Extreme Gradient Boosting Model Has a Better Performance in Predicting the Risk of 90-Day Readmissions in Patients with Ischaemic Stroke

Yuan Xu,*,[1] Xinlei Yang,*,[1] Hui Huang,† Chen Peng,‡ Yanqiu Ge,‡ Honghu Wu,§ Jiajing Wang,‡ Gang Xiong,* and Yingping Yi,*

Object: Ischemic stroke readmission within 90 days of hospital discharge is an important quality of care metric. The readmission rates of ischemic stroke patients are usually higher than those of patients with other chronic diseases. Our aim was to identify the ischemic stroke readmission risk factors and establish a 90-day readmission prediction model for first-time ischemic stroke patients. *Methods:* The readmission prediction model was developed using the extreme gradient boosting (XGboost) model, which can generate an ensemble of classification trees and assign a predictive risk score to each feature. The patient data were split into a training set (5159) and a validation set (911). The prediction results were evaluated with the receiver operating characteristic (ROC) curve and time-dependent ROC curve, which were compared with the outputs from the logistic regression (LR) model. *Results:* A total of 6070 adult patients (39.6% female, median age 67 years) without any ischemic attack (IS) history were included, and 520 (8.6%) were readmitted within 90 days. The XGboost-based prediction model achieved a standard area under the curve (AUC) value of .782 (.729-.834), and the best time-dependent AUC value was .808 in 54 days for the validation set. In contrast, the LR model yielded a standard AUC value of .771 (.714-.828) and best time-dependent AUC value of .797. *Conclusions:* The XGboost model obtained a better risk prediction for 90-day readmission for first-time ischemic stroke patients than the LR model. This model can also reveal the high risk factors for stroke readmission in first-time ischemic stroke patients.

Key Words: Ischemic stroke—90-day readmission—time-dependent ROC—XGboost

Background

Stroke is one of the most common causes of death and disability worldwide.[1-3] It has already surpassed heart disease as the leading cause of death and disability.[4-6] In China, the occurrence of strokes has been increasing over the past 30 years. Stroke survivors that are readmitted have higher mortality rates, increased disability, poorer functional outcomes, and heavier economic burdens.

Readmission after an ischemic stroke is a common and costly event. Approximately 7%-15% of ischemic stroke patients experience short-term readmission.[7-10] Compared with 30-day readmissions, few studies have paid attention

From the *Medical Big-Data Center, The Second Affiliated Hospital of Nanchang University, Nanchang, Jiangxi China; †Department of Neurosurgery, The Second Affiliated Hospital of Nanchang University, Nanchang, Jiangxi China; ‡School of Public Health, Medical School, Nanchang University, Nanchang, Jiangxi China; and §Biobank Center , The Second Affiliated Hospital of Nanchang University, Nanchang, Jiangxi China.

Received July 15, 2019; revision received September 3, 2019; accepted September 22, 2019.

Funding: Funding for this study was provided by the Ministry of Science and Technology of the People's Republic of China (2018YFC1312902) and the Department of Science and Technology of Jiangxi Province (20181BBE58021, 20171BBH80025, 20171BCD40024).

Address correspondence to Yingping Yi, Medical Big-Data Center, The Second Affiliated Hospital of Nanchang University, No. 1 MinDe Road, Nanchang 330006, Jiangxi, China. E-mails: yyp66@126.com, 13807089678@126.com.

[1]Yuan Xu, Xinlei Yang contribute equally in this study.

1052-3057/$ - see front matter

https://doi.org/10.1016/j.jstrokecerebrovasdis.2019.104441

to 90-day readmissions. Almost half of the first readmissions occur within 90 days; unfortunately, little is known about the factors that contribute to this phenomenon.[11] Previous studies reported the risk factors associated with long-term readmission, including age, hypertension, pneumonia, atherosclerosis, and length of index admission.[12] Despite many studies on ischemic stroke, there are no standardized models for predicting the risk of 90 days readmissions after ischemic stroke.

With advances in big data solutions, data mining, and machine learning techniques can use all the variables in a dataset to identify the potentially highly predictive variables. Compared to traditional classification methods, machine learning techniques are more convenient as they do not presume any associations between variables or the predictive power of any particular variables.[13-15] Although deep learning techniques perform with a high level of predictive accuracy,[13,15] overfitting, and high computational costs still have present major challenges. In addition, deep learning is difficult to explain in terms of calculating the importance of the feature variables. Therefore, in this study, we aim to use a machine learning algorithm to develop a 90-day recurrent risk prediction model for ischemic attack (IS) patients to identify and rank the potential risk factors for IS recurrence. According to our investigation, this is one of the first studies to apply machine learning methods to stroke readmission prediction. If this model can be proven to be effective, it can help to assess the patient discharge status, control rising hospitalization costs and reduce the readmission rate for short-term stroke patient management.

Methods

Data Source

This retrospective study used de-identified data from the database of Jiang Xi Province Medical Big Data Engineering & Technology Research Center, which gathers all original clinical data from 2006 to 2019 of 4.97 million patients. There are 13.1 million outpatient records, .87 million inpatients and 6.1 million medical records in this database. All data was de-identified by removal of 18 identifiers, includes names, birth dates, contact information, ID numbers, hospital numbers, etc.

Patients and Methods

This was a retrospective study used de-identified data from the Jiang Xi Province Medical Big Data Engineering & Technology Research Center. It was approved by the ethics committee of the Second Affiliated Hospital of Nanchang University. All patients 18 years and older with primary ischemic stroke or Transient ischemic attack (TIA) admitted to the second affiliated hospital of Nanchang university from 2007 February first to 2017 July first were identified in this study. Ischemic stroke was defined as a neurologic deficit lasting more than 24 hours because of ischemic lesions or transient ischemic attack where computed tomography or magnetic resonance imaging showed infarctions related to the clinical findings, while TIA was defined as a clinical diagnosis with transient focal cerebral dysfunction lasting less than 24 hours with no objective evidence of brain infarction on imaging. And in this study we conformed it with the discharging diagnosis using the International Classification of Disease, Tenth Edition (ICD 10). The baseline information includes the demographics, history of IS or TIA, laboratory data, complications, comorbidity. There were 21,366 adult patients (>18 years) admitted to the second affiliated hospital of Nanchang University with IS (ICD 10: I63.0-9) or TIA (ICD 10: G45.901).[16] Patients were excluded based on the following criteria: (1) death, transfer or 48 hour readmission; (2) previous stroke or TIA record in the admission records; (3) outpatient follow-up <90 days; (4) readmission in 90 days without stroke recurrence; (5) outpatient interview interval larger than 180 days. Finally, 6070 patients were eligible for this study, of which 520 (8.6%) patients were readmitted within 90 days (Fig 1).

Variables

The main outcome variable was the occurrence of ischemic stroke or TIA readmission within 90-day after discharge. The patients baseline information included age, gender, current smoke, current alcohol, history of IS or TIA, and length of stay (LOS).Pre-existing comorbidities conditions could potentially influence in the admission or readmisssion. The comorbidities included the presence of hypertension, diabetes mellitus, coronary heart disease, hyperlipidemia, atrial fibrillation, renal insufficiency, and renal cyst before the initial admission. Complications included pneumonia, urinary tract infection. Laboratory data contained complete blood count (sysmex, Japan), comprehensive metabolic panel testing (olympus, Japan), kidney and liver function in all groups. All the blood testing values were taking within 24 hours after admission before treatment.

Statistical Analysis

All the statistical analysis was completed using IBM SPSS 23.0. The continuous variables were presented as the mean (SD) if they were normally distributed and as the median (IQR) if they were not normally distributed. The dichotomous variables were presented as percentages (n). The categorical comparisons between groups were made by the Pearson χ^2 test. Student's t test was applied for the comparison of sufficiently normal data, and the Mann-Whitney U test or Kruskal-Wallis test was applied in the case of nonnormally distributed data. The baseline factors that showed significant differences ($P < .05$ according to the χ^2 test and Student's t test) between the 2 groups were selected in LR model as adjusted variables.

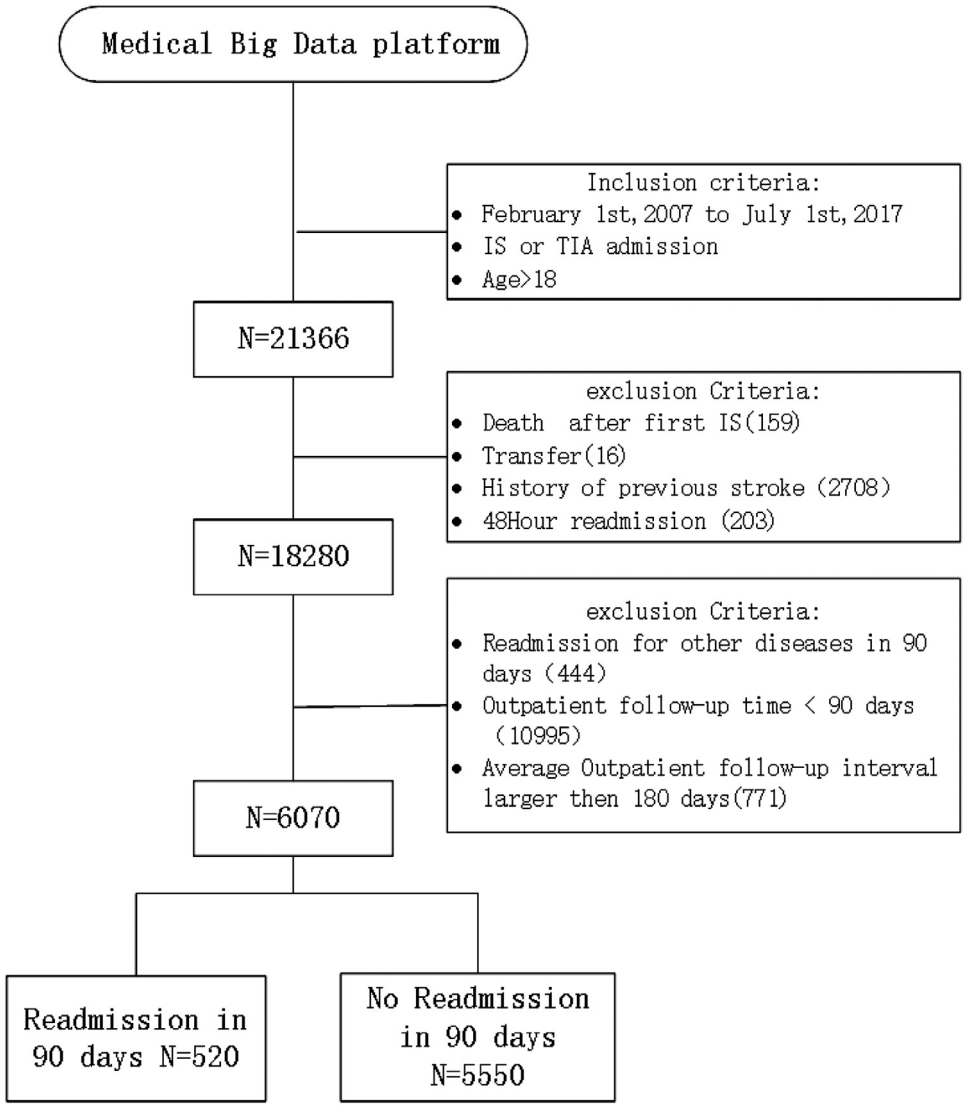

Figure 1. *Patients' flowchart. Hospital discharge diagnose according to International Classification of Diseases ICD 10. Abbreviations: IS, ischemic stroke; TIA, transient ischemic attack.*

Logistic regression (LR) model was performed by the forward stepwise procedure to assess the odds ratio (OR) and 95% confidence interval (CI) for the association between significant variables and 90-day readmission with ischaemic stroke.[17] Model were constructed as follows after adjust steps: age, LOS, Red blood cell distribution width (coefficient of variation (CV)), neutrophil percentage, alkaline phosphatase, hypertension, and pneumonia. The LR model was performed using R version 3.5.1 in R-studio with a statistics package.

XGboost Model

Supervised Learning is a kind of classification algorithm, which is to finding the parameters θ that best fit the training data x_i and labels y_i In order to train the supervised learning model, we need to define the objective function to measure how well the model fit the training data. Objective functions consist of 2 parts: training loss and regularization term:

$$Obj(\theta) = L(\theta) + \Omega(\theta) \tag{1}$$

where L is the training loss function, and Ω is the regularization term. The training loss measures how predictive our model is with respect to the training data.

Extreme gradient boosting (XGboost) is an improved supervised learning algorithm based on the Gradient Boosting Decision Tree algorithm. It makes use of 3 terms of Taylor expansion to make an approximation. It can be seen clearly that the final objective function only depends on the first and second derivatives of each data point on the error function.

$$Obj^{(t)} \cong \sum_{i=1}^{n} \left[l\left(y_i, \hat{y}^{(t-1)}\right) + g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) \right] + \Omega(f_t) + CONSTANT \tag{2}$$

where l is the training loss function (In this study, we use used logistic loss as loss function) and $\Omega(f_t)$ is the regularization term,

As an integrated learning algorithm, it combines the predictions from an ensemble of weak regression trees, which are added sequentially to the model to maximize the predictive performance and minimize the model complexity. At the same time, XGboost adds a complexity control model and learns from the random forests to reduce the computational load, making the model hard to over-fit. With the above characteristics, XGboost has drawn more attention for prediction model construction and risk identification in the medical field.[18,19] The XGboost model was implemented in Python (v.3.6) using the XGboost (v.0.7) package. In this study we randomly select 5159 patients as training data and 911 patients as validate data. Use learning curve and grid search method to find out best parameters with best accuracy and optimal generalization error of model.

Model Comparison

To avoid model overfitting, the patients were randomly split into a training set and a validation set. There were 5159 (85%) patients in the training set and 911 (15%) patients in the validation set. In the standard ROC analysis, the individual's status and event-time are fixed for the entire study period. In contrast, the status of an individual is observed and updated at each time point in a time-dependent ROC curve analysis. Thus, a ROC curve can be constructed at several time points, and the markers' predictive ability can be compared.[20] Compared to the standard ROC curve, the time-dependent ROC curve is an efficient tool to measure the performance of a candidate marker given the true disease status of individuals at every time point. In general, a baseline marker value is used to compute the predictive ability, but it can become weaker as the target time becomes farther from the baseline. Therefore, in this study, we use a time-dependent ROC curve to evaluate the effectiveness of each prediction model, which can reveal the effective differences at every time point for each model. The time-dependent ROC curve was generated using R version 3.5.1 in Rstudio with the "timeROC" version 0.3 package.

Results

Population and Characteristics

Among the 6070 eligible patients in this study, 520 (8.6%) were readmitted in 90 days. The median age was 67 years, and 2404 (39.6%) were female. The mean readmission interval was 33.3 days, and the median readmission interval was 26.5 days.

Factors Associated with 90-Day Readmission

Patients in the readmission group were more likely to have complications and longer hospital stays when compared with the nonreadmission group. There was no significant difference in sex, age, history of smoking, and alcohol intake between these 2 groups (Table 1). This study included 30 lab test variables, 7 of which had significant differences between the 2 groups, including the neutrophil absolute/percentage, red blood cell distribution width (CV), and alkaline phosphatase, lactate dehydrogenase and homocysteine levels.

The Predictors for a Stroke Patient's Readmission Using a Logistic Regression

The variables were selected in the LR models according to the χ^2 test and Student's t test ($P < .05$), as shown in Table 1. The LR analysis revealed that age (OR .99, 95% CI .98-1.00), hospital days (OR 1.015, 95% CI 1.007-1.022), RDW (CV) (OR 1.17, 95% CI 1.11-1.24), neutrophil percentage (OR 1.01, 95% CI 1.00-1.02), ALP (OR 1.003, 95% CI1.00-1.005), hypertension (OR 4.60, 95% CI 3.80-5.58), and pneumonia (OR 1.46, 95% CI 1.07-1.97) were independently associated with stroke patient readmission.

Predictors for a Stroke Patient's Readmission Using the XGboost Model

The XGboost model is insensitive to missing data, and the final algorithm used a total of 44 variables after data cleaning. The optimal parameter settings for the model were obtained by a grid search and an iterative algorithm. The optimal parameter settings are max-deep = 4, min-child-weight = 2, gamma = 1, subsample = 1, and colsample-bytree = .6, where the training set area under the curve = .865 and the test set AUC = .792. As shown in Figure 2, the top 10 factors were hypertension, RDW (CV), direct bilirubin, LOS, pneumonia, ALP, C-reactive protein, Aspartate aminotransferase, diabetes mellitus, and glycosylated haemoglobin. The specific importance scores are shown in Table 3.

Comparison of Model Performance

Using a validated set to compare the prediction efficiency of the 2 models, the AUC values of the XGboost model and logistic model for predicting 90-day readmission were .782 (.729-.834) and .771 (.714-.828), respectively (Fig 3). Net Reclassification Improvement of XGboost model is .03, the prediction ability of the new model is improved, and the correct classification proportion is increased by 3% compared to LR model. The time-dependent ROC curve from 7 to 90 days shows significant differences between the 2 models' AUC values using Student's t test ($P < .0001$). The maximum difference in the AUC values between the 2 models is .047, and the mean AUC difference is .013. The readmission prediction efficiency is basically the same from 14 to 34 days. The best prediction efficiency time for the XGboost and logistic models is 54 days, and the best AUC value for each model is .8075 and .7973, respectively. (Fig 4)

Table 1. *Baseline characteristics of patients with acute ischemic stroke according to the readmission group*

Variable	Readmission group (N = 520)	Nonreadmission group (N = 5550)	F score	*P* value
Female, no. (%)	206 (39.6)	2198 (39.6)	.000	.996
Age (y), mean (SD)	66.7 (12.2)	66.2 (11.6)	.912	.362
Smoke, no. (%)	50 (9.6)	553 (10.0)	.065	.799
Alcohol, no. (%)	73 (14)	678 (12.2)	1.456	.228
LOS (d), median (IQR)	12 [9,17]	11 [8,14]	-4.744	.000
Hypertension, no. (%)	361 (69.4)	1744 (31.4)	303.074	.000
Diabetes mellitus, no. (%)	146 (28.1)	822 (14.8)	62.425	.000
Pneumonia, no. (%)	63 (12.1)	300 (5.4)	38.073	.000
Urinary tract infection, no. (%)	15 (2.9)	54 (1.0)	15.461	.000
CHD, no. (%)	51 (9.8)	295 (5.3)	17.851	.000
Hyperlipidaemia, no. (%)	37 (7.1)	268 (4.8)	5.209	.022
Atrial fibrillation, no. (%)	25 (4.8)	132 (2.4)	11.136	.001
Renal insufficiency, no. (%)	17 (3.3)	92 (1.7)	7.002	.008
Renal cyst, no. (%)	20 (3.8)	207 (3.7)	.018	.894
HbA1c (%) median (IQR)	5.9 [5.5,6.6]	5.9 [5.5,6.6]	-.750	.453
Neutrophil percentage (%)	68.7 (11.1)	66.6 (10.9)	3.869	.000
Neutrophil absolute (10^9/L), mean (SD)	5.2 (2.6)	4.9 (2.5)	2.717	.007
Fg (g/l)	3.0 (1.0)	3.0 (.9)	1.474	.141
RBC (10^12/L)	4.2 (.7)	4.3 (.6)	-2.267	.024
RDW (SD) (fl)	44.0 (8.2)	43.4 (6.9)	1.533	.126
RDW (CV) (%)	13.2 (1.7)	12.9 (1.23)	3.912	.000
Globulin (g/l)	26.1 (5.2)	26.1 (4.4)	.017	.987
Lp(a) (mg/dl) median (IQR)	19.4 [10.9,34]	19.2 [10.8,33.0]	-.857	.392
ApoB (g/l)	.9 (.3)	.9 (.3)	.530	.596
LDL (mmol/l)	2.8 (.9)	2.8 (.9)	-.210	.834
HDL (mmol/l)	1.1 (.3)	1.1 (.3)	-.316	.752
ALP (u/l)	93.0 (35.5)	88.8 (33.3)	2.504	.013
CRP (mg/l)	11.6 (24.9)	9.8 (24.9)	1.114	.265
GGT (lu/l), median (IQR)	25.2 [17.7,38.1]	23.8 [16.9,36.9]	-1.664	.096
LD (lu/l), median (IQR)	181.9 [154.1,218.6]	174.9 [151.4,207.4]	-3.057	.002
AST (u/l), median (IQR)	21.4 [17.1,28.0]	21.2 [17.5,26.5]	-.765	.444
ALT (u/l), median (IQR)	15.3 [11.0,23.2]	15.2 [11.3,22.2]	-.012	.991
CK (lu/l), median (IQR)	81.4 [55.4,119.6]	84.9 [60.3,122.9]	-1.429	.153
CK-MB(lu/l), median (IQR)	11.9 [8.8,15.7]	11.5 [9.0,15.1]	-.161	.872
TB (μmol/l), median (IQR)	12 [8.8,16.8]	11.5 [8.7,15.8]	-1.173	.241
IBIL (μmol/l)	8.0 (5.5)	7.6 (5.2)	1.668	.095
DBIL (μmol/l)	5.6 (3.8)	5.6 (4.1)	-.172	.863
Hcy (μmol/l), median (IQR)	14 [10.9,34]	13.3 [10.7,17.0]	-1.967	.049
TC (mmol/l)	4.6 (1.1)	4.6 (1.0)	.298	.766
Triglyceride (mmol/l), median (IQR)	1.3 [1,1.8]	1.3 [1,1.9]	-.541	.588
K (mmol/l)	3.9 (.5)	3.9 (.4)	-.992	.321
Uric acid (μmol/l)	337.1 (110.8)	335.6 (106.0)	.286	.775
Creatinine (μmol/l), median (IQR)	75.3 [62.0,90.9]	74.0 [61.3,89.5]	-1.121	.262
GLU (mmol/l)	6.0 (2.9)	5.9 (2.5)	1.045	.296

Abbreviations: ALP, Alkaline phosphatase; ALT, Alanine aminotransferase; ApoB, Apolipoprotein B; AST, Aspartate aminotransferase; CHD, Coronary heart disease; CK, Creatine kinase; CRP, C reactive protein; DBIL, Direct bilirubin; Fg, Fibrinogen concentration; GGT, Gamma-Glutamyl Transferase; GLU, Glucose; HbA1c, Glycosylated haemoglobin; HDL, High density lipoprotein; IBIL, Indirect bilirubin; LD, Lactate dehydrogenase; LDL, Low density lipoprotein; LOS, Length of stay, hospital days; Lp(a), Lipoprotein(a); MB, creatine kinase isoenzymes; RBC: Red blood cell; RDW: Red blood cell distribution width; TB, Total bilirubin; TC, total cholesterol; IQR, Inter Quartile Range.

Discussion

In the previous studies, the values of blood testing were rarely included in the readmission model. In this study we want to figure out whether blood testing before treatment make sense in the readmission within 90-day. In the present retrospective study, the percentage of readmissions due to ischaemic stroke was 8.6%. Result showed that the LOS, RDW(CV), ALP, hypertension and pneumonia were consistently highly associated with the readmission event in each prediction model (Tables 2 and 3).

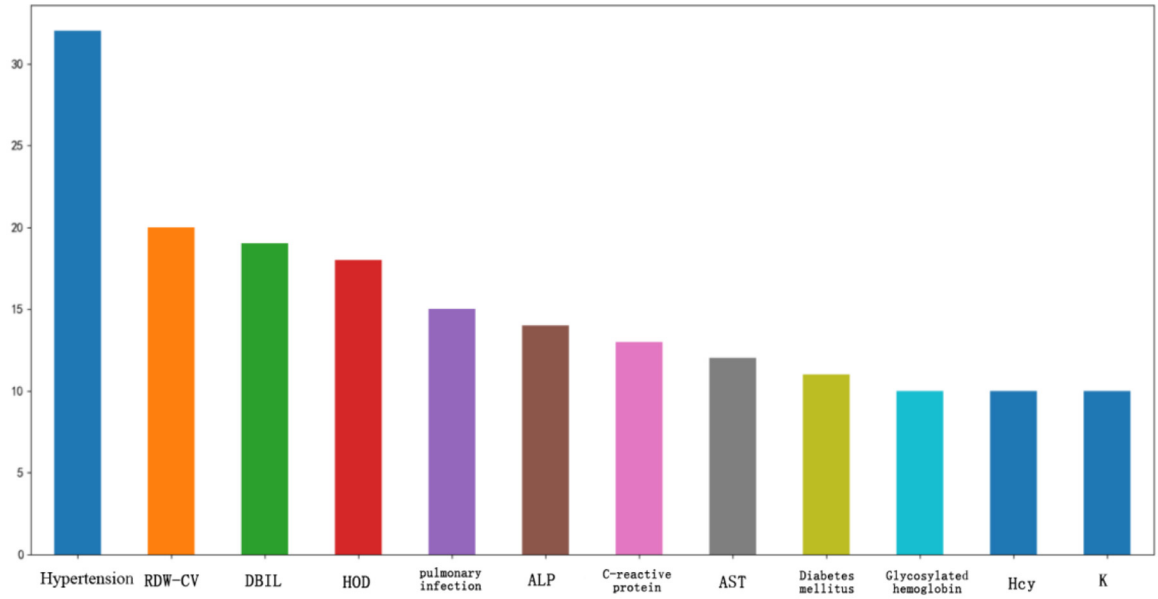

Figure 2. *Feature importance score of 90 days recurrence XGboost model. The features with the higher importance score(Y-axis) will be selected by more Boosting Decision Trees. Abbreviation: XGboost, extreme gradient boosting.*

The LOS was a significant predictor of ischemic stroke recurrence within 90 days, which was consistent with previous studies.[11] The LOS may be used as a crude proxy for stroke severity and can reflect the readmission rate. Hypertension has been reported as an independent risk factor for ischemic stroke and is associated with readmission among stroke patients. The RDW was significantly correlated with the stroke prognosis, and patients with higher RDW values showed poorer outcomes. The association between ALP and readmission needs to be studied in the future. Many studies have found that infection is the most common reason for readmission, and we confirmed this finding in our study. Some lab test features in the prediction model have been validated in previous studies. The association between the serum ALP level and stroke recurrence and other poor functional outcomes after stroke were explored.[21,22] The glycosylated haemoglobin level acted as an independent predictor in stroke recurrence.[23] A lower tHcy level was significantly associated with a reduction in first-time stroke risk in Chinese adults with hypertension.[24] These studies proved the effectiveness and robustness of this model from another aspect. In addition, this model exposes some lab test features that have not been rigorously studied, such as AST and DBIL, which might inspire us to consider other research directions for stroke recurrence.

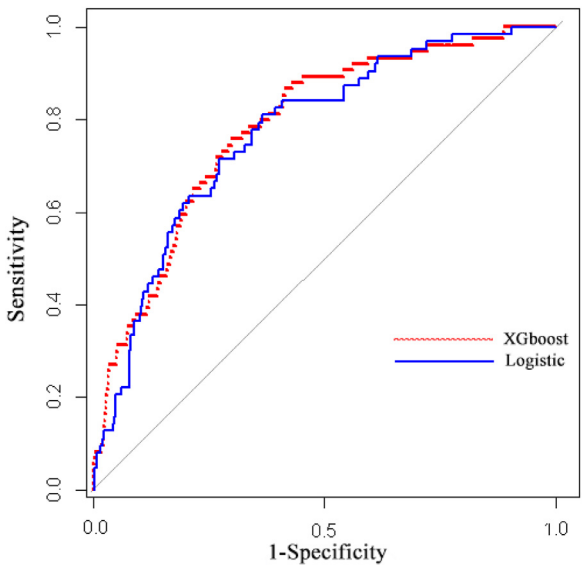

Figure 3. *ROC curves for each model predictions of stroke recurrence in 90 days. Abbreviation: ROC, receiver operating characteristic.*

The XGboost algorithm was applied to structured and unstructured patient data from the EMR to develop a prediction model for stroke readmission within 90 days. The standard AUC of the XGboost model is better than that of the traditional LR model in first-time ischemic stroke patients. The time-dependent ROC shows that the optimal performance is .808, and the AUC curve of the LR drops faster in the short term (<15 days), indicating the difference between the 2 models for risk prediction. Although the convergence speed of the LR model is faster, XGboost, as a boosting model, is a nonlinear model and has the characteristics of infinite approximation to the Vapnik-Chervonenkis dimension of the sample. Therefore, its Vapnik-Chervonenkis dimension is much larger than the d+1 dimension of the linear LR classifier, which can better fit the sample data. At the same time, the XGboost model will search for other features, B, that can minimize the residual loss under the subtree of feature A because the XGboost model has the capability of finding the best

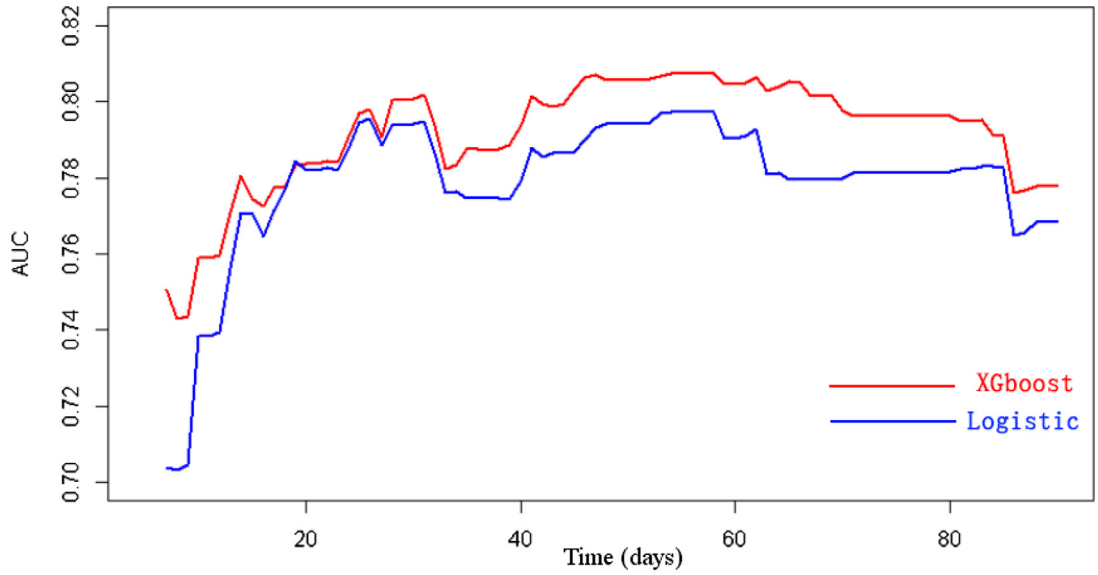

Figure 4. *Time-dependent AUC Curve of each recurrence model from 7 to 90 days. Abbreviation: AUC, area under the curve.*

Table 2. *Binary logistic regression analysis of predictors for stroke patient readmission*

Variables	B	OR	95% CI	*P* Value
Age	-.0098	.99	.98-1.00	.0444
LOS	.01494	1.015	1.007-1.022	.0009
RDW (CV)	.1586	1.17	1.11-1.24	<.0001
Neutrophil percentage	.0107	1.01	1.00-1.02	.0326
ALP	.0029	1.003	1.00-1.005	.0416
Hypertension	1.5254	4.60	3.80-5.58	<.0001
Pneumonia	.3797	1.46	1.07-1.97	.0419

Abbreviations: ALP,Alkaline phosphatase; CI,confidence interval; LOS, length of stay; OR, odds ratio; RDW, Red blood cell distribution width.

feature combination. The most relevant studies were based on the COX or LR analysis to explore a single variable hazard risk. There are a few studies that applied machine learning to predict cardiovascular outcomes.

Table 3. *Feature importance of the XGboost model*

No.	Variable	Importance score
1	Hypertension	32
2	RDW(CV)	20
3	DBIL	19
4	LOS	18
5	Pneumonia	15
6	ALP	14
7	C reactive protein	13
8	AST	12
9	Diabetes mellitus	11
10	HbA1c	9
11	Hcy	9
12	K	9

Abbreviations: ALP,alkaline phosphatase; AST, aspartate aminotransferase; DBIL, direct bilirubin; HbA1c, glycosylated haemoglobin; LOS, length of stay.

Ambale-Venkatesh et al applied several machine learning algorithms (random forest, LASSO-COX, and AIC-COX) to a prospective population-based observational cohort study, which was one of the few studies that used machine learning algorithms for CVD (cerebrovascular disease) event prediction. The study listed the top 20 variables ranked by the variable importance associated with 6 outcomes, and the best concordance index for stroke events is .77.[25] To our knowledge, the model developed in our study obtains one of the most comprehensive sets of variables and performs well in comparison with previously published models. To further interpret the readmission prediction models, we reveal the top 7 features measured by the LR model and the top 12 features measured by the XGboost model. As illustrated in Tables 2 and 3, these features include hypertension, age, RDW (CV), DBIL, LOS, pneumonia, ALP, C-reactive protein, AST, diabetes mellitus, glycosylated haemoglobin, Hcy, and K. Among these features, some complications (hypertension, pneumonia, and diabetes mellitus) have been validated as high-risk factors of stroke recurrence by experts and a large number of related studies.[26-29] Five features show high importance in

both models, including hypertension, pneumonia, LOS, RDW (CV), and ALP. Although age is considered significant in many studies, there is no significant difference in age between the trial group and the control group in this study; therefore, it is only significant in the LR model but less important in the XGboost model.

In addition, there are a number of validated features that are not included in this study that did not show significant importance. For example, Wangqin R et al. suggested that hypertension with poor blood pressure control, National Institute of Health stroke scale score, lipid-lowering therapy and intracranial arterial stenosis were risk factors associated with stroke.[10] Liu Q et al. found that marital status was associated with all adverse stroke outcomes in patients with acute ischemic stroke.[30]

Limitation of the Study

This study has several limitations. First, all the patients were from 1 single centre, and caution is advised in generalizing our findings. Second, a few patients may have been readmitted to other hospitals. Third, stroke subtypes were not included in this study. According to the Trial of Org classification, stroke is classified into large artery atherosclerosis, small artery occlusion, cardioembolism, other aetiology, and undetermined aetiology. The index stroke subtypes have been validated to be a specific risk for recurrent stroke.[31] We know that variables like medication at discharge or stroke severity were clinically important and should be included in this study, but it is really regrettably data from Jiang Xi Province Medical Big Data Engineering & Technology Research Center this part of data are seriously missing. This might have influences in the model, we will refine it in the future use other datasets.

Conclusions

We developed an XGboost-based model that simultaneously estimates the probabilities of IS recurrence readmission in 90 days and demonstrates the importance of the recurrence risk factors for hospitalized stroke patients. The experiments demonstrated that it is feasible to apply XGboost-based prediction models for stroke patient management, with optimal performance in comparison with classic LR methods. This model has good predictive power with more comprehensive risk factor information to aid in decision making at the point of stroke patient discharge.

Ethical Approval and Consent to Participate

The study protocol was reviewed and approved by the institutional review board of The Second Affiliated Hospital of Nanchang University, China; approval NO. [2017] 029.

Conflict of Interest

The authors declare that they have no competing interests.

References

1. Feigin VL, Norrving B, Mensah GA. Global burden of stroke. Circ Res 2017;120:439-448.
2. Feigin VL, Roth GA, Naghavi M, et al. Global burden of stroke and risk factors in 188 countries, during 1990-2013: a systematic analysis for the Global Burden of Disease Study 2013[J]. Lancet Neurol 2016;15:913-924.
3. Feigin VL, Mensah GA, Norrving B, et al. Atlas of the Global burden of stroke (1990-2013): The GBD 2013 Study. Neuroepidemiology 2015;45:230-236.
4. Chen Z, Jiang B, Ru X, et al. Mortality of stroke and its subtypes in China: results from a Nationwide Population-Based Survey. Neuroepidemiology 2017;48:95-102.
5. Wang W, Jiang B, Sun H, et al. Prevalence, incidence, and mortality of stroke in China: results from a Nationwide Population-Based Survey of 480 687 adults. Circulation 2017;135:759-771.
6. Liu L, Wang D, Wong K S, et al. Stroke and stroke care in China: huge burden, significant workload, and a national priority. Stroke 2011;42:3651-3654.
7. Bjerkreim ATM, Thomassen LMP, Brøgger JMP, et al. Causes and predictors for hospital readmission after ischemic stroke. J Stroke Cerebrovasc Disease 2015;24:2095-2101.
8. Vahidy FS, Donnelly JP, McCullough LD, et al. Nationwide estimates of 30-Day readmission in patients with ischemic stroke. Stroke 2017;48:1386-1388.
9. Terman SW, Reeves MJ, Skolarus LE, et al. Association between early outpatient visits and readmissions after ischemic stroke[J]. Circ Cardiovasc Qual Outcomes 2018;11:e4024.
10. Wangqin R, Wang X, Wang Y, et al. Risk factors associated with 90-day recurrent stroke in patients on dual antiplatelet therapy for minor stroke or high-risk TIA: a subgroup analysis of the CHANCE trial[J]. Stroke Vasc Neurol 2017;2:176-183.
11. Lin HJ, Chang WL, Tseng MC. Readmission after stroke in a hospital-based registry: risk, etiologies, and risk factors[J]. Neurology 2011;76:438-443.
12. Waters MF, Hoh BL, Lynn MJ, et al. Factors associated with recurrent ischemic stroke in the medical group of the SAMMPRIS trial[J]. JAMA Neurol 2016;73:308-315.
13. Golas S B, Shibahara T, Agboola S, et al. A machine learning model to predict the risk of 30-day readmissions in patients with heart failure: a retrospective analysis of electronic medical records data[J]. BMC Med Inform Decis Mak 2018;18:44.
14. Maali Y, Perez-Concha O, Coiera E, et al. Predicting 7-day, 30-day and 60-day all-cause unplanned readmission: a case study of a Sydney hospital[J]. BMC Med Inform Decis Mak 2018;18:1.
15. Schmidhuber J. Deep learning in neural networks: an overview[J]. Neural Netw 2015;61:85-117.
16. Geary L, Aronius J, Wettermark B, et al. Sociodemographic factors are associated with utilisation of statins after ischaemic stroke/TIA[J]. Int J Clin Pract 2017;71:e12936.
17. Wang S, Zhan T, Chen Y, et al. Multiple sclerosis detection based on biorthogonal wavelet transform, RBF Kernel Principal Component Analysis, and Logistic Regression. IEEE Access 2016;4:7567-7576. 4.
18. Shimoda A, Ichikawa D, Oyama H. Using machine-learning approaches to predict non-participation in a nationwide general health check-up scheme. Comput Method Progr Biomed 2018;163:39-46.
19. Hong WS, Haimovich AD, Taylor RA. Predicting hospital admission at emergency department triage using machine learning. PLoS One 2018;13:e201016.

20. Kamarudin AN, Cox T, Kolamunnage-Dona R. Time-dependent ROC curve analysis in medical research: current methods and applications. BMC Med Res Methodol 2017;17:53.
21. Zong L, Wang X, Li Z, et al. Alkaline phosphatase and outcomes in patients with preserved renal function: results from China National Stroke Registry. Stroke 2018;49: 1176-1182.
22. Zhong C, You S, Chen J, et al. Serum alkaline phosphatase, phosphate, and in-hospital mortality in acute ischemic stroke patients. J Stroke Cerebrovasc Dis 2018;27: 257-266.
23. Wu S, Shi Y, Wang C, et al. Glycated hemoglobin independently predicts stroke recurrence within one year after acute first-ever non-cardioembolic strokes onset in a Chinese cohort study. PLoS One 2013;8:e80690.
24. Huang X, Li Y, Li P, et al. Association between percent decline in serum total homocysteine and risk of first stroke. Neurology 2017;89:2101-2107.
25. Ambale-Venkatesh B, Yang X, Wu CO, et al. Cardiovascular event prediction by machine learning: The Multi-Ethnic Study of Atherosclerosis. Circ Res 2017;121: 1092-1101.
26. Hirayama A, Goto T, Faridi M K, et al. Age-related differences in the rate and diagnosis of 30-day readmission after hospitalization for acute ischemic stroke[J]. Int J Stroke 2018;13:717-724.
27. Suda S, Aoki J, Shimoyama T, et al. Stroke-associated infection independently predicts 3-month poor functional outcome and mortality. J Neurol 2018;265:370-375.
28. Erdur H, Scheitz JF, Ebinger M, et al. In-hospital stroke recurrence and stroke after transient ischemic attack: frequency and risk factors. Stroke 2015;46:1031-1037.
29. Arsava EM, Kim GM, Oliveira-Filho J, et al. Prediction of early recurrence after acute ischemic stroke. JAMA Neurol 2016;73:396-401.
30. Liu Q, Wang X, Wang Y, et al. Association between marriage and outcomes in patients with acute ischemic stroke. J Neurol 2018;265:942-948.
31. Toni D, Di Angelantonio E, Di Mscio MT, et al. Types of stroke recurrence in patients with ischemic stroke: a substudy from the PRoFESS trial. Int J Stroke 2014;9:873-878.