

Original Article

Construction and validation of a clinical prediction model for deep vein thrombosis in patients with digestive system tumors based on a machine learning

Yunfeng Zhang^{1*}, Yongqi Ma^{2*}, Jie Wang¹, Qiang Guan¹, Bo Yu³

¹Department of Vascular Surgery, Shanxi Provincial People's Hospital (The Fifth Clinical Medical School of Shanxi Medical University), No. 29 Shuangtasi Street, Taiyuan 030012, Shanxi, China; ²Shanxi University of Chinese Medicine, No. 121 Daxue Street, Yuci District, Jinzhong 030619, Shanxi, China; ³Department of Operating Room, Affiliated Hospital of Hebei University, No. 212 Yuhua East Road, Lianchi District, Baoding 071000, Hebei, China. *Equal contributors.

Received October 31, 2023; Accepted December 13, 2023; Epub January 15, 2024; Published January 30, 2024

Abstract: This study developed a deep vein thrombosis (DVT) risk prediction model based on multiple machine learning methods for patients with digestive system tumors undergoing surgical treatment. Data of 1048 patients with digestive system tumors admitted to Shanxi Provincial People's Hospital (College of Shanxi Medical University) from January 2020 to January 2023 were retrospectively analyzed, and 845 cases were screened according to the inclusion and exclusion criteria. The patients were divided into a training group (586 patients), and a validation group (259 patients), then feature selection was performed using six models, including Lasso regression, XGBoost, Random Forest, Decision Tree, Support Vector Machine, and Logistics. Predictive models were subsequently constructed from column-line plots, and the predictive validity of the models was assessed using receiver operating characteristic curves, precision-recall curves, and decision-curve analysis. In the model comparison, the XGBoost model showed the largest area under the curve (AUC) on the validation set ($P < 0.05$), demonstrating excellent predictive performance and generalization ability. We selected the common characteristic factors in the six models to further develop the column line plots to assess the DVT risk. The model performed well in clinical validation and effectively differentiated high-risk and low-risk patients. The differences in BMI, procedure time, and D-dimer were statistically significant between patients in the thrombus group and those in the non-thrombus group ($P < 0.05$). However, the AUC of the Xgboost model was found to be greater than that of the column chart model by the Delong test ($P < 0.05$). BMI, procedure time, and D-dimer are critical predictors of DVT risk in patients with digestive system tumors. Our model is an adequate assessment tool for DVT risk, which can help improve the prevention and treatment of DVT.

Keywords: Machine learning, Lasso regression, XGBoost, random forest, decision tree, support vector machine, digestive tumor, deep vein thrombosis, risk prediction

Introduction

Venous thromboembolism (VTE) is a serious vascular disease, with an incidence ranging from 0.75 to 2.69 per thousand, and the incidence of VTE disease is second only to that of acute myocardial infarction and cerebrovascular accident [1]. Virchow's triad theory explains that endothelial damage, altered hemodynamics, and blood hypercoagulability are the main factors leading to venous thrombosis [2]. The

interaction of these factors allows thrombus formation in deep veins, leading to partial or complete occlusion of vessels [3].

VTE has two main stages: deep vein thrombosis (DVT) and pulmonary thromboembolism (PE). DVT is the most common stage after surgery, especially in the lower extremities [4]. The incidence of lower extremity DVT is 10-40% after general surgery, 15-19% after abdominal surgery, and up to 40% in colorectal surgery

patients [5]. Alarming, only 50% of patients with DVT present with obvious signs and symptoms, such as lower extremity swelling, and deep and localized tenderness [6]. This makes the majority of initial DVTs undetectable, thereby complicating the diagnosis and treatment. Without timely diagnosis and intervention, the embolus can dislodge and travel with the vein to the lung, leading to fatal PE [7]. In patients underwent colorectal surgery, 5% of those with lower extremity DVT will develop PE [8]. A study in the United States identified VTE as one of the leading causes of prolonged hospitalization, increased mortality, and healthcare costs [9]. However, VTE is preventable, and evidence has suggested that appropriate prevention and treatment measures can reduce over 50% of the incidence of postoperative lower extremity DVT and 67% of the risk of PE [10, 11]. Therefore, it is important to identify the risk factors of DVT after gastrointestinal tumor surgeries and implement preventive and curative measures to improve patient recovery and ensure postoperative safety.

Although the main risk factors for DVT formation vary across specialties depending on the type of disease and individual differences, there are several common key risk factors in the same disease types, especially in patients with digestive system tumors [12]. These risk factors may vary with disease progression and perioperative stage, but their core elements are somewhat similar [13]. Therefore, this study aims to build a DVT risk prediction model for patients with digestive system tumors by integrating these common risk factors using machine learning. Such a model can help clinicians identify high-risk patients more accurately and take timely preventive measures, thereby improving patient outcomes.

Although there are studies focusing on the risk factors for DVT after surgery, most of them focused on patients undergoing surgery in general or specific types of surgery, such as patients undergoing knee replacement or hip replacement surgery. However, there are only a few studies on DVT risk prediction models for patients undergoing surgery for digestive system tumors, and even fewer ones used machine learning methods. In addition, most existing studies use traditional statistical methods rather than advanced machine learning techniques,

which limits the models' predictive ability and application range. The special characteristics of patients undergoing surgery for digestive system tumors, such as differences in tumor type, surgical approach, and perioperative management, make it possible that their DVT risk factors may be different from those of general surgical patients. Therefore, this study aims to fill this research gap by integrating these common risk factors and constructing a DVT risk prediction model for patients with digestive system tumors using machine learning methods. Such a model will help clinicians identify high-risk patients more accurately and provide individualized and precise guidance for DVT prevention and treatment.

Materials and methods

Sample size calculation

An observational study was planned to evaluate the incidence of postoperative DVT in patients with digestive system tumors. According to literature, the incidence of postoperative DVT in patients with digestive system diseases is 20%. The significance level was 0.05 for alpha and 0.95 for efficacy $1-\beta$. Based on the sample size estimation formula, 271 patients were needed when the expected absolute precision was 5%. Considering that some sample information may be missing, the planned sample size was increased by 10% (300 patients) to compensate for the missing information. During the study, the sample size may be adjusted if the interim assessment reveals that the actual incidence rate is significantly different from the expected rate or if the rate of missing sample information is higher than expected.

Ethics statement

The study was conducted with the approval of the Medical Ethics Committee of Shanxi Provincial People's Hospital (The Fifth Clinical Medical School of Shanxi Medical University).

Sample sources

Retrospectively, 1048 patients with digestive system tumors who were treated in Shanxi Provincial People's Hospital (The Fifth Clinical Medical School of Shanxi Medical University)

Machine learning-based clinical DVT prediction in gastrointestinal cancer

from January 2020 to January 2023 were collected as the subjects of this study.

Inclusion exclusion criteria

Inclusion criteria: 1. Patients who were treated at Shanxi Provincial People's Hospital (The Fifth Clinical Medical School of Shanxi Medical University) for gastrointestinal tumors, specifically gastric or colorectal cancer; 2. Patients with no history of prolonged bed rest, no history of DVT, and no history of using drugs that affect coagulation function within the three months prior to the surgery; 3. Patients with complete medical records; 4. Patients who underwent surgical treatment for their condition.

Exclusion criteria: 1. Patients with activity limitations due to trauma, fracture, or other diseases or those requiring long-term bed rest; 2. Patients with abnormal coagulation function; 3. Patients suffering from hematologic diseases; 4. Patients with severe liver and kidney function damage affecting coagulation function; 5. Patients with malignant diseases other than the specified gastrointestinal tumors; 6. Patients with lower limb DVT prior to surgery.

Diagnostic criteria for DVT

Patients were tested 14 days after the surgery following the diagnostic Criteria for Lower Extremity DVT [14]. Specifically, color Doppler ultrasonography reveals a solid mass of uneven echogenicity in the lower extremities, with diminished or absent color flow and spectral signals, and no collapse of the venous lumen after venous compression, and the veins are not compressible.

Sample screening and grouping

In this study, we collected 1,048 eligible samples according to the inclusion criteria, excluded 203 patients following the exclusion criteria, and finally included 845 eligible samples. Among them, 152 patients had DVT, and the patients were divided into a thrombotic group and a non-thrombotic group. In order to verify the generalizability of our model, we divided the 845 patients into a training group (n = 586) and a validation group (n = 259) according to a ratio of 7:3. There were 101 thrombotic patients and 485 non-thrombotic patients in the training group, and 51 thrombotic patients and 208 non-thrombotic patients in the validation group.

Clinical data collection

Relevant clinical data of the patients were collected, including age, gender, body mass index (BMI), history of hypertension, history of diabetes, history of coronary artery disease, history of smoking, history of alcohol consumption, type of tumor, duration of surgery, total cholesterol (TC), triglycerides (TG), albumin (Alb), total bilirubin (TB), creatinine (Cr), C-reactive protein (CRP), white blood cell count (WBC), Hemoglobin (Hb), Platelet count (PLT), Prothrombin time (PT), International Normalized Ratio (INR), D-dimer (DD) and the occurrence of complications at 14 d postoperatively (mainly including infection and hypoalbuminemia). All data except for complications were collected from patients one day before surgery. We have created a flowchart to make it easier for the reader to read (**Figure 1**).

Model construction

In this study, we used several machine learning methods to construct a risk prediction model for DVT in patients with digestive system tumors. First, feature selection and model optimization were performed by the Lasso regression model, whose performance was assessed by receiver operating characteristic curves and area under the curve (AUC) [15]. Subsequently, Random Forest and Support Vector Machine (SVM) models [16] were applied to further analyze the data, and both models were validated for their predictive ability by ROC curves and AUC values [17]. In addition, we also used Extreme Gradient Boosting (XGBoost) [18] and Decision Tree models [19], the feature importance of which was obtained and demonstrated through the corresponding functions. Finally, logistic regression models were utilized to construct nomograms [20] to provide clinicians an intuitive risk assessment tool. Each model was constructed with the aim of accurately identifying the key factors for DVT in order to improve clinical decision-making and patient outcomes.

Statistical analysis

The study began with data preprocessing and fundamental statistical analysis using SPSS 26.0 software. Count data were expressed as percentage (%) and subjected to a chi-square test. For measurement data, the distribution

Machine learning-based clinical DVT prediction in gastrointestinal cancer

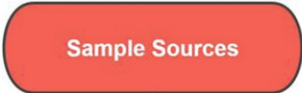


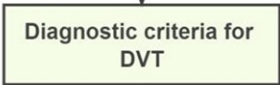
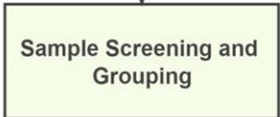
Screening process	Main content
	Retrospectively, 1048 patients with digestive system tumors who were treated in our hospital from January 2020 to January 2023 were collected as the subjects of this study.
	<ul style="list-style-type: none"> ① Patients treated at Shanxi Provincial People's Hospital (The Fifth Clinical Medical School of Shanxi Medical University) for gastrointestinal tumors, specifically gastric cancer or colorectal cancer. (n=984) ② Patients with no history of prolonged bed rest, no history of deep vein thrombosis (DVT), and no history of using drugs that affect coagulation function in the three months prior to surgery. (n=935) ③ Patients with complete medical records. (n=917) ④ Patients who underwent surgical treatment for their condition. (n=901)
	<ul style="list-style-type: none"> ① Patients with activity limitations due to trauma, fracture, other diseases, or those requiring long-term bed rest. (n=891) ② Patients with abnormal coagulation function. (n=880) ③ Patients suffering from hematologic diseases. (n=875) ④ Patients with serious liver and kidney function damage affecting coagulation function. (n=868) ⑤ Patients with malignant diseases other than the specified gastrointestinal tumors. (n=859) ⑥ Patients with lower limb DVT prior to surgery. (n=845)
	Diagnostic criteria for lower limb DVT[4]: Color Doppler ultrasound examination shows a solid mass with uneven echoes in the lower limbs, color blood flow signals and spectrum signals weaken or disappear, the lumen does not collapse after venous pressurization, and the veins cannot be compressed.
	A total of 203 patients were excluded according to the exclusion criteria, and a total of 845 samples were screened to meet the criteria. 152 of the 845 patients developed DVT, and the patients were divided into thrombotic and non-thrombotic groups according to the occurrence of DVT. In order to verify the generalizability of our model, we divided the 845 patients into a training group (n=586) and a validation group (n=259) according to the ratio of 7:3.

Figure 1. Flow chart of inclusion and exclusion criteria.

was first assessed using the Kolmogorov-Smirnov test. Data that conformed to a normal distribution were expressed as mean \pm standard deviation (Mean \pm SD) and analyzed by independent samples t-test. Data that did not conform to a normal distribution were expressed as quartiles P50 (P25, P75). In addition, we performed further statistical analyses using the R language (version 4.2.1) and utilized R packages, including rms (version 6.4.0), ResourceSelection (version 0.3-5), and part data. We employed functions from Table, ggplot2, and pROC for these analyses [21-25]. A binary logistic regression model was constructed via the glm function, and the rms

package was used to construct and visualize the nomogram model. We also used the part package to construct a decision tree model and the pROC package to plot ROC curves to evaluate the predictive performance of the models. Delong test was conducted to compare the AUCs between different models. $P < 0.05$ indicates a statistical difference.

Results

General patient information

In terms of clinical data, it was found that age, gender, BMI, history of coronary heart disease,

Machine learning-based clinical DVT prediction in gastrointestinal cancer

Table 1. Comparison of patients' general information

Items	Thrombosis group (n = 152)	Non-thrombotic group (n = 693)	$\chi^2/t/Z$ value	P-value
Age (years)	65.26±6.82	59 [54, 64]	76688.500	< 0.001
Gender				
Male	92	485	5.150	0.023
Female	60	208		
BMI (kg/m ²)	21.27±2.45	24.42±2.08	-14.691	< 0.001
History of hypertension				
Yes	38	180	0.062	0.804
No	114	513		
History of diabetes mellitus				
Yes	12	70	0.692	0.405
No	140	623		
History of coronary heart disease				
Yes	8	73	3.996	0.046
No	144	620		
History of smoking				
Yes	55	266	0.256	0.613
No	97	427		
History of alcohol consumption				
Yes	34	177	0.670	0.413
No	118	516		
Tumor type				
Gastric cancer	92	431	0.147	0.702
Colorectal cancer	60	262		
Duration of surgery (h)	4 [3, 5]	3 [2, 3]	79978	< 0.001
Postoperative complications				
Yes	51	69	56.962	< 0.001
No	101	624		
TC (mmol/L)	2.13±0.51	2.17±0.4	53886.500	0.655
TG (mmol/L)	31.63±6.69	29.75±5.13	-0.985	0.326
Alb (g/L)	14.99±4.21	14.65±3.69	3.267	0.001
TB (μmol/L)	85.39±9.85	82.69±10.72	0.917	0.360
Cr (μmol/L)	3.12±0.58	3.28±0.78	3.014	0.003
CRP (mg/L)	7.84±2.33	7.58±2.2	-2.873	0.004
WBC (10 ⁹ /L)	143.45±12.71	144.29±12.05	1.272	0.205
Hb (g/L)	179.32±21.24	184.28±21.28	-0.745	0.457
PLT (10 ⁹ /L)	2.13±0.51	2.17±0.4	-2.609	0.010
PT (s)	12 [11, 14]	13 [11, 15]	46166.500	0.016
INR	1.265 [1.02, 1.52]	1.31±0.3	47313.500	0.049
DD (mg/L)	3.85±1.14	1.18±0.4	28.617	< 0.001

Note: BMI, Body Mass Index; TC, Total Cholesterol; TG, Triglycerides; Alb, Albumin; TB, Total Bilirubin; Cr, Creatinine; CRP, C-Reactive Protein; WBC, White Blood Cell count; Hb, Hemoglobin; PLT, Platelet count; PT, Prothrombin Time; INR, International Normalized Ratio; DD, D-dimer.

duration of surgery, postoperative bedtime, postoperative complications, Alb, and DD were statistically different between the patients in

the thrombus group and the non-thrombus group (P < 0.05, **Table 1**). The two groups also differed in Cr, CRP, PLT, PT, INR, and DD (P <

0.05, **Table 1**). The rest of the general data were not statistically different ($P > 0.05$).

Machine learning models screening feature factors

To predict the occurrence of DVT in patients with digestive system tumors, samples in the training group were used to screen the feature factors through six machine learning models: Lasso, Xgboost, Random Forest, Decision Tree, SVM, and logistics regression. As a result, it was found that 8, 3, 23, 6, 23, and 4 eigenvectors were screened in the six models (**Figure 2A-F**).

Comparison of the predictive efficacy of machine learning models

ROC curves were plotted to evaluate the prediction effectiveness of the six models. The results showed that the AUCs of Lasso, Xgboost, Decision Tree, SVM and logistic were all greater than 0.97. Only the AUC of Random Forest model was 0.95 (**Tables 2, 3; Figure 3**). The Xgboost and SVM models showed significant predictive capability for the occurrence of DVT in patients with digestive system tumors.

Nomogram model construction and validation in patients with digestive system tumors

In this study, we constructed the nomogram model based on the common factors of the six machine learning models, including BMI, operation time and DD (**Figure 2G**), and found that no difference in the three feature factors between the training and validation sets ($P > 0.05$, **Table 4**). We then successfully constructed a nomogram model based on the training set data and these three factors (**Figure 4**). The risk formula was constructed based on the beta coefficient of the nomogram model: $= -2.370257263 + -3.221501831 * BMI + 2.30534465 * Duration\ of\ surgery + 7.397767547 * DD$. Risk scores were computed for each patient in the training and validation sets. By analyzing the ROC versus PR curves, it was found that the AUC for the training and validation sets in predicting DVT exceeded 0.99, indicating great predictive performance (**Figure 5A, 5B**). In addition, the decision-curve analysis (DCA) for both internal and external validation revealed that the predictive model exhibited a superior net clinical benefit

in predicting DVT probability across various threshold probabilities (training set: 0% to 99%; validation set: 1% to 98%), confirming its utility (**Figure 5C**). The calibration curves displayed C-index and AIC values of 0.993 (0.988-0.998) and 95.851 for the training set, and 0.983 (0.956-1.010) and 45.048 for the validation set, respectively. These results suggest a strong correlation between the actual and predicted probabilities of DVT, indicating good agreement between the two (**Figure 5D**).

Comparison of predictive efficacy of nomogram model with 6 machine learning models

At the end of the study, we compared the AUC of the Nomogram constructed based on the training group with the AUCs of the six machine learning models. It was found that there was no significant difference in the AUCs between Nomogram and Lasso, Decision tree, SVM, and logistics models ($P > 0.05$, **Table 5**). However, the AUC of the Xgboost model was greater than that of the Nomogram model, and the AUC of the Nomogram model was greater than that of the SVM model, with statistically significant differences ($P < 0.05$, **Table 5**).

Discussion

Lower extremity DVT dislodging and traveling through the circulatory system into the pulmonary artery is a significant trigger for fatal pulmonary embolism [26]. The characteristics of disease onset and progression in different specialties contribute to the differences in the incidence of lower extremity DVTs [27]. Currently, there is a lack of effective evidence-based research on the risk factors, clinical characteristics, and targeted preventive and therapeutic measures for the lower extremity DVT after gastrointestinal surgery.

Machine learning models are flexible enough to handle nonlinear and complex data structures and can also effectively deal with the challenges of high-dimensional data and missing values. They improve the accuracy of prediction and classification by training the models on large amounts of data and continuously optimizing their performance [28-32]. Another significant advantage of machine learning is its ability to recognize and exploit. In this study, we first constructed a prediction model for DVT in patients with digestive system tumors using six

Machine learning-based clinical DVT prediction in gastrointestinal cancer

Table 2. Receiver operating characteristic curve parameters of 6 machine learning models for predicting thrombus

Considerations	AUC	95% CI	Cut-off	Sensitivity	Specificity	Youden index
Lasso	0.994	0.985-1.000	0.24979	97.73%	98.02%	95.75%
Xgboost	1.000	0.999-1.000	0.48603	99.59%	99.01%	98.60%
Random Forest	0.945	0.914-0.975	0.5	99.79%	89.11%	88.90%
Decision tree	0.991	0.980-1.000	0.34318	98.56%	98.02%	96.58%
SVM	0.985	0.969-1.000	0.5	100.00%	97.03%	97.03%
Logistics	0.992	0.985-0.999	-1.765	95.46%	97.03%	92.49%

Note: AUC, area under the curve; Lasso, Least Absolute Shrinkage and Selection Operator; SVM, Support Vector Machine; XGBoost, eXtreme Gradient Boosting.

Table 3. Comparison of the AUCs of 6 machine learning models for predicting thrombus

The results of the test are very important for the	Z-value	P-value	AUC Difference	Standard error value	95% CI	
					Lower limit	Limit
Lasso - Xgboost	-1.247	0.212	-0.006	0.070	-0.014	0.003
Lasso - Random_Forest	3.410	0.001	0.050	0.142	0.021	0.078
Lasso - Decision_tree	0.494	0.621	0.003	0.100	-0.009	0.016
Lasso - SVM	0.972	0.331	0.009	0.114	-0.009	0.027
Lasso - logistics	0.703	0.482	0.002	0.089	-0.003	0.007
Xgboost - Random_Forest	3.554	0.000	0.055	0.126	0.025	0.085
Xgboost - Decision_tree	1.678	0.093	0.009	0.077	-0.001	0.019
Xgboost - SVM	1.743	0.081	0.014	0.094	-0.002	0.031
Xgboost - logistics	2.175	0.030	0.007	0.062	0.001	0.014
Random_Forest - Decision_tree	-3.118	0.002	-0.046	0.145	-0.076	-0.017
Random_Forest - SVM	-3.001	0.003	-0.041	0.155	-0.067	-0.014
Random_Forest - logistics	-3.352	0.001	-0.048	0.138	-0.076	-0.020
Decision_tree - SVM	0.800	0.424	0.006	0.118	-0.008	0.020
Decision_tree - logistics	-0.216	0.829	-0.001	0.095	-0.013	0.011
SVM - logistics	-0.794	0.427	-0.007	0.109	-0.025	0.010

Note: AUC, area under the curve; Lasso, Least Absolute Shrinkage and Selection Operator; SVM, Support Vector Machine; XGBoost, eXtreme Gradient Boosting.

machine-learning models. Lasso, Xgboost, random forest, decision tree, SVM, and logistic models found filtered out 8, 3, 23, 6, 23, and 4 feature factors, respectively. The ROC curves constructed from these eigenvectors showed that the AUC values of Lasso, Xgboost, decision tree, SVM, and logistics were all greater than 0.97, and only the AUC value of Random Forest was 0.95. These results suggest that the Lasso, Xgboost, and SVM models are high clinical value in predicting the occurrence of DVT in patients with gastrointestinal tumors. In contrast, in the study by Wang et al. [33], the AUC value of the DVT model constructed by logistic regression was only 0.780, with a sensitivity of 66.7% and a specificity of 77.7%. These comparative results highlight the superiority of the

machine learning model in this study, which provides a powerful tool for accurately predicting DVT in patients with gastrointestinal tumors, suggesting that machine learning techniques have high application potential in clinical research.

A column-line diagram is a graphical computational tool that visually represents the relationship between multiple variables and how they can be used to predict a particular outcome [34, 35]. In the medical field, it is commonly used to help physicians and researchers estimate disease risk or predict patient prognosis based on multiple clinical variables [36]. In this study, we employed six machine learning models to screen common factors and selected

Machine learning-based clinical DVT prediction in gastrointestinal cancer

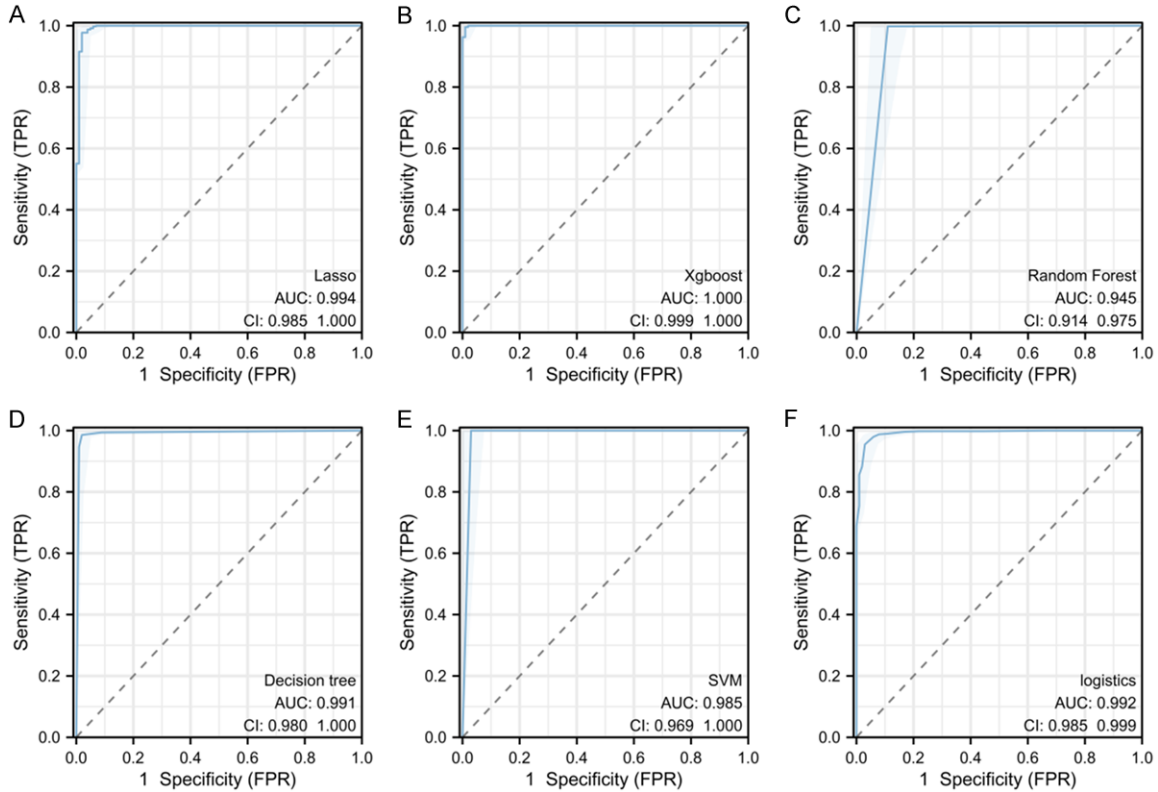


Figure 3. ROC curves of 6 machine learning models in predicting patient thrombosis. A. ROC curve of Lasso model for predicting the thrombosis. B. ROC curve of Xgboost model for predicting the thrombosis. C. ROC curve of Random forest model for predicting the thrombosis. D. ROC curve of Decision tree model for predicting the thrombosis. E. ROC curve of SVM model for predicting the thrombosis. F. ROC curve of Logistics model for predicting the thrombosis. Note: ROC, Receiver operating characteristic; Lasso, Least Absolute Shrinkage and Selection Operator; SVM, Support Vector Machine; XGBoost, eXtreme Gradient Boosting.

Table 4. Comparison of the 3 characteristic factors between the training and validation groups

Considerations	Training group (n = 586)	Validation group (n = 259)	Z-value	P-value
BMI (kg/m ²)	24.07 [22.24, 25.53]	24.14 [22.39, 25.36]	76962	0.743
Duration of surgery (h)	3 [3, 4]	3 [2, 4]	79898	0.190
DD (mg/L)	1.295 [0.9725, 1.68]	1.32 [1, 1.745]	73407.5	0.449

Note: BMI, Body Mass Index; DD, D-dimer.

three: BMI, surgery time, and DD. We reconstructed column-line plots based on the six characterized factors. We chose the column line graph model because of its interpretability, feature representation, and applicability advantages. Column line graphs visualize the relationship between critical features and predicted outcomes, simplifying the model's complexity and making it easy for non-specialists to understand and apply. These graphs highlight the key features influencing predicted outcomes, providing clinicians with a clear and concise prediction logic that helps the model to generalize in clinical practice.

In contrast, the six machine learning models exhibit a more complex structure with less explanatory power and involve a more technical validation and comparison process. The Nomogram model constructed a risk formula using beta coefficients and calculated risk scores for each patient in the training and validation sets. The results show that it has an extremely high predictive accuracy, with areas under the ROC and PR curves greater than 0.99. Meanwhile, internally and externally validated DCA and calibration curves confirm the model's consistency in net clinical benefit and predictive probability, demonstrating its high

Machine learning-based clinical DVT prediction in gastrointestinal cancer

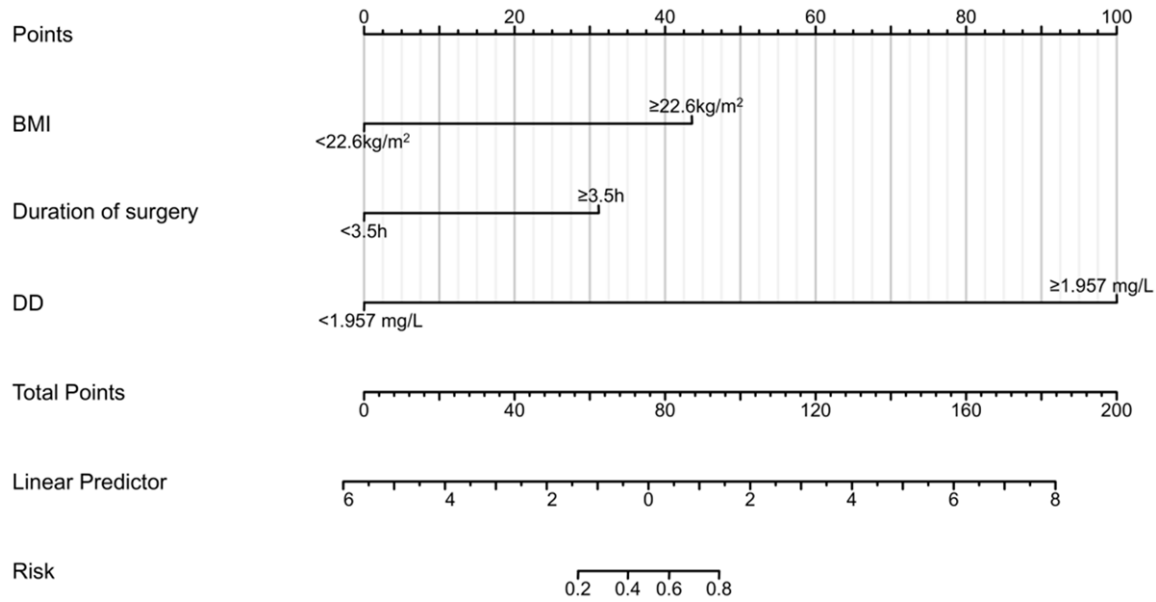


Figure 4. Nomogram model construction based on 3 characterization factors. Note: BMI, Body Mass Index; DD, D-dimer.

value in clinical prediction. Comparing the AUC values of the Nomogram with the six machine learning models, we found that the AUC of Xgboost was significantly higher than the Nomogram, proving that the Xgboost model has the same value in clinical prediction. Literature [37] has demonstrated that the high predictive value of column line graph prediction model for deep vein thrombotic fractures. In addition, Gao et al. [38] constructed a model using a column chart to predict early postoperative DVT in patients after open wedge high tibial osteotomy. In their study, the C index and Brier score of the histogram in the training cohort were 0.832 and 0.036, respectively, and the calibrated values after internal validation were 0.795 and 0.038, respectively. In addition, the ROC curves, calibration curves, the Hosmer-Lemeshow test, and the DCA suggested good performance in both the training and validation cohorts. The performance of our study on these evaluation metrics is similar to that of Gao et al., demonstrating the validity and reliability of our model. In this study, we also referenced other relevant studies to further confirm our findings and the validity of our model. For example, a study in a primary care setting developed a DVT prediction model that included factors such as D-dimer level, Wells score, gender, anticoagulant use, age, and familial venous thrombosis factors, showing

82% sensitivity and specificity [39]. Another study created a nomogram model for patients over 60 years of age with non-mild acute pancreatitis, which included factors such as age, gender, number of surgeries, and D-dimer, and achieved consistency indices of 0.827 and 0.803 in the training and validation sets, respectively [40]. In addition, for the risk of pulmonary embolism in patients with lower extremity DVT, another study developed a predictive model/scoring system based on seven risk factors, which performed well in calibration and discriminative ability [41]. These studies not only highlight the potential application of machine learning and statistical modeling in healthcare prediction but also provide clinicians with more accurate tools for assessing and managing patients' risk of DVT.

This study successfully applied multiple machine learning methods, especially synthesizing six different machine learning models, to identify the key features affecting the occurrence of DVT in patients with digestive system tumors and constructed a nomogram model accordingly. The model demonstrated high predictive accuracy and clinical value in both the training and validation sets, showing excellent interpretability and intuitive feature presentation, which provide clinicians a predictive tool that is easy to understand and apply. However,

Machine learning-based clinical DVT prediction in gastrointestinal cancer

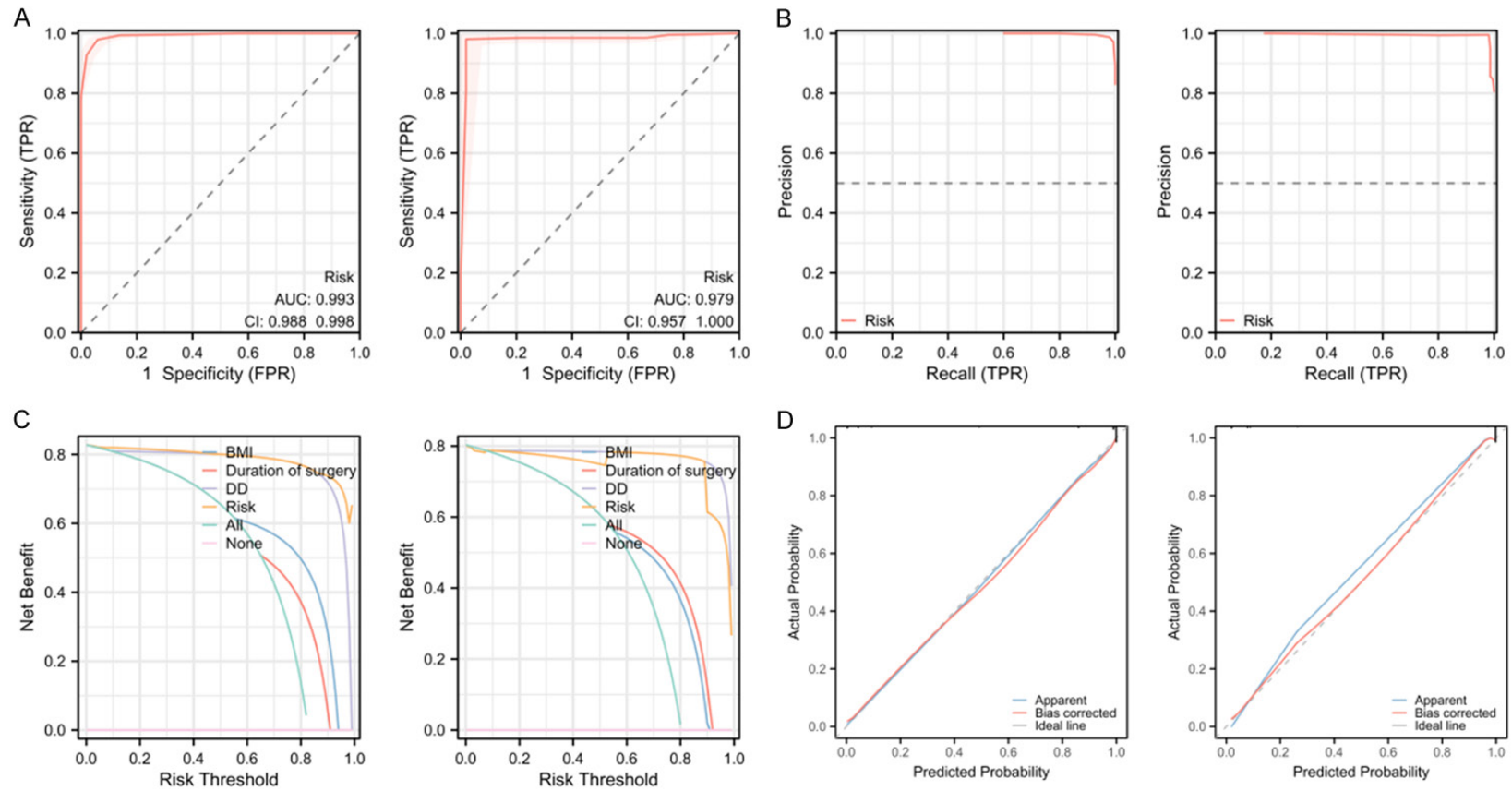


Figure 5. Internal and external validation of Nomogram model constructed based on 3 feature factors. A. ROC curve assessing the predictive efficacy of the validated training set and validation set models. B. PR curve assessing the precision and recall performance of the training set and validation set models under different thresholds. C. DCA assessing the benefits and losses under different thresholds. D. Diagnostic calibration curve assesses the predictive accuracy and reliability of the models. Note: ROC, Receiver Operating Characteristic; PR, Precision-Recall; DCA, Decision Curve Analysis.

Table 5. Comparison of the prediction effectiveness of the Nomogram model and the 6 machine learning models in the training group

The results of the tests were very positive for the	Z-value	P-value	AUC Difference	Standard error value	95% CI	
					Upper	Lower
Lasso - Nomogram	0.317	0.751	0.001	0.084	-0.006	0.008
Xgboost - Nomogram	2.653	0.008	0.007	0.055	0.002	0.012
Random Forest - Nomogram	-3.352	0.001	-0.048	0.135	-0.077	-0.02
Decision tree - Nomogram	-0.359	0.720	-0.002	0.090	-0.013	0.009
SVM - Nomogram	-0.917	0.359	-0.008	0.105	-0.024	0.009
logistics - Nomogram	-0.353	0.724	-0.001	0.078	-0.004	0.003

Note: AUC, Area under the curve; Lasso, Least Absolute Shrinkage and Selection Operator; SVM, Support Vector Machine; XGBoost, eXtreme Gradient Boosting.

there are some limitations in this study, including single-center data sources, selection bias due to retrospective design, and confounding factors that may have yet to be considered. Future research directions include multicenter validation, prospective studies, exploration of new data sources and features, and continuous iteration and optimization of the model to improve its generalizability, accuracy, and clinical applicability.

In summary, this study constructed a machine learning-based Nomogram model, which can accurately predict the risk of DVT in patients with digestive system tumors. Also, the identified critical features can provide a reference for understanding the pathogenesis of DVT.

Acknowledgements

Bethune Public Welfare Fund (XJ-2020-020); and Beijing Medical Award Fund (YXJL-2021-0353-0611).

Disclosure of conflict of interest

None.

Address correspondence to: Bo Yu, Department of Operating Room, Affiliated Hospital of Hebei University, No. 212 Yuhua East Road, Lianchi District, Baoding 071000, Hebei, China. E-mail: zss09287@163.com

References

- [1] Jackson CD, Cifu AS and Burroughs-Ray DC. Antithrombotic therapy for venous thromboembolism. *JAMA* 2022; 327: 2141-2142.
- [2] Verhamme P, Yi BA, Segers A, Salter J, Bloomfield D, Büller HR, Raskob GE and Weitz JI; ANT-005 TKA Investigators. Abrelcimab for prevention of venous thromboembolism. *N Engl J Med* 2021; 385: 609-617.
- [3] Weitz JI, Strony J, Ageno W, Gailani D, Hylek EM, Lassen MR, Mahaffey KW, Notani RS, Roberts R, Segers A and Raskob GE; AXIOMATIC-TKR Investigators. Milvexian for the prevention of venous thromboembolism. *N Engl J Med* 2021; 385: 2161-2172.
- [4] Galanaud JP, Trujillo-Santos J, Bikdeli B, Bertolotti L, Di Micco P, Poénu G, Falgá C, Zdravetska M, Lima J, Rivera-Civico F, Muixi JF and Monreal M; RIETE Investigators. Clinical presentation and outcomes of patients with cancer-associated isolated distal deep vein thrombosis. *J Clin Oncol* 2023; [Epub ahead of print].
- [5] Bikdeli B, Caraballo C, Trujillo-Santos J, Galanaud JP, di Micco P, Rosa V, Cusidó GV, Schellong S, Mellado M, Del Valle Morales M, Gavín-Sebastián O, Mazzolai L, Krumholz HM and Monreal M; RIETE Investigators. Clinical presentation and short- and long-term outcomes in patients with isolated distal deep vein thrombosis vs proximal deep vein thrombosis in the RIETE registry. *JAMA Cardiol* 2022; 7: 857-865.
- [6] Behrendt CA, Twerenbold R and Blankenberg S. The everlasting challenge to identify deep vein thrombosis in both clinical practice and research. *Eur Heart J* 2022; 43: 1882-1883.
- [7] Navarrete S, Solar C, Tapia R, Pereira J, Fuentes E and Palomo I. Pathophysiology of deep vein thrombosis. *Clin Exp Med* 2023; 23: 645-654.
- [8] Di Nisio M, van Es N and Büller HR. Deep vein thrombosis and pulmonary embolism. *Lancet* 2016; 388: 3060-3073.
- [9] Moss JL, Klok FA, Vo UG and Richards T. Controversies in the management of proximal deep vein thrombosis. *Med J Aust* 2023; 218: 61-64.
- [10] Schellong S, Ageno W, Casella IB, Chee KH, Schulman S, Singer DE, Desch M, Tang W, Voc-

Machine learning-based clinical DVT prediction in gastrointestinal cancer

- cia I, Zint K and Goldhaber SZ. Profile of patients with isolated distal deep vein thrombosis versus proximal deep vein thrombosis or pulmonary embolism: RE-COVERY DVT/PE study. *Semin Thromb Hemost* 2022; 48: 446-458.
- [11] Valeriani E, Di Nisio M, Porceddu E, Agostini F, Pola R, Spoto S, Donadini MP, Ageno W and Porfidia A. Anticoagulant treatment for upper extremity deep vein thrombosis: a systematic review and meta-analysis. *J Thromb Haemost* 2022; 20: 661-670.
- [12] He Y, Liu S and Su Y. Risk factors of deep vein thrombosis in children with osteomyelitis. *Ann Med* 2023; 55: 2249011.
- [13] Fujioka S, Ohkubo H, Kitamura T, Mishima T, Onishi Y, Tadokoro Y, Araki H, Matsushiro T, Yakuwa K, Miyamoto T, Torii S and Miyaji K. Risk factors for progression of distal deep vein thrombosis. *Circ J* 2020; 84: 1862-1865.
- [14] Tritschler T, Kraaijpoel N, Le Gal G and Wells PS. Venous thromboembolism: advances in diagnosis and treatment. *JAMA* 2018; 320: 1583-1594.
- [15] Lu J, Wang X, Sun K and Lan X. Chrom-Lasso: a lasso regression-based model to detect functional interactions using Hi-C data. *Brief Bioinform* 2021; 22: bbab181.
- [16] Mughal H, Bell EC, Mughal K, Derbyshire ER and Freundlich JS. Random forest model predictions afford dual-stage antimalarial agents. *ACS Infect Dis* 2022; 8: 1553-1562.
- [17] Youssef Ali Amer A. Global-local least-squares support vector machine (GLocal-LS-SVM). *PLoS One* 2023; 18: e0285131.
- [18] Li Y, Zou Z, Gao Z, Wang Y, Xiao M, Xu C, Jiang G, Wang H, Jin L, Wang J, Wang HZ, Guo S and Wu J. Prediction of lung cancer risk in Chinese population with genetic-environment factor using extreme gradient boosting. *Cancer Med* 2022; 11: 4469-4478.
- [19] Luo X, Ye L, Liu X, Wen X, Zhou M and Zhang Q. Interpretability diversity for decision-tree-initialized dendritic neuron model ensemble. *IEEE Trans Neural Netw Learn Syst* 2023; 6: 1-11.
- [20] Lv J, Liu YY, Jia YT, He JL, Dai GY, Guo P, Zhao ZL, Zhang YN and Li ZX. A nomogram model for predicting prognosis of obstructive colorectal cancer. *World J Surg Oncol* 2021; 19: 337.
- [21] Li G, Xu S, Yang S, Wu C, Zhang L and Wang H. An immune infiltration-related long non-coding RNAs signature predicts prognosis for hepatocellular carcinoma. *Front Genet* 2022; 13: 1029576.
- [22] Richards SM, Guo F, Zou H, Nigsch F, Baiges A, Pachori A, Zhang Y, Lens S, Pitts R, Finkel N, Loureiro J, Mongeon D, Ma S, Watkins M, Polus F, Albillos A, Tellez L, Martinez-González J, Bañares R, Turon F, Ferrusquía-Acosta J, Perez-Campuzano V, Magaz M, Fornis X, Badman M, Sailer AW, Ukomadu C, Hernández-Gea V and Garcia-Pagán JC. Non-invasive candidate protein signature predicts hepatic venous pressure gradient reduction in cirrhotic patients after sustained virologic response. *Liver Int* 2023; 43: 1984-1994.
- [23] Román Palacios C, Wright A and Uyeda J. treedata.table: a wrapper for data.table that enables fast manipulation of large phylogenetic trees matched to data. *PeerJ* 2021; 9: e12450.
- [24] Cao T, Li Q, Huang Y and Li A. plotnineSeq-Suite: a Python package for visualizing sequence data using ggplot2 style. *BMC Genomics* 2023; 24: 585.
- [25] Isaacs A and Lindenmann J. Pillars article: virus interference. I. The interferon. *Proc R Soc Lond B Biol Sci.* 1957. 147: 258-267. *J Immunol* 2015; 195: 1911-1920.
- [26] Barrosse-Antle ME, Patel KH, Kramer JA and Baston CM. Point-of-care ultrasound for bedside diagnosis of lower extremity DVT. *Chest* 2021; 160: 1853-1863.
- [27] Panpikoon T, Chuntaroj S, Treesit T, Chansanti O and Bua-Ngam C. Lower-extremity venous ultrasound in dvt-unlikely patients with positive D-Dimer test. *Acad Radiol* 2022; 29: 1058-1064.
- [28] Chen K, Shiomi A, Kagawa H, Hino H, Manabe S, Yamaoka Y, Kato S, Hanaoka M, Saito K, Maeda C, Kojima T, Shioi I, Nanishi K, Tanaka Y and Kasai S. Efficacy of a robotic stapler on symptomatic anastomotic leakage in robotic low anterior resection for rectal cancer. *Surg Today* 2022; 52: 120-128.
- [29] Mponponsuo K, Leal J, Spackman E, Somayaji R, Gregson D and Rennert-May E. Mathematical model of the cost-effectiveness of the BioFire FilmArray Blood Culture Identification (BCID) Panel molecular rapid diagnostic test compared with conventional methods for identification of *Escherichia coli* bloodstream infections. *J Antimicrob Chemother* 2022; 77: 507-516.
- [30] Johnson PM, Lin DJ, Zbontar J, Zitnick CL, Sriram A, Muckley M, Babb JS, Kline M, Ciavarra G, Alaia E, Samim M, Walter WR, Calderon L, Pock T, Sodickson DK, Recht MP and Knoll F. Deep learning reconstruction enables prospectively accelerated clinical knee MRI. *Radiology* 2023; 307: e220425.
- [31] Aromolaran O, Aromolaran D, Isewon I and Oyelade J. Machine learning approach to gene essentiality prediction: a review. *Brief Bioinform* 2021; 22: bbab128.
- [32] Garriga R, Mas J, Abraha S, Nolan J, Harrison O, Tadros G and Matic A. Machine learning

Machine learning-based clinical DVT prediction in gastrointestinal cancer

- model to predict mental health crises from electronic health records. *Nat Med* 2022; 28: 1240-1248.
- [33] Wang X, Jiang Z, Li Y, Gao K, Gao Y, He X, Zhou H and Zheng W. Prevalence of preoperative Deep Venous Thrombosis (DVT) following elderly intertrochanteric fractures and development of a risk prediction model. *BMC Musculoskelet Disord* 2022; 23: 417.
- [34] Zhang J, Ma F, Yao J, Hao B, Xu H, Guo X, Gao H and Yang T. Development and validation of a clinical prediction model for post thrombotic syndrome following anticoagulant therapy for acute deep venous thrombosis. *Thromb Res* 2022; 214: 68-75.
- [35] Li X, Wang Y and Xu J. Development of a machine learning-based risk prediction model for cerebral infarction and comparison with nomogram model. *J Affect Disord* 2022; 314: 341-348.
- [36] Du AX, Ali Z, Ajgeiy KK, Dalager MG, Dam TN, Egeberg A, Nissen CVS, Skov L, Thomsen SF, Emam S and Gniadecki R. Machine learning model for predicting outcomes of biologic therapy in psoriasis. *J Am Acad Dermatol* 2023; 88: 1364-1367.
- [37] Zhang L, He M, Jia W, Xie W, Song Y, Wang H, Peng J, Li Y, Wang Z and Lin Z. Analysis of high-risk factors for preoperative DVT in elderly patients with simple hip fractures and construction of a nomogram prediction model. *BMC Musculoskelet Disord* 2022; 23: 441.
- [38] Guo H, Wang T, Li C, Yu J, Zhu R, Wang M, Zhu Y and Wang J. Development and validation of a nomogram for predicting the risk of immediate postoperative deep vein thrombosis after open wedge high tibial osteotomy. *Knee Surg Sports Traumatol Arthrosc* 2023; 31: 4724-4734.
- [39] Shekarchian S, Notten P, Barbaty ME, Van Laanen J, Piao L, Nieman F, Razavi MK, Lao M, Mees B and Jalaie H. Development of a prediction model for deep vein thrombosis in a retrospective cohort of patients with suspected deep vein thrombosis in primary care. *J Vasc Surg Venous Lymphat Disord* 2022; 10: 1028-1036, e1023.
- [40] Yang DJ, Li M, Yue C, Hu WM and Lu HM. Development and validation of a prediction model for deep vein thrombosis in older non-mild acute pancreatitis patients. *World J Gastrointest Surg* 2021; 13: 1258-1266.
- [41] Zhao B, Hao B, Xu H, Premaratne S, Zhang J, Jiao L, Zhang W, Wang S, Su X, Sun L, Yao J, Yu Y and Yang T. Predictive model for pulmonary embolism in patients with deep vein thrombosis. *Ann Vasc Surg* 2020; 66: 334-343.