Original Research

Machine learning prediction of symptomatic intracerebral hemorrhage after stroke thrombolysis: a cross-cultural validation in Caucasian and Han Chinese cohort

Junfeng Liu [ID], Xinyue Chen, Xiaonan Guo, Renjie Xu, Yanan Wang and Ming Liu;
on behalf of the VISTA Collaborators#

Abstract

Background: Previous studies found that Asians seemed to have higher risk of HT after thrombolysis than Caucasians due to its race differences in genetic polymorphism. Whether the model developed by Caucasians could predict risk of symptomatic intracerebral hemorrhage (sICH) in Asians was unknown.

Objectives: To develop a machine learning–based model for predicting sICH after stroke thrombolysis in Caucasians and externally validate it in an independent Han Chinese cohort.

Design: The derivation Caucasian sample included 1738 ischemic stroke (IS) patients from the Virtual International Stroke Trials Archive (VISTA) data set, and the external validation Han Chinese cohort included 296 IS patients who were treated with intravenous thrombolysis.

Methods: Twenty-eight variables were collected across both samples. According to their properties, we classified them into six distinct clusters (ie, demographic variables, medical history, previous medication, baseline blood biomarkers, neuroimaging markers on initial CT scan and clinical characteristics). A support vector machine (SVM) model, which consisted of data processing, model training, testing and a 10-fold cross-validation, was developed to predict the risk of sICH after stroke thrombolysis. The receiving operating characteristic (ROC) was used to assess the prediction performance of the SVM model. A domain contribution analysis was then performed to test which cluster had the highest influence on the performance of the model.

Results: In total, 85 (4.9%) patients developed sICH in the Caucasians, and 29 (9.8%) patients developed sICH in the Han Chinese cohort. Eight features including age, NIHSS score, SBP (systolic blood pressure), DBP (diastolic blood pressure), ALP (alkaline phosphatase), ALT (alanine transaminase), glucose, and creatine level were included in the final model, all of which were from demographic, clinical characteristics, and blood biomarkers clusters, respectively. The SVM model showed a good predictive performance in both Caucasians (AUC = 0.87) and Han Chinese patients (AUC = 0.74). Domain contribution analysis showed that inclusion/exclusion of clinical characteristic cluster (NIHSS score, SBP, and DBP), had the highest influence on the performance of predicting sICH in both Caucasian and Han Chinese cohorts.

Conclusion: The established SVM model is feasible for predicting the risk of sICH after thrombolysis quickly and efficiently in both Caucasian and Han Chinese cohort.

Keywords: Caucasians, Han Chinese, ischemic stroke, machine learning, support vector machine, symptomatic intracerebral hemorrhage, thrombolysis

Correspondence to:
Ming Liu
Centre of Cerebrovascular
Diseases, Department
of Neurology, West
China Hospital, Sichuan
University, No. 37, Guo Xue
Xiang, Chengdu 610041,
China.
wyplmh@hotmail.com

Junfeng Liu
Yanan Wang
Department of Neurology,
West China Hospital,
Sichuan University,
Chengdu, China

Xinyue Chen
CT Collaboration, Siemens
Healthineers, Chengdu,
China

Xiaonan Guo
School of Information
Science and Engineering,
Yanshan University,
Qinhuangdao, China

Renjie Xu
Department of Respiratory
Medicine, West China
Hospital, Sichuan
University, Chengdu, China

Members of VISTA-Acute
steering Committee:
KR Lees (Chair), A
Alexandrov, PM Bath,
E Berge, E Bluhmki, N
Bornstein, C Chen, L
Claesson, SM Davis, G
Donnan, HC Diener, M
Fisher, M Ginsberg, B
Gregson, J Grotta, W
Hacke, MG Hennerici,
M Hommel, M Kaste, P
Lyden, J Marler, K Muir,
N Venketasubramanian,
R Sacco, A Shuaib, P Teal,
NG Wahlgren, S Warach,
and C Weimar.

Introduction

Symptomatic intracranial hemorrhage (sICH), which is the most feared complication after thrombolysis, leads to increased mortality and morbidity in patients with ischemic stroke.[1] The risk of sICH after thrombolysis varies from 2% to 7%, with higher rates in East Asia. In addition, in a pooled analysis of 6756 patients[2] from multiple randomized trials comparing alteplase with placebo, the risk of fatal sICH after thrombolysis increased with National Institutes of Health Stroke Scale (NIHSS) score on admission, from 1.6% with a baseline NIHSS score of 5–10 to 6.8% with an NIHSS score >21. Therefore, it raised the question whether patients with severe stroke could benefit from lower doses of alteplase, especially in Asians with higher susceptibility to bleeding compared to Caucasians.[3–5]

One large randomized controlled trial[5] investigated whether a lower dose of intravenous alteplase improved the outcomes of ischemic stroke and reduce the risk of sICH compared to the standard dose, and the results did not show a benefit of reducing 3-month death/disability though it reduced the risk of sICH. Similar findings were reported in the secondary analysis according to age, ethnicity (Asian *vs* non-Asian), and neurological deficits.[6] It is likely that only a small percentage of patients with high risk of sICH may benefit from low-dose alteplase. In this context, a reliable tool to identify these patients may be useful for the design of future randomized trials.

Several risk models were developed to assess the risk of sICH after stroke thrombolysis,[7,8] but the performance of these models is limited. Moreover, most prediction models were constructed using conventional a logistic regression method, the limitation of which includes the collinearity of variables and overfitting of the model. Machine learning (ML) algorithms, which allow more accurate prediction by modeling linear and nonlinear interactions among many variables,[9] have been applied in clinical neurosciences showing good performance for stroke diagnosis and prognosis.[10]

So far, few study[11–15] developed models for predicting stroke using ML approaches, of which only three studies[11–13] investigated the prediction of sICH after thrombolysis. However, none of the above studies were externally validated in an independent sample with a different culture and ethnicity, which may make the models overfitting and reduce the generalizability of predictive models in diverse populations (eg, with cultural and ethnic differences). In addition, previous studies[16,17] found that Asians seemed to have higher risk of hemorrhagic transformation (HT) after thrombolysis than Caucasians due to its race differences in genetic polymorphism. The GWTG-Stroke sICH risk (GRASPS) score even included Asian race as an independent predictor for sICH after thrombolysis.[18] It raised the question whether the model developed by Caucasians could predict risk of sICH in populations with different culture and ethnicity, such as Asians.

Support vector machine (SVM), which is a supervised ML algorithm, can be used both for binary class label prediction and regression-based prediction of property values.[19] It is powerful at predicting the classification of unseen new data sets with higher accuracy and is the most common technique among data analysts due to high accuracy provided under low computational power.[20] Therefore, we aimed to develop an ML model using SVM method to predict sICH after thrombolysis in a Caucasian sample from the acute section of the Virtual International Stroke Trials Archive (VISTA; vista.gla.ac.uk) data set. After that, we tested whether this ML model predicts sICH after thrombolysis when applied to an independent Han Chinese cohort from the West China hospital, Sichuan University, China. Finally, we also performed a domain contribution analysis using the identified ML algorithm to test whether Caucasian and Han Chinese cohort have the similar pathogenesis of sICH after thrombolysis. In the domain contribution analysis, different clusters of the predictors we used were modeled separately to measure the prediction accuracy in the presence and absence of each domain-specific set of predictors.

Methods

Participants

Derivation sample. We used data from the acute section of the VISTA data set as a derivation cohort, and only included Caucasians in this study. Patients' data used to derive any published score or model for predicting sICH was excluded in this study. VISTA is an anonymized data repository of completed stroke trials. All included trials were performed under appropriate institutional review board and regulatory approvals, and only fully anonymized data are held by VISTA.[21]

Validation sample. We consecutively reviewed Han Chinese patients with ischemic stroke (≥18 years old) treated with thrombolysis at the West China hospital, Sichuan University between 1 January 2012 and 31 December 2020 as a derivation cohort. Acute ischemic stroke was diagnosed based on the World Health Organization criteria,[22] and further confirmed by a computed tomography (CT) or magnetic resonance imaging (MRI) scan. This study complied with the principles of the Declaration of Helsinki and was approved by Biomedical Research Ethics Committee of our hospital [2016 (339)].

Predictors. The study included six distinct clusters of 28 predictors across both samples: (1) demographic variables (i.e. age and gender), (2) medical history [i.e. myocardial (MI), atrial (AF), hypertension, diabetes mellitus, previous stroke, previous TIA], (3) previous medication (i.e. previous anticoagulation and previous antiplatelet), (4) baseline blood biomarkers [i.e. blood glucose, neutrophils, platelet counts, total bilirubin (TBIL), creatinine, alkaline (ALP), alanine (ALT), aspartate aminotransferase (AST)], (5) neuroimaging markers on initial CT scan [i.e. Alberta Stroke Program Early (ASPECT) Score, visible hypodensity in middle cerebral artery (MCA) territory on CT; the extent of hypodensity in the MCA territory >1/3 of the MCA territory, and midline shift], and (6) clinical characteristics[i.e. onset to treatment (OTT), initial National Institutes of Health Stroke (NIHSS) score, systolic blood pressure, diastolic blood pressure on admission, and cardioembolic stroke]. The midline shift was defined as midline shift of more than 5 mm at the septum pellucidum level.[23]

Outcomes. We defined sICH as any deterioration on the neurological exam with ICH on CT brain imaging within 36 hours after stroke onset based on the definition of the National Institute of Neurological Disease and Stroke study (NINDS) criteria,[24] which is consistent with the definition used in Caucasians. In Han Chinese cohort, the presence of sICH was determined by two trained neurologists, and the inter-rater reliability (Kappa) for assessing the presence/absence of sICH in Han Chinese cohort was 0.87.

Data pre-processing and feature selection. The missing values were imputed by 5-nearest neighbor model. And Caucasian was defined as training cohort, while Han's cohort as testing cohort. Before the model training process, the imbalance between two classes in Caucasian cohort was resolved by synthetic minority oversampling technique (SMOTE) technique, which[25] is an oversampling approach that creates synthetic minority class samples. It potentially performs better than simple oversampling and it is widely used. Subsequently, a random forest model was built using all predictors we collected for the aims of sICH prediction. Then based on mean decrease in GINI index of this model, the importance of the features was ranked. We then included the top-8 features from this list for the next step.

Model building and performance evaluation. An SVM algorithm with built-in recursive feature elimination was used for final predictive model building. With this build-in function, the algorithm would select optimal subset of feature combination from those eight features at the mean time of model tuning. The objective function for the model tuning was the area under the curve (AUC). In order to get balance between the bias and variation of a model, the objective function was calculated based on averaged AUCs of internal 50 re-samples using five repeats 10-fold cross-validation. Then the SVM model with optimal features combination and tuning parameter of highest AUC was defined as the final predictive model. The performance of the models was evaluated by the AUC derived from receiver operating characteristics curves and other diagnostic statistics (sensitivity and specificity). Calibration curve was also performed to visually assessed the agreement between model predictive and actual probability for original data set. Then decision curve analysis (DCA) was used to evaluate the clinical value of our model independently on the basis of calculating the net benefit for patients at each threshold probability. To rule-out the possibility of over-fitting, the final model was externally validated in our Han Chinese people without re-calibration. Finally, we performed a domain contribution analysis using the SVM algorithm. In this process, clusters of predictors were modeled separately to measure the prediction accuracy in the presence and absence of each domain-specific set of predictors.

Statistical analysis. Continuous data were expressed as mean ± standard deviation or median with interquartile range, while categorical data were reported as frequencies and percentages. We used χ^2 test or Fisher's exact test to assess the differences in categorical data, while for continuous data, Student's *t*-test, analysis of
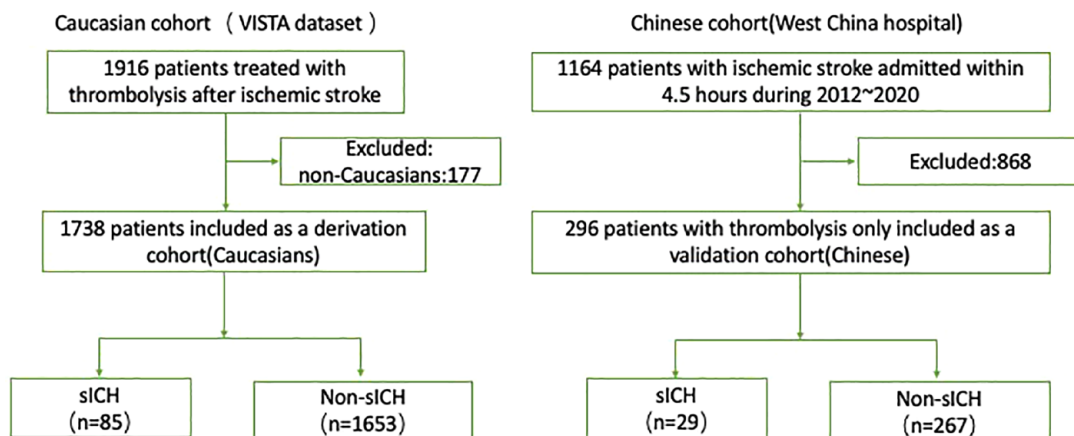
Figure 1. The flow chart of included patients in the study.
sICH, symptomatic intracerebral hemorrhage.

variance (ANOVA) or the Mann–Whitney *U*-test were performed. A two-sided $p < 0.05$ was considered statistically significant. All statistical analysis was performed using the statistical software R version 3.4.1 (http://www.R-project.org).

Results

Patient characteristics

In the final analysis, 1738 Caucasians (age: 68.37 ± 12.62; male: 1016 [58.5%]) from the VISTA data set were included as the derivation cohort, and 296 Han Chinese patients (age: 69.39 ± 13.37; male: 165 [55.7%]) as the validation cohort (Figure 1). The detailed descriptive information of the Caucasian and Han Chinese cohort in terms of baseline predictors and the outcome of interest are presented in Table 1. Distributions of sex and age were comparable between the Caucasian and Han Chinese cohorts ($p > 0.05$). Among the 1738 Caucasians, 85 (4.9%) patients developed sICH. In regard to the 296 Han Chinese patients, a total of 29 (9.8%) patients developed sICH. In addition, the frequency of missing data per variable is also shown in Table 1.

The SVM model construction and evaluation

The SVM model involved data processing, feature selection, model training, testing, and a 10-fold cross-validation (Figure 2). The feature importance for predicting sICH using random forest algorithm is showed in Figure 3. In the feature-selection step, age, NIHSS score, SBP, DBP, ALP, ALT, glucose, and creatine level were selected into the final SVM model, all of which were from the demographic, clinical characteristics, and blood biomarkers clusters, respectively. The model with the highest accuracy was selected as the final model for further evaluation. It showed a good performance for predicting sICH in the Caucasians, and the AUC was 0.87 [95% confidence interval (CI): 0.83–0.91, sensitivity: 0.67, specificity: 0.87; Figure 4(a)]. The SVM model was then externally validated in the Han Chinese cohort, which also showed a good discrimination ability for sICH prediction [AUC = 0.74, 95% CI: 0.64 to 0.83, sensitivity: 0.50, specificity: 0.87; Figure 4(b)].

In regard to clinical usability evaluation of the SVM model, we did calibration and DCA. Figure 5(a) illustrates the corresponding calibration curve of the SVM model in the Caucasian cohort, which suggested a favorable predictive performance satisfactorily consistent with the ideal curve. DCA was conducted to assess the clinical utility of the SVM model [Figure 5(b)]. According to the decision curve, it demonstrated that intervention on ischemic stroke patients on the basis of the SVM model leads to higher benefit.

Domain contribution analysis

To investigate which domain has the highest influence on the performance of the SVM model, a domain contribution analysis was performed. In this process, to calculate the prediction accuracy

Table 1. Descriptive information of Caucasian and Han Chinese cohorts at baseline.

	Caucasian samples (n = 1738)	Missing data (%)	Han Chinese samples (n = 296)	Missing data (%)	*p*
Demographic characteristics					
Age, [mean (SD)]	68.37 (12.62)	0 (0)	69.39 (13.37)	0 (0)	0.2
Male, *n* (%)	1016 (58.5)	0 (0)	165 (55.7)	0 (0)	0.408
Medical history					
Myocardial infarction, *n* (%)	255 (14.7)	0 (0)	4 (1.4)	0 (0)	<0.001
Atrial fibrillation, *n* (%)	426 (24.5)	0 (0)	81 (27.4)	0 (0)	0.309
Hypertension, *n* (%)	1218 (70.1)	0 (0)	175 (59.1)	0 (0)	<0.001
Diabetes mellitus, *n* (%)	316 (18.2)	0 (0)	70 (23.6)	0 (0)	0.03
Previous stroke, *n* (%)	238 (13.7)	0 (0)	36 (12.2)	0 (0)	0.52
Previous TIA, *n* (%)	142 (8.6)	89 (5.1)	2 (0.7)	0 (0)	<0.001
Smoker, *n* (%)	852 (49.1)	4 (0.2)	109 (36.8)	0 (0)	<0.001
Previous medication					
Previous anticoagulation, *n* (%)	81 (4.7)	0 (0)	17 (5.7)	0 (0)	0.462
Previous antiplatelet, *n* (%)	647 (37.2)	0 (0)	29 (9.8)	0 (0)	< 0.001
Clinical characteristics					
Onset to treatment time, (median [IQR])	3.42 [2.92, 3.88]	0 (0)	3.00 [2.00, 4.00]	0 (0)	< 0.001
NIHSS score on admission, (median [IQR])	14.00 [10.00, 18.00]	0 (0)	9.00 [4.75, 15.00]	0 (0)	< 0.001
Systolic blood pressure on admission, [mean (SD)]	152.82 (24.47)	0 (0)	148.96 (24.95)	0 (0)	0.013
Diastolic blood pressure on admission, [mean (SD)]	81.32 (16.31)	22 (1.3)	84.35 (15.76)	0 (0)	0.003
Cardioembolic stroke, *n*(%)	820 (47.5)	13 (0.7)	82 (27.7)	0 (0)	< 0.001
Baseline blood biomarkers					
Blood glucose, [mean (SD)]	7.38 (2.82)	253 (14.6)	8.10 (2.96)	0 (0)	< 0.001
Neutrophils, [mean (SD)]	7.43 (3.12)	489 (28.1)	6.25 (3.09)	0 (0)	< 0.001
Platelet count, [mean (SD)]	234.88 (74.92)	372 (21.4)	184.04 (60.34)	0 (0)	< 0.001
Total bilirubin, [mean (SD)]	7.87 (4.58)	171 (9.8)	11.52 (6.38)	0 (0)	< 0.001
Creatinine, [mean (SD)]	83.91 (23.89)	86 (4.9)	84.15 (57.13)	0 (0)	0.906
Alkaline phosphatase, [mean (SD)]	78.76 (35.92)	129 (7.4)	81.48 (24.97)	0 (0)	0.212
Alanine aminotransferase, [mean (SD)]	24.07 (23.69)	258 (14.8)	23.05 (17.20)	0 (0)	0.482

(Continued)

Table 1. (Continued)

	Caucasian samples (n = 1738)	Missing data (%)	Han Chinese samples (n = 296)	Missing data (%)	*p*
Aspartate aminotransferase, [mean (SD)]	25.13 (20.75)	375 (21.6)	25.02 (12.69)	0 (0)	0.928
Neuroimaging markers on initial CT scan					
ASPECT Score, (median [IQR])	10.00 [10.00, 10.00]	20 (1.2)	9.00 [8.00, 10.00]	0 (0)	< 0.001
Hypodensity, *n* (%)	131 (7.6)	20 (1.2)	178 (60.1)	0 (0)	< 0.001
>1/3 MCA territory, *n* (%)	3 (0.2)	20 (1.2)	71 (24.0)	0 (0)	< 0.001
Midline shift, *n* (%)	1 (0.1)	20 (1.2)	3 (1.0)	0 (0)	0.011
sICH, *n* (%)	85 (4.9)	0 (0)	29 (9.8)	0 (0)	0.001

ASPECT Score, Alberta Stroke Program Early CT score; MCA, middle cerebral artery; sICH, symptomatic intracerebral hemorrhage; TIA, transient ischemic attack.
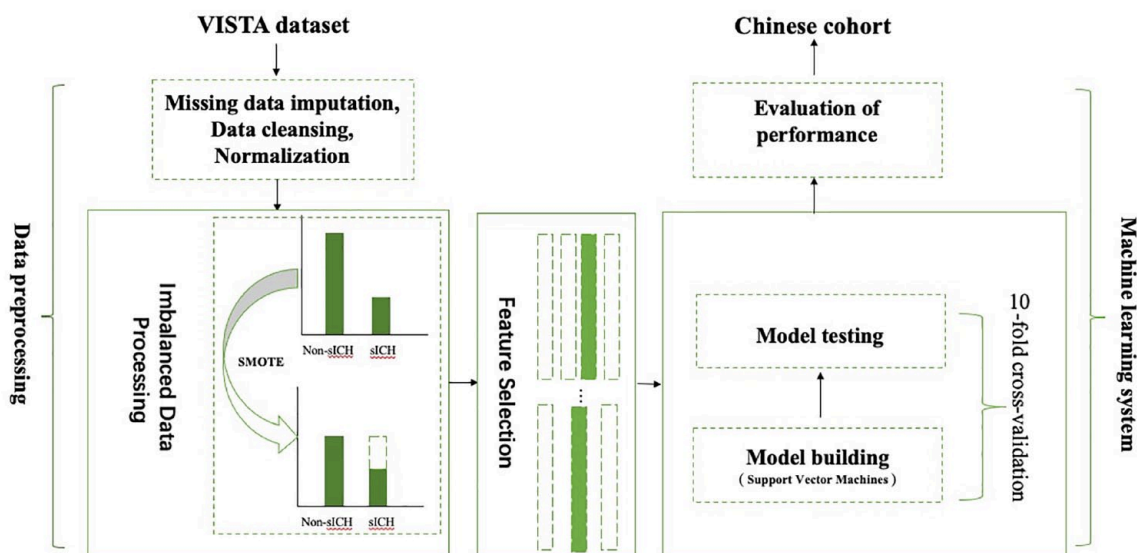
Figure 2. The machine learning model process. The VISTA data set first went through data pre-processing which included missing data imputation using 5-nearest neighbor model, data cleansing and normalization. For the categorized variables, the one hot encoding was used to cover all the possibilities, and for the continuous type of features, *Z* score normalization was applied. Then, the VISTA set went through imbalanced processing by using synthetic minority oversampling technique (SMOTE) technique. The SMOTE technique is an oversampling approach that creates synthetic minority class samples. It potentially performs better than simple oversampling and it is widely used. This process generated parameters, and the training data set was used to evaluate the accuracy of the model. In the end, our model was also external validated in an independent Chinese cohort.
sICH, symptomatic intracerebral hemorrhage.

for each set of features, models were evaluated including each cluster only (e.g. only the demographic cluster) and excluding that cluster (e.g. all features except for the demographic cluster).

The domain contribution analysis is shown in Table 2. In both the Caucasian and Han Chinese samples, inclusion or exclusion of clinical characteristic clusters had the highest influence on the

performance of predicting sICH. The models with each of these features only resulted in the highest AUC values and their absence in the model substantially decreased these measures.

Discussion

In this study, we developed a post-thrombolysis sICH prediction model using SVM algorithm in Caucasians, and the model was then externally validated in an independent Han Chinese cohort. To our knowledge, this is the first study to establish and validate a post-thrombolysis sICH prediction model by using a SVM algorithm in two independent samples with different cultures and ethnicities. Based on the prediction measures, the SVM model showed a good performance for predicting post-thrombolysis sICH in both Caucasian and Han Chinese cohorts, which suggests the generalizability of our predictive model.

The eight features included in the model were age, NIHSS score, SBP, DBP, ALP, ALT, glucose, and creatine level on admission, four of which are from the blood biomarkers domain. This approach to biomarker identification may be of particular benefit in intermediate risk groups of sICH where underlying subclinical risk is not apparent based on conventional risk factors. Age and NIHSS score are the two most commonly used variables employed in sICH prediction.[7,8] In addition, Bonkhoff et al.[26] also found that stroke severity (NIHSS score) was the overall most important predictor in their model for predicting severe complications after ischemic stroke including secondary intracerebral hemorrhage. It is consistent with our domain analysis, which showed clinical characteristics (NIHSS score, SBP and DBP) yielded the highest prediction measures in both the Caucasian and Han Chinese samples. In regard to our findings about blood pressure (eg, SBP and DBP), previous studies reported blood pressure to be a risk factor for sICH, sush as the Australian Streptokinase Trial,[27] SITS-MOST registry study,[28] and the NINDS trials.[29] It raises question whether intensive blood pressure management in patients with high risk of sICH would prevent sICH and improve the outcomes. More studies are needed to confirm this.

Our model developed by the Caucasian data set was able to predict sICH in the Han Chinese data set, which suggests the generalization of our model. It may indicate that a shared mechanism
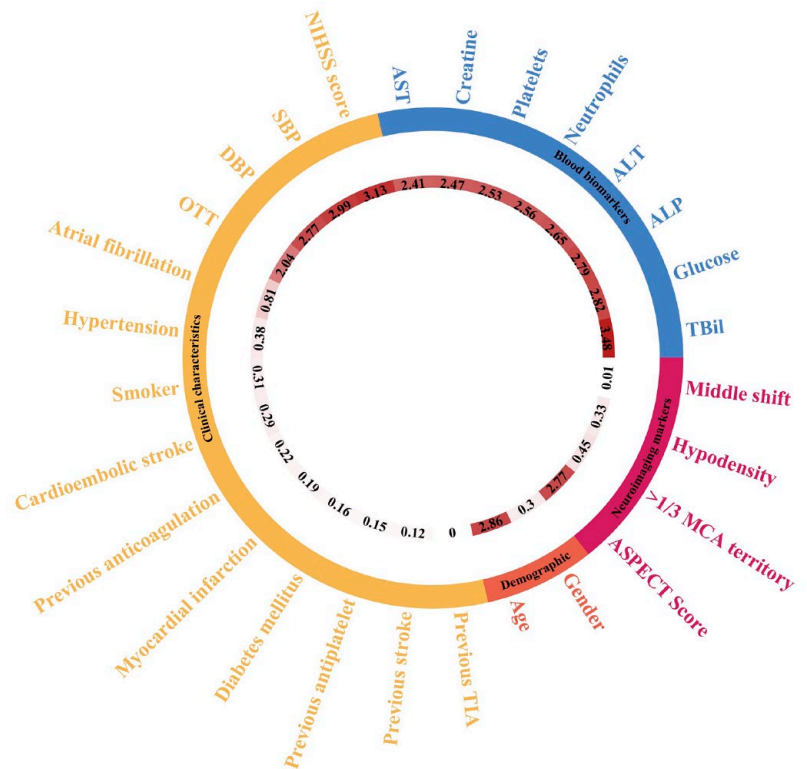

Figure 3. The feature importance of in the Caucasian cohort. The color arcs represent six distinct clusters of the 28 predictors we collected in the study. Numbers in the inner rings represent the importance of each feature for predicting sICH in the model.

TIA, transient ischemic attack; TBiL, total bilirubin; ALP, alkaline phosphatase; ALT, alanine aminotransferase; AST, aspartate aminotransferase; ASPECT score, Alberta Stroke Program Early CT Score; MCA, middle cerebral artery; OTT, onset to treatment time; NIHSS score, National Institutes of Health Stroke Scale score; SBP, systolic blood pressure; DBP, diastolic blood pressure.

of post-thrombolysis sICH between Caucasians and Han Chinese beyond the racial differences in genetic polymorphism reported before.[16] It may also explain that why previous clinical trials found no significant ethnic variation in the differential treatment effects of low-dose and standard-dose alteplase on the disability outcomes or sICH.[6] In the future, studies with genetic data are needed to investigate the effect of race differences on the presence of sICH.

One potential contribution of our study is the prospect of developing a versatile computerized test that enables early and quick screening of patients with a high risk for sICH in clinical practice. Another advantage of our model is the ability to recognize the best predictors within a domain (demographic, clinical characteristics, blood biomarkers, etc), as well as their importance with
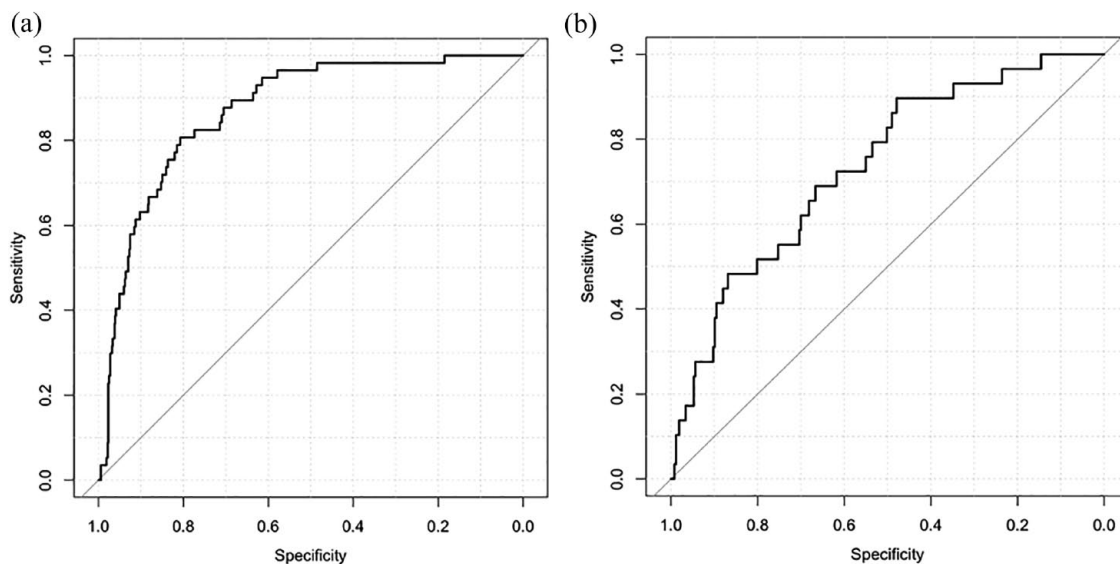
Figure 4. The ROC curves of the support vector machine (SVM) model in Caucasian and Chinese cohort. The AUCs of the SVM model for predicting sICH in Caucasian and Chinese cohort were 0.87 and 0.74, respectively: (a) Caucasian cohort and (b) Chinese cohort.
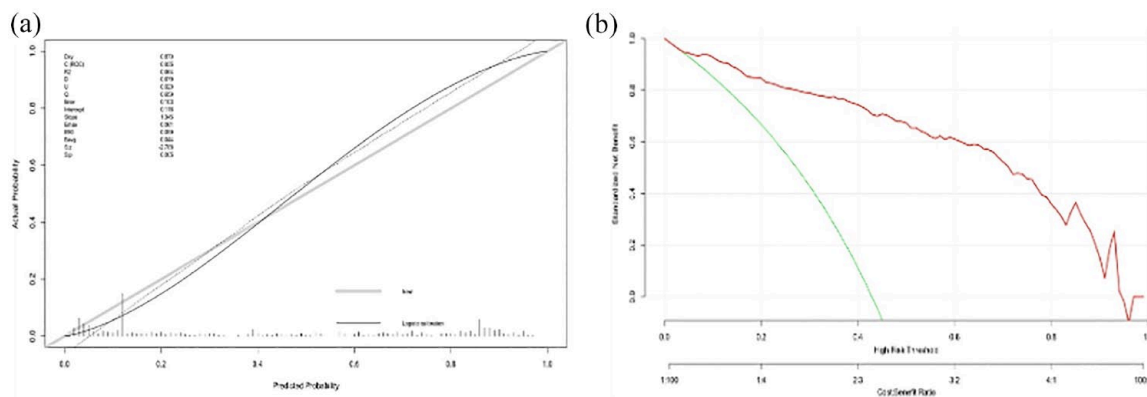ROC, receiver operating characteristics; sICH, symptomatic intracerebral hemorrhage.

Figure 5. Calibration curve and decision curve analysis of the support vector machine (SVM) model in Caucasian cohort: (a) Calibration curve of the SVM model and (b) decision curve analysis for the SVM model. The red line indicates the decision curve of the SVM model. The green line stands for the assumption that all patients developed hemorrhagic transformation, and the gray line represents the assumption that no patient had hemorrhagic transformation. The y axis measures the net benefit; the x-axis represents the predictive probability threshold.

respect to predictors from other domains. When we are developing the model, we considered the computation complexity, easy of prediction and interpretability. And this is the reason we performed feature selection rather than including all features in the model. Especially, in our case, this simpler model showed a good performance in both Caucasians and Chinese cohort. And

apparently, our model is less costly, in translation, as less features required to measure for a good model performance.

Moreover, the predictors in our model are part of routine clinical assessment for patients with acute ischemic stroke, all of which can be easily and rapidly obtained on admission. Since the calculation

Table 2. Domain-specific prediction performance for the support vector machine algorithm.

| | Caucasian cohort | | | | Han Chinese cohort | | | |
| | Only with cluster | | Excluding cluster | | Only with cluster | | Excluding cluster | |
	AUC	95% CI	AUC	95% CI	AUC	95% CI	AUC	95% CI
All	0.87	0.83–0.91	–	–	0.74	0.64–0.83	–	
Demographic (age)	0.62	0.54–0.69	0.85	0.80–0.90	0.49	0.36–0.60	0.73	0.64–0.83
Clinical characteristics (NIHSS score, SBP, DBP)	0.77	0.71–0.83	0.76	0.69–0.83	0.65	0.52–0.78	0.63	0.53–0.74
Baseline blood biomarkers (ALP, ALT, glucose, and creatine level)	0.74	0.67–0.82	0.81	0.76–0.86	0.63	0.53–0.73	0.68	0.58–0.79

ALP, alkaline phosphatase; ALT, alanine aminotransferase; AUC, area under the curve; CI, confidence interval; DBP, diastolic blood pressure; SBP, systolic blood pressure.

cannot be performed by humans, development of software is required before applying our ML model in clinical practice. After the development of automatic software, it may help a general internist to discuss thrombolysis with patients or families, even if the stroke physician is not available.

In addition, our decision curve (Figure 5b) suggests that intervention on ischemic stroke patients based on our model may lead to higher benefit. Although the thrombolysis is recommended by the current guidelines, informed consent is still required from the patient or family before thrombolysis in China. Whether applying our model would shorten the time and even affect the decision of thrombolysis need more study to further validation with a larger number of sICH cases. Another possible use of the model is that patients with a high risk of sICH after thrombolysis may most probably benefit from intensive monitoring (such as blood pressure management) or possibly a lower dose of intravenous alteplase. In addition, our SVM model may lead to lower sample sizes required to detect intervention effects with post-thrombolysis sICH.

Our findings should be considered in light of the following limitations. First, the Caucasians were from the VISTA clinical trials data set, which are not representative samples of patient population. However, the good performance of our model in an external Han Chinese cohort, suggests the generalizability of the model in different cultures and ethnicities. Notably, we constrained our model building by the Caucasians and did not

re-calibrate our model parameters to fit the Chinese cohort before evaluating the model by the datapoints in the set. Due to this, the result showed a subtle decrease in the model performance but still in a range to be regarded as useful in clinical practice. And it suggests that our model is reproducible and representative. Second, the Han Chinese cohort is from single hospital and the sample size is quite small. In the future, validation data from a multi-center study with larger sample size in China is required. Finally, the missing data of the predictors may generate bias. Although, the rate of missing data was less than 30% in the study, it still may be possible to affect the performance of the sICH prediction model. A large population with fewer missing data is needed.

Conclusion

In conclusion, we have shown that the SVM model can be powerful for prediction of sICH after stroke thrombolysis in both Caucasians and Han Chinese using readily available clinical data in a very short time. Our model has the potential for clinical research to identify subjects at high risk of sICH after thrombolysis and to evaluate therapeutic interventions for reducing post-thrombolytic sICH.

Declarations

Ethics approval and consent to participate
This study complies with the principles of the Declaration of Helsinki. VISTA is an anonymized

data repository of completed stroke trials. All included trials were performed under appropriate institutional review board and regulatory approvals, and only fully anonymized data are held by VISTA. The inclusion of the Han Chinese derivation cohort data was approved by Biomedical Research Ethics Committee of West China hospital, Sichuan University with the approval code [2016(339)].

Author contributions
Junfeng Liu: Writing – original draft.

Xinyue Chen: Formal analysis.

Xiaonan Guo: Methodology.

Renjie Xu: Methodology.

Yanan Wang: Data curation.

Ming Liu: Supervision.

ORCID iD
Junfeng Liu [iD] https://orcid.org/0000-0002-3858-0763

References

1. Yaghi S, Willey JZ, Cucchiara B, *et al.* Treatment and outcome of hemorrhagic transformation after intravenous alteplase in acute ischemic stroke a scientific statement for healthcare professionals from the American Heart Association/American Stroke Association. *Stroke* 2017; 48: e343–e361.

2. Emberson J, Lees KR, Lyden P, *et al.* Effect of treatment delay, age, and stroke severity on the effects of intravenous thrombolysis with alteplase for acute ischaemic stroke: a meta-analysis of individual patient data from randomised trials. *Lancet* 2014; 384: 1929–1935.

3. Yamaguchi T, Mori E, Minematsu K, *et al.* Alteplase at 0.6 mg/kg for acute ischemic stroke within 3 hours of onset: Japan Alteplase Clinical Trial (J-ACT). *Stroke* 2006; 37: 1810–1815.

4. Mori E, Minematsu K, Nakagawara J, *et al.* Effects of 0.6 mg/kg intravenous alteplase on vascular and clinical outcomes in middle cerebral artery occlusion: Japan Alteplase Clinical Trial II (J-ACT II). *Stroke* 2010; 41: 461–465.

5. Anderson CS, Robinson T, Lindley RI, *et al.* Low-dose versus standard-dose intravenous alteplase in acute ischemic stroke. *N Engl J Med* 2016; 374: 2313–2323.

6. Wang X, Robinson TG, Lee TH, *et al.* Low-dose vs standard-dose alteplase for patients with acute ischemic stroke: secondary analysis of the ENCHANTED randomized clinical trial. *JAMA Neurol* 2017; 74: 1328–1335.

7. Asuzu D, Nystrom K, Amin H, *et al.* Comparison of 8 Scores for predicting Symptomatic Intracerebral Hemorrhage after IV Thrombolysis. *Neurocrit Care* 2015; 22: 229–233.

8. Sung SF, Chen SCC, Lin HJ, *et al.* Comparison of risk-scoring systems in predicting symptomatic intracerebral hemorrhage after intravenous thrombolysis. *Stroke* 2013; 44: 1561–1566.

9. Lee SI, Celik S, Logsdon BA, *et al.* A machine learning approach to integrate big data for precision medicine in acute myeloid leukemia. *Nat Commun* 2018; 9: 42.

10. Kamal H, Lopez V and Sheth SA. Machine learning in acute ischemic stroke neuroimaging. *Front Neurol* 2018; 9: 945.

11. Dharmasaroja P and Dharmasaroja PA. Prediction of intracerebral hemorrhage following thrombolytic therapy for acute ischemic stroke using multiple artificial neural networks. *Neurol Res* 2012; 34: 120–128.

12. Yu Y, Guo D, Lou M, *et al.* Prediction of hemorrhagic transformation severity in acute stroke from source perfusion MRI. *IEEE Trans Biomed Eng* 2018; 65: 2058–2065.

13. Wang F, Huang Y, Xia Y, *et al.* Personalized risk prediction of symptomatic intracerebral hemorrhage after stroke thrombolysis using a machine-learning model. *Ther Adv Neurol Disord* 2020; 13: 1756286420902358.

14. Katsuki M, Kakizawa Y, Nishikawa A, *et al.* Postsurgical functional outcome prediction model using deep learning framework (Prediction One, Sony Network Communications Inc.) for hypertensive intracerebral hemorrhage. *Surg Neurol Int* 2021; 12: 203.

15. Katsuki M, Kakizawa Y, Nishikawa A, *et al.* Easily created prediction model using deep learning software (Prediction One, Sony Network Communications Inc.) for subarachnoid hemorrhage outcomes from small dataset at admission. *Surg Neurol Int* 2020; 11: 374.

16. Mehta RH, Cox M, Smith EE, *et al.* Race/ethnic differences in the risk of hemorrhagic complications among patients with ischemic stroke receiving thrombolytic therapy. *Stroke* 2014; 45: 2263–2269.

17. Ueshima S and Matsuo O. The differences in thrombolytic effects of administrated recombinant t-PA between Japanese and Caucasians. *Thromb Haemost* 2002; 87: 544–546.

18. Menon BK, Saver JL, Prabhakaran S, *et al.* Risk score for intracranial hemorrhage in patients with acute ischemic stroke treated with intravenous tissue-type plasminogen activator. *Stroke* 2012; 2293–299.

19. Noble WS. What is a support vector machine? *Nat Biotechnol* 2006; 24: 1565–1567.

20. Support vector machine applications in computational biology. In: Schoelkopf B, Tsuda K and Vert JP (eds) *Kernel methods in computational biology*. Cambridge, MA: MIT Press, 2019, pp. 71–92.

21. Ali M, Bath PMW, Curram J, *et al.* The Virtual International Stroke Trials Archive. *Stroke* 2007; 38: 1905–1910.

22. Aho K, Harmsen P, Hatano S, *et al.* Cerebrovascular disease in the community: results of a WHO collaborative study. *Bull World Health Organ* 1980; 58: 113–130.

23. Delgado P, Sahuquillo J, Poca MA, *et al.* Neuroprotection in malignant MCA infarction. *Cerebrovasc Dis* 2006; 21: 99–105.

24. Tissue plasminogen activator for acute ischemic stroke. *N Engl J Med* 1995; 1581–1587.

25. Wang S, Dai Y, Shen J, *et al.* Research on expansion and classification of imbalanced data based on SMOTE algorithm. *Sci Rep* 2021; 11: 24039.

26. Bonkhoff AK, Rubsamen N, Grefkes C, *et al.* Development and validation of prediction models for severe complications after acute ischemic stroke: a study based on the stroke registry of Northwestern Germany. *J Am Heart Assoc* 2022; 11: e023175.

27. Gilligan AK, Markus R, Read S, *et al.* Baseline blood pressure but not early computed tomography changes predicts major hemorrhage after streptokinase in acute ischemic stroke. *Stroke* 2002; 33: 2236–2242.

28. Wahlgren N, Ahmed N, Eriksson N, *et al.* Multivariable analysis of outcome predictors and adjustment of main outcome results to baseline data profile in randomized controlled trials: safe implementation of thrombolysis in Stroke-MOnitoring STudy (SITS-MOST). *Stroke* 2008; 39: 3316–3322.

29. Brott T, Lu M, Kothari R, *et al.* Hypertension and its treatment in the NINDS rt-PA stroke trial. *Stroke* 1998; 29: 1504–1504.

Visit SAGE journals online
journals.sagepub.com/home/tan

$SAGE journals