Chirurgia (2026) 121: 489-499
No. 5, September - October
Copyright© Celsius
http://dx.doi.org/10.21614/chirurgia.3379

Artificial Intelligence - Driven Risk Stratification in Emergency Surgery: Development and Validation of the PROMETHEUS-AI Model

Cătălin Cosma[1,2], Vlad Olimpiu Butiurca[1,2]*, Cosmin Nicolescu[1,2], Marian Botoncea[1,2]*, Călin Molnar[1,2]

[1]George Emil Palade University of Medicine, Pharmacy, Science and Technology of Târgu Mureș, Romania
[2]General Surgery, Emergency County Hospital of Târgu Mureș, Romania

***Corresponding author:**
Marian Botoncea, MD
George Emil Palade University of Medicine, Pharmacy, Science and Technology of Târgu Mureș, Romania
E-mail: marian.botoncea@umfsr.ro

Abbreviations:
ACS NSQIP: American College of Surgeons National Surgical Quality Improvement Program;
AI: Artificial Intelligence;
ASA: American Society of Anesthesiologists Physical Status Classification;
BMI: Body Mass Index;
CFS: Clinical Frailty Scale;
CRP: C-Reactive Protein;
EGS: Emergency General Surgery;
ICU: Intensive Care Unit;
NLR: Neutrophil-to-Lymphocyte Ratio;
POSSUM: Physiological and Operative Severity Score for the Enumeration of Mortality and Morbidity;
P-POSSUM: Portsmouth Physiological and Operative Severity Score for the Enumeration of Mortality and Morbidity;
PROMETHEUS-AI: Predictive Risk Optimization Model for Emergency and Threat Evaluation in Urgent Surgery;
PROBAST: Prediction Model Risk of Bias Assessment Tool;
ROC-AUC: Area Under the Receiver Operating Characteristic Curve;

Received: 10.06.2026
Accepted: 13.08.2026

Rezumat

Stratificarea riscului în chirurgia de urgență bazată pe inteligență artificială: dezvoltarea și validarea modelului PROMETHEUS-AI

Introducere: Chirurgia Generală de Urgență este asociată cu o morbiditate post-operatorie semnificativă, iar stratificarea precisă a riscului perioperator rămâne o provocare. Inteligența artificială şi tehnicile de învățare automată pot îmbunătăți predicția riscului prin integrarea simultană a multiple variabile demografice, clinice, biologice şi operatorii. Scopul acestui studiu a fost dezvoltarea şi validarea internă a PROMETHEUS-AI, un cadru predictiv bazat pe inteligență artificială pentru predicția complicațiilor postoperatorii în Chirurgia Generală de Urgență.
Materiale şi Metodă: A fost realizat un studiu observațional retrospectiv, unicentric, care a inclus 250 de pacienți adulți supuşi intervențiilor chirurgicale de urgență pentru apendicită acută, colecistită acută sau ocluzie intestinală non-oncologică în perioada 2020–2025. Variabilele demografice, clinice, biologice şi operatorii au fost integrate în modele de Logistic Regression, Random Forest şi Extreme Gradient Boosting (XGBoost). Validarea internă a fost realizată prin împărțirea lotului în seturi de antrenare şi validare în raport 80:20. Performanța modelelor a fost evaluată prin aria de sub curba ROC (ROC-AUC), acuratețe, sensibilitate, specificitate, precizie şi scor F1. Interpretabilitatea modelului a fost evaluată utilizând SHapley Additive exPlanations (SHAP).
Rezultate: Complicațiile postoperatorii au fost înregistrate la 48 de pacienți (19,2%), iar complicațiile postoperatorii majore (Clavien-Dindo grad IIIa-V) la 17 pacienți (6,8%). XGBoost a demonstrat cea mai bună performanță predictivă, cu un ROC-AUC de 0,88, o acuratețe de 85,0%, o sensibilitate de 83,0%, o specificitate de 86,0%, o precizie de 81,0% şi un scor F1 de 0,82. Random Forest a obținut un ROC-AUC de 0,84, iar Logistic Regression un ROC-AUC de 0,78. Analiza SHAP a identificat albumina serică, vârsta, clasificarea statusului fizic ASA, proteina C reactivă, lactatul seric şi raportul neutrofile/limfocite drept cei şase predictori cu cea mai mare importanță, urmați de durata intervenției chirurgicale, instabilitatea hemodinamică, sepsisul la internare şi Clinical Frailty Scale.
Concluzii: PROMETHEUS-AI a demonstrat o performanță promițătoare, validată intern, pentru stratificarea riscului de complicații postoperatorii în Chirurgia Generală de Urgență, XGBoost prezentând cea mai bună capacitate discriminatorie dintre algoritmii evaluați. Rezultatele susțin fezabilitatea utilizării învățării automate explicabile pentru evaluarea riscului perioperator. Totuşi, având în vedere

SHAP: SHapley Additive exPlanations; TRIPOD: Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis; XGBoost: Extreme Gradient Boosting.

designul retrospectiv unicentric, dimensiunea limitată a eşantionului şi absența validării externe, PROMETHEUS-AI trebuie considerat un model exploratoriu de tip proof-of-concept, care necesită validare prospectivă, multicentrică şi externă înainte de implementarea clinică.

Cuvinte-cheie: inteligență artificială, chirurgie de urgență, învățare automată, stratificarea riscului, complicații postoperatorii, XGBoost, chirurgie generală

Abstract

Background: Emergency General Surgery is associated with substantial postoperative morbidity, and accurate perioperative risk stratification remains challenging. Artificial intelligence and machine learning may improve risk prediction by integrating multiple demographic, clinical, laboratory, and operative variables. This study aimed to develop and internally validate PROMETHEUS-AI, an artificial intelligence–driven framework for predicting postoperative complications in Emergency General Surgery.

Methods: A retrospective single-center observational cohort study included 250 adult patients who underwent emergency surgery for acute appendicitis, acute cholecystitis, or non-oncological intestinal obstruction between 2020 and 2025. Demographic, clinical, laboratory, and operative variables were incorporated into Logistic Regression, Random Forest, and Extreme Gradient Boosting (XGBoost) models. Internal validation was performed using an 80:20 training–validation split. Model performance was assessed using receiver operating characteristic area under the curve (ROC-AUC), accuracy, sensitivity, specificity, precision, and F1-score. Model interpretability was evaluated using SHapley Additive exPlanations (SHAP).

Results: Postoperative complications occurred in 48 patients (19.2%), while major postoperative complications (Clavien-Dindo grade IIIa–V) occurred in 17 patients (6.8%). XGBoost demonstrated the highest predictive performance, with a ROC-AUC of 0.88, accuracy of 85.0%, sensitivity of 83.0%, specificity of 86.0%, precision of 81.0%, and an F1-score of 0.82. Random Forest achieved a ROC-AUC of 0.84, while Logistic Regression achieved a ROC-AUC of 0.78. SHAP analysis identified serum albumin, age, ASA physical status classification, C-reactive protein, serum lactate, and neutrophil-to-lymphocyte ratio as the six highest-ranked predictors, followed by operative duration, hemodynamic instability, sepsis on admission, and Clinical Frailty Scale.

Conclusions: PROMETHEUS-AI demonstrated promising internally validated performance for postoperative complication risk stratification in Emergency General Surgery, with XGBoost showing the highest discrimination among the evaluated algorithms. These findings support the feasibility of explainable machine learning for perioperative risk assessment. However, given the retrospective single-center design, limited sample size, and absence of external validation, PROMETHEUS-AI should be regarded as an exploratory proof-of-concept model requiring prospective multicenter and external validation before clinical implementation

Keywords: artificial intelligence, emergency surgery, machine learning, risk stratification, postoperative complications, XGBoost, predictive modeling

Introduction

Emergency General Surgery (EGS) includes an array of acute surgical disorders requiring urgent operative intervention and is one of the most demanding fields of modern surgical practice. However, accurate perioperative risk assessment continues to be an important part of emergency surgical management, especially in patients with physiological instability and multiple comorbidities. During the past three

decades, several risk stratification systems have been developed to estimate post-operative morbidity and mortality. The Physiological and Operative Severity Score for the Enumeration of Mortality and Morbidity (POSSUM) was one of the first and most comprehensive scoring systems to be developed for surgical audit and outcome prediction (1). Later, the Portsmouth modification (P-POSSUM) was proposed to improve the accuracy of mortality prediction (2). Further tools were added, such as the Surgical Apgar Score (3) and the American College of Surgeons National Surgical Quality Improvement Program (ACS NSQIP), Surgical Risk Calculator (4), expanding the options for perioperative risk assessment. Although these scoring systems are available, postoperative morbidity and mortality remain high in emergency surgical patients. High-risk surgical populations represent a disproportionate share of adverse postoperative outcomes and utilization of healthcare resources (5). Large national and international studies have shown considerable variation in mortality after emergency laparotomy, highlighting the need for better peri-operative risk stratification and clinical decision-making (6). Emergency General Surgery has been associated with significantly higher morbidity and mortality rates compared to elective surgical practice (7). In addition, EGS procedures impose a substantial healthcare burden and represent a significant proportion of national surgical activity, with marked differences in morbidity and mortality among individual emergency procedures (8). Patients who undergo urgent surgery have higher complication and mortality rates than those who undergo elective surgery (9). Population-based studies continue to show the large burden of emergency surgery on health care systems worldwide (10,11). The Lancet Commission on Global Surgery has highlighted the significance of surgery globally, and access to safe surgical care has been recognised as a crucial element in healthcare development (12). Postoperative complications are still a major cause of mortality worldwide. Recent evidence shows that death after surgery is one of the leading causes of preventable mortality worldwide (13,14). Other factors such as frailty have been shown to have a significant impact on postoperative outcomes and resource utilization in emergency surgical patients (15). Recent advances in artificial intelligence and machine learning have created new opportunities to improve perioperative risk prediction. Machine learning algorithms have the ability to analyze complex non-linear interactions between multiple variables and improve prediction performance when more data become available over time, unlike traditional statistical methods (16-20). The use of AI in surgery has grown rapidly and includes image analysis, clinical decision support systems, outcome prediction, and perioperative management (21-23). Machine learning techniques such as Random Forest (24), Gradient Boosting (25), and Extreme Gradient Boosting (XGBoost) (26) showed promising predictive performance in a number of medical specialties. Moreover, explainable AI techniques such as SHAP-based models have improved the transparency and interpretability of models, one of the major barriers to clinical implementation (27). This has resulted in further advances in the methodological quality of clinical prediction models, exemplified by the development of reporting and validation standards such as TRIPOD (28), PROBAST (29) and modern calibration frameworks (30). Although these scoring systems remain widely used, they rely predominantly on predefined statistical relationships and often demonstrate reduced predictive performance when applied to heterogeneous Emergency General Surgery populations. Recent advances in machine learning offer the opportunity to capture complex nonlinear interactions among demographic, clinical, laboratory, and operative variables, potentially improving individualized risk prediction. However, most currently available AI models remain disease-specific, lack explainability, or have not undergone adequate clinical validation, highlighting the need for transparent and reproducible predictive frameworks.

The present study aimed to develop and internally validate PROMETHEUS-AI (Predictive Risk Optimization Model for Emergency and Threat Evaluation in Urgent Surgery), an artificial intelligence-driven predictive model for postoperative risk stratification in patients undergoing emergency surgery for three frequently encountered conditions: acute appendicitis, acute cholecystitis, and non-oncological intestinal obstruction.

Material and Methods

Study Design and Setting

This retrospective observational cohort study was conducted at the General Surgery Clinic I, Emergency County Clinical Hospital Târgu Mureș, a tertiary referral center for emergency abdominal surgery in Romania. The study was designed as the pilot development phase of the PROMETHEUS-AI (Predictive Risk Optimization Model for Emergency and Threat Evaluation in Urgent Surgery) project, an artificial intelligence – driven framework developed to predict postoperative complications in Emergency General Surgery (EGS).

Adult patients undergoing emergency surgical treatment between January 2020 and December 2025 were retrospectively identified from the institutional electronic medical records and surgical databases. The study focused on three of the most frequently encountered emergency surgical pathologies: acute appendicitis, acute cholecystitis, and non-oncological intestinal obstruction.

The primary objective was to develop and internally validate a machine learning – based predictive model capable of identifying patients at increased risk of post-operative complications using routinely available demographic, clinical, laboratory, and operative variables. In addition to its clinical application, the PROMETHEUS-AI framework was designed to support perioperative resource allocation and improve decision-making in emergency surgical care.

Candidate predictors were prespecified before model training based on clinical relevance, prior literature, routine availability, and temporal availability at the time of prediction. No univariable significance testing or outcome-driven stepwise selection was used. All eligible predictors were entered into the candidate models, while algorithm-specific regularization and ensemble procedures determined their relative contribution.

A pilot development cohort consisting of 250 consecutive patients was used for model training and validation. The overall architecture of the PROMETHEUS-AI framework is presented in *Fig. 1.*

Patient Selection

Adult patients who underwent emergency surgical intervention at the General Surgery Clinic I, Emergency County Clinical Hospital Târgu Mureş, between January 2020 and December 2025 were retrospectively screened for eligibility. Patient identification was performed using institutional electronic medical records, operative registries, and discharge databases. The study focused on three common emergency surgical pathologies characterized by a high incidence and significant postoperative morbidity: acute appendicitis, acute cholecystitis, and non-oncological intestinal obstruction. Diagnoses were established based on a combination of clinical presentation, laboratory investigations, imaging findings, intraoperative observations, and histopathological examination when available.

Inclusion Criteria

Patients were eligible for inclusion if they met all of

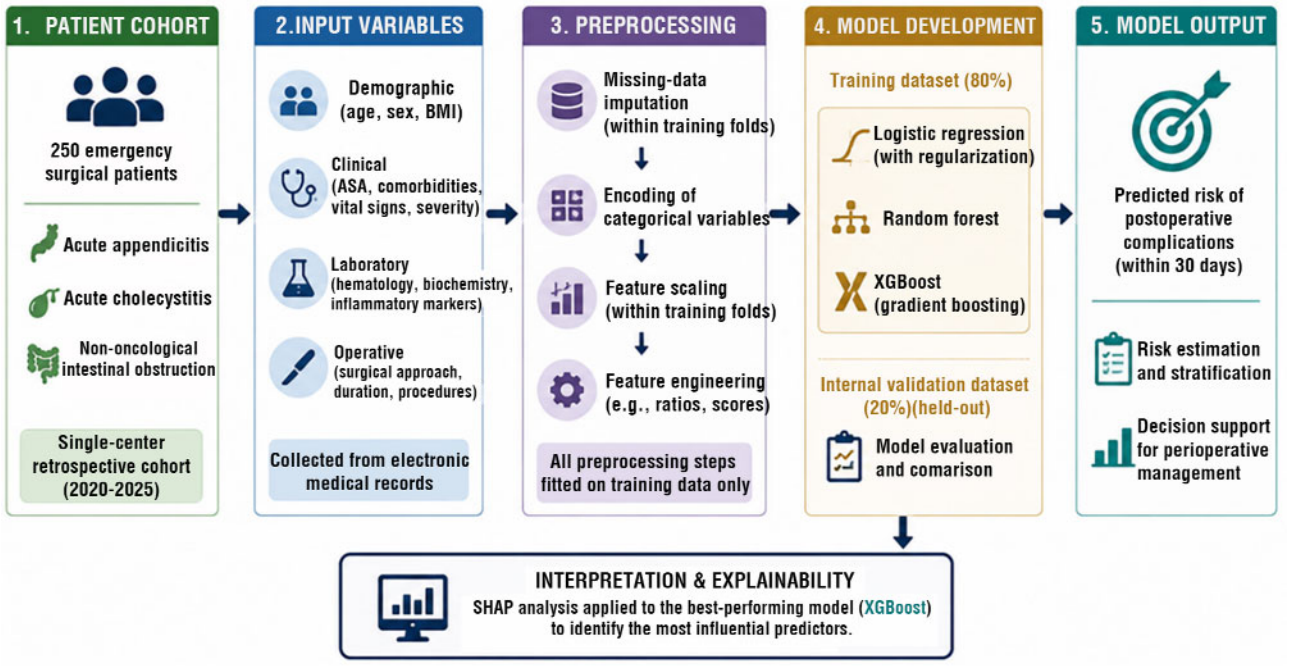

Figure 1. Overview of the PROMETHEUS-AI development workflow. Demographic, clinical, laboratory, and operative variables from 250 patients undergoing emergency surgery were processed using a reproducible machine-learning pipeline. Logistic Regression, Random Forest, and XGBoost models were developed using the training dataset and evaluated on an internally held-out validation dataset. SHAP analysis was applied to the best-performing model to identify the variables with the greatest influence on postoperative complication risk.

the following criteria:

- Age ≥18 years;
- Emergency surgical treatment performed during the study period;
- Primary diagnosis of acute appendicitis, acute cholecystitis, or non-oncological intestinal obstruction;
- Availability of complete perioperative clinical, laboratory, operative, and postoperative data;
- Minimum postoperative follow-up of 30 days or complete in-hospital outcome documentation.

Exclusion Criteria

Patients were excluded if they met one or more of the following criteria:

- Elective surgical procedures;
- Patients younger than 18 years of age;
- Malignant bowel obstruction or obstruction secondary to colorectal, gastric, gynecological, or other intra-abdominal malignancies;
- Non-operative management of the index pathology;
- Reinterventions performed for complications of a previous operation;
- Missing essential clinical, laboratory, operative, or outcome data;

For patients presenting with multiple hospital admissions during the study period, only the index emergency surgical episode was included in the final analysis to avoid data duplication and potential clustering effects. After application of the eligibility criteria, 250 consecutive patients were included in the pilot development cohort used for construction and internal validation of the PROMETHEUS-AI model. The final study population consisted exclusively of patients undergoing emergency surgical treatment for benign emergency abdominal pathology, thereby ensuring the development of a homogeneous predictive framework specifically tailored to Emergency General Surgery.

Data Collection and Variables

Clinical data were collected retrospectively from institutional electronic medical records, operative reports, anesthesia charts, laboratory databases, and discharge summaries using a standardized data collection protocol developed for the PROMETHEUS-AI project. All patient data were anonymized prior to analysis and entered into a dedicated research database. The PROMETHEUS-AI minimal core dataset was developed to include routinely available perioperative variables that can easily be integrated into real-world emergency surgical practice. Variables were selected based on clinical relevance, published evidence, availability in the institutional information system, and potential for use in machine learning-based prediction. Demographic variables were age, sex, body mass index (BMI) and date of admission. Clinical variables collected at hospital admission were the American Society of Anesthesiologists (ASA) Physical Status Classification, Clinical Frailty Scale (CFS), primary diagnosis, presence of sepsis on admission, hemodynamic instability, and the time interval between hospital admission and surgical intervention. The primary diagnosis was coded as acute appendicitis, acute cholecystitis or non-oncological intestinal obstruction. Laboratory investigations carried out preoperatively were hemoglobin, total leukocyte count, neutrophil count, lymphocyte count, C-reactive protein (CRP), serum lactate, serum albumin, and serum creatinine. In addition to absolute laboratory values, several derived variables were created through feature engineering procedures. The neutrophil-to-lymphocyte ratio (NLR) was calculated as an indicator of systemic inflammatory response. Nutritional status was assessed by serum albumin concentration, and hypoalbuminemia was defined as an albumin level < 3.5 g/dL. Preoperative anemia was defined as a hemoglobin concentration of less than 13 g/dL in males and less than 12 g/dL in females, according to World Health Organization criteria. Operative variables included surgical approach, operative duration, bowel resection, anastomosis formation, and stoma creation. Postoperative intensive care unit admission was recorded exclusively as a secondary outcome and was not included among the model predictors. These variables were selected as they represent operative complexity and have been linked to adverse postoperative outcomes in emergency surgical patients. Postoperative data included complications occurrence, Clavien-Dindo grade of complication, reoperation, length of hospital stay, intensive care unit utilization and 30-day mortality. All postoperative complications were graded by the Clavien–Dindo grading system and converted into binary outcome variables for machine learning analysis. The final dataset contained demographic, clinical, laboratory, operative, and outcome parameters, offering a comprehensive overview of the perioperative status of emergency surgical patients and aiding in the development of the PROMETHEUS-AI predictive framework.

Outcomes

The primary outcome of this study was the incidence of any postoperative complication within 30 days of

emergency surgical intervention. Postoperative complications were defined as any deviation from the expected postoperative course and were graded according to the Clavien-Dindo grading system. The primary endpoint was dichotomized for machine learning purposes with positive cases defined as patients with at least one postoperative complication and negative cases defined as patients with an uncomplicated postoperative course. Secondary outcomes were the occurrence of major postoperative complications (Clavien-Dindo grade IIIa–V), postoperative intensive care unit admission, reoperation, prolonged hospital stay, and 30-day mortality. Prolonged hospitalization was defined as length of stay > 10 days, indicative of increased healthcare resource utilization and postoperative recovery burden. Thirty-day mortality was defined as death during the index hospitalization or within 30 days of surgery from any cause. For exploratory analyses, we created additional outcome variables to assess the predictive performance of the PROMETHEUS-AI framework across different clinically relevant endpoints. These included major complications requiring surgical, endoscopic or radiological intervention, intensive care unit use and death. However, the main focus of the model development remained the prediction of any postoperative complication, as this end point provided the highest event rate and allowed more robust machine learning training in the pilot development cohort. All outcome variables were reviewed and confirmed independently with operative reports, discharge summaries, intensive care unit records, and follow-up notes when available. Outcome definitions were prespecified before model development to minimize information bias and maintain consistency throughout the analytical process.

Development of the PROMETHEUS-AI Model

PROMETHEUS-AI was developed as a supervised machine learning framework to predict postoperative complications of patients underwent emergency surgery for acute appendicitis, acute cholecystitis, and non-oncological intestinal obstruction. Before the model development all collected variables have been put through a standardized pre-processing workflow. Missing values, outliers, and data consistency of continuous variables were checked. Missing values of numerical variables were imputed by median based imputation. Missing values of categorical variables were imputed by most frequent category. Then, categorical features were one-hot encoded to add to machine learning algorithms. Where appropriate, continuous variables were normalized for optimal model performance. The neutrophil to lymphocyte ratio (NLR) was calculated as a marker of systemic inflammatory response, Other binary variables included hypoalbuminemia defined as serum albumin levels < 3.5 g/dL, and preoperative anemia according to World Health Organization criteria. These variables were kept in the final predictive dataset because of the known association of these variables with adverse postoperative outcomes and the potential for improvement in predictive accuracy. The primary end-point was the incidence of any postoperative complication within 30 days of surgery. We reformulated this endpoint as a binary classification problem, with positive cases defined as patients who developed at least one postoperative complication, and negative cases defined as patients who had an uneventful postoperative course. Finally, the dataset was randomly divided into training and validation cohorts at a ratio of 80:20 with stratified sampling to maintain the distribution of outcomes.All preprocessing operations were fitted exclusively on the training dataset and subsequently applied unchanged to the held-out validation dataset. Median values for numerical imputation, the most frequent categories for categorical imputation, one-hot encoding categories, and scaling parameters were therefore derived without access to validation data, thereby reducing the risk of information leakage. The proportion of missing observations was calculated for each predictor. The proportion of missing observations was assessed for each candidate predictor. No candidate predictor exceeded the prespecified missing-data threshold and therefore no variable was excluded on the basis of missingness. Any remaining missing numerical and categorical observations were handled within the training pipeline using median and most-frequent-category imputation, respectively. Model development was performed on training cohort and independent performance assessment was performed on validation cohort. During model development, three predictive algorithms were evaluated. We chose Logistic Regression as the reference statistical model because of its wide use in clinical prediction research and interpretability. The Random Forest was also chosen as an ensemble learning method that can capture the nonlinear interactions among variables while reducing overfitting through bootstrap aggregation. The main machine learning algorithm selected was Extreme Gradient Boosting (XGBoost) due to its demonstrated performance on structured clinical datasets and ability to model complex relationships between predictors and outcomes. The model was trained by iteratively optimizing the parameters of algorithm based on the training data set. After its construction, the predictive performance of the model

was tested in the validation cohort.The reported performance metrics are internally validated estimates from the PROMETHEUS-AI pilot development cohort. Subsequently, variable importance analyses were performed to determine the most influential factors contributing to postoperative complication risk. Explainable artificial intelligence techniques using SHapley Additive exPlanations (SHAP) were employed to quantify the individual contribution of each predictor to model output to increase transparency and interpretability. The developed PROMETHEUS-AI framework is not only envisioned as a predictive tool but also as a potential clinical decision-support system which can aid surgeons in perioperative risk stratification, patient counseling, resource allocation, and early identification of high-risk patients requiring intensified postoperative monitoring. The framework was developed using Python (Python Software Foundation, Wilmington, DE, USA) 3.11, an open-source programming language, which is widely used in machine learning and clinical predictive analytics. All the computational procedures were performed using the Google Colaboratory (Google Colab) cloud-based environment which provided the needed computational infrastructure for data preprocessing, model development, validation and explainability analyses. Data management and preprocessing were performed using Pandas 2.8 and Numpy 2.4 libraries. Implementation of missing value handling, feature transformation, dataset partitioning, model training and performance evaluation was done using scikit-learn machine learning framework 1.5.2. The Extreme Gradient Boosting algorithm was implemented using the XGBoost 2.4 library, and explainable artificial intelligence analyses were conducted utilizing the SHapley Additive exPlanations (SHAP) 0.47 framework. The machine learning workflow consisted of data cleaning, feature engineering, encoding of categorical variables, standardization of numerical variables, model training, internal validation and model interpretation. Predictor variables were automatically processed through dedicated preprocessing pipelines to guarantee reproducibility and consistency across all evaluated algorithms. Model performance was assessed by several complementary metrics including area under the receiver operating characteristic curve (ROC-AUC), accuracy, sensitivity, specificity, precision and F1-score. SHAP analyses and variable importance rankings were carried out to help interpretation of the model predictions and to identify the most important clinical, laboratory and operative variables associated with postoperative complications. The whole computational pipeline was developed following current guidelines for clinical prediction models based on machine learning, with emphasis on transparency, reproducibility and interpretability. This approach was designed to facilitate future external validation, multicenter implementation, and possible integration of the PROMETHEUS-AI framework into clinical decision-support systems for Emergency General Surgery.

Statistical Analysis

Statistical analyses were performed using Python and the Scikit-learn framework. Continuous variables were expressed as mean ± standard deviation or median with interquartile range, depending on data distribution, while categorical variables were presented as absolute frequencies and percentages. Comparisons between groups were performed using Student's t-test or Mann–Whitney U test for continuous variables and Chi-square or Fisher's exact test for categorical variables, as appropriate. Machine learning model performance was evaluated using area under the receiver operating characteristic curve (ROC-AUC), accuracy, sensitivity, specificity, precision, and F1-score. Internal validation was performed using an 80:20 training–validation split. A two-sided p-value <0.05 was considered statistically significant. Ninety-five percent confidence intervals for ROC-AUC, accuracy, sensitivity, specificity, precision, and F1-score were estimated using stratified bootstrap resampling of the held-out validation dataset with 2,000 iterations.

The study was conducted in accordance with the ethical principles of the Declaration of Helsinki and current national regulations governing biomedical research involving human subjects. Ethical approval for the study was obtained from the Institutional Ethics Committee of the Emergency County Clinical Hospital Târgu Mureş prior to data collection and analysis (Approval No. 29330/21.10.2025). Due to the retrospective observational design of the study and the use of anonymized clinical data, the requirement for individual informed consent was waived by the Ethics Committee. All patient information was de-identified before analysis, and data confidentiality was maintained throughout the study in accordance with applicable data protection regulations.

Results

A total of 250 consecutive patients undergoing emergency surgical treatment between January 2020 and December 2025 met the inclusion criteria and were included in the PROMETHEUS-AI pilot development cohort. The mean age of the study population

was 56.0 ± 21.4 years, with a slight predominance of male patients (52.0%). The mean body mass index was 26.9 kg/m². Acute appendicitis represented the most frequent diagnosis, accounting for 90 patients (36.0%), followed by acute cholecystitis in 80 patients (32.0%). Non-oncological intestinal obstruction accounted for the remaining cases and included adhesive small bowel obstruction, incarcerated or strangulated abdominal wall hernias, volvulus, gallstone ileus, internal hernias, and other benign mechanical causes of bowel obstruction. Regarding preoperative risk stratification, most patients were classified as ASA II or ASA III, reflecting the heterogeneous clinical severity typically encountered in Emergency General Surgery. The cohort incorporated a broad spectrum of demographic, clinical, laboratory, and operative characteristics, providing a representative pilot dataset for development and validation of the PROMETHEUS-AI predictive framework. Detailed baseline characteristics are summarized in *Table 1*.

The *Table 1* summarizes the demographic, clinical, laboratory, and operative characteristics of the PROMETHEUS-AI pilot development cohort. The study population included adult patients undergoing emergency surgical treatment for acute appendicitis, acute cholecystitis, and non-oncological intestinal obstruction between January 2020 and December 2025. Variables were selected based on their clinical relevance and potential contribution to postoperative risk prediction. Data are presented as mean ± standard deviation or number (percentage), as appropriate.

Postoperative complications occurred in 48 patients, corresponding to an overall complication rate of 19.2%. Major postoperative complications (Clavien-Dindo grade IIIa–V) were observed in 17 patients (6.8%). Intensive care unit admission was required in 8 patients (3.2%), while 8 patients (3.2%) underwent reoperation during the postoperative period. The mean postoperative length of hospital stay was 6.0 days. No 30-day mortality events were recorded within the pilot cohort. These findings demonstrate a clinically relevant burden of postoperative morbidity and support the need for accurate perioperative risk stratification tools capable of identifying patients at increased risk of adverse outcomes following emergency surgical intervention.

The predictive performance of the PROMETHEUS-AI framework was evaluated using Logistic Regression, Random Forest, and Extreme Gradient Boosting (XGBoost) algorithms. XGBoost demonstrated the highest overall predictive performance, achieving a ROC-AUC of 0.88, an accuracy of 85.0%, a sensitivity of 83.0%, a specificity of 86.0%, a precision of 81.0%, and an F1-score of 0.82. Random Forest achieved a ROC-AUC of 0.84, an accuracy of 81.0%, a sensitivity of 79.0%, a specificity of 82.0%, a precision of 75.0%, and an F1-score of 0.77. Logistic Regression, used as the reference model, achieved a ROC-AUC of 0.78, an accuracy of 75.0%, a sensitivity of 70.0%, a specificity of 77.0%, a precision of 65.0%, and an F1-score of 0.67.

Table 1. Baseline Characteristics of the Study Population (n = 250)

Variable	Value
Age, Years	56.0 ± 21.4
Sex, N (%)	
Male	130 (52.0)
Female	120 (48.0)
Body mass index (BMI), kg/m²	26.9 ± 4.3
ASA physical status classification, N (%)	
ASA I	54 (21.6)
ASA II	74 (29.6)
ASA III	67 (26.8)
ASA IV	55 (22.0)
Primary diagnosis, N (%)	
Acute appendicitis	90 (36.0)
Aacute cholecystitis	80 (32.0)
Adhesive small bowel obstruction	35 (14.0)
Incarcerated/strangulated inguinal hernia	15 (6.0)
Incarcerated/strangulated incisional hernia	12 (4.8)
Femoral hernia with obstruction	5 (2.0)
Volvulus	5 (2.0)
Gallstone ileus	3 (1.2)
Internal hernia	3 (1.2)
Other benign mechanical obstruction	2 (0.8)
Clinical variables	
Clinical frailty scale	3.7 ± 2.0
Sepsis on admission, N (%)	64 (25.6)
Hemodynamic instability, N (%)	30 (12.0)
Time from admission to surgery, H	17.4 ± 10.5
Laboratory variables	
Hemoglobin, G/DL	13.0 ± 1.8
Leukocyte count, x 10^9/l	14.8 ± 5.4
C-reactive protein, mg/l	125.5 ± 69.5
Lactate, mmol/l	3.1 ± 1.3
Albumin, g/dl	3.5 ± 0.5
Creatinine, mg/dl	1.44 ± 0.52
Operative variables, N (%)	
Laparoscopic approach	115 (46.0)
Open approach	135 (54.0)
Bowel resection performed	28 (11.2)
Anastomosis performed	66 (26.4)
Stoma creation	38 (15.2)

Abbreviations: ASA, American Society of Anesthesiologists; BMI, body mass index; CRP, C-reactive protein.

The superior performance of the ensemble-based machine learning algorithms suggests that nonlinear interactions among demographic, clinical, laboratory, and operative variables contribute substantially to postoperative risk prediction. Comparative model performance metrics are summarized in *Table 2*, while receiver operating characteristic curve analysis is presented in *Fig. 2*.

To improve model interpretability and facilitate clinical applicability, explainable artificial intelligence techniques based on SHapley Additive exPlanations (SHAP) were applied to the best-performing algorithm. SHAP-based feature importance analysis

Table 2. Clinical outcomes and predictive performance of the PROMETHEUS-AI models

A. Postoperative Outcomes (n = 250)

Outcome	Value
Any postoperative complication, N (%)	48 (19.2)
Major complication (Claven–Dindo IIIA or more), N (%)	17 (6.8)
ICU admission, N (%)	8 (3.2)
Reoperation, N (%)	8 (3.2)
Length of hospital stay, days	6.0 ± 3.2
30-day mortality, N (%)	0 (0.0)

B. Comparative Performance of the PROMETHEUS-AI Models

Performance metric	Logistic regression	Random forest	XGBoost
ROC-AUC (95% CI)	0.78 (0.63–0.93)	0.84 (0.71–0.97)	0.88 (0.74–1.00)
Accuracy (%)(95% CI)	75.0 (61.1–86.0)	81.0 (68.1–90.6)	85.0 (72.4–93.3)
Sensitivity (%)(95% CI)	70.0 (35.4–91.9)	79.0 (54.4–93.9)	83.0 (58.6–96.4)
Specificity (%)(95% CI)	77.0 (62.7–87.9)	82.0 (67.0–91.9)	86.0 (71.3–94.9)
Precision (%)(95% CI)	65.0 (35.0–87.2)	75.0 (50.9–91.3)	81.0 (56.5–94.3)
F1-Score (95% CI)	0.67 (0.48–0.86)	0.77 (0.59–0.89)	0.82 (0.65–0.92)

Data are presented as mean ± standard deviation or number (percentage), as appropriate.
Abbreviations: ICU, intensive care unit; ROC-AUC, area under the receiver operating characteristic curve.

identified serum albumin, age, ASA physical status classification, C-reactive protein, serum lactate, and neutrophil-to-lymphocyte ratio as the six highest-ranked predictors of postoperative complications, followed by operative duration, hemodynamic instability, sepsis on admission, and Clinical Frailty Scale. Conversely, younger age, preserved nutritional status, lower inflammatory burden, and less complex surgical procedures contributed to a reduced predicted risk. SHAP analysis provided both global and individual-level interpretability, allowing visualization of the relative contribution of each predictor to model output. This explainability framework enhances transparency and supports the potential integration of PROMETHEUS-AI into routine clinical decision-making. The ranking of the most influential variables is illustrated in *Fig. 3*.

Discussion

This paper reports on the development and internal validation of PROMETHEUS-AI, a machine learning framework for the prediction of postoperative complications in Emergency General Surgery. The model showed

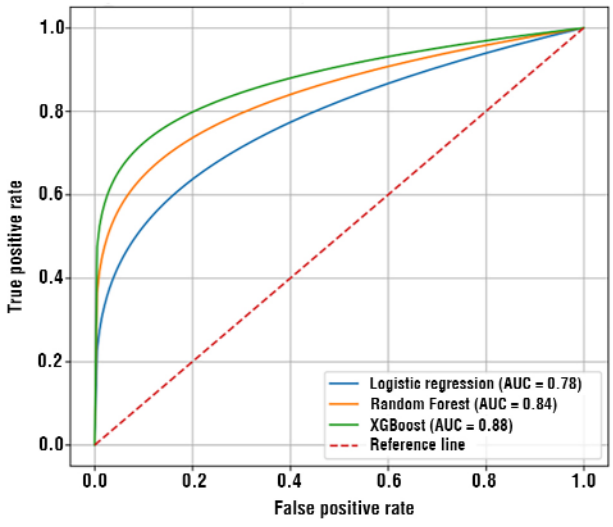

Figure 2. Receiver operating characteristic curves of the evaluated machine-learning models for postoperative complication prediction. XGBoost demonstrated the highest discrimination (ROC-AUC 0.88), followed by Random Forest (0.84) and Logistic Regression (0.78).

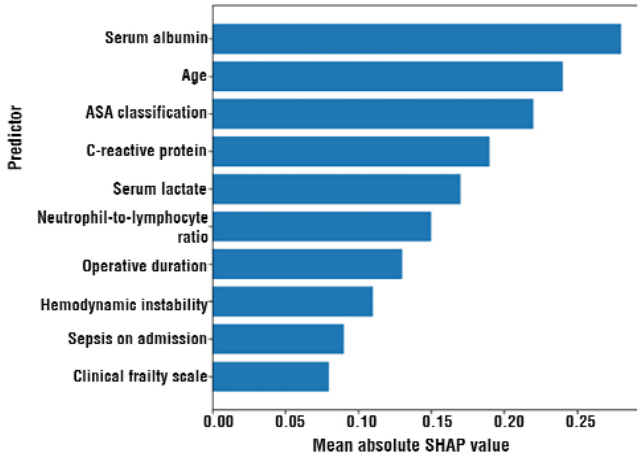

Figure 3. SHAP-based feature importance analysis of the PROMETHEUS-AI model. The figure presents the ten highest-ranked predictors according to their mean absolute SHAP values. Serum albumin, age, ASA physical status classification, C-reactive protein, serum lactate, and neutrophil-to-lymphocyte ratio were the six most influential predictors, followed by operative duration, hemodynamic instability, sepsis on admission, and Clinical Frailty Scale. Higher mean absolute SHAP values indicate a greater overall contribution to model predictions.

a promising predictive performance with routinely available demographic, clinical, laboratory and operative variables. XGBoost had the highest discriminative ability among the evaluated algorithms. These findings support the feasibility of the application of artificial intelligence techniques to the perioperative risk stratification in patients undergoing emergency surgery for acute appendicitis, acute cholecystitis and non oncological intestinal obstruction (16-23). Accurate risk prediction remains a key challenge in Emergency General Surgery Traditional scoring systems such as POSSUM, P-POSSUM, the Surgical Apgar Score and the ACS NSQIP Surgical Risk Calculator have enhanced peri-operative assessment but are largely based on pre-defined statistical relationships between variables (1-4). Conversely, machine learning algorithms have the capacity to identify complex non-linear interactions and continuously adapt to increasingly heterogeneous clinical datasets (18,19). The better performance of Random Forest and XGBoost in the current study is in line with previous reports highlighting the usefulness of artificial intelligence for surgical outcome prediction and perioperative decision support (23,31-38). The PROMETHEUS-AI framework has the important strength of integrating explainable artificial intelligence techniques. SHAP analysis revealed serum albumin, age, ASA classification, inflammatory biomarkers and operative complexity as the most important contributors for postoperative complication risk. These findings are biologically plausible and consistent with previous evidence showing the influence of nutritional status, frailty, systemic inflammation and operative burden on post-operative outcomes in emergency surgical patients (15,27,34,35). There are several limitations to the current study. First, the retrospective, single-center design may limit generalizability. Second, the pilot cohort included 250 patients, which constrained model complexity and external validity. Third, external validation was not performed and future multicenter studies will be required to evaluate the robustness of the model across different healthcare settings (39-43). Furthermore, no mortality prediction model could be developed because no 30-day mortality events occurred in the study cohort. Despite these limitations, PROMETHEUS-AI provides a practical proof-of-concept for the incorporation of artificial intelligence into Emergency General Surgery. Future work should involve larger multicenter datasets, prospective validation, incorporation of real-time electronic health record data, and use as a clinical decision-support tool. The methodological framework used in this study is aligned with the current recommendations for development, validation and reporting of machine learning based clinical prediction models such as TRIPOD-AI and DECIDE-AI guidance (44,45).

Conclusion

PROMETHEUS-AI demonstrated promising internally validated performance for predicting postoperative complications in Emergency General Surgery using routinely available demographic, clinical, laboratory, and operative variables. Among the evaluated algorithms, XGBoost achieved the highest discrimination, while SHAP analysis identified clinically plausible contributors to predicted risk. However, the retrospective single-center design, limited sample size, small number of outcome events, and absence of external validation restrict the generalizability of these findings. PROMETHEUS-AI should therefore be regarded as an exploratory proof-of-concept model rather than a tool ready for clinical use. External multicenter and prospective validation, further assessment of calibration across independent populations, and evaluation of clinical utility are required before implementation in routine emergency surgical practice.

Conflicts of Interest

The authors declare no conflicts of interest.

Ethics Approval

The institution's ethics committee approved the study 29330/21.10.2025.

References

1. Copeland GP, Jones D, Walters M. POSSUM: a scoring system for surgical audit. Br J Surg. 1991;78(3):355-360. doi:10.1002/bjs.1800780327.
2. Prytherch DR, Whiteley MS, Higgins B, Weaver PC, Prout WG, Powell SJ. POSSUM and Portsmouth POSSUM for predicting mortality. Br J Surg. 1998;85(9):1217-1220.
3. Gawande AA, Kwaan MR, Regenbogen SE, Lipsitz SA, Zinner MJ. An Apgar score for surgery. J Am Coll Surg. 2007;204(2):201-208.
4. Bilimoria KY, Liu Y, Paruch JL, Zhou L, Kmiecik TE, Ko CY, et al. Development and evaluation of the universal ACS NSQIP surgical risk calculator. J Am Coll Surg. 2013; 217(5):833-842.e3.
5. Pearse RM, Harrison DA, James P, Watson D, Hinds C, Rhodes A, et al. Identification and characterisation of the high-risk surgical population in the United Kingdom. Crit Care. 2006;10(3):R81.
6. Saunders DI, Murray D, Pichel AC, Varley S, Peden CJ. Variations in mortality after emergency laparotomy: the first report of the UK Emergency Laparotomy Network. Br J Anaesth. 2012;109(3):368-375.
7. Havens JM, Peetz AB, Do WS, Cooper Z, Kelly E, Askari R, et al. The excess morbidity and mortality of emergency general surgery. J Trauma Acute Care Surg. 2015; 78(2): 306-311.
8. Scott JW, Olufajo OA, Brat GA, Rose JA, Zogg CK, Haider AH, et al. Use of national burden to define operative emergency general surgery. JAMA Surg. 2016;151(6): e160480.
9. Mullen MG, Michaels AD, Mehaffey JH, Guidry CA, Turrentine FE, Hedrick TL, et al. Risk associated with complications and mortality after urgent surgery vs elective surgery. JAMA Surg. 2017;152(8):768-774.

10. Becher RD, Hoth JJ, Miller PR, Mowery NT, Chang MC, Meredith JW. A critical assessment of outcomes in emergency versus nonemergency general surgery using the American College of Surgeons National Surgical Quality Improvement Program database. Am Surg. 2011;77(7):951-959.
11. Kassahun WT, Babel J, Mehdorn M. Assessing differences in surgical outcomes following emergency abdominal exploration for complications of elective surgery and high-risk primary emergencies. Sci Rep. 2022;12(1):1349.
12. Meara JG, Leather AJM, Hagander L, Alkire BC, Alonso N, Ameh EA, et al. Global Surgery 2030: evidence and solutions for achieving health, welfare, and economic development. Lancet. 2015;386(9993):569-624.
13. Nepogodiev D, Martin J, Biccard B, Makupe A, Bhangu A. Global burden of post-operative death. Lancet. 2019;393(10170):401.
14. Ingraham AM, Cohen ME, Bilimoria KY, Ko CY, Hall BL, Russell TR, et al. Comparison of outcomes after emergency general surgery procedures: potential for targeted improvement. Surgery. 2010;148(2):217-238.
15. McIsaac DI, Moloo H, Bryson GL, van Walraven C. The Association of Frailty With Outcomes and Resource Use After Emergency General Surgery: A Population-Based Cohort Study. Anesth Analg. 2017;124(5):1653-1661.
16. Hashimoto DA, Rosman G, Rus D, Meireles OR. Artificial intelligence in surgery: promises and perils. Ann Surg. 2018;268(1):70-76.
17. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat Med. 2019;25(1):44-56.
18. Rajkomar A, Dean J, Kohane I. Machine learning in medicine. N Engl J Med. 2019; 380(14):1347-1358.
19. Esteva A, Robicquet A, Ramsundar B, Kuleshov V, DePristo M, Chou K, et al. A guide to deep learning in healthcare. Nat Med. 2019;25(1):24-29.
20. Jiang F, Jiang Y, Zhi H, Dong Y, Li H, Ma S, et al. Artificial intelligence in healthcare: past, present and future. Stroke Vasc Neurol. 2017;2(4):230-243.
21. Egert M, Steward JE, Sundaram CP. Machine learning and artificial intelligence in surgical fields. Indian J Surg Oncol. 2020;11(4):573-577.
22. Morris MX, Rajesh A, Asaad M, Hassan A, Saadoun R, Butler CE. Deep learning applications in surgery: current uses and future directions. Am Surg. 2023;89(1): 36-42.
23. Wang J, Tozzi F, Ganjouei AA, Romero-Hernandez F, Feng J, Calthorpe L, et al. Machine learning improves prediction of postoperative outcomes after gastrointestinal surgery: a systematic review and meta-analysis. J Gastrointest Surg. 2024;28(6):956-965.
24. Breiman L. Random forests. Mach Learn. 2001;45(1):5-32.
25. Friedman JH. Greedy function approximation: a gradient boosting machine. Ann Stat. 2001;29(5):1189-1232.
26. Chen T, Guestrin C. XGBoost: a scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: Association for Computing Machinery; 2016. p. 785-794.
27. Lundberg SM, Erion G, Chen H, DeGrave A, Prutkin JM, Nair B, et al. From local explanations to global understanding with explainable AI for trees. Nat Mach Intell. 2020;2(1):56-67.
28. Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis: the TRIPOD Statement. BMC Med. 2015;13:1.
29. Wolff RF, Moons KGM, Riley RD, Whiting PF, Westwood M, Collins GS, et al. PROBAST: a tool to assess the risk of bias and applicability of prediction model studies. Ann Intern Med. 2019;170(1):51-58.
30. Van Calster B, McLernon DJ, van Smeden M, Wynants L, Steyerberg EW. Calibration: the Achilles heel of predictive analytics. BMC Med. 2019;17(1):230.
31. Bertsimas D, Dunn J, Velmahos GC, Kaafarani HMA. Surgical risk is not linear: derivation and validation of a novel, user-friendly, and machine-learning-based Predictive OpTimal Trees in Emergency Surgery Risk (POTTER) calculator. Ann Surg. 2018; 268(4):574-583.
32. Gao J, Merchant AM. A machine learning approach in predicting mortality following emergency general surgery. Am Surg. 2021;87(9):1409-1415.
33. Chen B, Sheng W, Wu Z, Ma B, Cao N, Li X, et al. Machine learning based perisurgical risk calculator for abdominal related emergency general surgery: a multicenter retrospective study. Int J Surg. 2024;110(6):3527-3535.
34. Bihorac A, Ozrazgat-Baslanti T, Ebadi A, Motaei A, Madkour M, Pardalos PM, et al. MySurgeryRisk: development and validation of a machine-learning risk algorithm for major complications and death after surgery. Ann Surg. 2019;269(4):652-662.
35. Datta S, Loftus TJ, Ruppert MM, Giordano C, Upchurch GR Jr, Rashidi P, et al. Added value of intraoperative data for predicting postoperative complications: the MySurgeryRisk PostOp extension. J Surg Res. 2020;254:350-363.
36. Loftus TJ, Tighe PJ, Filiberto AC, Efron PA, Brakenridge SC, Mohr AM, et al. Artificial intelligence and surgical decision-making. JAMA Surg. 2020;155(2):148-158.
37. Loftus TJ, Upchurch GR Jr, Bihorac A. Building an Artificial Intelligence-Competent Surgical Workforce. JAMA Surg. 2021;156(6):511-512.
38. van de Sande D, van Genderen ME, Huiskens J, Gommers D, van Bommel J. Moving from bytes to bedside: a systematic review on the use of artificial intelligence in the intensive care unit. Intensive Care Med. 2021;47(7):750-760.
39. Damen JAAG, Hooft L, Schuit E, Debray TPA, Collins GS, Tzoulaki I, et al. Prediction models for cardiovascular disease risk in the general population: systematic review. BMJ. 2016;353:i2416.
40. Siontis GCM, Tzoulaki I, Castaldi PJ, Ioannidis JPA. External validation of new risk prediction models is infrequent and reveals worse prognostic discrimination. J Clin Epidemiol. 2015;68(1):25-34.
41. van Smeden M, Moons KGM, de Groot JAH, Collins GS, Altman DG, Eijkemans MJC, et al. Sample size for binary logistic prediction models: beyond events per variable criteria. Stat Methods Med Res. 2019;28(8):2455-2474.
42. Riley RD, Snell KIE, Ensor J, Burke DL, Harrell FE Jr, Moons KGM, et al. Minimum sample size for developing a multivariable prediction model: Part I - continuous outcomes. Stat Med. 2019;38(7):1262-1275.
43. Riley RD, Snell KIE, Ensor J, Burke DL, Harrell FE Jr, Moons KGM, et al. Minimum sample size for developing a multivariable prediction model: Part II — binary and time-to-event outcomes. Stat Med. 2019;38(7):1276-1296.
44. Collins GS, Moons KGM, Dhiman P, Riley RD, Beam AL, Van Calster B, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ. 2024;385:e078378.
45. Vasey B, Nagendran M, Campbell B, Clifton DA, Collins GS, Denaxas S, et al. Reporting guideline for the early-stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. Nat Med. 2022;28(5):924-933.