BMC Medical Informatics and Decision Making

RESEARCH

Open Access

A new risk assessment model of venous thromboembolism by considering fuzzy population

Xin Wang[1,3†], Yu-Qing Yang[5†], Xin-Yu Hong[2,3], Si-Hua Liu[2,3], Jian-Chu Li[1], Ting Chen[4,6] and Ju-Hong Shi[2*]

Abstract

Background Inpatients with high risk of venous thromboembolism (VTE) usually face serious threats to their health and economic conditions. Many studies using machine learning (ML) models to predict VTE risk overlook the impact of class-imbalance problem due to the low incidence rate of VTE, resulting in inferior and unstable model performance, which hinders their ability to replace the Padua model, a widely used linear weighted model in clinic. Our study aims to develop a new VTE risk assessment model suitable for Chinese medical inpatients.

Methods 3284 inpatients in the medical department of Peking Union Medical College Hospital (PUMCH) from January 2014 to June 2016 were collected. The training and test set were divided based on the admission time and inpatients from May 2016 to June 2016 were included as the test dataset. We explained the class imbalance problem from a clinical perspective and defined a new term, "fuzzy population", to elaborate and model this phenomenon. By considering the "fuzzy population", a new ML VTE risk assessment model was built through population splitting. Sensitivity and specificity of our method was compared with five ML models (support vector machine (SVM), random forest (RF), gradient boosting decision tree (GBDT), logistic regression (LR), and XGBoost) and the Padua model.

Results The 'fuzzy population' phenomenon was explained and verified on the VTE dataset. The proposed model achieved higher specificity (64.94% vs. 63.30%) and the same sensitivity (90.24% vs. 90.24%) on test data than the Padua model. Other five ML models couldn't simultaneously surpass the Padua's sensitivity and specificity. Besides, our model was more robust than five ML models and its standard deviations of sensitivities and specificities were smaller. Adjusting the distribution of negative samples in the training set based on the 'fuzzy population' would exacerbate the instability of performance of five ML models, which limited the application of ML methods in clinic.

Conclusions The proposed model achieved higher sensitivity and specificity than the Padua model, and better robustness than traditional ML models. This study built a population-split-based ML model of VTE by modeling the class-imbalance problem and it can be applied more broadly in risk assessment of other diseases.

Keywords Venous thromboembolism, Risk assessment, Machine learning, Fuzzy population

†Xin Wang and Yu-Qing Yang contributed equally to this work.

*Correspondence:
Ju-Hong Shi
shijh@pumch.cn
[1]Department of Ultrasound, Peking Union Medical College Hospital, Beijing, China
[2]Department of Respiration, Peking Union Medical College Hospital, No.1, Shuaifuyuan, Dongcheng District, Beijing 100730, China
[3]Chinese Academy of Medical Sciences, Peking Union Medical College, Beijing, China
[4]Computer Science and Technology, Tsinghua University, Beijing, China
[5]State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications, Beijing, China
[6]Beijing, China

Introduction

Venous thromboembolism (VTE), comprising deep venous thrombosis (DVT) and pulmonary embolism (PE), is a life-threatening disease associated with more than one-half million hospitalizations in the United States each year, and is a contributing cause in 100,000 or more deaths [1, 2]. As a common cardiovascular disease, VTE often leads to complications including recurrent VTE, post-pulmonary embolism syndrome, chronic thromboembolic pulmonary hypertension, and post-thrombotic syndrome, causing heavy burden to both life quality and economy [3].

Prophylaxis against VTE such as anticoagulant drugs, graduated compression stockings and venous foot pump can reduce mortality efficiently. Studies have shown that appropriate prevention can lower patients' VTE incidence from 10.5 to 14.9% to 5.5–5.6% in medical department and also reduce VTE events in surgical departments [4–6]. Since hypercoagulability is one of the most important VTE risk factors, anticoagulant drugs realize the disease prevention by changing coagulation status for high VTE risk patients [7]. However, it may also cause bleeding events and even death, especially for patients with low VTE risk, who is not hypercoagulabe. Therefore, to recognize patients with high VTE risk clinically is vital and a precise VTE risk assessment model is needed to guide prevention.

The American College of Chest Physicians recommended Padua risk assessment model, a linear model consisting of 11 VTE risk factors, to stratify VTE risk for medical inpatients [8, 9]. Inpatients with Padua score of no less than 4 points were considered as high-risk and recommended to receive prophylaxis. However, due to the close correlations between VTE and race, genetic background and disease spectrum, studies have shown that the Padua model, which was derived based on the Western population, is not suitable for Chinese inpatients [10, 11]. Thus, it is necessary to establish a model suitable for Chinese inpatients.

With the rapid development of artificial intelligence technology, machine learning (ML) models are increasingly employed in medical research [12, 13]. Support vector machine (SVM), random forest (RF), gradient boosting decision tree (GBDT), logistic regression (LR) and XGBoost have been proposed to do VTE risk assessment, but most of them trained models by proportional random selection of VTE and non-VTE patients (e.g. 1:1 or K-fold cross-validation) [14–16]. Wang, et al. compared multiple ML models by training them on 188 VTE and 188 non-VTE patients, and showed that performances of ML models were instable and their sensitivities were lower than the Padua [17]. In addition, interpretability of ensemble-based models such as RF, GBDT and XGBoost was limited though their predictive performances were relatively well.

Due to the low incidence rate of VTE, the number of VTE patients is significantly lower than non-VTE patients. This class imbalance problem greatly impacts the construction and performance of ML models. Previous studies usually adjusted the ratio of positive and negative samples through simple oversampling or undersampling without analyzing and handling the class imbalance from a clinical perspective [14–17]. These approaches lead to unstable model performance, resulting in models lacking robustness and making them challenging to apply in clinical practice.

Furthermore, we attempt to explain the reason of class imbalance problem from the perspective of clinical medicine, and refer to it as the "fuzzy population" phenomenon. The 'fuzzy population' means that the doctor cannot recognize the patient population from the normal population precisely based on the characteristics of patients, due to the low incidence rate and the mechanism complexity of specific disease (Fig. 1). There are some individuals with the same or very similar clinical characteristics to those of patients, but they don't suffer from the disease. Thus, the probability statistical method is widely used to do the risk stratification for patient and normal populations according to values of multiple clinical variables, and within the groups with relative high risk, there is still a certain percentage of normal individuals. It is important to note that the existence of 'fuzzy population' phenomenon is not due to the insufficient number of clinical features we use or the lack of modeling ability, but to the limitation of our understanding of the disease. The existence of this phenomenon implies that within a certain period of time, only a minority of patients visiting the hospital will develop the disease, while the majority of patients are negative samples, leading to a class imbalance issue in the dataset.

Considering the problem of Padua model and the 'fuzzy population' phenomenon, to build a new VTE risk assessment model for Chinese inpatients, this study proposed a population-split-based approach as shown in Fig. 2. This approach first split patients into different groups according to their values of feature vectors, then filtered out trustless groups, and finally trained ML models in the unit of groups. It explores relationship between VTE events and combinations of clinical variables by constructing different groups, which shows advantages of robustness and good performance. Then our model was compared with Padua and multiple traditional ML models on real clinical dataset to verify its efficiency, indicating the potential to help clinicians evaluate VTE risk and guide prevention.

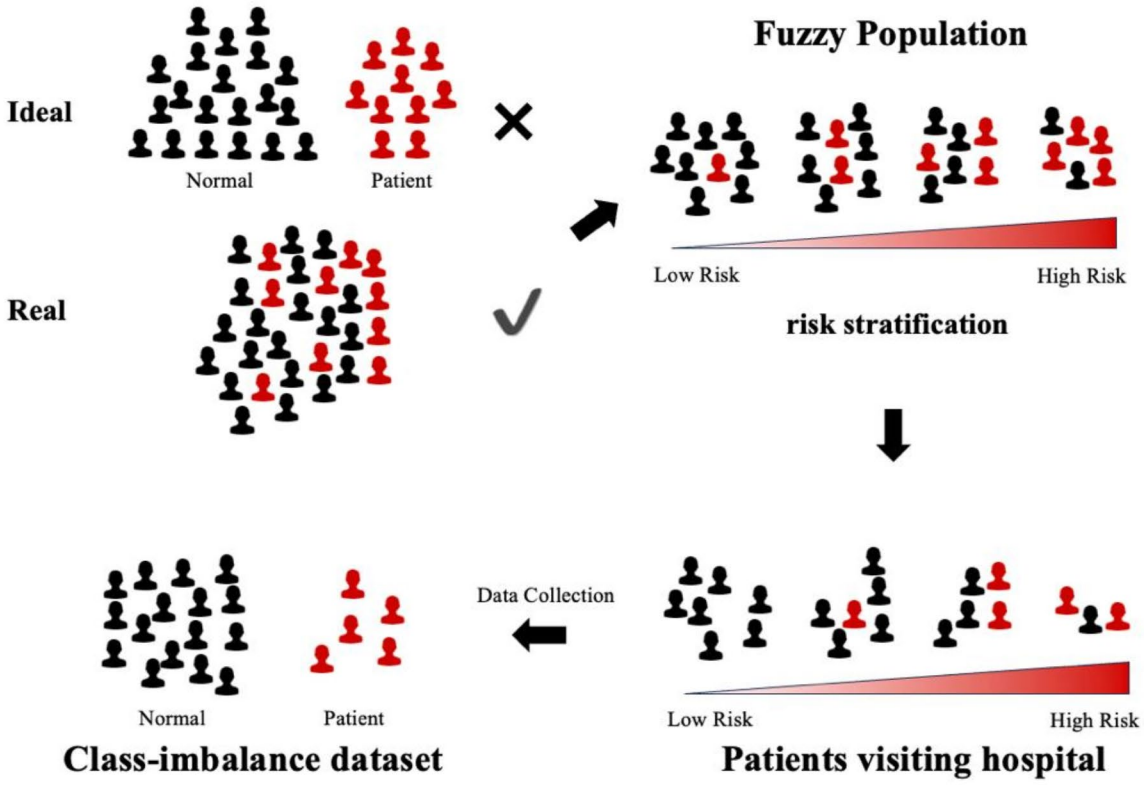

Fig. 1 The visualization of "fuzzy population" and class imbalance problem

Materials and methods

Population

This study analyzed inpatients who developed VTE, including DVT and PE, and partially non-VTE patients in medical department of Peking Union Medical College Hospital (PUMCH) from January 2014 to June 2016. All these inpatients from May 2016 to June 2016 were included as a test dataset for model verification and the other patients formed the training dataset. All the enrolled patients met the following inclusion/exclusion criteria: inclusion criteria: over 18 years old, hospital stay≥72 h; exclusion criteria: receiving anticoagulation medicine (e.g., therapeutic dose of low-molecular-weight heparin for treatment of acute myocardial infarction) other than the anticoagulation regimen for VTE diagnosed during the hospitalization.

DVT was diagnosed as the presence of intraluminal blocking or filling defects in the deep veins of the upper or lower limbs evidenced by venography or deep vein thrombogenesis illustrated by color Doppler ultrasonography. PE was diagnosed either as the presence of intraluminal blocking and/or filling defects in the pulmonary arteries by pulmonary angiography, computed tomographic pulmonary arteriography or magnetic resonance, or by radionuclide lung ventilation-perfusion scans showing multiple pulmonary segmental perfusion defects. This study was approved by the Ethics Committee of PUMCH in Chinese Academy of Medical Sciences (reference number for ethics approval: B164).

Variable selection

The categorical variables involved in modeling are VTE risk factors, including active cancer, previous VTE, reduced mobility, thrombophilia, recent trauma and/or surgery, age≥70 years, heart and/or respiratory failure, acute myocardial infarction or ischemic stroke, acute infection and/or rheumatologic disorder, obesity, ongoing glucocorticoid treatment, hormone replacement therapy including estrogen or progesterone, mechanical ventilation [11].

The fuzzy population and its effect on model construction

The 'fuzzy population' can greatly affect the construction of ML models. Generally, positive or negative samples, namely patients or normal individuals, in the training dataset of ML models were random selected proportionally from all data. When we sample only positive or negative samples, or change the original ratio of positive and negative samples in the local sample space, these will affect model's risk prediction for the samples within this local mathematic space, which fluctuate the sensitivity and specificity of the trained ML model (Fig. 3). For

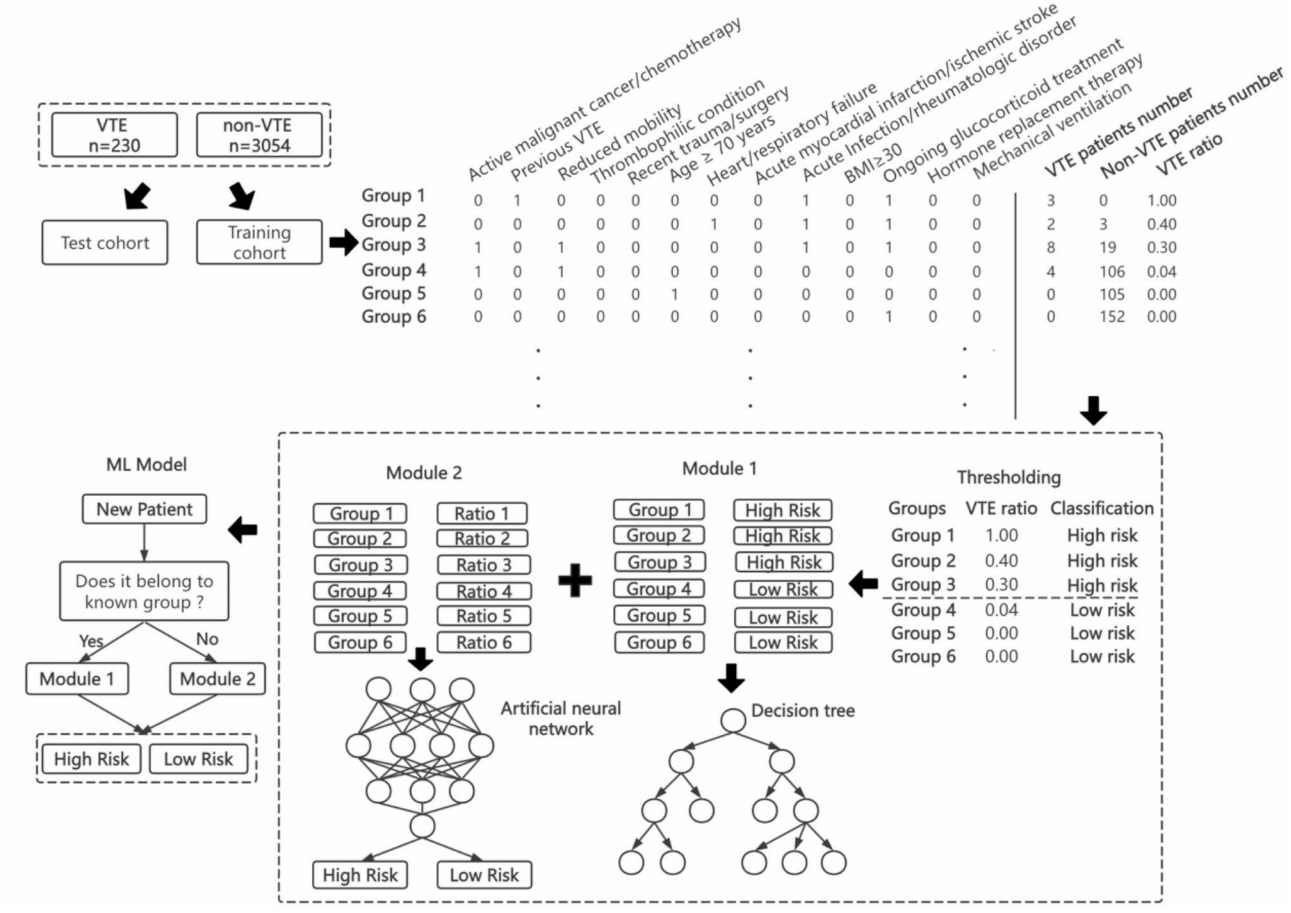

Fig. 2 The schema of proposed VTE risk assessment ML approach. Firstly, training and test cohorts were constructed and patients in training data were split into different groups according to values of VTE-related clinical variables. For different group $C^{(i)}$ and $C^{(j)}$, their corresponding feature vectors $v^{(i)}$ and $v^{(j)}$ satisfied $v^{(i)} \neq v^{(j)}$. Then *VTE risk ratio* was calculated in every group and groups were sorted accordingly. Next probability of distribution of patients in each group was estimated using VTE incidence rate and only groups with probability < 0.05 were saved. Based on sorted result, accumulated sensitivities and specificities were calculated for every group and groups were classified into high and low risks by thresholding, which formed a new training set based on groups. Using this training set, the proposed model consists of two modules, group-memory module for patients in known groups and group-prediction module for the unknown. Decision tree was used in group-memory module. For group-prediction module, VTE ratios for groups were used instead of high or low risk label, and artificial neural network was fitting

example, for each patient and surrounding normal individuals with the same or very similar features, if only normal individuals are included in the training dataset due to the sampling bias, the model will tend to predict individuals with such characteristics as the negative or low risk during model training, which leads to a decrease in the model's sensitivity and increasing in specificity in this local sample space.

Taking the VTE as an example to further explain the 'fuzzy population'. Risk factors in Padua model are all categorical variables. VTE and non-VTE individuals can share the same values of all risk factors. Let $x_i = (x_{i0}, x_{i1}, x_{i2}, \ldots, x_{iK-1})$ be the feature vector of the i^{th} patient with K clinical variables where $x_{ij} \in \{0,1\}$ represents value of the j^{th} variable, and $y_i \in \{0,1\}$ indicates whether the individual has VTE. The 'fuzzy population' phenomenon means that there are the i^{th} and j^{th} samples,

$$x_{ik} = x_{jk}, \text{ for k} = 0, \ldots, \text{K} - 1 \text{ but } y_i \neq y_j,$$

and it results in many groups with different combinations of values of clinical variables. Assuming that by analyzing all data in a hospital for a certain period, there are n individuals with the same feature vector $\text{v} = (v_0, v_1, \ldots, v_{K-1})$ and,

$$x_{0k} = x_{1k} = \ldots = x_{n-1\,k} = v_k, \; for\; k = 0, \ldots, K-1.$$

VTE events occurred in m persons,

$$\sum_{i=0}^{n-1} y_i = \text{m}, \text{ for i} \in \text{C}, \; |\text{C}| = \text{n},$$

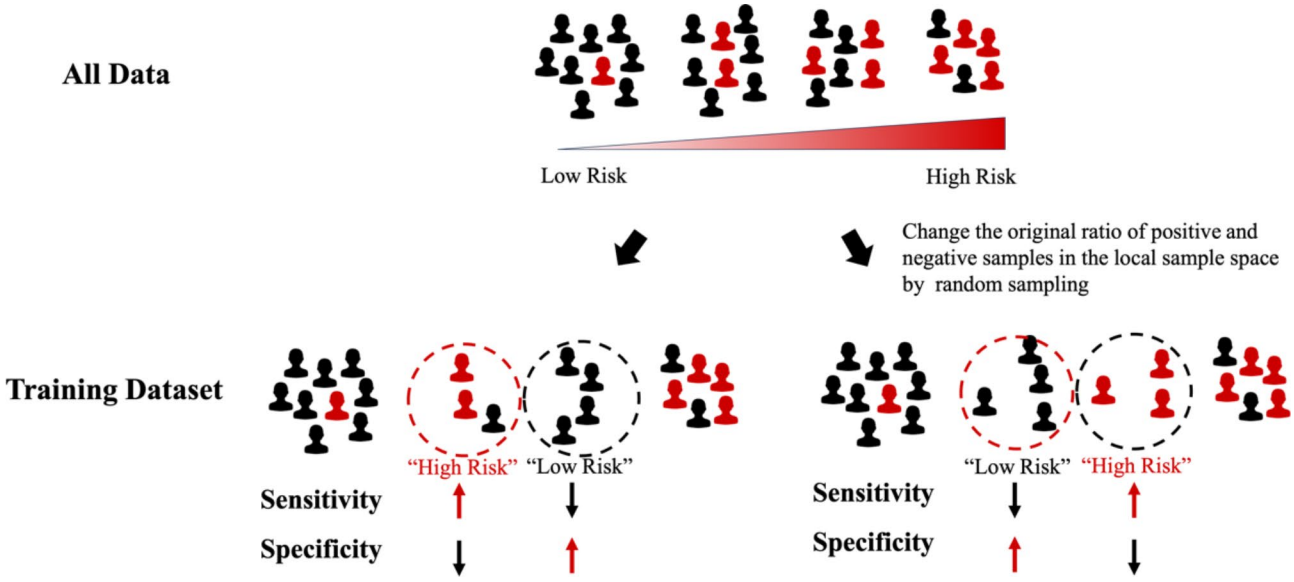

Fig. 3 Influence of the "fuzzy population" on the construction of ML model

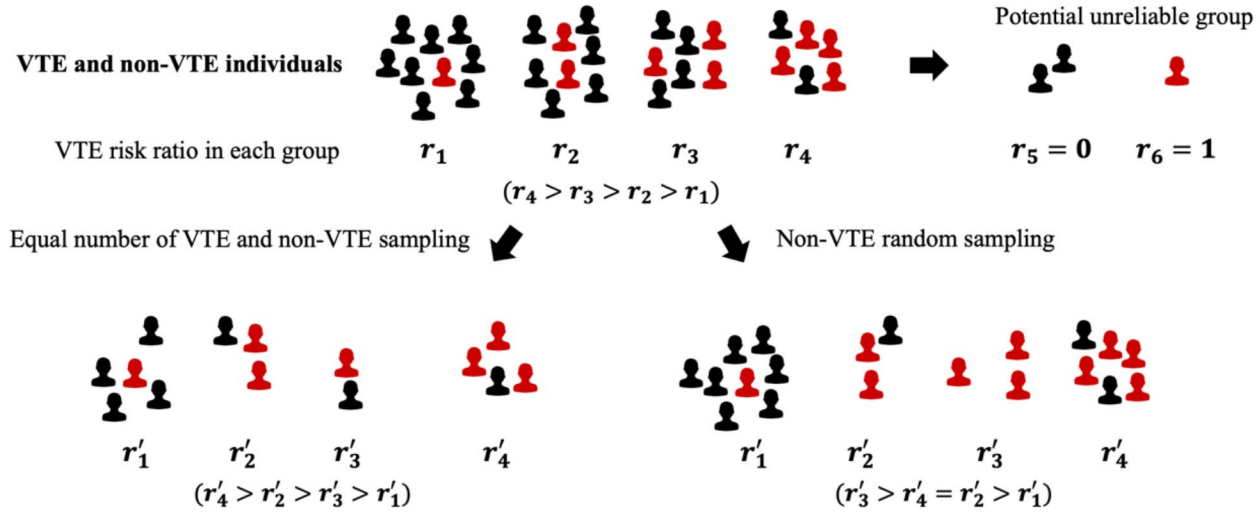

Fig. 4 The inconsistent distribution of VTE risk ratios of groups and unreliable groups by random sampling for ML model

where n persons with identical feature vector v make up the unique group C. Then we can define the *VTE risk ratio* for the group C as $r = m/n$, which represents the risk of disease in individuals with similar clinical characteristics.

The existence of this phenomenon limits the performance of ML model in two aspects (Fig. 4). On one hand, most reported VTE ML models were trained using equal number of VTE and non-VTE individuals or all VTE patients and random sampling non-VTE individuals, which was not consistent with distribution of the real clinical dataset. In these training set, the estimation of *VTE risk ratio* of the group C, denoted as r', may deviate the real value r, which represents the *VTE risk ratio* in the real population. When $r' \gg r$, the ML model tends to predict the sample with feature vector v to be high risk and vice versa. Besides, when the training set includes more samples from the group C with feature vector v, the number of samples from other groups with vector v' $(v' \neq v)$ tends to be less, which will influence the prediction of samples with feature vector v'. Therefore, it is unreasonable to construct the training set by traditional approaches for ML model. On the other hand, when the number of collected samples in group C is small, estimation of its real *VTE risk ratio* r may be unreliable. Especially, due to the low incidence rate of VTE, non-VTE individuals are more likely to be observed in a group. If there is a group C with relatively high *VTE risk ratio* r but only a small number of non-VTE samples from it were

collected, ML model based on it will predict patients of group C as low risk, which reduces model's sensitivity.

Population clustering analysis and population split

In order to learn patterns of distribution of VTE patients, population clustering analysis was performed using 16 features, including 13 VTE risk factors, Padua score, Padua high risk, and the number of non-zero risk factors. Inspired by the clustering analysis results, patients were split into different groups based on values of 13 risk factors. Denote the dataset consisted of N patients as $X = \{(x_i, y_i)\}_{i=0}^{N-1}$, where $x_i = (x_{i1}, x_{i2}, x_{i3}, \ldots, x_{iK})$ represented feature vector of the i^{th} individual, x_{ik} is binary variable, and $y_i \in \{0,1\}$. Now the dataset X was split into T groups $\mathrm{C} = \left\{C^{(t)}\right\}_{t=0}^{T-1}$, and for the t^{th} group, $n^{(\mathrm{t})} = \left|C^{(t)}\right|$ represented its number of individuals. For every sample belonged to $C^{(t)}$, their values of feature vector were the same $v^{(t)} = (v_0^{(t)}, v_1^{(t)}, \ldots, v_K^{(t)})$. For any two groups $C^{(t)}$ and $C^{(q)}$, we have

$$v^{(t)} \neq v^{(q)},\ for\ any\ C^{(t)} \in C,\ C^{(q)} \in C,\ and\ t \neq q.$$

Calculation of *VTE risk ratio* and group filtering

After splitting the dataset into T groups, *VTE risk ratio* for each group was calculated. For group $C^{(t)}$, let $m^{(\mathrm{t})} = \sum_{j \in C^{(t)}} y_j$ be the number of VTE patients, and its *VTE risk ratio* was $r^{(t)} = \frac{m^{(\mathrm{t})}}{n^{(\mathrm{t})}}$. To remove potential unreliable groups which couldn't represent real *VTE risk ratio*, the probability of including $m^{(\mathrm{t})}$ VTE patients among $n^{(\mathrm{t})}$ patients for group $C^{(t)}$ was computed using the incidence rate of VTE,

$$P^{(\mathrm{t})} = R^{m^{(\mathrm{t})}}(1-R)^{n^{(\mathrm{t})}-m^{(\mathrm{t})}},$$

where R was the VTE incidence rate in whole population. Then groups with $P^{(\mathrm{t})} \geq \mathrm{threshold}$ (e.g. 0.05) were filtered out and the remaining groups were saved to train model.

Training dataset construction

Considering the effect of the 'fuzzy population', training set was built in unit of group $C^{(t)}$ instead of patient (x_i, y_i). Firstly, all saved groups were sorted by *VTE risk ratio* $r^{(t)}$, number of VTE patients $m^{(\mathrm{t})}$, and [-1* number of non-VTE patients $(n^{(\mathrm{t})} - m^{(\mathrm{t})})$]. Then accumulated sensitivities and specificities were calculated from group $C^{(0)}$ to $C^{(T-1)}$. Groups after filtering were denoted as $C_f = \left\{C_f^{(t)}\right\}_{t=0}^{T_f-1}$ and $T_f = |C_f|$ was the number of groups. For the t^{th} group $C_f^{(t)}$, its accumulated sensitivity and specificity were

$$\mathrm{Sen}^{(t)} = \frac{\sum_{i=0}^{t} m^{(\mathrm{i})}}{\sum_{h=0}^{T_f-1} m^{(\mathrm{h})}} \text{ and } \mathrm{Spec}^{(t)} = 1 - \frac{\sum_{i=0}^{t} n^{(\mathrm{i})} - m^{(\mathrm{i})}}{\sum_{h=0}^{T_f-1} n^{(\mathrm{h})} - m^{(\mathrm{h})}}.$$

Next groups were classified into high and low VTE risks by thresholding values of $\mathrm{Spec}^{(t)}$ (e.g. 75%). The groups with $\mathrm{Spec}^{(t)} \geq threshold$ were recognized as the high risk with $y^{(t)} = 1$, the other groups were the low risk with $y^{(t)} = 0$. Finally, model training set $X_f = \{(v^{(t)}, y^{(t)})\}_{t=0}^{T_f-1}$ with T_f samples was constructed.

Derivation of new model

In the training set, labels of groups were assigned based on statistical analysis, and they were regarded as the ground truth, or known knowledge. For patients from known groups $C_f^{(t)}$, the VTE risk could be obtained simply by looking up a table consisted of all groups $(v^{(t)}, y^{(t)})$. For patients from unknown groups $C^{unknown}$,

$$v^{(unknown)} \neq v^{(t)},\ for\ any\ C_f^{(t)} \in C_f,\ t \in \{0,1,\ldots,T_f-1\}.$$

Reasonable and accurate VTE risk prediction for these unknown patients was needed based on results of known groups, which was the goal of training a ML model.

Thus, the proposed VTE risk assessment model consisted of two modules, the group-memory module for patients from known groups and group-prediction module for patients from unknown groups. For group-memory module, a decision tree model was used to record all pairs $(v^{(t)}, y^{(t)})$ from X_f, and contributions of risk factors could be analyzed by comparing feature weights. For group-prediction module, an artificial neural network (ANN) was used to fit the relationship between feature vector of group $v^{(t)}$ and *VTE risk ratio* $r^{(t)}$. By comparing goodness of fit, the optimal ANN with the highest R^2 was selected, and patient with predicted VTE ratio ≥ 0.5 was recognized as the high VTE risk.

Model evaluation and comparison with other ML methods

To verify the proposed model's efficiency, five traditional ML models including SVM, RF, GBDT, LR [18], and XGBoost [19], and Padua model (Table 1), were compared. Five ML models were trained in the popular method, on the same training patients as the proposed model, and considering that the number of non-VTE patients were larger than the number of VTE, non-VTE patients equal to the number of VTE patients were randomly selected to construct 1:1 training set. For five ML models, 10-fold cross validation was used and the

Table 1 Padua risk assessment model

Risk factor	Score
Active malignant cancer/chemotherapy	3
Previous VTE	3
Reduced mobility	3
Thrombophilic condition	3
Recent trauma/surgery	2
Age > = 70	1
Heart/respiratory failure	1
Acute myocardial infarction/ischemic stroke	1
Acute infection/rheumatologic disorder	1
BMI > = 30 kg/m^2	1
Ongoing glucocorticoid treatment	1

A Padua score ≥ 4 is classified as high risk

optimal ML models were chosen with the highest Youden index [20]. Model's sensitivity, specificity, and Youden index were computed to evaluate their predictive validity, and the training process was repeated five times to calculate mean values and standard deviations. The Youden index is commonly utilized as a summary measure of the ROC (Receiver Operating Characteristic) curve and it can be calculated as below,

$$Youden\ index = Sensitivity + Specificity - 1,$$

which enables the selection of an optimal cutoff point.

Results

Characteristics of distribution of VTE patients

230 VTE patients and 3054 non-VTE patients were included in this study. Clustering analysis with 13 VTE risk factors and 3 Padua-score-related features on these patients. Figure 5 showed that VTE patients didn't get together and were scattered among non-VTE patients. The distribution of VTE patients with a Padua score ≥ 6 points (accounting for 49.13% of overall VTE patients) is more intensely concentrated, while VTE patients with Padua score < 6 points were poorly characterized by the Padua model and distributed over a wide area (accounting for 50.87% of overall VTE patients). 14.35% VTE patients had a Padua score under 4 points and were stratified incorrectly as low risk. In addition, 85.87% of the high-risk patients recognized by the Padua were non-VTE patients. Importance of 13 risk factors on VTE prediction was evaluated via the Shapley Additive Explanations (SHAP) summary plot. The top three factors including previous VTE, reduced mobility, and heart/respiratory failure contributes more to the VTE than other variables.

Inspired by the clustering analysis results, patients with same values of feature vectors were grouped and the ratio of VTE patients were calculated. Some representative groups were shown at Table 2. It could be seen that there were VTE and non-VTE patients with identical values of risk factors and different groups had distinct *VTE risk ratios*, which proved the existence of 'fuzzy population'. Four groups in Table 2 had both

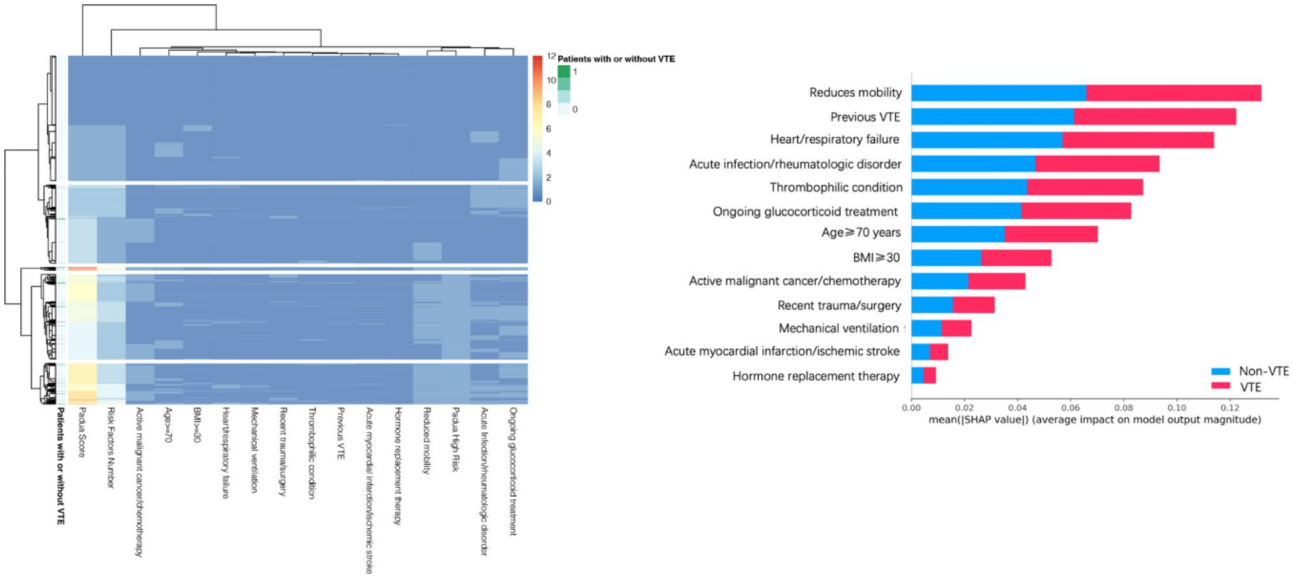

Fig. 5 Population clustering and feature importance analysis on inpatients from PUMCH. The clustering analysis with 13 VTE risk factors and 3 Padua-score-related features (Padua score, the number of VTE risk factors, and Padua high risk) was conducted on 3284 inpatients from PUMCH including 230 VTE and 3054 non-VTE patients. Each row represented a patient with (labeled with dark green color) or without (labeled with light green color) VTE in this heatmap. Features listed in the columns were labeled dark blue color as a lower value and red color as a higher value. Besides, feature importance of 13 VTE risk factors was evaluated with the SHAP summary plot

Table 2 Representative groups and their VTE ratio in PUMCH data

Group Index	Feature vector of group	Number of VTE Patients	Number of Non-VTE	VTE Risk Ratio
0	0,1,0,0,0,0,0,0,1,0,1,0,0	3	0	100%
1	0,0,1,0,0,1,1,0,1,0,0,0,0	4	2	66.67%
2	1,0,1,0,0,0,0,0,1,0,1,0,0	8	30	21.05%
3	0,0,1,0,0,0,0,0,0,0,0,0,0	3	168	1.75%
4	0,0,0,0,0,0,0,0,0,0,0,0,0	2	637	0.31%
5	1,0,1,0,0,0,0,0,1,1,0,0,0	0	2	0%
6	1,0,1,0,1,0,0,0,0,0,0,0,0	0	13	0%

Feature vector of every group was a 13-dimension 0/1 vector of which the elements corresponded to values of 13 VTE risk factors orderly as Fig. 4 and were separated by commas

VTE and non-VTE patients and there were two groups with only non-VTE patients. The 4^{th} group with feature vector $\mathrm{v} = (0,0,0,0,\ldots,0)$ had more patients than any other groups, which meant that most of patients didn't have non-zero VTE risk factor. The 3^{th} group with $\mathrm{v} = (0,0,1,0,\ldots,0)$ was the second largest group, which showed that there were many patients with just one risk factor, the reduced mobility. The 1^{th} group with $\mathrm{v} = (0,0,1,0,\ldots,0)$ had more VTE than the non-VTE patients, but the 2^{th} group with $\mathrm{v} = (1,0,1,0,\ldots,0)$ had less VTE than the non-VTE patients.

Mean predictive validity of VTE risk assessment models

The training dataset consisted of 189 VTE patients and 1531 non-VTE patients, while test dataset included 41 VTE patients and 1523 non-VTE patients. Mean values of sensitivities and specificities of five ML models, Padua and the proposed model on all training and test patients were listed in Table 3. Compared to the result of Padua model, generally five ML models had relatively higher specificities but lower sensitivities. Within five ML models, RF was the only model with sensitivities > 0.80 on both training and test data and average performance of specificities (> 0.80) of XGBoost were the best. There was no model with both higher sensitivity and specificity than the Padua among five ML models. However, on the training data, the proposed model achieved advantages on both sensitivity and specificity over the Padua by considering the 'fuzzy population'. On the test data, mean values of sensitivities of the proposed model were very similar with the Padua and specificities of the proposed were higher. In addition, standard deviations of predictive validity of the proposed model were far less than the five ML models.

The optimal predictive validity of VTE risk assessment models

Further, the optimal ML models were selected by cross validation for five ML models and the proposed model, and their performances on training and test data were shown in Table 4. In general, patterns of ML models were

Table 3 Comparison of mean predictive validity of five ML models, Padua and proposed model

Model Name	Training set		Test set	
	Sensitivity	Specificity	Sensitivity	Specificity
SVM	0.7894 ± 0.0220	0.7240 ± 0.0431	0.8341 ± 0.0420	0.7032 ± 0.0424
RF	0.8413 ± 0.0150	0.7757 ± 0.0138	0.8195 ± 0.0452	0.7349 ± 0.0156
GBDT	0.7883 ± 0.0404	0.8135 ± 0.0394	0.8146 ± 0.0396	0.7825 ± 0.0441
LR	0.7397 ± 0.0304	0.7960 ± 0.0230	0.8195 ± 0.0293	0.7869 ± 0.0230
XGBoost	0.7524 ± 0.0316	0.8328 ± 0.0185	0.8293 ± 0.0218	0.8064 ± 0.0239
Padua	**0.8466**	**0.6127**	**0.9024**	**0.6330**
Proposed method	**0.8995 ± 1.110E-16**	**0.6741 ± 0.0056**	**0.9024 ± 1.110E-16**	**0.6453 ± 0.0033**

Values of sensitivity and specificity were represented with 'mean value ± standard deviation'. The model training process was repeated five times to calculate the predictive validity. Note that sensitivities and specificities on training process were computed using all patients from training data

Table 4 Comparison of predictive validity of the optimal ML models, proposed model and Padua

Model Name	Training set			Test set	
	Sensitivity	Specificity	Youden	Sensitivity	Specificity
SVM	0.8042	0.7511	0.5554	0.8292	0.7104
RF	0.8307	0.7975	**0.6282**	0.8780	0.7643
GBDT	0.8148	0.8223	**0.6372**	0.8780	0.7787
LR	0.7725	0.7864	0.5589	0.8537	0.7708
XGBoost	0.7883	0.8302	**0.6185**	0.8537	0.7919
Padua	**0.8466**	**0.6127**	**0.4593**	**0.9024**	**0.6330**
Proposed method	**0.8995**	**0.6786**	**0.5781**	**0.9024**	**0.6481**

The optimal ML models and proposed model were selected by maximizing the value of Youden index on training data. Note that metrics of predictive validity on training process were computed using all patients from training data

consistent with the results in Table 3. Five ML models had higher specificities by sacrificing the sensitivities. Value of Youden index of the optimal proposed model was not as high as the RF, GBDT and XGBoost, but better than the SVM, LR, and Padua. However, the optimal proposed model had the highest sensitivity and its specificity was better than the Padua on both training and test data, which verified its excellent consistency (Table 5).

Discussion

In this study, we proposed a new VTE risk assessment model which split the population into groups based on values of risk factors, and established group-memory and group-prediction modules respectively in order to consider the effect of 'fuzzy population' and describe VTE patients' characteristics better. By comparing with five traditional ML models and Padua model on patients from PUMCH, effectiveness of our proposed model was validated and it showed better robustness than traditional ML models trained on equal number of VTE and non-VTE patients. The proposed model was the only one which showed advantages on both sensitivity and specificity over Padua model.

For five ML models trained on equal number of VTE and non-VTE patients, results in Table 4 were calculated by using default threshold 0.5 for predictive probability to classify patients into the high or low risks. Due to the fact that sensitivity and specificity of model can be different by changing values of thresholds, the relationship between predictive validity, namely sensitivity and specificity, and the threshold was explored further and plotted at Fig. 6. RF, GBDT, and XGBoost were selected typically because they achieved higher values of Youden index than the proposed model. From Fig. 6 it could be seen that, on training and test data, for GBDT and XGBoost, there was not a threshold that had higher sensitivity and specificity simultaneous than the proposed model. For RF, thresholds with better predictive validity than the proposed model only existed on test data. In summary three ML models couldn't obtain higher sensitivities and specificities than the proposed by changing predictive thresholds, which proved our model's efficiency again.

One notable result from Table 3 was that standard deviations of sensitivities and specificities of five ML models trained on equal number of VTE and non-VTE patients were larger than our proposed model, which showed that ML models, lacked robustness. Due to the neglect of 'fuzzy population', within training set, randomly selected non-VTE patients would disturb the correct estimation of *VTE risk ratio*s of groups, which lead to instability of model's performance. To elaborate the influence of 'fuzzy population' on ML models' predictive performance, changing of sensitivities and specificities of ML models by strengthening the effect of 'fuzzy population' was visualized in Figure S1. For group $C^{(t)} \in C_{fuzzy}$ which included less VTE patients than non-VTE patients, namely *VTE risk ratio* $r^{(t)} = \frac{m^{(t)}}{n^{(t)}} \leq 0.5$, by increasing the number of non-VTE patients of group $C^{(t)}$ in the training set, VTE ratio $r^{(t)'}$ could be changed from >0.5 to ≤ 0.5. When $r^{(t)'} \leq 0.5$, the ML model tended to predict individuals from group $C^{(t)}$ as the low VTE risk, which would affect model's performance. It could be seen that with the increasing of risk ratio of groups with $r^{(t)'} \leq 0.5$ in training set, sensitivities and specificities of three ML models varied dramatically. Therefore, taking the 'fuzzy population' into account was crucial to model's robustness. Actually, 'fuzzy population' is very common in medical area, since the incidence of most diseases is relatively low and the pathogenesis of patients is complex. This method can be widely used in many aspects, especially disease screening or risk prediction.

Currently our study still needs to be improved in several aspects. Firstly, due to the low incidence of VTE, the sample size of this research center is still limited. To assess statistical differences of predictive validities between the proposed model and Padua model, studies with larger sample sizes are still required. Multi-center and prospective researches are also needed to validate and promote the model further. Secondly, with increasing number of VTE samples, the deep learning methods [21, 22] maybe can replace the ANN model to further improve our model.

Table 5 Detailed comparison between the optimal proposed model and Padua

Statistic	Padua model used in clinic	Proposed best model
Sensitivity	0.9024(0.7687, 0.9728)	0.9024(0.7687, 0.9728)
Specificity	0.6330(0.6082, 0.6572)	**0.6481(0.6235, 0.6721)**
Youden Index	0.5354(0.3769, 0.6300)	**0.5505(0.3922,0.6448)**
PPV	0.0620(0.0440, 0.0846)	0.0646(0.0459,0.0879)
NPV	0.9959(0.9895, 0.9989)	0.9960(0.9897,0.9989)
PLHR	2.4587(2.1800, 2.7731)	2.5642(2.2707,2.8956)
NLHR	0.1541(0.0607, 0.3913)	0.1505(0.0593,0.3822)

PPV, positive predictive value; NPV, negative predictive value; NLHR, negative likelihood ratio; PLHR, positive likelihood ratio

Conclusion

Based on population clustering analysis and the Padua model, by considering the effect of 'fuzzy population', this study proposes a new VTE risk assessment model using data from Chinese medical inpatients. The 'fuzzy population' is not limited to the VTE but is prevalent in other diseases with lower incidence rates as well. By considering this issue during the construction of prediction models for other diseases, the performance of models may be further enhanced. Our proposed VTE prediction model

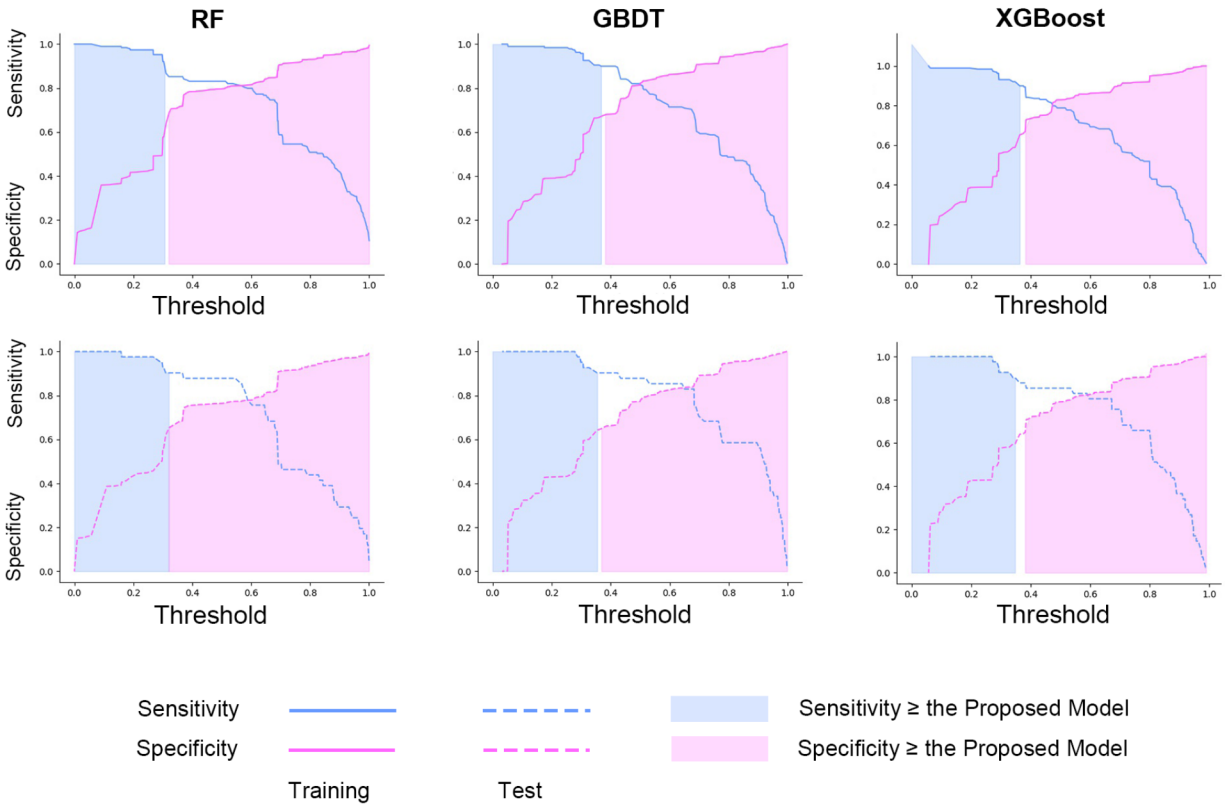

Fig. 6 Changing of sensitivities and specificities of three ML models with the increasing of predictive thresholds. 3 ML models with higher Youden index (RF, GBDT and XGBoost) than the proposed model were selected. Results in training and test data were shown in upper and lower figures respectively. For each figure, thresholds with higher specificities than the proposed model were marked with pink, and with higher sensitivities were marked with blue

exhibits strong predictive capabilities in VTE risk assessment, offering clinicians a valuable tool for risk stratification and effective prevention strategies in clinical practice.

Abbreviations

ANN	Artificial neural network
DVT	Deep venous thrombosis
GBDT	Gradient boosting decision tree
LR	Logistic regression
ML	Machine learning
PE	Pulmonary embolism
PUMCH	Peking Union Medical College Hospital
RF	Random forest
SVM	Support vector machine
VTE	Venous thromboembolism

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12911-024-02834-3.

Supplementary Material 1

Acknowledgements
We thank the Peking Union Medical College Hospital for supporting the research.

Author contributions
Xin Wang & Yu-Qing Yang: Conceptualization, Methodology, Formal analysis, Writing - original draft. Xin-Yu Hong & Si-Hua Liu: Investigation, Data curation. Jian-Chu Li & Ting Chen: Supervision, Writing - review & editing. Ju-Hong Shi: Resources, Supervision, Project Administration, Funding acquisition, Writing - review & editing.

Funding
Supported by the National High Level Hospital Clinical Research Funding (2022-PUMCH-C-017), the National Natural Science Foundation of China (62203060, 62403492,82202299), and R&D Program of Beijing Municipal Education Commission (KM202310005030).

Data availability
Data is available upon reasonable request to the corresponding author. The code related to model building is available on GitHub (https://github.com/YangLab-BUPT/VTE_fuzzy_population.git).

Declarations

Ethics approval and consent to participate
This study was approved by the Ethics Committee of PUMCH in Chinese Academy of Medical Sciences (reference number for ethics approval: B164). All methods were carried out according to the Declaration of Helsinki. All participants gave informed consent.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Received: 27 May 2023 / Accepted: 19 December 2024
Published online: 30 December 2024

References

1. Grosse SD, Nelson RE, Nyarko KA, Richardson LC, Raskob GE. The economic burden of incident venous thromboembolism in the United States: a review of estimated attributable healthcare costs. Thromb Res. 2016;137:3–10.
2. Giordano NJ, Jansson PS, Young MN, Hagan KA, Kabrhel C, Epidemiology. Pathophysiology, stratification, and natural history of pulmonary embolism. Techniques Vascular Interventional Radiol. 2017;20(3):135.
3. Bartholomew JR. Update on the management of venous thromboembolism. Cleve Clin J Med. 2017;84(12 suppl 3):39–46.
4. Agnelli G. Prevention of venous thromboembolism in surgical patients. Circulation. 2004;110(24_suppl_1):IV-4-IV-12.
5. Cohen AT, Davidson BL, Gallus AS, et al. Efficacy and safety of fondaparinux for the prevention of venous thromboembolism in older acute medical patients: randomised placebo controlled trial. BMJ. 2006;332(7537):325–9.
6. Samama MM, Cohen AT, Darmon J-Y, et al. A comparison of enoxaparin with placebo for the prevention of venous thromboembolism in acutely ill medical patients. N Engl J Med. 1999;341(11):793–800.
7. Bagot CN, Arya R. Virchow and his triad: a question of attribution. Br J Haematol. 2008;143(2):180–90.
8. Kearon C, Akl EA, Ornelas J, et al. Antithrombotic therapy for VTE disease: CHEST Guideline and Expert Panel Report. Chest. 2016;149(2):315–52.
9. Barbar S, Noventa F, Rossetto V, et al. A risk assessment model for the identification of hospitalized medical patients at risk for venous thromboembolism: the Padua prediction score. J Thromb Haemostasis Jth. 2010;8(11):2450–7.
10. Liu X, Liu C, Chen X, Wu W, Lu G. Comparison between Caprini and Padua risk assessment models for hospitalized medical patients at risk for venous thromboembolism: a retrospective study. Interact Cardiovasc Thorac Surg. 2016;23(4):ivw158.
11. Wang X, Hong XY, Li JY, et al. Value of Padua risk assessment model in evaluating venous thromboembolism of hospitalized patients in the department of internal medicine. Med J Peking Union Med Coll Hosp. 2018;9(3):48–55.
12. Waljee AK, Lipson R, Wiitala WL, et al. Predicting hospitalization and outpatient corticosteroid use in inflammatory bowel disease patients using machine learning. Inflamm Bowel Dis. 2017;24(1):45–53.
13. Montazeri M, Montazeri M, Montazeri M, Beigzadeh A. Machine learning models in breast cancer survival prediction. Technol Health Care Official J Eur Soc Eng Med. 2016;24(1):31.
14. Ferroni P, Zanzotto FM, Scarpato N, et al. Risk assessment for venous thromboembolism in chemotherapy-treated ambulatory cancer patients: a machine learning approach. Med Decis Making. 2017;37(2):234–42.
15. James S, Suguness A, Hill A, Shatzel J. Novel algorithms to predict the occurrence of in-hospital venous thromboembolism: machine learning classifiers developed from the 2012 national inpatient sample. Chest. 2015;148(4):492A.
16. Sabra S, Malik KM, Alobaidi M. Prediction of venous thromboembolism using semantic and sentiment analyses of clinical narratives. Comput Biol Med. 2018;94:1–10.
17. Wang X, Yang YQ, Liu SH, Hong XY, Sun XF, Shi J. Comparing different venous thromboembolism risk assessment machine learning models in Chinese patients. J Eval Clin Pract. 2020;26(1):26–34.
18. Swami A, Jain R. Scikit-learn: machine learning in Python. J Mach Learn Res. 2013;12(10):2825–30.
19. Chen T, Guestrin C. XGBoost: a scalable tree boosting system. 2016.
20. Youden WJ. Index for rating diagnostic tests. Cancer. 1950;3(1):32–5.
21. Ching T, Himmelstein DS, Beaulieujones BK, et al. Opportunities and obstacles for deep learning in biology and medicine. J Royal Soc Interface. 2018;15(141):20170387.
22. Yang Y, Wang X, Huang Y, Chen N, Shi J, Chen T. Ontology-based venous thromboembolism risk assessment model developing from medical records. BMC Med Inf Decis Mak. 2019;19(4):151.

Publisher's note