



Development and validation of machine learning models for postoperative venous thromboembolism prediction in colorectal cancer inpatients: a retrospective study

Li Qin[#], Zhikun Liang^{#^}, Jingwen Xie, Guozeng Ye, Pengcheng Guan, Yaoyao Huang, Xiaoyan Li

Department of Pharmacy, the Sixth Affiliated Hospital, Sun Yat-sen University, Guangzhou, China

Contributions: (I) Conception and design: L Qin, Z Liang; (II) Administrative support: X Li; (III) Provision of study materials or patients: J Xie, Y Huang; (IV) Collection and assembly of data: G Ye, P Guan; (V) Data analysis and interpretation: Z Liang; (VI) Manuscript writing: All authors; (VII) Final approval of manuscript: All authors.

[#]These authors contributed equally to this work.

Correspondence to: Dr. Xiaoyan Li. Department of Pharmacy, the Sixth Affiliated Hospital, Sun Yat-sen University, 26 Erheng Road of Yuan Village, Tianhe District, Guangzhou 510655, China. Email: lixian5@mail.sysu.edu.cn.

Background: Colorectal cancer (CRC) is a heterogeneous group of malignancies distinguished by distinct clinical features. The association of these features with venous thromboembolism (VTE) is yet to be clarified. Machine learning (ML) models are well suited to improve VTE prediction in CRC due to their ability to receive the characteristics of a large number of features and understand the dataset to obtain implicit correlations.

Methods: Data were extracted from 4,914 patients with colorectal cancer between August 2019 and August 2022, and 1,191 patients who underwent surgery on the primary tumor site with curative intent were included. The variables analyzed included patient-level factors, cancer-level factors, and laboratory test results. Model training was conducted on 30% of the dataset using a ten-fold cross-validation method and model validation was performed using the total dataset. The primary outcome was VTE occurrence in postoperative 30 days. Six ML algorithms, including logistic regression (LR), random forest (RF), extreme gradient boosting (XGBoost), weighted support vector machine (SVM), a multilayer perception (MLP) network, and a long short-term memory (LSTM) network, were applied for model fitting. The model evaluation was based on six indicators, including receiver operating characteristic curve-area under the curve (ROC-AUC), sensitivity (SEN), specificity (SPE), positive predictive value (PPV), negative predictive value (NPV), and Brier score. Two previous VTE models (Caprini and Khorana) were used as the benchmarks.

Results: The incidence of postoperative VTE was 10.8%. The top ten significant predictors included lymph node metastasis, C-reactive protein, tumor grade, anemia, primary tumor location, sex, age, D-dimer level, thrombin time, and tumor stage. In our results, the XGBoost model showed the best performance, with a ROC-AUC of 0.990, a SEN of 96.9%, a SPE of 96.1% in training dataset and a ROC-AUC of 0.908, a SEN of 77.5%, a SPE of 93.7% in validation dataset. All ML models outperformed the previously developed models (Caprini and Khorana).

Conclusions: This study developed postoperative VTE predictive models using six ML algorithms. The XGBoost VTE model might supply a complementary tool for clinical VTE prophylaxis decision-making and the proposed risk factors could shed some light on VTE risk stratification in CRC patients.

Keywords: Surgical colorectal cancer patient; venous thromboembolism (VTE); machine learning model

[^] ORCID: 0000-0002-4417-6011.

Submitted Dec 12, 2022. Accepted for publication Feb 02, 2023. Published online Feb 15, 2023.

doi: 10.21037/jgo-23-18

View this article at: <https://dx.doi.org/10.21037/jgo-23-18>

Introduction

Venous thromboembolism (VTE), consisting of deep vein thrombosis (DVT) and pulmonary embolism (PE), is a common complication after surgery in cancer patients (1). The impact of specific cancers on venous thromboembolism has been studied for years (2). The risk of VTE occurrence in cancer patients varies greatly due to differences in tumor site, therapies, or other risk factors (3). Thus, the clinical benefit of VTE prophylaxis for cancer patients depends mainly on accurate and individual-appropriate predictive Models (4).

Colorectal cancer (CRC) is the third most common cancer worldwide (5,6). Current VTE risk screening guidelines for CRC patients, which are developed from incomplete CRC patient cohorts and are extrapolated from data of other cancer types, have low sensitivity and specificity. The Khorana score was initially developed using multivariate logistic regression method in ambulatory cancer patients and was further validated in hospitalized cancer patients (7). The score is based on five parameters: site of the cancer, obesity, platelet count, hemoglobin, and white blood cell count; colorectal cancer is scored as 0 for

‘site of cancer’. Several validation studies of the Khorana model in patients with gastrointestinal cancer have shown conflicting results (8). Due to the limitations that some potential laboratory biomarkers (such as D-dimer) were not involved as predictors, the receiver operating characteristic curve-area under the curve (ROC-AUC) values of the Khorana model were previously in a range of 0.5–0.7, and a value of over 0.8 is expected (9). D-dimer and soluble P-selectin levels were subsequently added into the Vienna model, which led to improvement in VTE prediction (10). Nevertheless, the soluble P-selectin test has been less clinically applied than other laboratory predictors due to its high cost (10). The Caprini model, which is recommended by the guidelines of the American College of Chest Physicians (ACCP), is the most widely used model in surgical patients. Several studies have been performed to validate the predictive ability of the Caprini model for surgical patients with CRC, the ROC-AUC values were in a range of 0.6–0.7 (11,12). Despite the acceptable prediction performance of this model, the Caprini score was rightfully criticized for its complexity and difficulty in interviewing patients for all risk factors (more than 30 factors). It is worth noting that Caprini score is developed by a summary of risk factors from 538 patients not statistical method (13).

Colorectal cancer is a heterogeneous group of malignancies distinguished by distinct clinical, biological, and genetic features. Although the particular primary location of a tumor in the large bowel, tumor stages and chemotherapy regimens have prognostic significance for patients, the association of these risk factors with venous thromboembolism is yet to be clarified (14–16).

Recently, various machine learning (ML) methods, including decision tree-based algorithms, support vector machines (SVMs), and artificial neural networks (ANNs), have been developed for risk prediction in both diseases and disease-associated clinical complications (17,18). Due to the powerful computational learning ability without reliance on rule-based preprogramming, we hypothesized that ML might provide a powerful alternative approach for CRC patient-specific VTE predictive tool development. Another major advantage of ML techniques lies in their ability to handle the highly complex and uncertain error structure of clinical datasets, giving their ability to explain how much a

Highlight box

Key findings

- This study developed a XGBoost model with excellent performance in the prediction of venous thromboembolism (VTE) occurrence in colorectal cancer (CRC) surgical patients.

What is known and what is new?

- The current strategy of VTE risk assessment among CRC inpatients is to use different VTE models, but neither the current generic nor cancer-specific models have adequate sensitivity and specificity.
- In this study, we selected the widely available clinical features for machine learning (ML) models to predict the occurrence of VTE in surgical CRC patients. All established ML models outperformed the Caprini and Khorana model.

What is the implication, and what should change now?

- This study shows that machine learning is a novel approach to accurately predict the occurrence of VTE in surgical CRC patients. The proposed risk factors through model interpretation could shed light on VTE risk stratification in CRC.

given input feature contributes to a model output.

Therefore, the aim of this study was to develop different binary classification VTE predictive models for surgically hospitalized CRC patients using different ML methods and to compare the performance of these ML models with that of previous risk models (Khorana and Caprini). We present the following article in accordance with the TRIPOD reporting checklist (available at <https://jgo.amegroups.com/article/view/10.21037/jgo-23-18/rc>).

Methods

Study design and participants

This is a single-center, retrospective observational study. The Institutional Review Board of the Sixth Affiliated Hospital, Sun Yat-sen University (approval number: 2021ZSLYEC-420) approved this retrospective study with a waiver of informed consent due to its retrospective nature. This study was carried out with adherence to the stipulations of the Declaration of Helsinki (as revised in 2013). This study was limited to inpatients with nonemergent surgery (Table S1). All patients underwent CRC surgery on the primary tumor site with curative intent. The eligibility criteria included: (I) at least 18 years of age at enrollment; (II) at least 7 days of hospital stay length; and (III) patients with a histopathologic diagnosis of malignant tumor before being diagnosed with VTE. The exclusion criteria consisted of the following items: (I) patients admitted for palliative care; (II) patients with recently diagnosed VTE who were actively receiving anticoagulation treatment; and (III) patients who died during hospitalization.

The data acquisition took place between August 2019 and August 2022. Approaches used to identify VTE cases in patients with colorectal cancer have been reported (19). The clinical data from a total of 4,914 surgically hospitalized patients were recorded in an IRB-approved prospectively maintained colorectal cancer database. The variables included patient-level factors (sex, age at diagnosis, body mass index (BMI), comorbidities, cardiovascular and thromboembolic risk factors), cancer-level factors (tumor stages and grades, primary tumor location), and treatment-level factors (Table S2). The laboratory data during the patient's hospital stay were collected repeatedly at different time intervals. Static features are defined as statistics including the mean, standard deviation, minimum, maximum, and median of laboratory data with multiple repeated measurements. Dynamic temporal features were generated based on the original laboratory data with a time

interval of 24 hours. Each patient had 11 time points (from preoperative 3 days to postoperative 7 days). The details of all these variables are shown in Table S2.

The primary outcome was 30-day, non central venous catheter (CVC)-associated VTE in patients with either DVT or PE. Cases were defined as patients with new diagnoses of VTE [within 30 days of the surgical procedure using a postoperative imaging study (CT and ultrasound)]. In our cancer center, systemic ultrasound and CT examinations were performed routinely for every surgical patient. Imaging studies were also performed if patients developed new-onset postoperative symptoms, such as edema of the limb, unexplained pain and fever, skin ulceration, gait disorders, or abnormal laboratory findings during hospitalization. Patients with CVC-associated VTE were considered to have only VTE if VTE was also present at other sites. For the patients with a stay of less than 30 days, two formally trained case reviewers were required to make patient contact via phone call or a *WeChat* message to conduct a thorough review of medical records to identify postoperative VTE diagnosed or managed at other institutions.

Khorana score and Caprini score

Each included patient was assessed retrospectively by two previous VTE models for VTE risk, including one general VTE risk model [Caprini (20)] and one cancer-specific VTE risk model [Khorana (21)]. The stratification of VTE risk was based on the cutoff points recommended in the corresponding derivation cohorts of the different models. In Khorana, the patients were categorized into three risk groups based on the score: "low" (score 0), "intermediate" (score 1–2), and "high" (score ≥ 3) (21). The Caprini score also produces a cumulative risk score based on 39 risk factors. According to the modified version of the Caprini RAM by the ACCP (the most widely used version of the Caprini RAM), patients are classified as follows: "very low risk" (score 0), "low risk" (1–2), "moderate risk" (3–4), and "high risk" (≥ 5) (18). The risk factors identified by these two models and the points assigned for each factor are shown in Table S3.

Machine learning model

Prior to training our model, the continuous variables were normalized, and the categorical variables were encoded as dummy variables. ML models including logistic regression

(LR), random forest (RF), extreme gradient boosting (XGBoost), weighted support vector machine (SVM), a multilayer perception (MLP) network, and a long short-term memory (LSTM) network were employed in this study for VTE prediction. Briefly, the LSTM model accepts dynamic data, while the other ML models accept static features. A grid search strategy based on 10-fold cross-validation was applied for hyperparameter tuning and model training, which was conducted on the training dataset (30% of the total). The total cohort was used as the test dataset to compare the performance of all models based on six indicators, including the ROC-AUC, sensitivity (SEN), specificity (SPE), positive predictive value (PPV), negative predictive value (NPV), and Brier score.

Model interpretation

To enable model interpretability, a SHapley Additive exPlanations (SHAP) analysis was implemented. For explanation of ML models based on the static features, the SHAP values of individual patients were calculated to estimate the variable's contribution to predict the class label in the model. For an explanation of the LSTM model based on dynamic features, SHAP values were calculated by fractions using 24-hour intervals. A global ranking of how each variable contributed to the predicted VTE outcome at the group level was derived from the mean absolute SHAP values in ML models except LSTM. For the LSTM model, the global ranking of each continuous variable was calculated at different time points.

Statistical analysis

All statistical analyses and graphs were realized using packages in the Python platform. Continuous variables were described using their median values and interquartile ranges. Categorical variables were described using frequency counts and percentages. Comparisons between the VTE and non-VTE groups were conducted by ANOVA, nonparametric Student's *t* test, or chi-square test in different situations. Missing data were imputed using multivariate imputation by the chained equations method in Python. All tests were two-sided; *P* values less than 0.05 were considered statistically significant. The model performance was considered excellent for ROC-AUC values 0.9–1, good for ROC-AUC values 0.8–0.9, fair for ROC-AUC values 0.6–0.8, and poor for AUC values 0.5–0.6.

Results

Characteristics of the study cohort

Of the 4,914 patients who were admitted to the colorectal cancer center between August 2019 and August 2022, 1191 surgical patients who met the eligibility criteria during the study were included. A brief description of the patient characteristics and the detailed description of all variables are shown in *Table 1* and *Table S2*, respectively. All patients with CRC were Chinese, with a median age of 63 years, a median hospitalization duration of 32 days, and an American Society of Anesthesiologists (ASA) grade II to III. The overall VTE rate of our study population was 10.8%. The surgery-related information of the patients is shown in *Table S1*.

Results of the model performance

Six ML models were established (LR, RF, XGBoost, SVM, MLP, and LSTM). *Table 2* shows the prediction performance of the VTE models in the CRC patients. Six indicators, including ROC-AUC, sensitivity, specificity, PPV, NPV, and Brier score, were applied to assess the candidate models. Additionally, the model performance was visualized in ROC curves, precision-recall (PR) curves, and with binary classification performance (*Figure 1* and *Figure S1*). The previously developed Caprini score and Khorana score were applied to compare and evaluate model performance. Our results indicated that the XGBoost model achieved the overall best prediction, with an ROC-AUC of 0.908 (95% CI: 0.870–0.941). Despite a lower ROC-AUC of 0.868 (95% CI: 0.818–0.915), the MLP model had the highest PPV of 64.8% (95% CI: 55.9–73.3%) among all the candidate models. Overall, all the ML models performed better than the previously developed models (Caprini and Khorana).

Interpretation and evaluation of machine learning models

The XGBoost model and MLP model were selected for further interpretation due to their better performance. We additionally interpreted the LSTM model, as time-series laboratory data analysis was only available in this model. The results of the SHAP analysis of XGBoost, MLP, and LSTM are shown in *Figures 2–4*, respectively. The SHAP values were used to represent the local contribution of each feature to the individual predictions made by the

Table 1 Patient characteristics at baseline

Patient characteristics	All patients, n (%) or median (IQR), n=1,191	All patients, n (%) or median (IQR), VTE, n=129	All patients, n (%) or median (IQR), non VTE, n=1,062	P value
Patient-related factors				
Age (years)	63 (54 to 70)	64 (56 to 72)	63 (54 to 70)	0.169
Females	419 (35.1%)	61 (47.3%)	358 (33.6%)	0.003
BMI ≥ 25	224 (18.8%)	30 (23.3%)	194 (18.2%)	0.166
BMI ≥ 28	53 (4.4%)	7 (5.4%)	46 (4.3%)	0.564
Hypertension	315 (26.4%)	32 (24.8%)	283 (26.6%)	0.667
Diabetes mellitus	149 (12.5%)	13 (10.1%)	136 (12.8%)	0.382
Dyslipidemia	119 (10.0%)	8 (6.2%)	111 (10.4%)	0.131
Liver cirrhosis	7 (0.6%)	2 (1.6%)	5 (0.5%)	0.129
Hepatic dysfunction	46 (3.9%)	3 (2.3%)	43 (4.0%)	0.340
Chronic lung disease	30 (2.5%)	5 (3.9%)	25 (2.3%)	0.295
Heart failure	5 (0.4%)	–	5 (0.5%)	0.435
History of a myocardial infarction	70 (5.9%)	4 (3.1%)	66 (6.2%)	0.157
History of a stroke	61 (5.1%)	7 (5.4%)	54 (5.15%)	0.862
Atrial fibrillation	16 (1.3%)	1 (0.8%)	15 (1.4%)	0.555
Varicose vein	6 (0.5%)	2 (1.6%)	4 (0.4%)	0.075
History of VTE	2 (0.2%)	2 (1.6%)	–	<0.001
History of major bleeding	37 (3.1%)	3 (2.3%)	34 (3.2%)	0.592
Cancer-related factors				
Tumor stage I-II	570 (47.7%)	51 (39.5%)	519 (48.7%)	0.048
Tumor stage III-IV	624 (52.3%)	78 (60.5%)	546 (51.3%)	0.048
Metastasis disease	288 (24.1%)	39 (30.2%)	249 (23.4%)	0.086
The site of tumor				
Right colon	316 (26.5%)	33 (25.6%)	283 (26.6%)	0.810
Transverse colon	92 (7.7%)	4 (3.1%)	88 (8.3%)	0.038
Left colon	221 (18.5%)	16 (12.4%)	205 (19.2%)	0.059
Sigmoid colon/rectum	563 (47.2%)	74 (57.4%)	489 (45.9%)	0.018
Appendix/cecum	2 (0.2%)	2 (1.6%)	–	<0.001

BMI, body mass index; IQR, interquartile range; VTE, venous thromboembolism.

corresponding models.

Based on the mean absolute SHAP value, ten of the most significant clinically associated features for VTE prediction, including lymph node metastasis (N class 0), C-reactive protein (CRP over 10 mg/L), tumor grade (IIa/IIb), anemia, primary tumor location (sigmoid/rectum),

sex, age (age group: 60 to 75), D-dimer (over 0.5 $\mu\text{g/mL}$), and tumor stage (I/II), were identified in the XGBoost model (*Figure 2A,2B*). Similarly, calculation of the SHAP values in the MLP model interpretation revealed that anemia, gender, lymph node metastasis (N class 0), tumor grade (IIa), tumor class 4a (T class 4a), primary tumor

Table 2 VTE occurrence prediction performance of the VTE models in CRC patients

Variables (model)	AUC/C-index ^a (95% CI)	Brier score	Sensitivity (%, 95% CI)	Specificity (%, 95% CI)	Positive predictive value (%, 95% CI)	Negative predictive value (%, 95% CI)
Previous VTE-RAMs						
Caprini score						
Cutoff 5 point ^{a,b}	0.769 (0.711–0.821)	–	96.9 (93.1–99.6)	100.0	10.5 (8.6–12.5)	100.0
Cutoff 9 points ^c		–	59.7 (50.5–69.0)	85.0 (82.5–87.3)	32.6 (26.5–39.3)	94.6 (93.1–96.1)
Khorana score						
Cutoff 3 points ^b	0.646 (0.598–0.699)	–	10.9 (5.6–17.3)	98.0 (97.0–98.9)	40.0 (22.2–58.3)	90.1 (88.2–91.9)
Cutoff 1 points ^c		–	62.0 (52.5–70.8)	63.8 (60.6–67.0)	17.2 (13.5–21.2)	93.3 (91.2–95.2)
Machine learning models						
LR						
Training cohort	0.937 (0.898–0.971)	0.046	82.8 (72.7–92.5)	90.8 (72.7–92.5)	51.9 (41.2–62.5)	97.8 (96.4–99.0)
Testing cohort	0.894 (0.856–0.929)	0.052	76.0 (67.8–84.7)	87.9 (85.9–90.0)	43.4 (36.6–50.3)	96.8 (95.5–98.0)
RF						
Training cohort	0.912 (0.866–0.953)	0.066	78.1 (66.7–89.1)	88.1 (85.1–91.2)	44.2 (33.9–55.3)	97.1 (95.5–98.7)
Testing cohort	0.866 (0.822–0.908)	0.070	72.1 (63.4–80.9)	87.2 (84.9–89.3)	40.6 (33.5–47.5)	96.3 (94.8–97.5)
SVM						
Training cohort	0.943 (0.921–0.961)	0.068	100.0	90.6 (87.9–93.3)	56.1 (46.7–66.7)	100.0
Testing cohort	0.879 (0.847–0.910)	0.071	79.1 (72.1–86.3)	88.8 (86.7–90.8)	46.2 (39.5–53.5)	97.2 (96.1–98.3)
XGBoost						
Training cohort	0.990 (0.980–0.997)	0.029	96.9 (91.5–100.0)	96.1 (94.2–97.8)	74.7 (64.5–84.2)	99.6 (98.9–100.0)
Testing cohort	0.908 (0.870–0.941)	0.047	77.5 (69.3–85.4)	93.7 (92.1–95.3)	59.9 (51.9–68.2)	97.2 (96.0–98.2)
MLP						
Training cohort	0.967 (0.917–99.8)	0.013	92.2 (83.9–98.2)	99.4 (98.7–99.8)	95.2 (88.9–99.8)	99.1 (97.9–99.9)
Testing cohort	0.868 (0.818–0.915)	0.066	71.3 (63.2–79.6)	95.3 (93.8–96.6)	64.8 (55.9–73.3)	96.5 (95.1–97.6)
LSTM						
Training cohort	0.822 (0.812–0.916)	0.118	78.3 (70.1–86.3)	85.9 (83.7–88.2)	38.5 (31.5–45.6)	97.2 (96.1–98.3)
Testing cohort	0.803 (0.783–0.885)	0.122	74.2 (65.5–82.5)	86.5 (84.2–88.7)	39.7 (33.3–46.7)	96.5 (95.2–97.7)

^a, The value of the C-index is the same as that of the AUC in the logistic regression model; ^b, Recommended cutoff points based on derivation studies; ^c, Calculated cutoff points based on ROC curves. Data are ROC-AUCs and (95% CI). ROC, receiver operating characteristic; AUCs, areas under the curve; CI, confidence interval; CRC, colorectal cancer; RAM, risk assessment model; LR, logistic regression; RF, random forest; SVM, support vector machine; MLP, multilayer perception network; LSTM, long short-term memory; VTE, venous thromboembolism.

location (sigmoid/rectum), body mass index (BMI ≥ 25), thrombin time (TT < 14 s), D-dimer (over 0.5 $\mu\text{g/mL}$), and C-reactive protein (CRP over 10 mg/L) had the most substantial contribution to the model output (*Figure 3A,3B*). *Figure 2A* and *Figure 3A* also demonstrate the direction of

the correlation between the feature and model output. For example, primary tumor location (sigmoid/rectum) had an asymmetric distribution of SHAP values, with primary tumor location (sigmoid/rectum) indicating an increasing association with VTE occurrence.

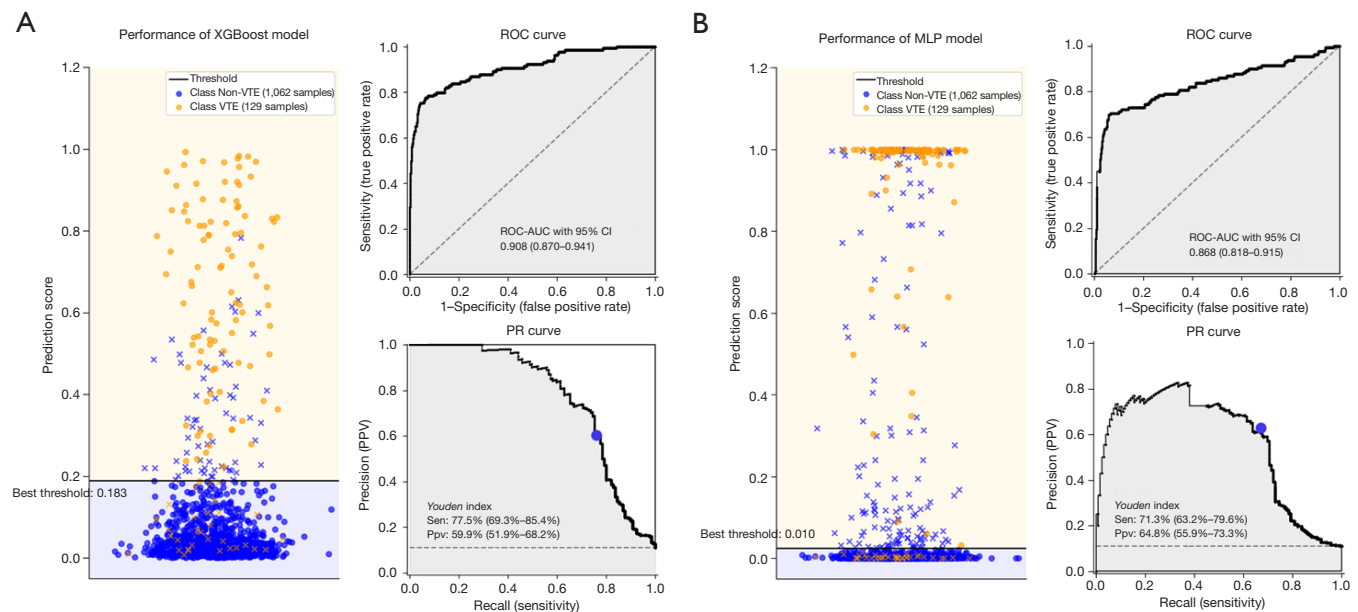


Figure 1 The model performance of the XGBoost model (plot A) and MLP model (plot B). The classification based on the best threshold, the ROC curve and PR curve were plotted to measure the performance of the two machine learning models, and the AUCs were also calculated with 95% CIs. The best threshold points of these PR curves were plotted with corresponding sensitivities and positive predictive values. AUC, area under the curve; CI, confidence interval; MLP, multilayer perception network; PPV, positive predictive value; PR, precision-recall curve; ROC, receiver operating characteristic curve; SEN, sensitivity; VTE, venous thromboembolism; XGBoost, extreme gradient boosting; *Youden index*: = sensitivity + specificity – 1.

In the LSTM model, the variable contributing to the predicted VTE outcome at the continuous variable level was derived from mean absolute SHAP values, which were calculated at different time points. The ten highest performing clinically relevant variables in the LSTM model are exhibited, including creatine kinase (CK), creatinine (Cr), C-reactive protein (CRP), α -hydroxybutyrate dehydrogenase (α -HBDH), lactate dehydrogenase (LDH), lipoprotein (a) [Lp (a)], myoglobin (MYO), platelet (PLT), uric acid (UA), and white blood cells (WBC). *Figure 4* shows that WBC level was the most impactful variable. The trends of these ten variables over time are also depicted in *Figure 4*. For example, the impact of WBC on the model output seems to increase gradually over time from preoperative Day 2 to postoperative Day 6.

For a single patient prediction, the final model output and prediction confidence are formed by the sum of contributions made by each of their features (22). The XGBoost model predictions for three patients, including a patient with a significant positive outcome, a patient with an indeterminate outcome, and a patient with a strong negative

outcome, are shown in *Figure 2C–2E*, respectively. Similarly, the MLP model predictions for three patients (positive, indeterminate, and negative) are depicted in *Figure 3C–3E*, respectively.

Discussion

This 1191-sample retrospective cohort study developed multiple prediction models for VTE in surgical patients with CRC. Six different types of supervised machine learning algorithms (LR, RF, XGBoost, SVM, MLP, and LSTM) were applied to examine the features in our cohort. Two widely used VTE models, the Caprini model and the Khorana model, were used as the benchmark models to compare the performance of the ML models. We found that the XGBoost model achieved the best classification performance with the highest ROC-AUC, and the MLP model achieved the highest PPV in our cohort. The performance of the two benchmark models was “moderate” (ROC-AUC of Caprini: 0.769 and ROC-AUC of Khorana: 0.646), but all the ML models achieved an ROC-AUC over 0.8.

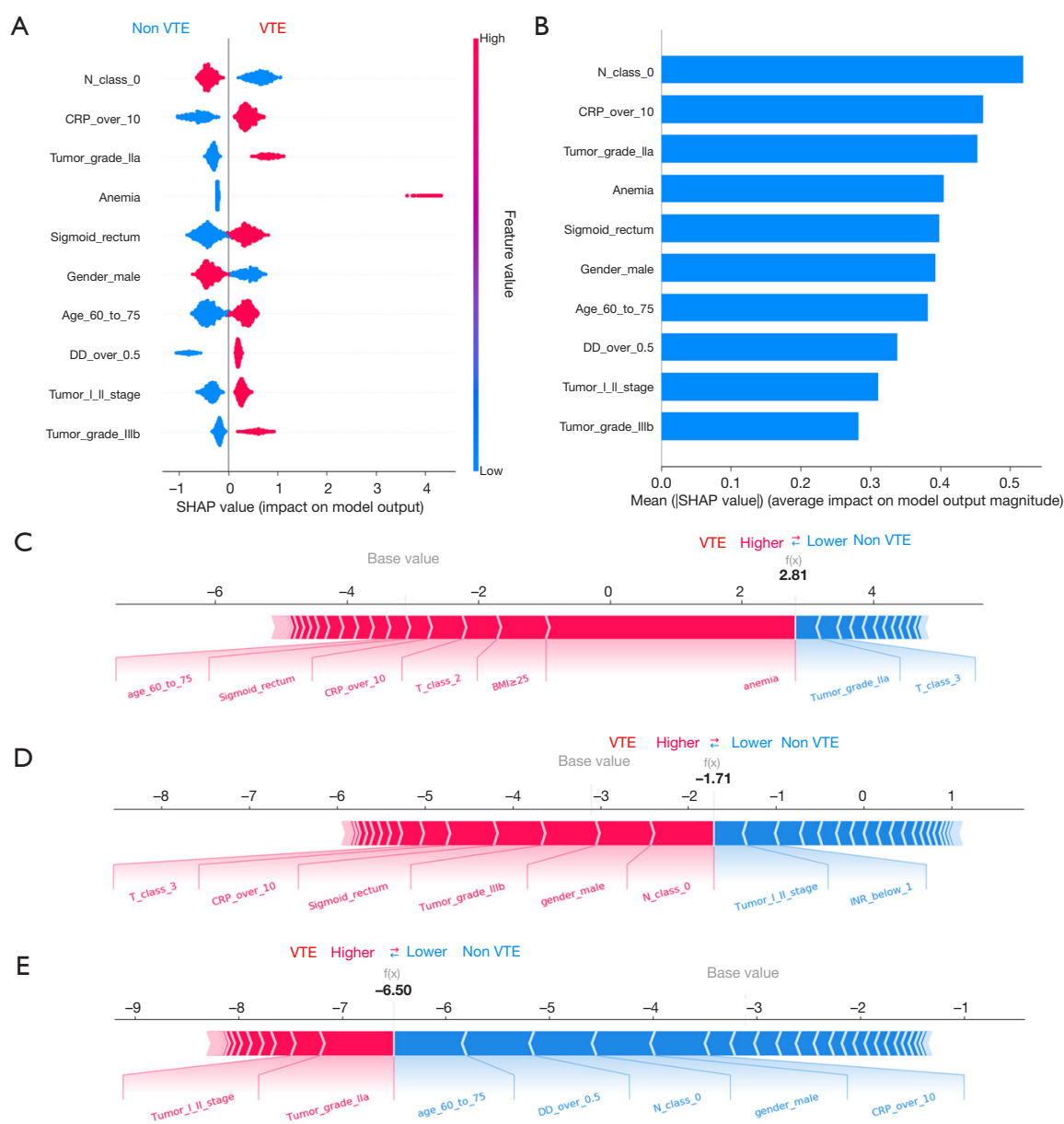


Figure 2 Interpretation and evaluation of the XGBoost model. Plot (A) reports the result of the SHAP analysis on the dataset. The study variables are described using mean absolute SHAP values. The variables ranked among the top ten are shown as distributions across individual patients. Each point in the figure represents the SHAP value of a single patient. The y-axis indicates the rank of the variable contribution to model prediction. The x-axis represents the mean absolute SHAP value. Blue and red colors indicate lower and higher values of the variables, respectively. Plot (B) shows the amount of a feature contributing to the model output indicated by the SHAP analysis. Individual predictions in the XGBoost model for (C) a patient with a strong positive outcome prediction, (D) a patient with an indeterminate prediction, and (E) a patient with a strong negative outcome prediction. The sum of the expected SHAP value (base value) and the calculated SHAP value of all individual variables is defined as the model output for a single patient. A model output value greater than the expected SHAP value indicates a positive prediction (VTE occurrence), while an output value less than the expected SHAP value indicates a negative prediction (no VTE occurrence). Blue and red colors indicate negative and positive effects of variables, respectively. The visual size represents the magnitude of the effect. BMI, body mass index; CRP, C-reactive protein; DD, D-dimer; INR, international normalized ratio; SHAP, SHapley Additive exPlanations; VTE, venous thromboembolism; XGBoost, extreme gradient boosting.

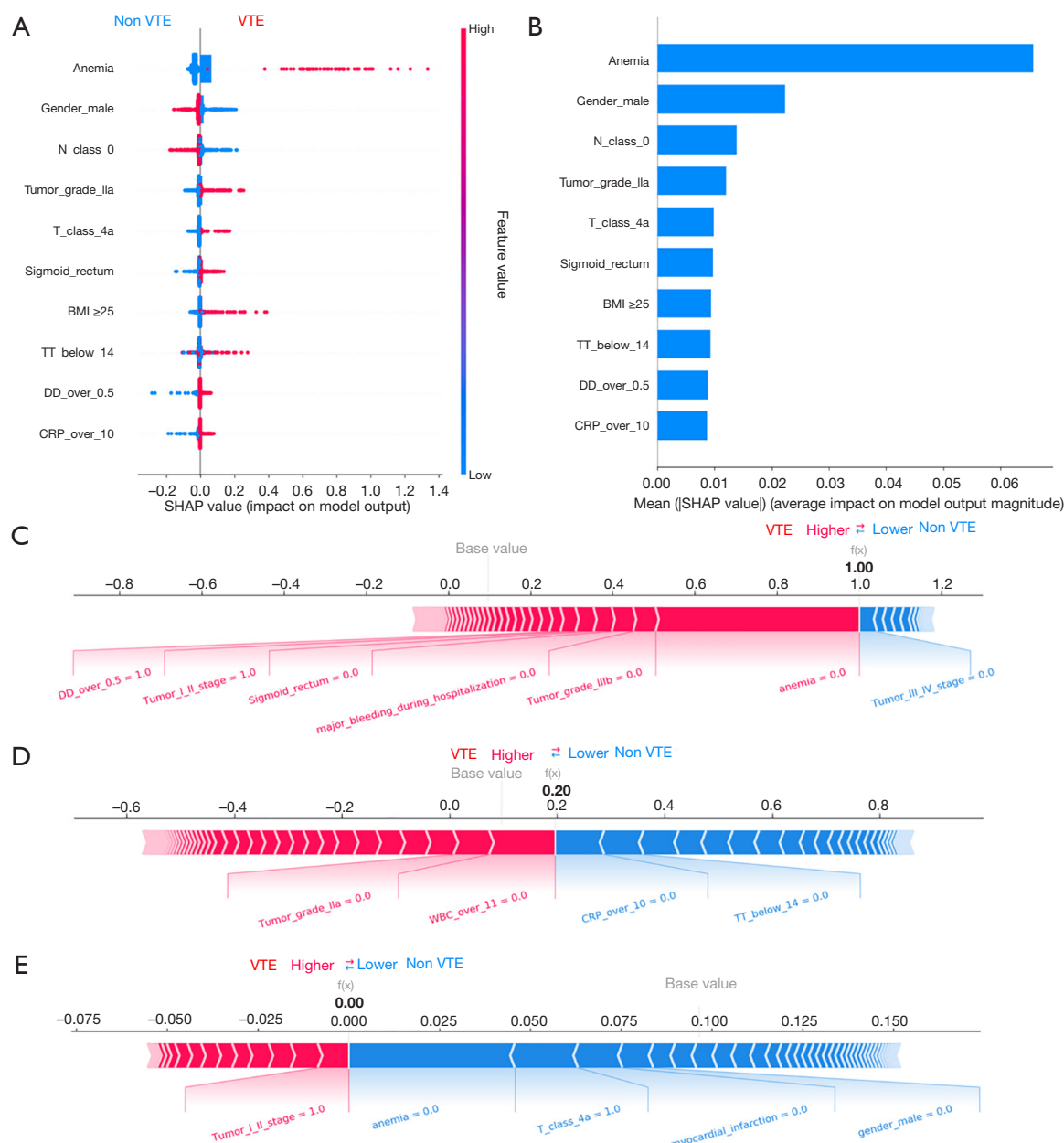


Figure 3 Interpretation and evaluation of the MLP model. Plot (A) reports the result of the SHAP analysis on the dataset. The study variables are described using mean absolute SHAP values. The variables ranked among the top ten are shown as distributions across individual patients. Each point in the figure represents the SHAP value of a single patient. The y-axis indicates the rank of the variable contribution to model prediction. The x-axis represents the mean absolute SHAP value. Blue and red colors indicate lower and higher values of variables, respectively. Plot (B) shows the amount of a feature contributing to the model output indicated by SHAP analysis. Individual predictions in the MLP model for (C) a patient with a strong positive outcome prediction, (D) a patient with an indeterminate prediction, and (E) a patient with a strong negative outcome prediction. The sum of the expected SHAP value (base value) and the calculated SHAP value of all individual variables is defined as the model output for a single patient. A model output value greater than the expected SHAP value indicates a positive prediction (VTE occurrence), while an output value less than the expected SHAP value indicates a negative prediction (no VTE occurrence). Blue and red colors indicate negative and positive effects of variables, respectively. The visual size represents the magnitude of the effect. BMI, body mass index; CRP, C-reactive protein; DD, D-dimer; MLP, multilayer perceptron network; TT, thrombin time; SHAP, SHapley Additive exPlanations; VTE, venous thromboembolism; WBC, white blood cell.

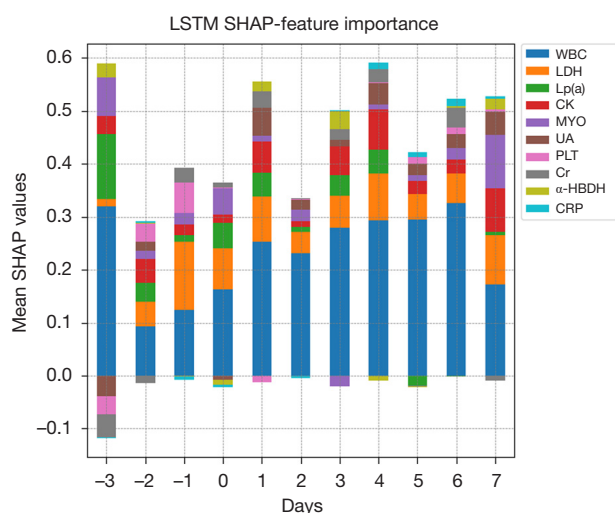


Figure 4 Interpretation and evaluation of the LSTM model. In the LSTM model, the variable contributing to the predicted VTE outcome at the continuous variable level was derived from mean absolute SHAP values, which were calculated at different time points. The ten highest performing clinically relevant variables in the LSTM model are exhibited according to the mean absolute SHAP values at different time intervals. The position on the x-axis indicates days before surgery (negative numbers) and after surgery (positive numbers); Day 0 indicates the day of surgery; the rank on the y-axis is determined by the mean absolute contribution of the variable to the model's output. CK, creatine kinase; Cr, creatinine; CRP, C-reactive protein; α -HBDH, α -hydroxybutyrate dehydrogenase; LDH, lactate dehydrogenase; Lp (a), lipoprotein (a); LSTM, long short-term memory; MYO, myoglobin; PLT, platelet; UA, uric acid; WBC, white blood cells.

In our study, the Caprini model and the Khorana model had acceptable NPV (>90%) but a low PPV (<40%). Thus, these models are highly effective at identifying patients with CRC who are at low risk for VTE, but the models may not be as predictive in individual specific cancers because they were designed as a tool for a mixed solid tumor population. Of note, although the Caprini model exhibited a satisfactory sensitivity (over 90% by the recommended cutoff points based on derivation studies), it stratified more than 90% of patients as high risk, and its low PPV (approximately 10%) suggested that a large portion of patients may be exposed to unnecessary risks associated with VTE prophylaxis.

In general, an ideal prediction model should simultaneously achieve excellent sensitivity and specificity; however, there is a trade-off between these two desirable properties. It

is unclear how this trade-off affects the clinical benefit of VTE prophylaxis. In fact, patients with colorectal cancer who have high VTE risk frequently present with venous cavernomas and/or signs of portal hypertension, which can lead to complications such as variceal bleeding (23). These patients are routinely seen in clinical practice, but they are underrepresented in clinical trials, making VTE prophylaxis and treatment decisions difficult. Taking the increased risk of bleeding in CRC patients into account, we believe that the VTE prediction model for this population should have high sensitivity and positive predictive value. Our results show that the XGBoost model has the highest ROC-AUC and the second highest PPV. Despite the fact that the MLP model's performance was comparable to that of the XGBoost model, the sum of sensitivity and positive predictive value of the XGBoost model was higher. Therefore, we considered that this model was the optimal model for postoperative VTE prediction in CRC patients.

Machine learning models can receive the characteristics of a large number of features and understand the dataset to obtain implicit correlations to serve complex binary/multiclass classification. SHAP is an explainable framework that allows researchers to comprehend and trust the results and output created by ML algorithms. Seven common impactful predictors in both the XGBoost model and MLP model were identified, including N class, sex, CRP, D-dimer, anemia, primary tumor location (sigmoid/rectum), and tumor grade. Tumor Nodes Metastases (TNM) stages of tumor lesions, sex, D-dimer and CRP are known risk factors for VTE, and there are more data to support their use in various cancer types (24). However, primary tumor location in the large bowel is a unique risk factor in CRC. While the prognostic value of primary tumor location for overall survival in CRC seems clear and consistent in reports in recent decades, the association between primary tumor location and VTE remains unclear. Our results showed that patients with a primary tumor site at the sigmoid/rectum had a higher risk of VTE than patients with other primary tumor sites. Since all patients in this study underwent CRC surgery at the primary tumor site with curative intent, we hypothesized that surgical factors may be implicated in postoperative VTE occurrence.

A previous study also suggested that postoperative venous thrombosis development in patients receiving colon and rectal surgery could be affected by surgical features, including resection site, use of laparoscopic surgery, and procedure time (25). In this study, more patients with VTE underwent

laparoscopic surgery, with more tumors located in the right colon and sigmoid/rectum. To our knowledge, few studies have reported the association between surgical factors and postoperative VTE in CRC patients. However, an animal study and a human study suggested that laparoscopic surgery might promote portomesenteric venous thrombosis (PMVT) due to portal and mesenteric venous flow decline through insufflation of the abdomen and increased intra-abdominal pressure (26,27). Furthermore, a previous 1,224-patient study reported a higher incidence (10.8%) of PMVT in patients undergoing restorative proctocolectomy than in patients undergoing left (3.9%) and right (1.9%) colectomy (28). In our study, the incidence of VTE in patients undergoing proctectomy/proctocolectomy was 13.2%, higher than the incidence in patients undergoing left (7.2%), transverse (4.3%), and right (10.5%) colectomy, which was consistent with the previous study (28). Considering the different underlying mechanisms between PMVT and VTE (including DVT and PE), further investigation is warranted to validate the interaction between VTE and surgical factors in CRC patients.

Generally, SHAP works well on static features but not on dynamic time series features. In ML models accepting static features, each predictor was grouped as model input and model interpretation, which show correlations but not causality. In the present study, we attempt to apply the SHAP framework in the LSTM model. The dynamic data of the CRC patients were transformed into a $n \times 11 \times 78$ size 3D matrix before entering the model. The contribution of continuous variables to model prediction was quantified by the mean absolute SHAP values, which were calculated at 11 time points (from preoperative 3 days to postoperative 7 days). The impact of WBC on the model output gradually increased over time from preoperative Day 2 to postoperative Day 6, which is consistent with previous reports in the literature (29). However, the influence of other dynamic variables was not significant, and there was no obvious trend.

Our study had potential limitations: (I) The study was performed in a single center. (II) Our study was a retrospective analysis of a consecutively collected colorectal cancer database. However, because of incorrect coding, there could have been missing individuals when identifying patients with VTE. (III) Although we found that some potential surgical factors had the potential to be used in postoperative VTE prediction in CRC patients, further research is required among different populations utilizing a larger study sample size.

Conclusions

This study shows that machine learning for predictive modeling is a novel approach to accurately predict the occurrence of VTE in surgical CRC patients. Furthermore, we developed an XGBoost model with high sensitivity and positive predictive value in the prediction of VTE occurrence, which might supply a complementary tool for clinical VTE prophylaxis decision-making in colorectal cancer. The proposed risk factors through model interpretation, such as surgical factors, could shed some light on VTE risk stratification in CRC patients.

Acknowledgments

Funding: None.

Footnote

Reporting Checklist: The authors have completed the TRIPOD reporting checklist. Available at <https://jgo.amegroups.com/article/view/10.21037/jgo-23-18/rc>

Data Sharing Statement: Available at <https://jgo.amegroups.com/article/view/10.21037/jgo-23-18/dss>

Peer Review File: Available at <https://jgo.amegroups.com/article/view/10.21037/jgo-23-18/prf>

Conflicts of Interest: All authors have completed the ICMJE uniform disclosure form (available at <https://jgo.amegroups.com/article/view/10.21037/jgo-23-18/coif>). The authors have no conflicts of interest to declare.

Ethical Statement: The authors are accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved. This study was carried out with adherence to the stipulations of the Declaration of Helsinki (as revised in 2013). The Institutional Review Board of the Sixth Affiliated Hospital, Sun Yat-sen University (approval number: 2021ZSLYEC-420) approved this retrospective study. Informed consent was waived due to the retrospective nature of this study.

Open Access Statement: This is an Open Access article distributed in accordance with the Creative Commons

Attribution-NonCommercial-NoDerivs 4.0 International License (CC BY-NC-ND 4.0), which permits the non-commercial replication and distribution of the article with the strict proviso that no changes or edits are made and the original work is properly cited (including links to both the formal publication through the relevant DOI and the license). See: <https://creativecommons.org/licenses/by-nc-nd/4.0/>.

References

1. Khorana AA, Mackman N, Falanga A, et al. Cancer-associated venous thromboembolism. *Nat Rev Dis Primers* 2022;8:11.
2. Timp JF, Braekkan SK, Versteeg HH, et al. Epidemiology of cancer-associated venous thrombosis. *Blood* 2013;122:1712-23.
3. Eichinger S. Cancer associated thrombosis: risk factors and outcomes. *Thromb Res* 2016;140 Suppl 1:S12-7.
4. Spyropoulos AC, Raskob GE. New paradigms in venous thromboprophylaxis of medically ill patients. *Thromb Haemost* 2017;117:1662-70.
5. Patel SG, Karlitz JJ, Yen T, et al. The rising tide of early-onset colorectal cancer: a comprehensive review of epidemiology, clinical features, biology, risk factors, prevention, and early detection. *Lancet Gastroenterol Hepatol* 2022;7:262-74.
6. Douaiher J, Ravipati A, Grams B, et al. Colorectal cancer: global burden, trends, and geographical variations. *J Surg Oncol* 2017;115:619-30.
7. Mulder FI, Candeloro M, Kamphuisen PW, et al. The Khorana score for prediction of venous thromboembolism in cancer patients: a systematic review and meta-analysis. *Haematologica* 2019;104:1277-87.
8. Ay C, Dunkler D, Marosi C, et al. Prediction of venous thromboembolism in cancer patients. *Blood* 2010;116:5377-82.
9. van Es N, Ventresca M, Di Nisio M, et al. The Khorana score for prediction of venous thromboembolism in cancer patients: An individual patient data meta-analysis. *J Thromb Haemost* 2020;18:1940-51.
10. Patell R, Rybicki L, McCrae KR, et al. Predicting risk of venous thromboembolism in hospitalized cancer patients: Utility of a risk assessment tool. *Am J Hematol* 2017;92:501-7.
11. Lu X, Zeng W, Zhu L, et al. Application of the Caprini risk assessment model for deep vein thrombosis among patients undergoing laparoscopic surgery for colorectal cancer. *Medicine (Baltimore)* 2021;100:e24479.
12. Yao J, Lang Y, Su H, et al. Construction of risk assessment model for venous thromboembolism after colorectal cancer surgery: a Chinese single-center study. *Clin Appl Thromb Hemost* 2022;28:10760296211073748.
13. Caprini JA, Arcelus JJ, Hasty JH, et al. Clinical assessment of venous thromboembolic risk in surgical patients. *Semin Thromb Hemost* 1991;17:304-12.
14. Khorana AA, Francis CW, Culakova E, et al. Risk factors for chemotherapy-associated venous thromboembolism in a prospective observational study. *Cancer* 2005;104:2822-9.
15. Haddad TC, Greeno EW. Chemotherapy-induced thrombosis. *Thromb Res* 2006;118:555-68.
16. Wei Q, Wang Y, An YB, et al. Rationale and design of a prospective, multicenter, cohort study on the evaluation of postoperative Venous ThromboEmbolic incidence in patients with ColoRectal Cancer (CRC-VTE trial). *Transl Cancer Res* 2022;11:1406-12.
17. Boehm KM, Aherne EA, Ellenson L, et al. Multimodal data integration using machine learning improves risk stratification of high-grade serous ovarian cancer. *Nat Cancer* 2022;3:723-33.
18. Jin S, Qin D, Liang BS, et al. Machine learning predicts cancer-associated deep vein thrombosis using clinically available variables. *Int J Med Inform* 2022;161:104733.
19. Vardi M, Ghanem-Zoubi NO, Zidan R, et al. Venous thromboembolism and the utility of the Padua Prediction Score in patients with sepsis admitted to internal medicine departments. *J Thromb Haemost* 2013;11:467-73.
20. Caprini JA. Thrombosis risk assessment as a guide to quality patient care. *Dis Mon* 2005;51:70-8.
21. Khorana AA, Kuderer NM, Culakova E, et al. Development and validation of a predictive model for chemotherapy-associated thrombosis. *Blood* 2008;111:4902-7.
22. Lundberg S, Lee SI. A unified approach to interpreting model predictions. Paper presented at: 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, California, United States. December 4-9, 2017:1-10.
23. Parikh S, Shah R, Kapoor P. Portal vein thrombosis. *Am J Med* 2010;123:111-9.
24. Neeman E, Liu V, Mishra P, et al. Trends and risk factors for venous thromboembolism among hospitalized medical patients. *JAMA Netw Open* 2022;5:e2240373.
25. Moghadamyeghaneh Z, Hanna MH, Carmichael JC, et al. A nationwide analysis of postoperative deep vein thrombosis and pulmonary embolism in colon and rectal

- surgery. *J Gastrointest Surg* 2014;18:2169-77.
26. James AW, Rabl C, Westphalen AC, et al. Portomesenteric venous thrombosis after laparoscopic surgery: a systematic literature review. *Arch Surg* 2009;144:520-6.
 27. Jakimowicz J, Stultiëns G, Smulders F. Laparoscopic insufflation of the abdomen reduces portal venous flow. *Surg Endosc* 1998;12:129-32.
 28. Robinson KA, O'Donnell ME, Pearson D, et al. Portomesenteric venous thrombosis following major colon and rectal surgery: incidence and risk factors. *Surg Endosc* 2015;29:1071-9.
 29. Zakai NA, Callas PW, Repp AB, et al. Venous thrombosis risk assessment in medical inpatients: the medical inpatients and thrombosis (MITH) study. *J Thromb Haemost* 2013;11:634-41.

Cite this article as: Qin L, Liang Z, Xie J, Ye G, Guan P, Huang Y, Li X. Development and validation of machine learning models for postoperative venous thromboembolism prediction in colorectal cancer inpatients: a retrospective study. *J Gastrointest Oncol* 2023;14(1):220-232. doi: 10.21037/jgo-23-18

Table S1 Surgical variables of the whole cohort

Surgical variables	All patients, n (%), n=1,191	All patients, n (%), VTE, n=129	All patients, n (%), Non VTE, n=1,062	P value
ASA classification				
2	759 (63.6%)	86 (66.7%)	673 (63.2%)	0.681
3	435 (36.4%)	43 (33.3%)	392 (36.8%)	0.514
Type of resection				
Right colon	316 (26.5%)	33 (25.6%)	283 (26.6%)	0.810
Transverse colon	92 (7.7%)	4 (3.1%)	88 (8.3%)	0.038
Left colon	221 (18.5%)	16 (12.4%)	205 (19.2%)	0.059
Proctectomy/Proctocolectomy	563 (47.2%)	74 (57.4%)	489 (45.9%)	0.018
Length of surgery				
0–59 min	6 (0.5%)	1 (0.8%)	5 (0.5%)	0.643
60–119 min	38 (3.2%)	2 (1.6%)	36 (3.4%)	0.263
120–179 min	227 (19.0%)	20 (15.5%)	207 (19.4%)	0.282
≥180 min	923 (77.3%)	106 (82.2%)	817 (76.7%)	0.265
Surgical approach				
Open	257 (21.5%)	25 (19.5%)	232 (21.8%)	0.868
Laparoscopic/Robotic	937 (78.5%)	104 (80.6%)	833 (78.2%)	0.890

ASA, American Society of Anesthesiologists; CRC, colorectal cancer patient; VTE, venous thromboembolism.

Table S2 The characteristics of the whole cohort

Patient characteristics	All patients, n (%) or median (IQR), n=1,191	All patients, n (%) or median (IQR), VTE, n=129	All patients, n (%) or median (IQR), Non VTE, n=1,062	P value
Patient-related factors				
Age	63 [54 to 70]	64 [56 to 72]	63 [54 to 70]	0.169
Females	419 (35.1%)	61 (47.3%)	358 (33.6%)	0.003
BMI ≥ 25	224 (18.8%)	30 (23.3%)	194 (18.2%)	0.166
BMI ≥ 28	53 (4.4%)	7 (5.4%)	46 (4.3%)	0.564
Hypertension	315 (26.4%)	32 (24.8%)	283 (26.6%)	0.667
Diabetes mellitus	149 (12.5%)	13 (10.1%)	136 (12.8%)	0.382
Dyslipidemia	119 (10.0%)	8 (6.2%)	111 (10.4%)	0.131
Liver cirrhosis	7 (0.6%)	2 (1.6%)	5 (0.5%)	0.129
Hepatic dysfunction	46 (3.9%)	3 (2.3%)	43 (4.0%)	0.340
Chronic lung disease	30 (2.5%)	5 (3.9%)	25 (2.3%)	0.295
Heart failure	5 (0.4%)	–	5 (0.5%)	0.435
History of a myocardial infarction	70 (5.9%)	4 (3.1%)	66 (6.2%)	0.157
History of a stroke	61 (5.1%)	7 (5.4%)	54 (5.15%)	0.862
Atrial fibrillation	16 (1.3%)	1 (0.8%)	15 (1.4%)	0.555
Varicose vein	6 (0.5%)	2 (1.6%)	4 (0.4%)	0.075
History of VTE	2 (0.2%)	2 (1.6%)	–	< 0.001
History of major bleeding	37 (3.1%)	3 (2.3%)	34 (3.2%)	0.592
Cancer-related Factors				
Tumor stage I-II	570 (47.7%)	51 (39.5%)	519 (48.7%)	0.048
Tumor stage III-IV	624 (52.3%)	78 (60.5%)	546 (51.3%)	0.048
T stage				
1	54 (4.5%)	6 (4.7%)	48 (4.5%)	0.941
2	148 (12.4%)	11 (8.5%)	137 (12.9%)	0.158
3	694 (58.1%)	73 (56.6%)	621 (58.3%)	0.708
4a	227 (19.0%)	31 (24.0%)	196 (18.4%)	0.124
4b	71 (5.9%)	8 (6.2%)	63 (5.9%)	0.897
N stage				
0	624 (52.3%)	51 (39.5%)	573 (53.8%)	0.001
1a	210 (17.6%)	32 (24.8%)	178 (16.7%)	0.023
1b	93 (7.8%)	13 (10.1%)	80 (7.5%)	0.304
1c	24 (2.0%)	2 (1.6%)	22 (2.1%)	0.694
2a	162 (13.6%)	20 (15.5%)	142 (13.3%)	0.497
2b	81 (6.8%)	11 (8.5%)	70 (6.6%)	0.404
M stage				
0	978 (81.9%)	100 (77.5%)	878 (82.4%)	0.170
1a	177 (14.8%)	23 (17.8%)	154 (14.5%)	0.309
1b	13 (1.1%)	1 (0.8%)	12 (1.1%)	0.716
1c	26 (2.2%)	5 (3.9%)	21 (2.0%)	0.162
Metastasis disease	288 (24.1%)	39 (30.2%)	249 (23.4%)	0.086
The site of tumor				
Right colon	316 (26.5%)	33 (25.6%)	283 (26.6%)	0.810
Transverse colon	92 (7.7%)	4 (3.1%)	88 (8.3%)	0.038
Left colon	221 (18.5%)	16 (12.4%)	205 (19.2%)	0.059
Sigmoid colon/Rectum	563 (47.2%)	74 (57.4%)	489 (45.9%)	0.018
Appendix/cecum	2 (0.2%)	2 (1.6%)	–	< 0.001
Treatment-related Factors				
Radiotherapy	46 (3.9%)	4 (3.1%)	42 (3.9%)	0.639
Neoadjuvant chemotherapy	227 (19.0%)	25 (19.4%)	202 (19.0%)	0.910
Preoperative chemotherapy regimens				
FLOT ^a	1 (0.1%)	1 (0.8%)	–	0.207
FOLFIRI ^b	4 (0.3%)	1 (0.8%)	3 (0.3%)	0.368
FOLFIRI ^b + targeted drug ^k	2 (0.2%)	–	2 (0.2%)	0.622
mFOLFOX6 ^c	145 (12.2%)	19 (14.7%)	126 (11.9%)	0.347
mFOLFOX6 ^c + targeted drug ^k	11 (0.9%)	2 (1.6%)	9 (0.8%)	0.431
FOLFOXIRI ^d	30 (2.5%)	2 (1.6%)	28 (2.6%)	0.457
FOLFOXIRI ^d + bevacizuma ^b ⁱ	8 (0.7%)	4 (3.1%)	4 (0.4%)	< 0.001
FOLFOXIRI ^d + cetuximab ^j	4 (0.3%)	–	4 (0.4%)	0.485
XELIRI ^e	1 (0.1%)	–	1 (0.1%)	0.727
XELO ^f + targeted drug ^k	1 (0.1%)	–	1 (0.1%)	0.727
XELOX ^g	24 (2.0%)	2 (1.6%)	22 (2.1%)	0.691
PD-1/PD-L1 inhibitors	19 (1.6%)	3 (2.3%)	16 (1.5%)	0.483
Postoperative chemotherapy	43 (3.6%)	13 (10.1%)	30 (2.8%)	< 0.001
Postoperative chemotherapy regimens				
FOLFIRI ^b	2 (0.2%)	/	2 (0.2%)	0.622
FOLFIRI ^b + targeted drug ^k	2 (0.2%)	1 (0.8%)	1 (0.1%)	0.074
mFOLFOX6 ^c	10 (0.8%)	1 (0.8%)	9 (0.8%)	0.932
mFOLFOX6 ^c + targeted drug ^k	19 (1.6%)	1 (0.8%)	18 (1.7%)	0.431
FOLFOXIRI ^d	22 (1.8%)	3 (2.3%)	19 (1.8%)	0.669
FOLFOXIRI ^d + bevacizumab ⁱ	8 (0.7%)	2 (1.6%)	6 (0.6%)	0.196
FOLFOXIRI ^d + cetuximab ^j	1 (0.1%)	–	1 (0.1%)	0.727
XELIRI ^e	1 (0.1%)	–	1 (0.1%)	0.727
XELO ^f + targeted drug ^k	3 (0.3%)	1 (0.8%)	2 (0.2%)	0.209
XELOX ^g	2 (0.2%)	–	2 (0.2%)	0.622
De Gramont ^h	3 (0.3%)	–	3 (0.3%)	0.546
PD-1/PD-L1 inhibitors	19 (1.6%)	2 (1.6%)	17 (1.6%)	0.966
Patients receiving prophylaxis (LMWH)	297 (24.9%)	36 (27.9%)	261 (24.5%)	0.806
Laboratory-level factors (dynamic temporal data)				
RBC (< 4 × 10 ¹² /L)	1133 (95.1%)	123 (96.9%)	1009 (94.9%)	0.332
HCT (< 0.37)	1184 (99.4%)	127 (99.2%)	1057 (99.4%)	0.762
HGB (< 110 g/L)	1042 (87.5%)	120 (93.8%)	922 (86.7%)	0.023
MCH (< 27 pg)	235 (19.7%)	21 (16.4%)	214 (20.1%)	0.317
MCHC (< 310 g/L)	30 (2.5%)	6 (4.7%)	24 (2.3%)	0.097
MCV (< 80 fL)	114 (9.6%)	7 (5.5%)	107 (10.1%)	0.095
MPV (> 11 fL)	209 (17.5%)	27 (21.1%)	182 (17.1%)	0.264
PDW-SD (< 9.8)	19 (1.6%)	3 (2.3%)	16 (1.5%)	0.474
RDW-CV (> 14.5)	687 (57.7%)	80 (62.5%)	607 (57.1%)	0.243
RDW-SD (< 35 fL)	3 (0.3%)	1 (0.8%)	2 (0.2%)	0.206
NRBC/100 WBC (> 0.1)	985 (82.7%)	106 (82.8%)	879 (82.7%)	0.972
NRBC (> 0.05 × 10 ⁹ /L)	94 (7.9%)	6 (4.7%)	88 (8.3%)	0.155
FPG (> 7.9 mmol/L)	513 (43.1%)	60 (46.9%)	453 (42.6%)	0.358
CO ₂ /CP (> 29 mmol/L)	14 (1.2%)	0	14 (1.3%)	0.192
CO ₂ /CP (< 22 mmol/L)	18 (1.5%)	1 (0.8%)	17 (1.6%)	0.474
PLT (> 350 × 10 ⁹ /L)	137 (11.5%)	14 (10.9%)	123 (11.5%)	0.815
D-dimer (> 0.5 µg/mL)	927 (77.6%)	128 (99.2%)	799 (75.0%)	< 0.001
FIB (> 4 g/L)	171 (14.3%)	25 (19.4%)	146 (13.7%)	0.082
PT (< 10 s)	14 (1.2%)	–	14 (1.3%)	0.190
PT-INR (< 1)	282 (23.6%)	20 (15.5%)	262 (24.6%)	0.022
aPTT (< 20 s)	3 (0.3%)	3 (2.3%)	–	< 0.001
TT (< 14 s)	387 (32.4%)	55 (42.6%)	332 (31.2%)	0.009
A/G (> 2.0)	15 (1.3%)	4 (3.1%)	11 (1.0%)	0.045
GLb (< 20 g/L)	123 (10.3%)	15 (11.7%)	108 (10.2%)	0.584
ALT (> 40 U/L)	361 (30.3%)	44 (34.4%)	317 (29.8%)	0.290
AST (> 40 U/L)	482 (40.5%)	61 (47.7%)	421 (39.6%)	0.080
γ-GT (> 50 U/L)	610 (51.2%)	69 (53.9%)	541 (50.9%)	0.519
AFP (> 400 µg/L)	23 (1.9%)	1 (0.8%)	22 (2.1%)	0.317
AKP (> 135 U/L)	12 (1.0%)	2 (1.6%)	10 (0.9%)	0.506
ALb (> 51 g/L)	–	–	–	–
PA (> 0.35 g/L)	–	–	–	–
TP (< 60 g/L)	931 (78.2%)	101 (78.9%)	830 (78.1%)	0.831
TG (> 1.81 mmol/L)	437 (36.7%)	53 (41.4%)	384 (36.1%)	0.241
CH (> 5.68 mmol/L)	–	–	–	–
Lp (a) (> 300 mg/L)	626 (52.6%)	67 (52.3%)	559 (52.6%)	0.958
ApoA1 (> 2.36 g/L)	–	–	–	–
ApoB (> 1.28 g/L)	–	–	–	–
LDL (> 3.36 mmol/L)	–	–	–	–
HDL (< 0.78 mmol/L)	248 (20.8%)	21 (16.4%)	227 (21.4%)	0.193
TBA (> 10.0 µmol/L)	5 (0.4%)	2 (1.6%)	3 (0.3%)	0.034
TBIL (> 17.1 µmol/L)	270 (22.7%)	42 (32.8%)	228 (21.4%)	0.004
DBIL (> 6.8 µmol/L)	365 (30.6%)	58 (45.3%)	307 (28.9%)	< 0.001
IBIL (> 10.2 µmol/L)	506 (42.5%)	47 (36.7%)	459 (43.2%)	0.162
U-BLD (+, µmol/L)	143 (12.0%)	13 (10.2%)	130 (12.2%)	0.495
U-BIL (> 14 µmol/L)	906 (76.1%)	96 (75.0%)	810 (76.2%)	0.764
URO (–, µmol/L)	7 (0.6%)	2 (1.6%)	5 (0.5%)	0.127
KET (+, mg/L)	340 (28.5%)	35 (27.3%)	305 (28.7%)	0.750
PRO (+, g/L)	39 (3.3%)	2 (1.6%)	37 (3.5%)	0.249
LEU (+,µ/L)	47 (3.9%)	6 (4.7%)	41 (3.9%)	0.648
BUN (> 8.8 mmol/L)	65 (5.5%)	5 (3.9%)	60 (5.6%)	0.413
USG (> 1.03)	4 (0.3%)	1 (0.8%)	3 (0.3%)	0.357
SQEP (> 21.4/µL)	–	–	–	–
BYST (µL)	86.4 ± 28.2	90.7 ± 22.8	85.92 ± 28.8	0.068
HYAL (µL)	2.1 ± 0.5	2.3 ± 0.5	2.1 ± 0.5	< 0.001
Cr (> 133.0 µmol/L)	674 (56.6%)	53 (41.4%)	621 (58.4%)	< 0.001
CYS-C (> 1.03 mg/L)	1129 (94.8%)	117 (91.4%)	1012 (95.2%)	0.068
RBP (> 57.9 mg/L)	5 (0.4%)	–	5 (0.5%)	0.437
UA (> 416 µmol/L)	5 (0.4%)	–	5 (0.5%)	0.437
WBC (> 11 × 10 ⁹ /L)	215 (18.0%)	40 (31.0%)	175 (16.4%)	< 0.001
LYM (> 3.5 × 10 ⁹ /L)	11 (0.9%)	–	11 (1.0%)	0.248
LYM (< 0.8 × 10 ⁹ /L)	37 (3.1%)	2 (1.6%)	35 (3.3%)	0.287
LYMR (> 0.4)	–	–	–	–
NEU (> 7 × 10 ⁹ /L)	1151 (96.6%)	125 (97.7%)	1026 (96.5%)	0.793
NEUR (> 0.7)	1186 (99.6%)	127 (99.2%)	1059 (99.6%)	0.503
MONO (> 0.8 × 10 ⁹ /L)	1017 (85.4%)	115 (89.8%)	902 (84.9%)	0.131
MONOR (> 0.08)	274 (23.0%)	37 (28.9%)	237 (22.3%)	0.093
EOS (> 0.5 × 10 ⁹ /L)	188 (15.8%)	20 (15.6%)	168 (15.8%)	0.958
EOSR (> 0.05)	8 (0.7%)	1 (0.8%)	7 (0.7%)	0.872
BASO (> 0.1 × 10 ⁹ /L)	299 (25.1%)	20 (15.6%)	279 (26.2%)	0.009
BASOR (> 0.005)	292 (24.5%)	37 (28.9%)	255 (24.0%)	0.222
PCT (> 0.5 ng/mL)	887 (74.5%)	97 (75.8%)	790 (74.3%)	0.720
CRP (> 10 mg/L)	778 (65.2%)	103 (79.8%)	675 (63.4%)	< 0.001
hs-CRP (> 3 mg/L)	1191 (100%)	–	–	–
NSE (> 5.4 µg/L)	–	–	–	–
MYO (> 75 ng/mL)	182 (15.2%)	22 (17.1%)	160 (15.0%)	0.544
HsTnI (> 0.2 ng/mL)	38 (3.2%)	10 (7.8%)	28 (2.6%)	0.002
CK (> 174 U/L)	150 (12.6%)	14 (10.9%)	136 (12.8%)	0.535
CKMB (> 25 U/L)	106 (8.9%)	18 (14.0%)	88 (8.3%)	0.032
LDH (> 245 U/L)	179 (15.0%)	29 (22.5%)	150 (14.1%)	0.012
α-HBDH (> 220 U/L)	48 (4.0%)	7 (5.4%)	41 (3.8%)	0.389

^a cycle of FLOT chemotherapy consisted of the following: Day 1: Intravenous (IV) leucovorin 200mg/m² in 2 h; IV oxaliplatin 85mg/m² in 120 min; IV docetaxel 50mg/m²; 5-fluorouracil (5-FU) 400 mg/m² bolus IV then 2400 mg/m² perfusion IV over 46 h. The next chemotherapy cycle was repeated on the 15th day. ^b A cycle of FOLFIRI chemotherapy consisted of the following: Day 1: IV irinotecan 180 mg/m² in 90 min; IV leucovorin 200mg/m² in 2 h; 5-FU 400 mg/m² bolus IV then 2400 mg/m² perfusion IV over 46 h. The next chemotherapy cycle was repeated on the 15th day. ^c A cycle of mFOLFOX6 consists of the following: Day 1: IV oxaliplatin 85mg/m² in 120 min; IV leucovorin 200 mg/m² in 2 h; 5-FU 400 mg/m² bolus IV then 2,400 mg/m² perfusion IV over 46 h. The next chemotherapy cycle was repeated on the 15th day. ^d A cycle of FOLFOXIRI consisted of the following: Day 1: IV irinotecan 165 mg/m² in 90 min; IV oxaliplatin 85mg/m² in 120 min; IV leucovorin 200mg/m² in 2 h; 5-FU 400 mg/m² bolus IV then 2,400 mg/m² perfusion IV over 46 h. The next chemotherapy cycle was repeated on the 15th day. ^e A cycle of XELIRI consists of the following: Day 1: IV irinotecan 165 mg/m² in 90 min; capecitabine 1,000 mg/m² PO twice daily for 14 out of 21 days. ^f A cycle of XELO consists of the following: Day 1: IV oxaliplatin 130mg/m² in 120 min; capecitabine 1,000 mg/m² PO twice daily for 14 out of 21 days. ^g A cycle of XELOX consists of the following: capecitabine 1,000 mg/m² PO twice daily for 14 out of 21 days. ^h A cycle of De Gramont consists of the following: Day 1: IV leucovorin 200mg/m² in 2 h; 5-FU 400 mg/m² bolus IV then 1,200 mg/m² perfusion IV over 46 h. The next chemotherapy cycle was repeated on the 15th day. ⁱ Bevacizumab (5 mg/kg every 14 days) and standard fluoropyrimidine-based chemotherapy. ^j Cetuximab (500 mg/m² IV every 14 days) and standard fluoropyrimidine-based chemotherapy. ^k Targeted drug: oral molecular target drugs including sorafenib, regorafenib, and fruquintinib. AFP, alpha-fetoprotein; A/G, albumin/globulin; AKP, alkaline phosphatase; ALb, albumin; ALT, alanine aminotransferase; ApoA1, apolipoprotein A1; ApoB, apolipoprotein B; aPTT, activated partial thromboplastin time; AST, aspartate aminotransferase; BASO, basophils; BASOR, basophils ratio; BMI, body mass index; BUN, blood urea nitrogen; BYST, Budding yeast cells; CH, cholesterol; CK, creatine kinase; CKMB, creatine kinase isoenzyme; CO₂/CP, carbon dioxide combining power; Cr, creatinine; CRP, C-reactive protein; hs-CRP, high sensitivity-C-reactive protein; CYS-C, cystatin C; DBIL, direct bilirubin; EOS, eosinophils; EOSR, eosinophils ratio; FPG, fasting glucose proxy measure; FIB, fibrinogen; GLb, globulin; γ-GT, γ-glutamyltranspeptidase; α-HBDH, α-hydroxybutyrate dehydrogenase; HCT, hematocrit; HGB, hemoglobin; HDL, high-density lipoprotein; HsTnI, high-sensitivity troponin I; HYAL, hyalinecasts; IBIL, indirect bilirubin; IQR, interquartile range; KET, ketone bodies; LDH, lactate dehydrogenase; LDL, low-density lipoprotein; LEU, urine leukocytes; LMWH, low molecular weight heparin; Lp (a), lipoprotein (a); LYM, lymphocytes; LYMR, lymphocytes ratio; MON, mean corpuscular hemoglobin; MCHC, mean cell hemoglobin concentration; MCV, mean corpuscular volume; MONO, monocyte; MONOR, monocyte ratio; MPV, mean platelet volume; MYO, myoglobin; NEU, neutrophils; NEUR, neutrophils ratio; NRBC, nucleated red blood cells; NSE, neuron-specific enolase; PA, prealbumin; PCT, Serum procalcitonin; PDW-SD, platelet distribution width-standard deviation; PLT, platelet; PRO, urinary protein; PT, prothrombin time; PT-INR, prothrombin time-international normalized ratio; RBC, red blood cells; RBP, plasma-retinol binding protein; RDW-CV, red blood cell distribution width-coefficient of variability; RDW-SD, red blood cell distribution width-standard deviation; SQEP, squamous epithelial cells; TBA, total bile acid; TBIL, total bilirubin; TG, total triglycerides; TP, serum total protein; TT, thrombin time; UA, uric acid; U-BLD, urine routine occult blood; U-BIL, urine routine bilirubin; URO, urobilinogen; USG, urine specific gravity; VTE, venous thromboembolism; WBC, white blood cells.

Table S3 Previous VTE risk assessment models and risk factors used as predictors

Risk factors in Caprini score	Relative risk score	Risk factors in Khorana score	Relative risk score
Age, 41–60	1	Site of cancer ^b	3
Age, 61–74	2	Very high risk (stomach, pancreas)	3
Age, 75+	3	High risk (lung, lymphoma, gynecologic, bladder, testicular)	3
Acute myocardial infarction	1	Prechemotherapy platelet count 350×10 ⁹ /L or more	3
Heart failure	1	Hemoglobin level less than 100 g/L or use of red cell growth factors	2
Varicose veins	1	Prechemotherapy leukocyte count more than 11×10 ⁹ /L	1
Obesity (BMI >25)	1	BMI ≥35	1
Inflammatory bowel disease	1		
Sepsis (within 1 month)	1		
COPD or abnormal pulmonary function	1		
Severe lung disease, including pneumonia (within 1 month)	1		
Oral contraceptives or hormone replacement therapy	1		
Pregnancy or postpartum (within 1 month)	1		
History of unexpected stillborn infant, recurrent spontaneous abortion (≥3), premature birth with toxemia or growth-restricted infant	1		
Medical patient currently at bed rest	1		
Minor surgery planned	1		
History of prior major surgery (within 1 month)	1		
Swollen legs	1		
Central venous catheter	2		
Arthroscopic surgery	2		
Major surgery (>45 min)	2		
Malignancy (present or previous)	2		
Laparoscopic procedure >45 min	2		
Patient confined to bed (>72 h)	2		
Immobilizing plaster cast (within 1 month)	2		
History of VTE ^c	3		
Positive Factor V Leiden; positive prothrombin G20210A; elevated serum homocysteine ^a	3		
Positive lupus anticoagulant ^a	3		
Heparin-induced thrombocytopenia	3		
Family history of VTE ^c	3		
Elevated anticardiolipin antibodies	3		
Stroke (within 1 month)	5		
Multiple trauma (within 1 month)	5		
Elective major lower extremity arthroplasty	5		
Hip, pelvis, or leg fracture (within 1 month)	5		
Acute spinal cord injury (paralysis) (within 1 month)	5		

^a, These risk factors cannot be tested at our site. ^b, Cancer patients with local or distant metastases and/or in whom chemotherapy or radiotherapy had been performed in the past 6 months. ^c, All types of VTE are included with the exception of superficial vein thromboembolism. BMI, body mass index; COPD, chronic obstructive pulmonary disease; VTE, venous thromboembolism.

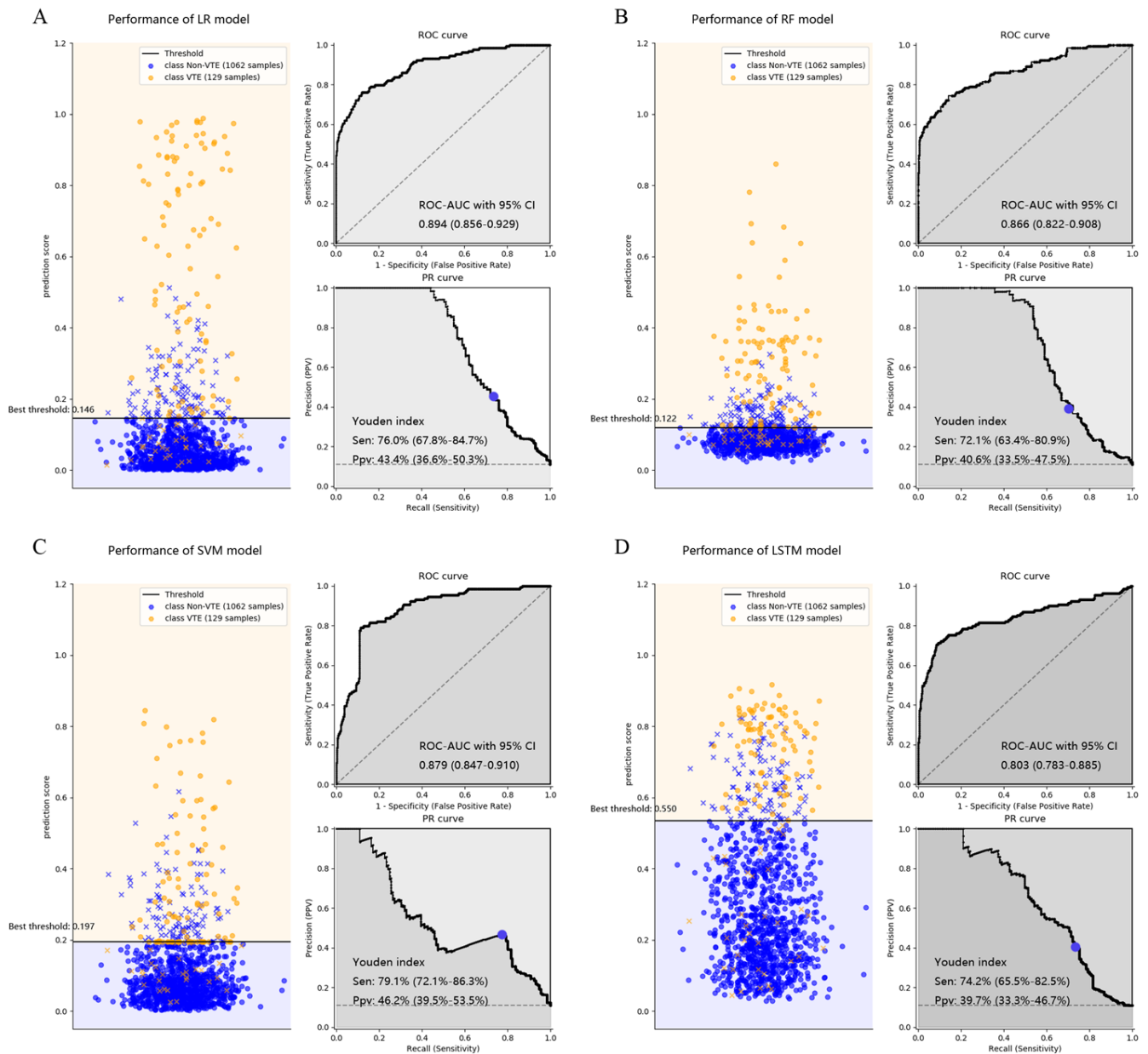


Figure S1 The model performance of the LR model (plot A), RF model (plot B), SVM (plot C), and LSTM (plot D). The classification based on the best threshold, the ROC curve and PR curve were plotted to measure the performance of the four machine learning models, and the AUCs were also calculated with 95% CIs. The best threshold points of these PR curves were plotted with corresponding sensitivities and positive predictive values. AUC, area under the curve; CI, confidence interval; LR, logistic regression; LSTM, long short-term memory; PPV, positive predictive value; PR, precision-recall curve; RF, random forest; ROC, receiver-operating characteristic curve; SEN, sensitivity; SVM, support vector machine; VTE, venous thromboembolism; *Youden* index = sensitivity + specificity - 1.