-0.871	0.798	0.746	0.813	0.528	0.919	0.618	0.119
	Validation cohort	0.780	0.729-0.830	0.738	0.602	0.776	0.431	0.874	0.502	0.143
Random forest	Training cohort	0.996	0.994-0.998	0.959	0.982	0.952	0.852	0.995	0.912	0.055
	Validation cohort	0.800	0.751-0.850	0.780	0.643	0.819	0.500	0.891	0.562	0.134
Support vector machine	Training cohort	0.946	0.927-0.966	0.922	0.877	0.935	0.791	0.964	0.832	0.073
	Validation cohort	0.746	0.690-0.803	0.724	0.633	0.750	0.416	0.879	0.502	0.146
XGBoost	Training cohort	0.924	0.904-0.943	0.837	0.877	0.825	0.585	0.960	0.702	0.089
	Validation cohort	0.805	0.758-0.852	0.760	0.714	0.773	0.470	0.906	0.567	0.134
LightGBM	Training cohort	0.919	0.900-0.939	0.844	0.860	0.840	0.601	0.955	0.708	0.093
	Validation cohort	0.809	0.762-0.857	0.776	0.653	0.810	0.492	0.892	0.561	0.132

Classification metrics were calculated using model-specific thresholds determined by the Youden index in the training cohort and then applied unchanged to the validation cohort. The Brier score was used to evaluate overall prediction error, with lower values indicating better performance. The paired DeLong test showed no statistically significant difference between the validation AUCs of XGBoost and LightGBM ($P=0.647$). Although LightGBM achieved a marginally higher validation AUC and a lower Brier score, XGBoost showed higher sensitivity, NPV, and F1 score and was therefore selected as the final model in accordance with the clinical objective of identifying high-risk women and minimizing missed high-risk cases. Random forest showed near-perfect performance in the training cohort but a marked decline in the validation cohort, suggesting potential overfitting. AUC, area under the receiver operating characteristic curve; CI, confidence interval; PPV, positive predictive value; NPV, negative predictive value; XGBoost, Extreme Gradient Boosting.

assigned to these variables because their SHAP distributions did not demonstrate a sufficiently clear monotonic pattern. These findings show that demographic, perioperative and early postoperative nursing assessment variables contributed to individualized risk predictions. SHAP values represent predictive contributions rather than causal effects. The dependence plot further showed that higher postoperative pain NRS scores at 2 h were generally associated with higher SHAP values and greater predicted risk (Fig. 3).

Discussion

In the present single-center retrospective cohort study, five prediction models were developed and internally validated for failure of first ambulation within 12 h after cesarean delivery using routinely collected perioperative nursing assessment data. The overall incidence of ambulation failure was 21.9%, indicating that ~1/5 women did not complete first ambulation within the early postoperative window. Among the five models, LightGBM achieved a marginally higher validation AUC and a lower Brier score, whereas XGBoost achieved higher sensitivity, negative predictive value and F1 score. The paired DeLong test showed no statistically significant difference between the validation AUCs of XGBoost and LightGBM, indicating comparable discrimination. Because the clinical purpose of the model was to identify women at high risk and minimize missed high-risk cases, XGBoost was selected as the final model.

Early ambulation is a core element of enhanced recovery pathways and is closely related to postoperative functional recovery. Although enhanced recovery programs were initially developed in general surgery, their principles have been increasingly adopted in obstetric care to improve maternal recovery, reduce unnecessary immobilization and promote earlier return to self-care and infant care (15). Previous studies have suggested that structured enhanced recovery pathways after cesarean delivery can shorten postoperative length of stay, improve pain control and increase maternal satisfaction when implemented with multidisciplinary support (16,17). However, early mobilization after cesarean delivery differs from mobilization after numerous other surgical procedures because women may experience residual neuraxial anesthesia effects, lower limb weakness or numbness, uterine contraction pain, incision pain, urinary catheter-related discomfort and orthostatic symptoms within the first postoperative hours. Therefore, predicting failure of early ambulation is clinically meaningful because it reflects the interaction between surgical recovery, anesthesia recovery, pain control, hemodynamic stability and bedside nursing assessment.

The incidence of failure of first ambulation in the present study was 21.9%, which appears clinically plausible for a 12-h postoperative endpoint. Previous enhanced recovery studies have reported improvements in early mobilization, but the exact timing and definition of ambulation vary considerably across institutions (18). Some studies define early ambulation as sitting out of bed, standing, walking a short distance or walking independently, whereas others consider mobilization as a broader process that includes turning, sitting, dangling, standing and walking (19). In the present study, the outcome was defined as failure to complete first ambulation within 12 h

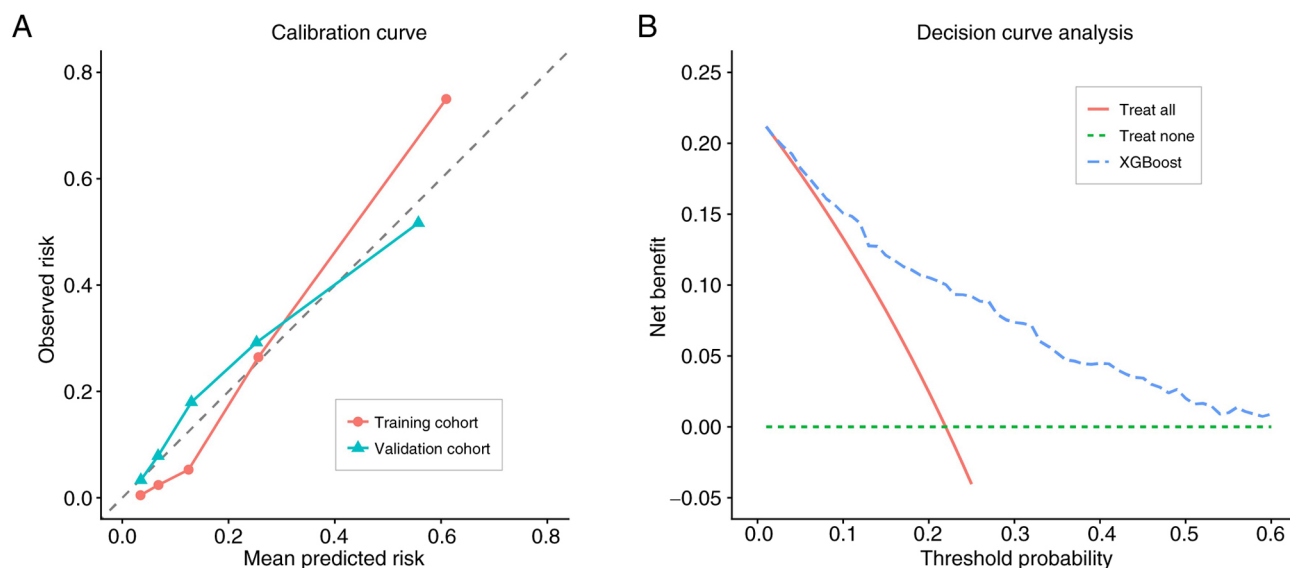


Figure 2. Calibration performance and decision curve analysis of the final XGBoost model. (A) Calibration curves of the XGBoost model in the training and internal validation cohorts. The dashed diagonal line represents perfect agreement between predicted and observed risks. The curves showed general agreement, with some deviation from ideal calibration at higher predicted probabilities in the internal validation cohort. (B) Decision curve analysis of the XGBoost model across the threshold probabilities displayed in the figure. The XGBoost model showed higher estimated net benefit than the treat-all and treat-none strategies across much of the evaluated range. These findings represent preliminary decision-analytic evidence within the source population and do not establish clinical utility or improved patient outcomes. XGBoost, Extreme Gradient Boosting.

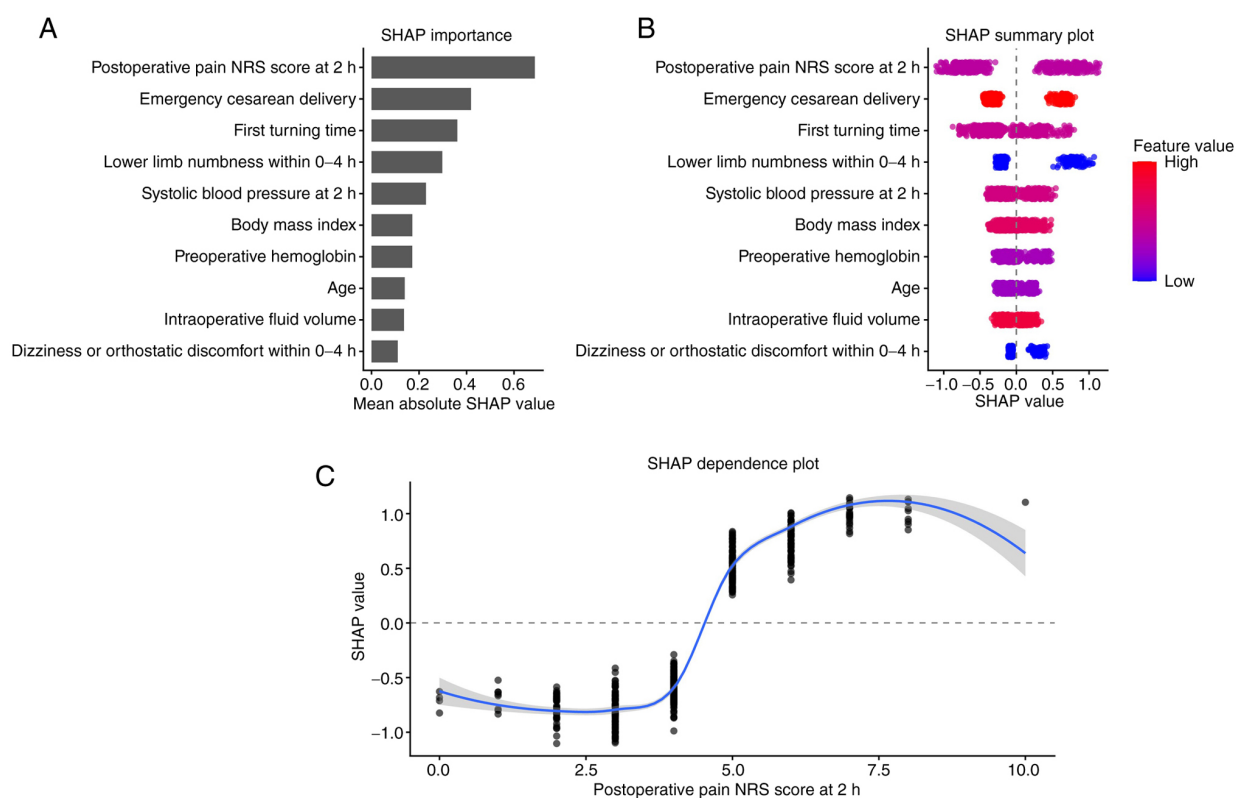


Figure 3. SHAP-based interpretation of the final XGBoost model. (A) SHAP importance plot showing the 10 highest-ranked predictors according to mean absolute SHAP value, in descending order: postoperative pain NRS score at 2 h, emergency cesarean delivery, time to first documented turning, lower limb numbness within 0–4 h, systolic blood pressure at 2 h, body mass index, preoperative hemoglobin, age, intraoperative fluid volume and dizziness or orthostatic discomfort within 0–4 h. (B) SHAP summary plot showing the distribution and direction of feature contributions to model predictions. Higher postoperative pain scores, emergency cesarean delivery, later time to first documented turning, lower limb numbness, higher body mass index, lower preoperative hemoglobin, lower systolic blood pressure and dizziness or orthostatic discomfort generally shifted predictions toward a higher estimated risk of failure of first ambulation. Age and intraoperative fluid volume contributed to model predictions, although their SHAP distributions did not indicate a simple monotonic direction of association. SHAP values represent predictive contributions and should not be interpreted as causal effects. In particular, time to first documented turning was a dynamically updated early recovery marker rather than an independent causal determinant of ambulation failure. (C) SHAP dependence plot for postoperative pain NRS score at 2 h, showing that higher pain scores were generally associated with higher SHAP values and greater predicted risk. NRS, numeric rating scale; SHAP, SHapley Additive exPlanations.

or discontinuation of an ambulation attempt because of documented intolerance. This definition is practical for nursing documentation because it captures both absence of ambulation and failed ambulation attempts. It also better reflects real-world bedside decision-making, where the reason for stopping ambulation is often as important as the timing itself.

Among all predictors, postoperative pain NRS score at 2 h was the most important contributor to the final XGBoost model. This finding is consistent with the clinical observation that inadequate pain control is one of the most direct barriers to turning, sitting, standing and walking after cesarean delivery. Acute post-cesarean pain is affected by incision pain, uterine involution pain, visceral discomfort, anxiety and individual pain sensitivity. Poorly controlled acute pain can limit respiratory effort, reduce willingness to move, delay breastfeeding positioning and impair maternal functional recovery (20). In addition, severe acute pain after childbirth has been associated with persistent pain and postpartum psychological morbidity, suggesting that early pain assessment should not be regarded only as a comfort indicator but also as a functional recovery marker (21). The present finding supports the importance of pain-guided mobilization planning. Women with high pain scores at 2 h may benefit from timely analgesic optimization, nonpharmacological comfort measures, abdominal support and stepwise assisted mobilization rather than routine encouragement alone.

Emergency cesarean delivery was the second most important predictor. Compared with elective cesarean delivery, emergency cesarean delivery is often associated with greater preoperative anxiety and psychological stress, reflecting the more urgent and less predictable clinical circumstances surrounding emergency delivery (22). These factors may contribute to postoperative fatigue, anxiety, pain perception, hemodynamic instability, and delayed readiness for mobilization. Previous obstetric anesthesia and recovery studies have shown that emergency cesarean delivery is associated with higher perioperative stress and less predictable postoperative recovery compared with scheduled procedures (23). In the present study, the high SHAP ranking of emergency cesarean delivery suggests that procedure context should be incorporated into nursing assessment. Women after emergency cesarean delivery may need more proactive postoperative education, reassurance, pain assessment and supervised mobilization.

First turning time was another influential predictor in the present XGBoost model. It was defined as the elapsed time from completion of surgery to the earliest documented spontaneous or assisted turning in bed and was treated as a dynamically updated indicator of early functional recovery. Turning in bed is an early nursing-observed activity that generally occurs before sitting, standing and ambulation in progressive postoperative mobilization pathways (24). However, delayed first turning may itself reflect pain, residual motor or sensory block, dizziness, fatigue, reduced activity tolerance or the need for additional nursing assistance. It may therefore lie on the recovery pathway leading to ambulation and should not be interpreted as an independent causal determinant of ambulation failure. Its SHAP importance indicates that it contributed to model predictions but does not demonstrate that delayed turning caused the outcome or that accelerating turning alone would reduce the risk of ambulation failure. Clinically, delayed

first turning may instead serve as an early warning signal prompting reassessment of pain, lower limb recovery, hemodynamic tolerance and readiness for progressive mobilization.

Lower limb numbness within 0-4 h was also notably associated with ambulation failure in the present study. This finding is clinically reasonable because most cesarean deliveries are performed under neuraxial anesthesia, and residual sensory or motor block may impair balance, confidence and safety during standing or walking. Although early ambulation is encouraged in enhanced recovery pathways, mobilization should occur only after adequate recovery of lower limb strength and sensation. Inappropriate ambulation before full recovery may increase the risk of falls or injury. Postoperative mobility assessment after neuraxial anesthesia therefore requires careful evaluation of lower limb sensation, muscle strength, dizziness and hemodynamic status (25). The importance of lower limb numbness in the present SHAP analysis highlights the need to distinguish between unwillingness to ambulate and temporary physiological inability to ambulate. For women with residual numbness, nursing care should focus on reassessment, fall prevention and delayed but supervised mobilization when safe.

Systolic blood pressure at 2 h and dizziness or orthostatic discomfort within 0-4 h were also important predictors in the present study, indicating that hemodynamic tolerance is a key determinant of successful ambulation. Orthostatic intolerance after surgery may manifest as dizziness, nausea, blurred vision, weakness, sweating or hypotension during sitting or standing. These symptoms may be related to neuraxial anesthesia, blood loss, fasting, fluid shifts, vasodilation, pain and limited oral intake. Previous perioperative studies have shown that orthostatic intolerance is a frequent barrier to early mobilization and may delay recovery even when surgical complications are absent (26,27). In the present study, lower systolic blood pressure and dizziness or orthostatic discomfort contributed to higher predicted risk, supporting the need for structured pre-ambulation assessment. Before first ambulation, nurses should evaluate blood pressure, symptoms during position change, lower limb strength, pain control and availability of assistance.

Body mass index and preoperative hemoglobin were also included among the top SHAP predictors in the present analysis. Higher body mass index may increase the physical difficulty of turning, standing and walking after abdominal surgery and may also be associated with longer operative time, higher anesthetic complexity and greater postoperative discomfort. Obesity has been associated with increased cesarean-related morbidity and delayed functional recovery in several obstetric studies (28). Lower preoperative hemoglobin may reflect reduced physiological reserve, anemia-related fatigue and lower tolerance to blood loss, all of which may affect early mobilization capacity. Postpartum anemia has been associated with fatigue, reduced quality of life, impaired cognition and depressive symptoms, which may indirectly influence willingness and ability to ambulate (29). These findings suggest that women with high BMI or anemia may require more individualized recovery planning, including early support, gradual mobilization, and close symptom monitoring.

A notable strength of the present study is that the model was developed using routinely available perioperative and nursing assessment data rather than specialized tests or

complex imaging. The 10 highest-ranked predictors included demographic characteristics, perioperative variables and early postoperative nursing observations, specifically postoperative pain NRS score at 2 h, emergency cesarean delivery, time to first documented turning, lower limb numbness within 0-4 h, systolic blood pressure at 2 h, body mass index, preoperative hemoglobin, age, intraoperative fluid volume and dizziness or orthostatic discomfort within 0-4 h. Of these variables, numerous are already recorded during routine perioperative care, which may facilitate future workflow integration. A risk estimate generated during the early postoperative recovery period could help nurses prioritize patients for reassessment, analgesic optimization, supervised mobilization, fall-risk prevention and individualized health education. However, the SHAP ranking reflects predictive importance and should not be interpreted as demonstrating that each variable is independently modifiable or causally related to ambulation failure.

The comparison among models also highlights the risk of optimism and overfitting in flexible prediction algorithms. The available sample size also warrants cautious interpretation. Although the training cohort contained 228 outcome events for 26 prespecified candidate predictors, categorical encoding may have increased the effective dimensionality, and the event count may still be insufficient to exclude instability in flexible machine-learning models. The use of cross-validation and constrained boosting parameters reduced but did not eliminate the risk of overfitting. Random forest achieved a markedly high apparent training AUC of 0.996 but showed a notable decline in the internal validation cohort, where its AUC was 0.800. This pattern demonstrates why apparent training performance cannot be interpreted as expected performance in new patients and highlights the importance of adequate sample size and internal validation in prediction-model development (30). The other machine learning models may also retain some degree of optimism despite their smaller apparent-to-validation performance gaps, because all models were evaluated using an internal split from the same single-center source population rather than an independent temporal or geographic cohort. Among the two leading models, LightGBM achieved a marginally higher validation AUC and a lower Brier score, whereas XGBoost achieved higher sensitivity, negative predictive value and F1 score. The paired DeLong test showed no statistically significant difference between their validation AUCs ($P=0.647$), indicating comparable discrimination. False-negative predictions could lead to missed opportunities for reassessment, analgesic optimization, fall-risk prevention, or supervised mobilization, whereas false-positive predictions could increase nursing workload or unnecessary assistance. Because the initial intended use is risk-sensitive early screening, sensitivity, negative predictive value and F1 score were prioritized after confirming comparable discrimination. Therefore, XGBoost was retained as the final model, while recognizing that its apparent performance and clinical impact require temporal and external validation.

Calibration and decision curve analysis provided additional information beyond discrimination. Discrimination alone is insufficient for evaluating a prediction model because a model with a high AUC may still produce poorly calibrated risk estimates (31). In the present study, the calibration curve

showed acceptable agreement between predicted and observed risks within the source population, and decision curve analysis indicated a higher estimated net benefit than the treat-all and treat-none strategies across a range of threshold probabilities. Nevertheless, these results were derived from single-center internal validation and cannot establish clinical utility or readiness for bedside decision-making. Decision curve analysis estimates potential net benefit under specified assumptions but does not demonstrate that use of the model improves nursing decisions or patient outcomes. Appropriate thresholds, transportability, workflow feasibility and clinical impact should therefore be evaluated prospectively in independent external populations before implementation is considered.

The use of SHAP analysis provided transparent descriptions of how recorded variables contributed to individual model predictions. Explainability may help clinicians inspect whether a prediction is consistent with the available clinical record, but it does not establish causal effects, independent determinants, modifiable treatment targets or improved patient outcomes. This distinction is particularly important for time-updated recovery indicators such as first turning time, which may reflect the combined effects of pain, residual anesthesia, activity tolerance and nursing assistance. In the present study, SHAP-ranked variables should therefore be interpreted as predictive markers of early postoperative recovery status. They may prompt reassessment or safety checks, but the model does not establish that changing pain, dizziness, turning time or any other highly ranked feature will reduce ambulation failure. Interpretable machine learning may facilitate model review and error detection, but challenges involving usability, clinician trust, workflow integration and prospective evaluation remain important before implementation (32).

Transparent reporting and structured risk-of-bias assessment are important for evaluating prediction models but do not by themselves establish clinical implementation readiness (13,33). Interpretive 'spin' may lead preliminary or internally validated findings to be perceived as demonstrating greater clinical benefit than the study design supports (34). Prospective registration can improve transparency in future validation and implementation research (35). Broader biomarker research further illustrates that the clinical meaning of a predictive marker depends on its measurement context, target population, reproducibility and intended decision pathway (36-38). Accordingly, the SHAP-ranked variables in the present study should be interpreted as predictive markers of early postoperative recovery status rather than independently causal or immediately actionable intervention targets.

The present study has several limitations. First, it was conducted in a single tertiary hospital, and local nursing protocols, anesthesia practices, postoperative mobilization routines, case mix and documentation habits may limit transportability to other settings. The random training-validation split assessed performance only within the same source population and does not constitute temporal or geographic external validation. Consequently, the model should not be regarded as clinically ready, and its performance requires evaluation in independent multicenter cohorts before implementation can be considered. Accordingly, the model applies only to relatively stable postoperative women considered clinically

eligible for early mobilization and should not be generalized to all women undergoing cesarean delivery. In particular, it should not be extrapolated to women transferred to intensive care, women with severe postpartum hemorrhage, women with explicit medical restrictions on ambulation or women with severe postoperative complications. Second, the retrospective design may have introduced selection and information bias because eligibility, predictors and outcomes depended on routine documentation. Although the final analytic dataset contained no item-level missing values, undocumented symptoms or ambulation attempts could still have resulted in measurement error or outcome misclassification, and exclusion of records with insufficient information may have introduced selection bias. Third, time to first documented turning was a dynamically updated recovery indicator rather than a predictor available at a uniform fixed postoperative time point. Its availability depended on when turning occurred and was documented, which should be considered when interpreting the model and designing prospective validation studies. Fourth, some potentially relevant variables were unavailable, including detailed neuraxial block regression time, exact analgesic dosing, breastfeeding attempts, patient motivation, family support, sleep quality and nurse-to-patient ratio. Fifth, decision curve analysis estimates potential net benefit under specified assumptions but does not demonstrate improved care or patient outcomes. Finally, SHAP analysis does not establish causality. Predictors such as pain, time to first documented turning and dizziness should be interpreted as predictive markers of recovery status rather than proven causal factors or independently modifiable intervention targets. Future research should define a uniform prediction time, undertake temporal and geographic external validation, assess calibration and decision thresholds across settings, and prospectively determine whether model-guided assessment improves nursing decisions or postoperative recovery.

In summary, the present study developed and internally validated an explainable XGBoost model for predicting failure of first ambulation within 12 h after cesarean delivery using routinely collected perioperative nursing assessment data. The model demonstrated good discrimination and acceptable calibration within this single-center retrospective cohort, and its performance was comparable to or better than that of the other evaluated models according to clinically relevant classification metrics. The 10 highest-ranked predictors according to mean absolute SHAP value were postoperative pain NRS score at 2 h, emergency cesarean delivery, time to first documented turning, lower limb numbness within 0-4 h, systolic blood pressure at 2 h, body mass index, preoperative hemoglobin, age, intraoperative fluid volume and dizziness or orthostatic discomfort within 0-4 h. Time to first documented turning should be interpreted as a dynamically updated marker of functional recovery rather than a causal determinant of ambulation failure, and SHAP rankings represent predictive contributions rather than causal effects. These findings indicate that explainable machine learning may offer a useful framework for early risk stratification and individualized postoperative mobilization planning after cesarean delivery. External and prospective validation are needed to determine its transportability and clinical impact before routine implementation.

Acknowledgements

Not applicable.

Funding

No funding was received.

Availability of data and materials

The data generated in the present study may be requested from the corresponding author.

Authors' contributions

LX, YJ and WL designed the study and collected the clinical data. JX performed data preprocessing and statistical analyses. HD supervised the study, interpreted the results and revised the manuscript critically for important intellectual content. LX and YJ confirm the authenticity of all the raw data. All authors read and approved the final manuscript.

Ethics approval and consent to participate

The present study was conducted in accordance with the principles of The Declaration of Helsinki and was approved by the Clinical Trial Ethics Committee of The Affiliated Hospital of Southwest Medical University (approval no. KY2026314). As a retrospective study using routinely collected medical records, the requirement for written informed consent was waived by the Ethics Committee. All patient data were anonymized prior to analysis to ensure confidentiality and privacy.

Patient consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

References

1. Mackeen AD, Sullivan MV, Bender W, Di Mascio D and Berghella V: Evidence-based cesarean delivery: Postoperative care (part 10). *Am J Obstet Gynecol MFM* 7: 101549, 2025.
2. Lestari MI, Sari D, Chandra S, Purwoko P, Isngadi I and Umar TP: Enhanced recovery after cesarean (ERAC) versus conventional care: An expanded systematic review and meta-analysis of 18,368 subjects. *J Anaesthesiol Clin Pharmacol* 41: 48-61, 2025.
3. Zhao ZM, Yang JQ, Wang GX, Yu XJ and Feng Y: A meta-analysis of the effects of multimodal intervention measures on the recovery of postpartum women after cesarean section. *Front Med (Lausanne)* 12: 1690139, 2025.
4. Hooda R, Malik N, Pathak P, More H and Singh V: Impact of postoperative pain on early initiation of breastfeeding and ambulation after cesarean section: A randomized trial. *Breastfeed Med* 18: 132-137, 2023.
5. Sultan P, Monks DT, Sharawi N, Bamber J, Panelli DM, Sauro KM, Shah PS, Muraca GM, Metcalfe A, Wood SL, *et al*: Guidelines for postoperative care in cesarean delivery: Enhanced recovery after surgery society recommendations (part 3)-2025 update. *Am J Obstet Gynecol* 233: S184-S198, 2026.

6. Saelmans A, Seinen T, Pera V, Markus AF, Fridgeirsson E, John LH, Schiphof-Godart L, Rijnbeek P, Reys J and Williams R: Implementation and updating of clinical prediction models: A systematic review. *Mayo Clin Proc Digit Health* 3: 100228, 2025.
7. Giaxi P, Vivilaki V, Sarella A, Harizopoulou V and Gourounti K: Artificial intelligence and machine learning: An updated systematic review of their role in obstetrics and midwifery. *Cureus* 17: e80394, 2025.
8. Vasudevan L, Kibria MG, Kucirka LM, Shieh K, Wei M, Masoumi S, Balasubramanian S, Victor A, Conklin JL, Gurcan MN, *et al*: Machine learning models to predict risk of maternal morbidity and mortality from electronic medical record data: Scoping review. *J Med Internet Res* 27: e68225, 2025.
9. Graseh JL, Rojas JC, Sharifi M, McLaughlin MM, Bhamidipalli SS and Haas DM: Impact of enhanced recovery after surgery pathway for cesarean delivery on postoperative pain. *AJOG Glob Rep* 3: 100169, 2023.
10. Ponce-Bobadilla AV, Schmitt V, Maier CS, Mensing S and Stodtmann S: Practical guide to SHAP analysis: Explaining supervised machine learning model predictions in drug development. *Clin Transl Sci* 17: e70056, 2024.
11. Wiens M, Verone-Boyle A, Henscheid N, Podichetty JT and Burton J: A tutorial and use case example of the eXtreme gradient boosting (XGBoost) artificial intelligence algorithm for drug development applications. *Clin Transl Sci* 18: e70172, 2025.
12. Deng L, Lu K and Hu H: An interpretable LightGBM model for predicting coronary heart disease: Enhancing clinical decision-making with machine learning. *PLoS One* 20: e0330377, 2025.
13. Collins GS, Moons KGM, Dhiman P, Riley RD, Beam AL, Van Calster B, Ghassemi M, Liu X, Reitsma JB, van Smeden M, *et al*: TRIPOD+AI statement: Updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ* 385: e078378, 2024.
14. Zhao L, Leng Y, Hu Y, Xiao J, Li Q, Liu C and Mao Y: Understanding decision curve analysis in clinical prediction model research. *Postgrad Med J* 100: 512-515, 2024.
15. Ciechanowicz S, De Greef E, Van de Velde M and Sultan P: Optimising recovery after caesarean delivery. *Best Pract Res Clin Anaesthesiol* 38: 199-208, 2024.
16. Walker M, Sobel M, Siddiqi N, Carvalho JCA, Jahan N, Santini S, Watts N, Dart K, Wang S, Huszti E and Thomas J: Implementation of enhanced recovery after surgery for caesarean delivery: A quality improvement initiative. *BMJ Open Qual* 14: e003391, 2025.
17. de Oliveira TTC, Tapioca V, Pereira E, Ferreira B, Soares T, Amaral S, Assis ML, Carvalho G, Curi JS and Kumar K: Enhanced recovery after surgery for caesarean delivery: A systematic review and meta-analysis of randomized trials with trial sequential analysis. *Braz J Anesthesiol* 76: 844817, 2026.
18. Lin L, Chen P, Wan T, Xie G, Wang P, Yan X, Li R, Chen W and Zhang W: Nurse-led evidence-based practice to promote gastrointestinal function recovery in women after cesarean delivery. *Int J Womens Health* 17: 4399-4408, 2025.
19. Arslan E and Özkan S: Effects of preoperative mobilization education using the teach-back method on patient outcomes after gynecological surgery: A randomized-controlled study. *Nurs Health Sci* 27: e70151, 2025.
20. Gamez BH and Habib AS: Predicting severity of acute pain after cesarean delivery: A narrative review. *Anesth Analg* 126: 1606-1614, 2018.
21. Yurashevich M, Cooter Wright M, Sims SC, Tan HS, Berger M, Ji RR and Habib AS: Inflammatory changes in the plasma and cerebrospinal fluid of patients with persistent pain and postpartum depression after elective cesarean delivery: An exploratory prospective cohort study. *Can J Anaesth* 70: 1917-1927, 2023.
22. Ferde YA, Bizuneh YB, Workie MM and Admass BA: 'Prevalence and associated factors of preoperative anxiety among obstetric patients who underwent cesarean section': A cross-sectional study. *Ann Med Surg (Lond)* 74: 103272, 2022.
23. Orovou E and Antoniou E: Voices of women with emergency cesarean section experience: A qualitative approach. *Cureus* 16: e53429, 2024.
24. Akkaya H and Ayhan H: The effect of in-bed turning and mobilization training given to patients who undergo lumbar disk surgery on postoperative first mobility level: A single-blind, randomized controlled trial. *Int J Orthop Trauma Nurs* 50: 101022, 2023.
25. Landau R and Sultan P: Neuraxial anesthesia and pain management for cesarean delivery. *Am J Obstet Gynecol* 233: S135-S152, 2026.
26. Yi H, Tang W, Shen Y, Tan L, Zeng F and Yang S: Orthostatic intolerance during early mobilization following thoracoscopic lung resection: A prospective observational study. *BMC Surg* 24: 265, 2024.
27. Hardy PYJP, Tavano A, Jacquet SV, Monseur JJ, Bastin ML, Kohonen LP, Haumann AE and Joris JL: The impact of orthostatic intolerance on early ambulation following abdominal surgery in an enhanced recovery programme. *Acta Anaesthesiol Scand* 66: 454-462, 2022.
28. Zimmo K, Ching-Johnson J, Jones PM, Singh SI, Dobrowlanski A, Symons YT, de Vrijer B, Sultan P, Carvalho B and Sebbag I: Quality of recovery after cesarean delivery in patients with class III obesity: A prospective observational cohort study. *Int J Obstet Anesth* 61: 104312, 2025.
29. Neef V, Choorapoikayil S, Hof L, Meybohm P and Zacharowski K: Current concepts in postpartum anemia management. *Curr Opin Anaesthesiol* 37: 234-238, 2024.
30. Martin GP, Riley RD, Ensor J and Grant SW: Statistical primer: Sample size considerations for developing and validating clinical prediction models. *Eur J Cardiothorac Surg* 67: ezaf142, 2025.
31. Cabot JH and Ross EG: Evaluating prediction model performance. *Surgery* 174: 723-726, 2023.
32. Abbas Q, Jeong W and Lee SW: Explainable AI in clinical decision support systems: A meta-analysis of methods, applications, and usability challenges. *Healthcare (Basel)* 13: 2154, 2025.
33. Wolff RF, Moons KGM, Riley RD, Whiting PF, Westwood M, Collins GS, Reitsma JB, Kleijnen J and Mallett S; PROBAST Group†: PROBAST: A tool to assess the risk of bias and applicability of prediction model studies. *Ann Intern Med* 170: 51-58, 2019.
34. Boutron I, Haneef R, Yavchitz A, Baron G, Novack J, Oransky I, Schwitzer G and Ravaud P: Three randomized controlled trials evaluating the impact of 'spin' in health news stories reporting studies of pharmacologic treatments on patients'/caregivers' interpretation of treatment benefit. *BMC Med* 17: 105, 2019.
35. Banno M, Tsujimoto Y and Kataoka Y: Studies registered in non-ClinicalTrials.gov accounted for an increasing proportion of protocol registrations in medical research. *J Clin Epidemiol* 116: 106-113, 2019.
36. Pizzimenti C, Fiorentino V, Ieni A, Rossi ED, Germanà E, Giovannella L, Lentini M, Alessi Y, Tuccari G, Campennì A, *et al*: BRAF-AXL-PD-L1 signaling axis as a possible biological marker for RAI treatment in the thyroid cancer ATA intermediate risk category. *Int J Mol Sci* 24: 10024, 2023.
37. Fiorentino V, Pizzimenti C, Franchina M, Pepe L, Russotto F, Tralongo P, Micali MG, Militi GB and Lentini M: Programmed cell death ligand 1 immunohistochemical expression and cutaneous melanoma: A controversial relationship. *Int J Mol Sci* 25: 676, 2024.
38. Germanà E, Pepe L, Pizzimenti C, Ballato M, Pierconti F, Tuccari G, Ieni A, Giuffrè G, Fadda G, Fiorentino V and Martini M: Programmed cell death ligand 1 (PD-L1) immunohistochemical expression in advanced urothelial bladder carcinoma: An updated review with clinical and pathological implications. *Int J Mol Sci* 25: 6750, 2024.



Copyright © 2026 Xiong *et al*. This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International (CC BY-NC-ND 4.0) License.