Asia-Pacific Journal of Oncology Nursing 9 (2022) 100128

Contents lists available at ScienceDirect

Asia-Pacific Journal of Oncology Nursing

journal homepage: www.apjon.org

Original Article

Development and validation of a machine learning model to predict venous thromboembolism among hospitalized cancer patients

Lingqi Meng[a,b], Tao Wei[b], Rongrong Fan[a], Haoze Su[c], Jiahui Liu[b], Lijie Wang[a], Xinjuan Huang[a], Yi Qi[b], Xuying Li[a,b,*]

[a] *Xiangya School of Nursing, Central South University, Changsha, China*
[b] *Hunan Cancer Hospital, The Affiliated Cancer Hospital of Xiangya School of Medicine, Central South University, Changsha, China*
[c] *Nanjing University of Aeronautics and Astronautics, Nanjing, China*

ARTICLE INFO

Keywords:
Venous thromboembolism
Machine learning
Cancer
Hospitalization
Model

ABSTRACT

Objective: Hospitalized cancer patients are at high risk of venous thromboembolism (VTE). However, no predictive model has been specifically developed for this population. Machine learning (ML) is advantageous for model development. This study was aimed at developing predictive models using three different ML algorithms and logistic regression for VTE risk among hospitalized cancer patients and comparing their predictive performance.
Methods: A retrospective case–control study was conducted on hospitalized cancer patients at Hunan Cancer Hospital, China, between October 1, 2021, and February 30, 2022. Patients diagnosed with vein thrombosis before or after admission were excluded. Patient, tumor, treatment, and laboratory indicator information was obtained from the hospital information system. The data were randomly split into distributions of 80% for training and 20% for testing. Logistic regression and three ML algorithms—the support vector machine, random forest, and extreme gradient boosting (XGBoost)—were used to develop the models. Model performance was compared using F1, G-mean, area under the receiver operating characteristic curve (AUROC), accuracy, precision, recall rate, and specificity. Feature rankings were achieved based on the permutation scores of the selected features in the optimal model.
Results: A total of 1100 patients (mean [SD] age, 54.75 [11.08] years; 485 [44.09%] male) were included in this study. There were 340 patients (30.9%) in the VTE group. The XGBoost model achieved the best performance with the following evaluation metrics: F1 (0.750), G-mean (0.816), AUROC (0.818), accuracy (0.845), precision (0.750), recall rate (0.750), and specificity (0.888). D-dimer level, diabetes, hypertension, pleural metastasis, and hematological malignancies were identified as the five most significant features of the XGBoost model.
Conclusions: Four predictive models were developed using ML algorithms. The XGBoost model was the optimal predictive model compared with the other three models. This study indicates that ML may play an important role in VTE risk estimation among hospitalized patients with cancer and provides a reference for thromboprophylaxis.

Introduction

Venous thromboembolism (VTE), including deep venous thromboembolism (DVT) and pulmonary embolism (PE), is a common and potentially fatal disease. It is acknowledged that cancer is associated with an increased risk of VTE.[1] Cancer patients have a four- to seven-fold higher risk of developing VTE than patients without cancer.[1] VTE is the second leading cause of death in cancer patients, resulting in an increased negative impact on patients' prognosis and consumption of medical resources.[2,3] Moreover, hospitalization significantly increases the occurrence of VTE in cancer patients.[4] The absolute incidence of VTE in hospitalized cancer patients is estimated to be between 2% and 17%.[5]

The American Society of Hematology (ASH) guideline panel suggested using pharmacological thromboprophylaxis for hospitalized cancer patients without VTE.[6] However, it is not routinely applied owing to the increased risk of bleeding.[7] This implies that it is critical to weigh the

[*] Corresponding author.
E-mail address: 1259734926@qq.com (X. Li).

https://doi.org/10.1016/j.apjon.2022.100128
Received 28 May 2022; Accepted 30 July 2022

risks of VTE and bleeding. Thus, the National Institute for Health and Care Excellence (NICE) suggested assessing the VTE risks using VTE risk assessment tools and performing a personalized anticoagulant therapy.[8]

Many risk assessment tools have been developed to predict the risk of VTE in cancer patients. Two, in particular, are widely applied: the Khorana score (KS)[9] and the prospective Comparison of Methods for Thromboembolic Risk Assessment with Clinical Perceptions and AwareneSS in Real Life Patients-Cancer Associated Thrombosis (COMPASS-CAT) score.[10] Unfortunately, both have been derived from populations of outpatients with cancer and showed poor predictive performance in assessing the VTE risk of hospitalized cancer patients, such as the Chinese population.[11–13] We did not find a risk assessment model specifically developed for hospitalized cancer patients.[14] Therefore, it is necessary to develop a special predictive model to evaluate VTE risk in hospitalized cancer patients.

Machine learning (ML) is a subdiscipline of artificial intelligence that can help researchers handle big and complicated data, find regularities, and make predictions. It has advantages in data processing and induction because it can train a large amount of sample data compared to traditional statistical methods.[15] Recently, ML methods have been applied successfully to evaluate VTE risk among trauma patients and patients with peripherally inserted central catheters.[16,17] Support vector machine (SVM), random forest (RF), and extreme gradient boosting (XGBoost) are the most widely used ML algorithms for VTE prediction. Some studies have shown that SVM, RF, and XGBoost are the most efficient ML algorithms for classification problems of VTE prediction, respectively.[18–20] Additionally, they are classic and representative algorithms based on different machine learning methods. Therefore, we assumed that the three MLs were effective analytical methods for predicting VTE risk among hospitalized cancer patients. As in the literature, we used logistic regression as a baseline classifier for comparison.[18–20]

Given the above challenges and the advantages of ML methods, it has important clinical implications for developing a predictive model for hospitalized cancer patients using ML methods. It helps nurses to identify high-risk patients with VTE and provides a reference for thromboprophylaxis.

This study had two aims: (1) to develop four predictive models to evaluate VTE risk among hospitalized cancer patients using three ML algorithms (SVM, RF, and XGBoost) and logistic regression; (2) to validate and compare the predictive performance of the four models and identify the optimal model with the best performance.

Methods

Setting and data sources

This retrospective case–control study was conducted from October 1, 2021, to February 30, 2022. Data were obtained from Hunan Cancer Hospital in China. Data collection was performed using the Hospital Information System (HIS) from October 1 to December 15, 2021. The inclusion criteria were as follows: Participants (1) had valid hospital records between December 1, 2017, and November 30, 2020, (2) were aged 18 years and older, (3) were diagnosed with malignant cancers, and (4) were receiving treatment in the hospital, including surgery, chemotherapy, or radiotherapy. Participants were excluded if they had been diagnosed with VTE and/or superficial vein thrombosis before or upon admission. This study was conducted in accordance with the Declaration of Helsinki. The Ethics Committee of Hunan Cancer Hospital approved the study (Approval No. KYJJ-2021-291) and waived the requirement for informed consent owing to the retrospective design of the study. This study followed the transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD).

Variables

Potential predictors of VTE were selected on the basis of the results of previous studies.[21–24] There were 120 potential predictors (input variables) divided into four separate categories: patient, tumor, treatment, and laboratory indicator characteristics. Patient characteristics included age, body mass index, sex, performance status, blood type, comorbid diseases (such as lung disease, heparin-induced thrombocytopenia, peripheral vascular disease, and coronary artery disease), pathological symptoms and signs (such as fracture, acid-base poisoning, swollen legs, serious infection, and phlebitis), past medical history (eg., thrombosis, smoking, drinking, and tumor), time (admission, cancer diagnosis, and VTE diagnosis), pregnancy, and postpartum. Tumor characteristics included the primary site of cancer, pathological classification, metastasis site, and stage of the tumor. Treatment characteristics included course information with chemotherapy (times and regimen), surgery, puncture operation, admission to the ICU, endocrine therapy, hormone therapy, granulocyte colony-stimulating factor, transfusion of blood, hypertonic irritating drugs, vascular access, and plaster cast. Laboratory indicators included carcinoembryonic antigen, carbohydrate antigen-125, glycosylated hemoglobin, white blood cells, red blood cells, hemoglobin, platelets, prothrombin time, international normalized ratio, fibrinogen concentration, D-dimer, antithrombin III, fibrin degradation products, high-sensitivity C-reactive protein, total cholesterol, blood chlorine, and serum homocysteine. For nonsurgical inpatients, the first laboratory indicators after admission were used. For inpatients who underwent surgery, laboratory indicators were the first laboratory examination indices after the first surgery. Patients diagnosed with VTE before surgery were treated as nonsurgical patients. All the input variables are presented in Appendix A**.**

The outcome variable was VTE, defined as DVT and/or PE diagnosed at Hunan Cancer Hospital. The diagnosis of DVT was based on positive findings on vascular Doppler ultrasound. The diagnosis of PE was based on positive findings of computed tomography pulmonary angiography (CTPA), spiral computed tomography, and high probability ventilation/perfusion scanning.

All records were generated when patients were discharged. Data were extracted strictly and separately from medical records by two qualified training nurses.

Data preprocessing

The problem of highly imbalanced data existed between the non-VTE population (94,027 patients) and the VTE population (583 patients). Two approaches were taken to solve the expected classification bias. At the data level, we under-sampled the non-VTE population to balance the majority and minority class.[25,26] Stratified sampling was performed in all eligible non-VTE patients according to the month of the first admission.[27] The random sampling number in every stratification was identified as 2.5 times the number of participants in the VTE group in the same month. Random sampling was performed using a “random sample of cases” in SPSS (version 26.0). At the algorithm level, cost-sensitive learning techniques were performed. We added more weight to the minority class (VTE group) to equalize the VTE and non-VTE groups.[28] The class_weight function in the sklearn package and scale_pos_weight function in the xgboost package were used for this purpose. We deleted a column of features if more than 20% of the data was missing. The missing data for the remaining features were imputed using the mean value or mode.

Feature selection

The samples were split into a training set (80%) for model development and a testing set (20%) for model validation. Random assignment to the training or testing set was stratified according to VTE status. Feature selection was performed during the training. Important features were filtered using Pearson's chi-squared test or Student's *t*-test. The features were standardized to realize feature scaling. Recursive feature elimination (RFE) was then employed to screen the optimized variable combinations.

Development of models

The following four algorithms were chosen to develop the predictive

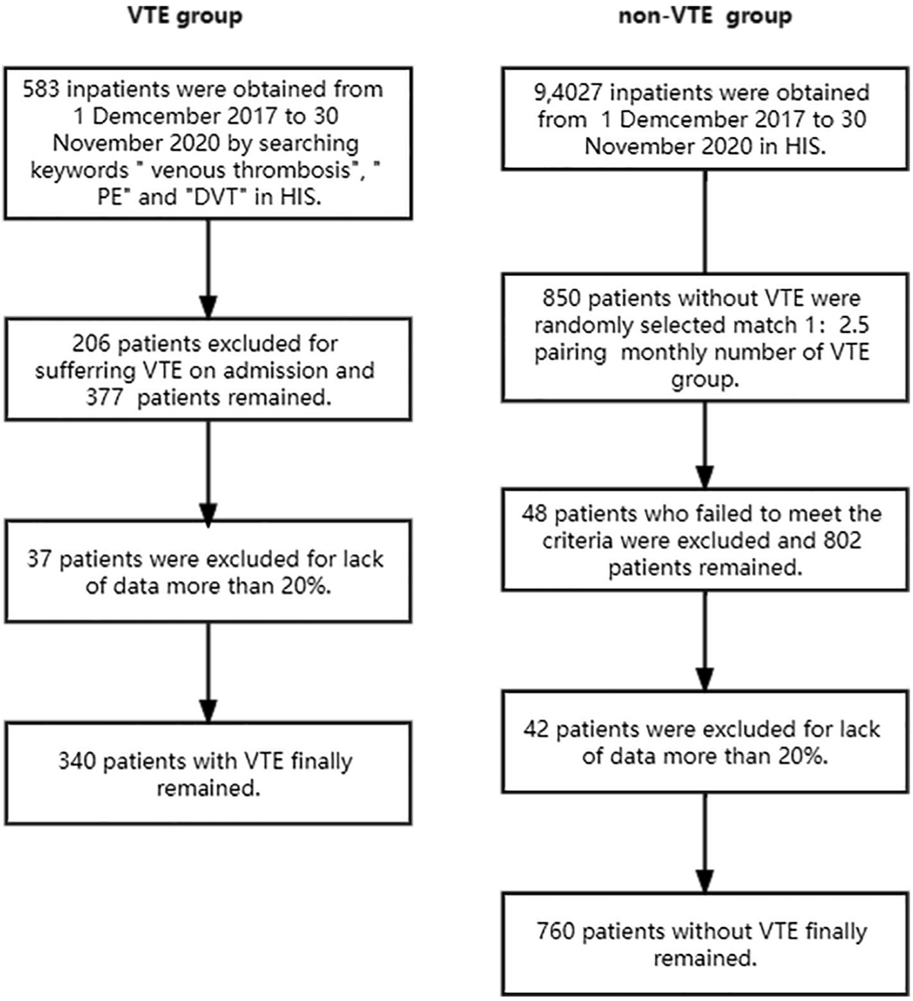
VTE group
non-VTE group
583 inpatients were obtained from 1 Demcember 2017 to 30 November 2020 by searching keywords " venous thrombosis", " PE" and "DVT" in HIS.
206 patients excluded for sufferring VTE on admission and 377 patients remained.
37 patients were excluded for lack of data more than 20%.
340 patients with VTE finally remained.
9,4027 inpatients were obtained from 1 Demcember 2017 to 30 November 2020 in HIS.
850 patients without VTE were randomly selected match 1: 2.5 pairing monthly number of VTE group.
48 patients who failed to meet the criteria were excluded and 802 patients remained.
42 patients were excluded for lack of data more than 20%.
760 patients without VTE finally remained.

Fig. 1. Flowchart diagram. VTE, venous thromboembolism; DVT, deep venous thromboembolism; HIS, Hospital Information System.

models: a classical multivariate statistical method (logistic regression) and three other classic MLs (SVM, RF, and XGBoost). Logistic regression is a conventional statistical method with good interpretable ability. The SVM is a representative algorithm based on a kernel. It exhibits superior performance in nonlinear classification problems.[29] RF is a representative ensemble bagging algorithm. It controls overfitting problems in the decision tree.[30] XGBoost is a representative algorithm based on ensemble boosting that complements the overfitting problem of the gradient boosting model.[31]

The optimal parameters in the four models were retrospectively identified using 10-fold cross-validation. The logistic regression, SVM, and RF models were implemented using logistic regression, SVR, and random forest classifier in the sklearn package, respectively. The XGBoost model was implemented using the xGBoost package.

Performance and validation of models

Data in the testing set were measured to assess the predictive performance of the four models. Considering the data imbalance, the predictive performances of the four models were compared based on the three evaluation metrics, including F1, G-mean, and the area under the receiver operating characteristic curve (AUROC) to select the optimal model.[32,33] Other evaluation metrics included confusion metrics, accuracy, precision, recall, and specificity. The confusion matrices included four indicators: true negative, false positive, false negative, and true positive.[32] Accuracy was calculated as follows [True positive + True negative]/[True positive + False negative + True negative + False positive]. Precision was calculated as True positive/[True positive + False positive]. The recall was calculated as True positive/[True positive + False negative]. Specificity was calculated using the formula: Ture negative/[True negative + False positive]. F1 was calculated as (2 × precision × recall)/(precision + recall). G-mean was calculated by $\sqrt{}$(Recall × Specificity).[32] The AUROC was calculated using the AUROC curve, which is a graphical plot showing the diagnostic capability of a binary classifier as its discrimination threshold changes.[34]

Feature rankings in the optimal model

Feature rankings were performed according to the permutation scores of the selected features. The absolute magnitude of a permutation score represents the effect of the feature on the model performance, determined by the difference in the AUROC before and after the alteration of the feature in the model. This process was repeated for each feature selected in the model. Permutation scores were obtained from the optimal model using data from the testing set.

Data analysis

Data were prepared using SPSS (version 26.0) from December 15, 2021, to January 1, 2022. Data processing and analysis were conducted using SPSS and the sklearn package in Python (version 3.7) from January 1 to February 15, 2022.

Table 1
Performance results of four models in the testing set.

Method	TN	FP	FN	TP	Acc	Pre	Rec	Spe	F1	G-mean	AUROC
Logistic regression	116	36	17	51	0.759	0.586	0.750	0.763	0.669	0.756	0.757
SVM	119	33	18	50	0.768	0.602	0.735	0.783	0.661	0.759	0.759
RF	136	16	28	40	0.800	0.714	0.588	0.895	0.644	0.725	0.743
XGBoost	135	17	17	51	0.845	0.750	0.750	0.888	0.750	0.816	0.818

TN, true negative; FP, false positive; FN, false negative; TP, true positive; Acc, accuracy; Pre, precision; Rec, recall rate; Spe, specificity; AUROC, the area under the receiver operating characteristic curve; SVM, support vector machine; RF, random forest; XGBoost, extreme gradient boosting.

Table 2
The area under the receiver operating characteristic curve (AUROC) of the training set and testing set.

Method	Training set (AUROC, 95%CI)	Testing set (AUROC, 95%CI)
Logistic regression	0.823 (0.796, 0.850)	0.757 (0.689, 0.816)
SVM	0.828 (0.801, 0.854)	0.759 (0.697, 0.818)
RF	1.0 (1.0, 1.0)	0.743 (0.678, 0.808)
XGBoost	0.976 (0.965, 0.987)	0.818 (0.762, 0.870)

SVM, support vector machine; RF, random forest; XGBoost, extreme gradient boosting.

Results

Study population characteristics

A total of 1100 patients (mean [SD] age, 54.75 [11.08] years; 485 [44.09%] male) were included in the study. A flowchart of patient enrollment is shown in Fig. 1. There were 340 patients who developed symptomatic, image-confirmed DVT and/or PE, with an average age of 55.54 ± 11.09. A total of 760 patients, with an average age of 54.40 ± 11.07, were not diagnosed with DVT or PE. In the VTE group, 142 (41.8%) patients were male. In the non-VTE group, 343 (45.1%) patients were male. Patient characteristics are shown in Appendix B. Missing data were available for 204 of the participants. Details of missing data processing are provided in Appendix A of the supplement.

Feature selection and development of models

We deleted invalid features and imputed missing data for some features (in Appendix A). After data pre-processing, 108 features were retained for feature selection. The top 90 (90/108) most important features were chosen using univariate analysis (in Appendix B and C). After RFE, 33, 24, 32, and 35 features were selected for the logistic regression, SVM, RF, and XGBoost models, respectively. Appendix D shows the results of the feature selection for the four classifiers. Appendix E shows the optimal parameters for model performance in the four models.

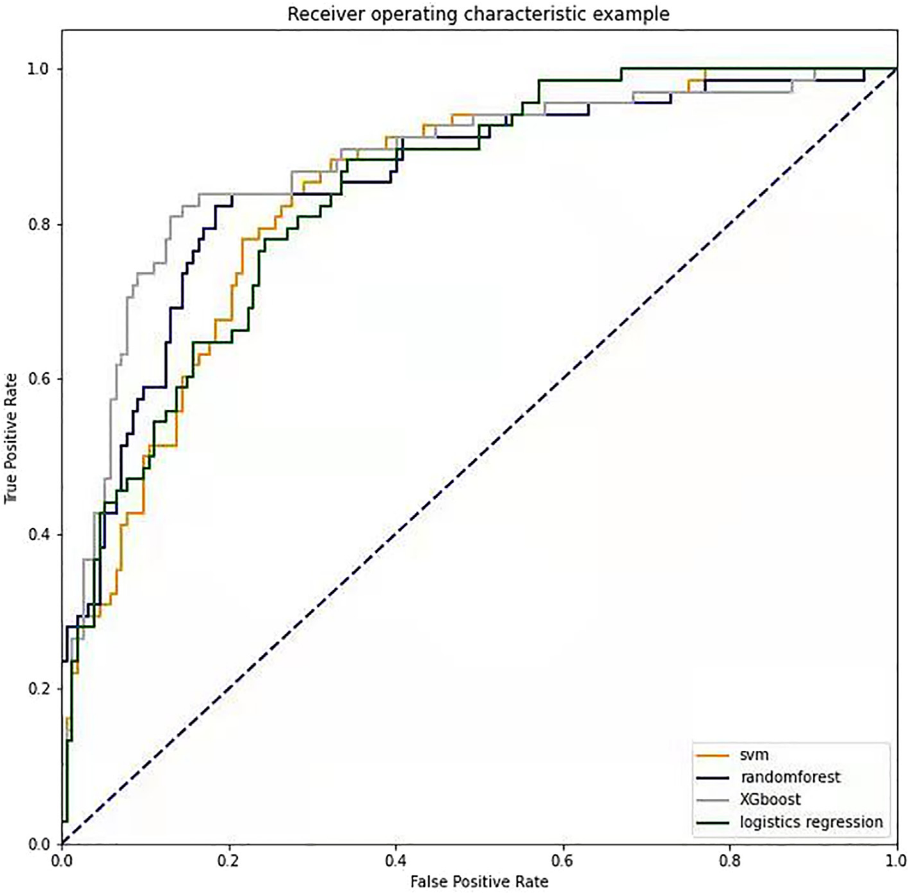

Fig. 2. The area under the receiver operating characteristic (AUROC) curve of four models in the testing set. SVM, support vector machine mode; RF, random forest model; XGBoost, extreme gradient boosting model.

Performance and validation of models

Table 1 presents the performance results of the four models. Overall, the XGBoost model achieved the best performance, with the highest values of F1 (0.750), G-mean (0.816), AUROC (0.818), accuracy (0.845), precision (0.750), and recall (0.750). Table 2 shows the AUROC of the training and testing sets. Overfitting occurred in the RF model in the training set. No overfitting occurred in the other models, which indicated that the models had a high generalization ability. Fig. 2 shows the AUROC curves generated using the four models for the testing set. Among these four models, the XGBoost model had the largest AUROC.

Feature rankings in XGBoost model

The relative scaled importance of the features in the XGBoost model is shown in Fig. 3 and Appendix F. After ranking the importance of features, the results showed that D-dimer level, diabetes, hypertension, pleural metastasis, and hematological malignancies were the top five important risk factors for VTE in hospitalized cancer patients.

Discussion

In this study, ML models were used to assess VTE risk in hospitalized cancer patients in the Chinese population. To our knowledge, this is one of the earliest studies to use ML to evaluate VTE risk in hospitalized cancer patients. According to the validation of the predictive performance of the models, all models could effectively predict the risk of VTE. Our results demonstrated that ML offers a novel solution for early forecasting and evaluation of VTE risk in inpatients with cancer. We also found that the XGBoost model had the best predictive performance among the four models based on the AUROC.

The logistic regression, SVM, RF, and XGBoost models included 33, 24, 32, and 35 features, respectively. This demonstrates that the ML model has advantages when dealing with numerous clinical features.

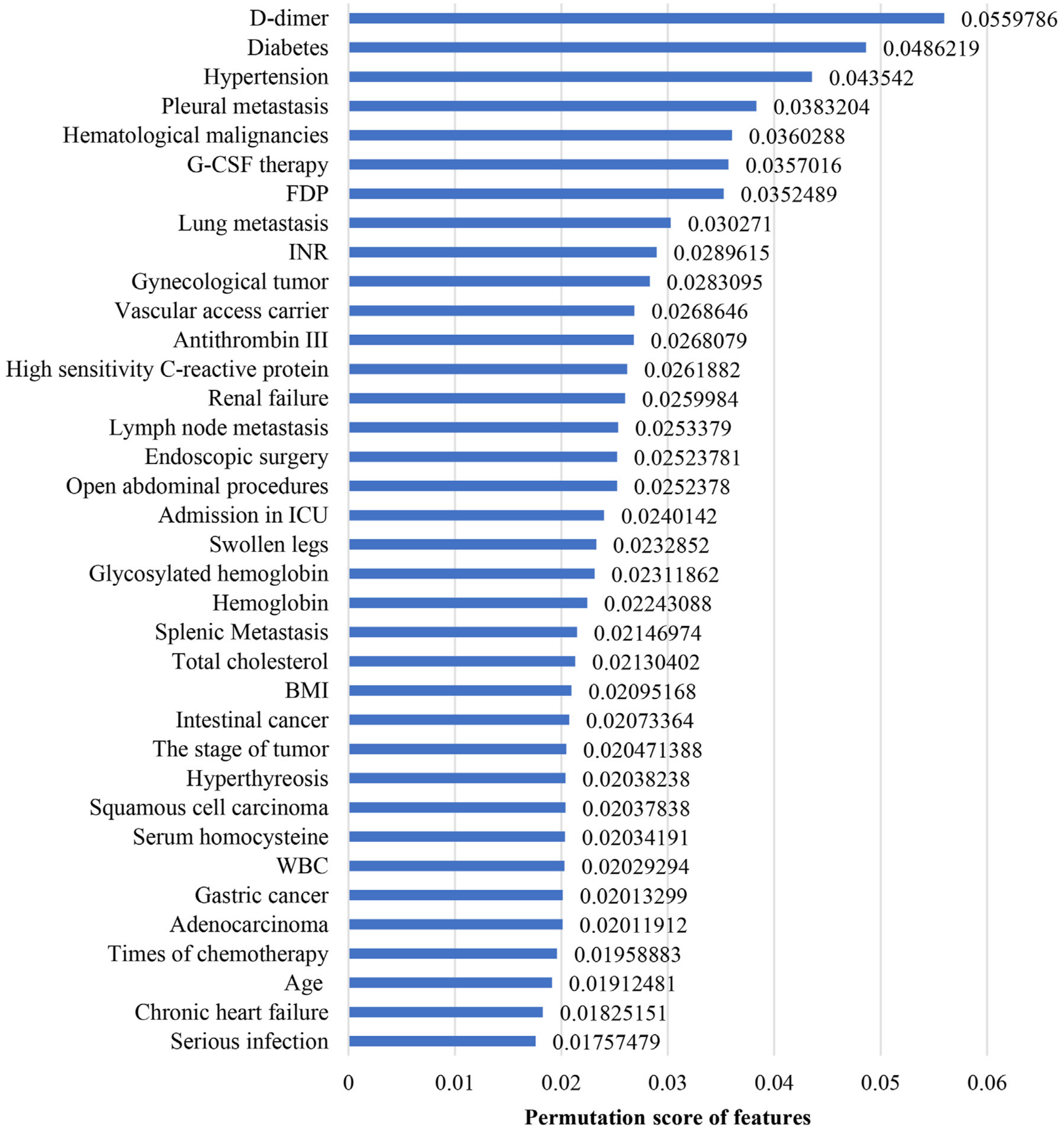

Fig. 3. Permutation score of the most important features in the XGBoost model in descending order. G-CSF, granulocyte colony-stimulating factor; FDP, fibrin degradation products; INR, international normalized ratio; BMI, body mass index; WBC, white blood cell.

Although there are many features in the model, most of them can be captured from the HIS. It is feasible to incorporate a predictive model into the HIS to automatically calculate the risk value. This reduces the clinical burden of assessing VTE risk, thus improving the efficiency of assessment.

The results provide some risk predictors for VTE in hospitalized patients with cancer. There were some common features in each ML model, including diabetes, renal failure, distant metastasis, lymph node metastasis, frequency of chemotherapy, D-dimer, fibrin degradation products, international normalized ratio, vascular access carrier, pleural metastasis, hematological malignancies, gynecological tumor, and squamous cell carcinoma. Diabetes and renal failure are essential risk factors for VTE, which is consistent with the previous evidence.[35,36] The results showed that cancer-related features, such as distant metastasis (particularly pleural metastasis), lymph node metastasis, type of tumor (particularly hematological malignancies and gynecological tumor), squamous cell carcinoma, and time of chemotherapy, influenced VTE risk in the cancer population. It's supported by Weitz et al.[37] Coagulation indicators (such as fibrin degradation products, international normalized ratio, and D-dimer) revealed the predictive performance for VTE, which is consistent with the findings of Posch et al.[38]

Several studies have demonstrated that the KS is a useful risk stratification tool for cancer inpatients in Canadian and American populations, etc.[39,40] Compared to KS, our XGBoost model has some advantages. First, our model includes more patient characteristics, such as surgery, anti-cancer treatment, and comorbidities, which are associated with the occurrence of VTE in hospitalized cancer patients, as supported by previous studies.[41–43] It is important to incorporate more associated variables to improve prediction modeling. Second, laboratory indicators on admission were collected as features rather than as those before chemotherapy. We evaluated the risk of cancer-associated VTE based on laboratory indicators on admission. This solved the problem of obtaining data before chemotherapy to calculate the KS. Third, ML techniques may be more effective in developing a prediction model compared to the logistic regression used in the KS. MLs can handle complex and nonlinear large datasets. Moreover, they have the unique ability to model data by appling Boolean logic, absolute conditionality, conditional probabilities, and other unconventional strategies while they still could draw heavily on statistics and probabilities.[44]

Limitations

First, several variables were deleted because their missing values exceeded 20%. The deletion of variables may neglect their potential effects on models, resulting in reduced predictive performance. Examples of these variables include hemorrhage and coagulation disorder, acute spinal cord injury, and carbohydrate antigen-125. It would be interesting to explore the potential effects of such features in future research. Second, ML models lack transparency, manifesting in interpretable algorithms, and invisible training set data. Thus, it is difficult to interpret the effects of selected features on the model. However, the effect of each feature on model performance can be visualized by permutation scores. Finally, the outcome variable was identified retrospectively using discharge diagnosis codes. In most cases, asymptomatic VTE events are not detected. Therefore, the predictive model may have a limited ability to identify asymptomatic VTE. External validation is required to evaluate the generalization ability of the models.

ML models can evaluate the VTE risk of hospitalized cancer patients manually, which provides support for practitioners' decisions. Ranking the importance of the selected features could help nurses and patients to understand the effect of such features on the occurrence of VTE. ML models can be conveniently applied in clinical practice if they are incorporated into the HIS. However, their application in clinical practice is suitable only when the ML models have extremely high stability and accuracy. Future efforts to optimize the performance of the models using different methods are needed, such as training the models with larger databases, adjusting the parameters in the models, and developing models with a combination of multiple ML algorithms.

Conclusions

Using ML methods, four predictive models were used to evaluate the VTE risk among hospitalized cancer patients. The XGBoost model was found to best predict VTE risk compared with the other three models. This study indicates that MLs may play an important role in risk estimation in this era of big data. ML models could provide a reference for clinicians and nurses to assess the risk levels of hospitalized patients with cancer.

Acknowledgments

The authors would like to acknowledge the funding support of the Scientific research projects of Chinese Nursing Association and the Project supported by Provincial Natural Science Foundation of Hunan.

Authors' contributions

Concept and design: Lingqi Meng, Xuying Li. Acquisition of data: Lijie Wang, Yi Qi, Jiahui Liu, Lingqi Meng. Analysis and interpretation of data: Lingqi Meng, Haoze Su. Drafting of the manuscript: Lingqi Meng, Xuying Li. Critical revision of the manuscript: All authors. Dr. Li had full access to all the data in the study and take responsibility for the integrity of the data and the accuracy of the data analysis.

Declaration of competing interest

None declared.

Funding

This work was supported by the Scientific research projects of Chinese Nursing Association (Grant No. ZHKY202106) and Project supported by Provincial Natural Science Foundation of Hunan (Grant No. 2020JJ8018).

Ethics statement

This study was approved by the Ethics Committee of Hunan Cancer Hospital (Approval No. KYJJ-2021-291).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.apjon.2022.100128.

References

1. Louzada ML, Majeed H, Dao V, Wells PS. Risk of recurrent venous thromboembolism according to malignancy characteristics in patients with cancer-associated thrombosis: a systematic review of observational and intervention studies. *Blood Coagul Fibrinolysis*. 2011;22:86–91.
2. Sorensen HT, Mellemkjaer L, Olsen JH, Baron JA. Prognosis of cancers associated with venous thromboembolism. *N Engl J Med*. 2000;343:1846–1850.
3. Lyman GH, Culakova E, Poniewierski MS, Kuderer NM. Morbidity, mortality and costs associated with venous thromboembolism in hospitalized patients with cancer. *Thromb Res*. 2018;164:S112–S118.
4. Khorana AA, Francis CW, Culakova E, Kuderer NM, Lyman GH. Frequency, risk factors, and trends for venous thromboembolism among hospitalized cancer patients. *Cancer-Am Cancer Soc*. 2007;110:2339–2346.
5. Patell R, Zwicker JI. Inpatient prophylaxis in cancer patients: where is the evidence? *Thromb Res*. 2020;191:S85–S90.
6. Lyman GH, Carrier M, Ay C. American Society of Hematology 2021 guidelines for management of venous thromboembolism: prevention and treatment in patients with cancer. ***Blood Adv***. 2021;5, 1953-1953.

7. Zhai Z, Kan Q, Li W, et al. VTE risk profiles and prophylaxis in medical and surgical inpatients: the identification of Chinese hospitalized patients' risk profile for venous thromboembolism (DissolVE-2)-A cross-sectional study. ***Chest***. 2019;155:114–122.
8. Treasure T, Hill J. NICE guidance on reducing the risk of venous thromboembolism in patients admitted to hospital. *J R Soc Med*. 2010;103:210–212.
9. Khorana AA, Kuderer NM, Culakova E, Lyman GH, Francis CW. Development and validation of a predictive model for chemotherapy-associated thrombosis. ***Blood***. 2008;111:4902–4907.
10. Gerotziafas GT, Taher A, Abdel-Razeq H, et al. A predictive score for thrombosis associated with breast, colorectal, lung, or ovarian cancer: the prospective COMPASS-cancer-associated thrombosis study. *Oncol*. 2017;22:1222–1231.
11. Hu Y, Li X, Zhou H, et al. Comparison between the Khorana prediction score and Caprini risk assessment models for assessing the risk of venous thromboembolism in hospitalized patients with cancer: a retrospective case control study. *Interact Cardiovasc Thorac Surg*. 2020;31:454–460.
12. Jin S, Qin D, Liang BS, et al. Machine learning predicts cancer-associated deep vein thrombosis using clinically available variables. *Int J Med Inf*. 2022;161:104733.
13. Xiong W, Zhao YF, Du H, Wang YM, Xu M, Guo XJ. Optimal authoritative risk assessment score of Cancer-associated venous thromboembolism for hospitalized medical patients with lung Cancer. *Thromb J*. 2021;19:95.
14. Khorana AA, DeSancho MT, Liebman H, Rosovsky R, Connors JM, Zwicker J. Prediction and prevention of cancer-associated thromboembolism. *Oncol*. 2021;26: E2–E7.
15. Na KS, Kim YK. The application of a machine learning-based brain magnetic resonance imaging Approach in major depression. ***Adv Exp Med Biol***. 2021;1305: 57–69.
16. Liu S, Zhang F, Xie L, et al. Machine learning approaches for risk assessment of peripherally inserted Central catheter-related vein thrombosis in hospitalized patients with cancer. ***Int J Med Inf***. 2019;129:175–183.
17. He L, Luo L, Hou X, et al. Predicting venous thromboembolism in hospitalized trauma patients: a combination of the Caprini score and data-driven machine learning model. ***BMC Emerg Med***. 2021;21:60.
18. Lei H, Zhang M, Wu Z, et al. Development and validation of a risk prediction model for venous thromboembolism in lung cancer patients using machine learning. ***Frontiers in Cardiovascular Medicine***. 2022:9845210.
19. Liu H, Yuan H, Wang YM, Huang WW, Xue H, Zhang XY. Prediction of venous thromboembolism with machine learning techniques in young-middle-aged inpatients. *Sci Rep*. 2021;11:12868.
20. Lu C, Song J, Li H, et al. Predicting venous thrombosis in osteoarthritis using a machine learning algorithm: a population-based cohort study. *J Personalized Med*. 2022;12:114.
21. Dhami SPS, Patmore S, O'Sullivan JM. Advances in the management of cancer-associated thrombosis. *Semin Thromb Hemost*. 2021;47:139–149.
22. Razouki ZA, Ali NT, Nguyen VQ, Escalante CP. Risk factors associated with venous thromboembolism in breast cancer: a narrative review. *Supportive Care in Cancer*. 2022.
23. Ryan L, Mataraso S, Siefkas A, Pellegrini E, Barnes G, Green-Saxena A, et al. A Machine Learning Approach to Predict Deep Venous Thrombosis Among Hospitalized Patients. *Clinical and Applied Thrombosis-Hemostasis*. 2021, 271076029621991185.
24. Wang P, Wang Y, Yuan Z, et al. Venous thromboembolism risk assessment of surgical patients in Southwest China using real-world data: establishment and evaluation of an improved venous thromboembolism risk model. *BMC Med Inf Decis Making*. 2022; 22:59.
25. Mazurowski MA, Habas PA, Zurada JA, Lo JY, Baker JA, Tourassi GD. Training neural network classifiers for medical decision making: the effects of imbalanced datasets on classification performance. *Neural Network*. 2008;21:427–436.
26. Zhang J, Chen L. Clustering-based undersampling with random over sampling examples and support vector machine for imbalanced classification of breast cancer diagnosis. ***Computer Assisted Surgery***. 2019;24:62–72.
27. Ramezan CA, Warner TA, Maxwell AE. Evaluation of sampling and cross-validation tuning strategies for regional-scale machine learning classification. *Rem Sens*. 2019; 11:185.
28. Li FL, Zhang XY, Zhang XQ, Du CL, Xu Y, Tian YC. Cost-sensitive and hybrid-attribute measure multi-decision tree over imbalanced data sets. *Inf Sci*. 2018;422:242–256.
29. Cao L, Shen H. *Ieee. Virtual Sample Generation Approach for Imbalanced Classification*. 2018:177–182.
30. George AAP, Lacerda M, Syllwasschy BF, Hopp MT, Wissbrock A, Imhof D. HeMoQuest: a webserver for qualitative prediction of transient heme binding to protein motifs. *BMC Bioinf*. 2020;21:124.
31. Chen TQ, Guestrin C. *Assoc Comp M. XGBoost: A Scalable Tree Boosting System*. 2016: 785–794.
32. Hasan N, Bao YK. Comparing different feature selection algorithms for cardiovascular disease prediction. *Health Technol*. 2021;11:49–62.
33. Wu JC, Shen J, Xu M, Shao ML. A novel combined dynamic ensemble selection model for imbalanced data to detect COVID-19 from complete blood count. *Comput Methods Progr Biomed*. 2021:211106444.
34. Liu WF, Wang SF, Ye ZH, Xu PP, Xia XT, Guo MG. Prediction of lung metastases in thyroid cancer using machine learning based on SEER database. *Cancer Med*. 2022; 11:2503–2515.
35. Bai J, Ding X, Du XH, Zhao XF, Wang ZQ, Ma ZQ. Diabetes is associated with increased risk of venous thromboembolism: a systematic review and meta-analysis. ***Thromb Res***. 2015;135:90–95.
36. Janus N, Mahe I, Launay-Vacher V, Laroche JP, Deray G. Renal function and venous thromboembolic diseases. ***J Mal Vasc***. 2016;**41**:389–395.
37. Weitz JI, Haas S, Ageno W, et al. Cancer associated thrombosis in everyday practice: perspectives from GARFIELD-VTE. ***J Thromb Thrombolysis***. 2020;50:267–277.
38. Posch F, Riedl J, Reitter EM, et al. Dynamic assessment of venous thromboembolism risk in patients with cancer by longitudinal D-Dimer analysis: a prospective study. ***J Thromb Haemostasis***. 2020;18:1348–1356.
39. Parker A, Peterson E, Lee AYY, et al. Risk stratification for the development of venous thromboembolism in hospitalized patients with cancer. *J Thromb Haemostasis*. 2018; 16:1321–1326.
40. Patell R, Rybicki L, McCrae KR, Khorana AA. Predicting risk of venous thromboembolism in hospitalized cancer patients: utility of a risk assessment tool. *Am J Hematol*. 2017;92:501–507.
41. Anderson DR, Morgano GP, Bennett C, et al. American Society of Hematology 2019 guidelines for management of venous thromboembolism: prevention of venous thromboembolism in surgical hospitalized patients. ***Blood Adv***. 2019;3:3898–3944.
42. Abdel-Rahman O, Wu C, Easaw J. Risk of arterial and venous thromboembolic events among patients with colorectal carcinoma: a real-world, population-based study. ***Future Oncol***. 2021;17, 3977-+.
43. Lyman GH, Bohlke K, Khorana AA, et al. Venous thromboembolism prophylaxis and treatment in patients with cancer: American society of clinical oncology clinical practice guideline update 2014. ***J Clin Oncol***. 2015;33, 654-U174.
44. Kourou K, Exarchos TP, Exarchos KP, Karamouzis MV, Fotiadis DI. Machine learning applications in cancer prognosis and prediction. *Comput Struct Biotechnol J*. 2015;13: 8–17.