Journal of Translational Medicine

RESEARCH Open Access

ICH-LR2S2: a new risk score for predicting stroke-associated pneumonia from spontaneous intracerebral hemorrhage

Jing Yan[1,2†], Weiqi Zhai[3,7,8,9†], Zhaoxia Li[1,2], LingLing Ding[1,2], Jia You[3,7,8,9], Jiayi Zeng[3], Xin Yang[2], Chunjuan Wang[1,2], Xia Meng[1,2], Yong Jiang[1,2], Xiaodi Huang[6], Shouyan Wang[3,7,8,9], Yilong Wang[1,2], Zixiao Li[1,2,4,5*], Shanfeng Zhu[3,7,8,9*], Yongjun Wang[1,2,5], Xingquan Zhao[1,2,5*] and Jianfeng Feng[3,7,8,9]

Abstract

Purpose: We develop a new risk score to predict patients with stroke-associated pneumonia (SAP) who have an acute intracranial hemorrhage (ICH).

Method: We applied logistic regression to develop a new risk score called ICH-LR2S2. It was derived from examining a dataset of 70,540 ICH patients between 2015 and 2018 from the Chinese Stroke Center Alliance (CSCA). During the training of ICH-LR2S2, patients were randomly divided into two groups – 80% for the training set and 20% for model validation. A prospective test set was developed using 12,523 patients recruited in 2019. To further verify its effectiveness, we tested ICH-LR2S2 on an external dataset of 24,860 patients from the China National Stroke Registration Management System II (CNSR II). The performance of ICH-LR2S2 was measured by the area under the receiver operating characteristic curve (AUROC).

Results: The incidence of SAP in the dataset was 25.52%. A 24-point ICH-LR2S2 was developed from independent predictors, including age, modified Rankin Scale, fasting blood glucose, National Institutes of Health Stroke Scale admission score, Glasgow Coma Scale score, C-reactive protein, dysphagia, Chronic Obstructive Pulmonary Disease, and current smoking. The results showed that ICH-LR2S2 achieved an AUC = 0.749 [95% CI 0.739–0.759], which outperforms the best baseline ICH-APS (AUC = 0.704) [95% CI 0.694–0.714]. Compared with the previous ICH risk scores, ICH-LR2S2 incorporates fasting blood glucose and C-reactive protein, improving its discriminative ability. Machine learning methods such as XGboost (AUC = 0.772) [95% CI 0.762–0.782] can further improve our prediction performance. It also performed well when further validated by the external independent cohort of patients (n = 24,860), ICH-LR2S2 AUC = 0.784 [95% CI 0.774–0.794].

Conclusion: ICH-LR2S2 accurately distinguishes SAP patients based on easily available clinical features. It can help identify high-risk patients in the early stages of diseases.

Introduction

As the major complication of a stroke, stroke-associated infections (SAIs) have resulted in increased mortality [1]. It is reported that approximately 30% of post-stroke patients have infections [2]. Among those with infections, stroke-associated pneumonia (SAP), the most acute type of SAI, has the worst impact on

*Correspondence: lizixiao2008@hotmai.com; zhusf@fudan.edu.cn; zxq@vip.163.com
†Jing Yan and Weiqi Zhai are contributed equally to this work.
[1] Vascular Neurology, Department of Neurology, Beijing Tiantan Hospital, Capital Medical University, Beijing 100070, China
[3] Institute of Science and Technology for Brain-Inspired Intelligence, Fudan University, Shanghai 200433, China
Full list of author information is available at the end of the article

functional outcomes [3, 4]. The incidence rate of SAP is approximately 10% among stroke patients but could be as high as 40% among high-risk populations [5]. Except for the relatively high incidence rate, SAP has serious consequences such as increased mortality, extended hospital stays, and deteriorated functional outcomes at discharge [6].

However, previous studies lacked effective prophylactic treatment for SAP in clinical practice [7, 8]. One of the reasons for such a failure of clinical trials of prophylactic antibiotics is the difficulty in selecting patients with the highest risk of SAP [9]. Consequently, clinical practice has no suitable routine measures to identify patients with the highest risk of developing SAP. Accurately selecting such patients could improve the results of future clinical trials. For prevention and treatment, it is, therefore, crucial to accurately identify those patients at risk during the acute phases of stroke.

SAP mainly includes acute ischemic stroke (AIS) and intracranial hemorrhage (ICH) stroke pneumonia. Rates of pneumonia are reportedly higher in patients with ICH than those with AIS [10]. However, most SAP research focuses on AIS, with relatively few studies examining ICH [11–13]. Therefore, it requires an objective and easily applicable model that predicts the probability of the development of pneumonia in ICH patients. Several recently developed clinical scores are available to predict SAP for stroke patients. Examples are the Pneumonia Score [14], Veteran's Health Administration cohort score [15], ICH-APS (Intracerebral Haemorrhage-Associated Pneumonia Score) A and B [16], Pneumonia (PNA) prediction score [17], ISAN (Prestroke Independence, Sex, Age, NIHSS) score [18], ACDD4 (Age, Dysarthria, Dysphagia, CHF) [19], and PASS (Preventive Antibiotics in Stroke Study) [20]. Although most of these scores are SAP scoring models for AIS, some are also suited to predicting ICH patient scores [21]. According to previous studies [22, 23], SAP is associated with various risk factors, including older age, male gender, dysphagia, stroke-induced immunodepression syndrome, and chronic obstructive pulmonary disease (COPD). Different clinical scores consider different risk factors, with their comparisons detailed in Additional file 1: Table S1. Except for this, these studies on clinical scores have two main drawbacks. First, they rely on relatively small datasets ranging from 286 to 11,551 for predicting SAP (see Additional file 1: Table S2). Second, some variables used in these risk scores are not easily accessible directly. For example, a recent risk score for ICH, ICH-APS-B uses hematoma volume, infratentorial location, and extension into ventricles. As such, their conclusions may limit generalizability. Therefore, a new risk score should be developed for predicting ICH-associated SAP by using large-scale, multi-center data and clinical variables with readily available values.

In this study, we developed an ICH risk score called ICH-LR2S2 to predict pneumonia for risk assessment. It has been evaluated by the two large-scale multi-center cohorts. Except for the risk score, our machine learning model can be extended by adding additional variables to further improve the accuracy of its predictions. Note that this is a prospective study on predicting intracerebral hemorrhage stroke-associated pneumonia. More importantly, using the external independent validation data cohort has further demonstrated the benefits for clinical practice from our methods.

Methods

Participants

This study collected data on over one million patients from the Chinese Stroke Center Alliance (CSCA), a national, hospital-based, multi-center program initiated in August 2015. The CSCA requires participating hospitals to only enroll patients who meet the following criteria: (1) over 18 years old; (2) had the primary diagnosis of acute stroke/transient ischemic attacks (TIA) confirmed by brain CT or MRI, including acute AIS, TIA, intracerebral hemorrhage, or subarachnoid hemorrhage (SAH); (3) within seven days of symptom onset; and (4) admitted to hospital either directly or through emergency departments. Patients with cerebral venous sinus thrombosis or non-cerebrovascular diseases were excluded. For ensuring the accuracy of diagnosis and the quality of stroke care, performance metrics were used over the whole controlling process by strictly following the national standards and guideline recommendations prespecified or updated by the Steering Committee of CSCA. Detailed information about the CSCA design and methodology can be found in previous publications [24]. This study had been approved by the Central Institutional Review Board of Beijing Tiantan Hospital.

Patients with intracranial hemorrhagic stroke were selected, resulting in a total of 83,063 patients as our study cohort. Among the selected patients, 61,869 patients had no pneumonia (74.47%), while 21,194 patients had pneumonia (25.52%). There are more than 500 characteristic variables, including clinical variables on admission such as blood pressure, blood sugar, uric acid, pneumonia, National Institute of Health stroke scale (NIHSS), and modified Rankin Scale (mRS), as well as external variables such as hospital level, education level, and family income status.

Definition and indicators of pneumonia

Pneumonia can be diagnosed by a typical chest X-ray, clinical symptoms, signs such as a cough, purulent sputum, fever, and laboratory tests such as white blood cell count. SAP after ICH can be diagnosed by a treating physician who uses clinical and laboratory indicators of respiratory infections such as fever, cough, and auscultation of respiratory cracks, new purulent sputum, or positive sputum culture, together with typical chest X-ray findings from PISCES (Pneumonia in Stroke ConsEnsuS) [25]. Hospital-acquired pneumonia was documented by excluding those cases that occurred before the stroke. Data on the development of SAP after ICH were prospectively collected.

Study procedure

To validate our model for predicting the likelihood of pneumonia, we used data collected from two multicenter cohorts in this study – the internal prospective research cohort of CSCA and the external independent verification cohort of CNSR II. In our experiments, we allocated the data from 2015 to 2018 for training with an internal verification ratio of 8:2, and the 2019 data for testing. After that, data records with missing values were filled in through data processing. The feature selection was then performed to select features that have important impacts on pneumonia. To this end, we first trained our model by using the two classic models—XGboost and logistic regression. ICH-LR2S2 was then calculated using the feature weight coefficients of logistic regression. After consulting doctors, the score interval was slightly modified according to the medical risk values to comply with the medical consensus [16, 26]. Additionally, we examined its performance on an external verification cohort. For the benefit of clinical practice, we stratified the patient population and analyzed the whole population cohort. The flow diagram is shown in Fig. 1.

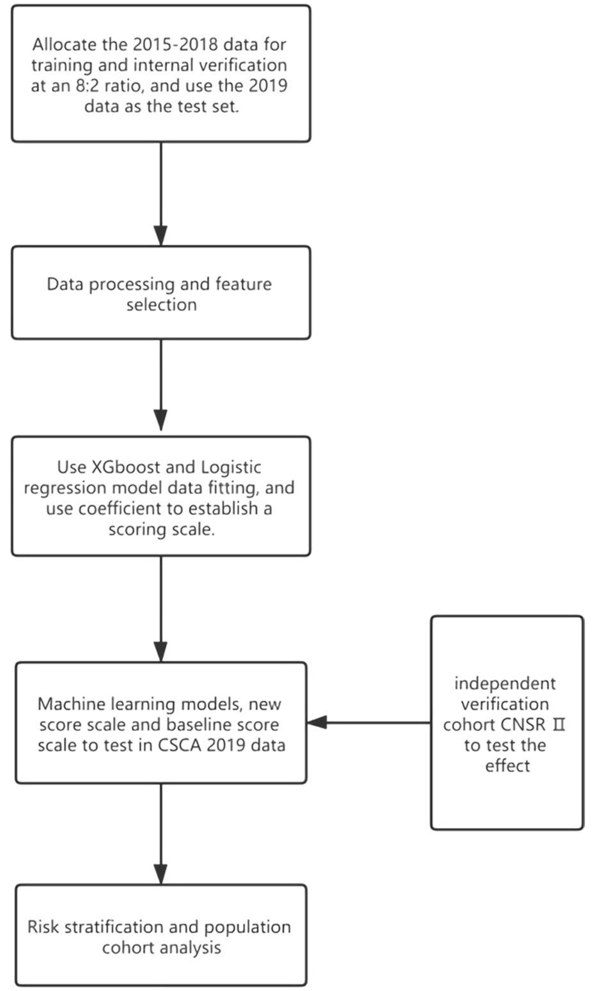
Allocate the 2015-2018 data for training and internal verification at an 8:2 ratio, and use the 2019 data as the test set.
Data processing and feature selection
Use XGboost and Logistic regression model data fitting, and use coefficient to establish a scoring scale.
Machine learning models, new score scale and baseline score scale to test in CSCA 2019 data
independent verification cohort CNSR Ⅱ to test the effect
Risk stratification and population cohort analysis

Fig. 1 Flow diagram for the derivation cohort, validation cohort, and the schema of our model

Data processing

Additional file 1: Table S3 shows the proportion of missing data for the selected variables. If a variable with a missing value was a continuous one, we filled it with the median value of that variable in the dataset. If it was binary, we filled it with 0, which means that there is no such disease history (our binary variables only include disease history and gender, and the gender variable is not missing). We finally obtained the data with the training set = 56,432, internal validation cohort = 14,108, and test set = 12,523 in our experiments.

Feature selection

Considering medical variables from the perspective of clinical practice, we focused on screening medical variables related to human physiological characteristics and disease history conditions. We tried to select as few variables as possible without reducing the prediction accuracy for pneumonia. Feature selection was performed using the permutation method [27], which is suitable for tree models. The importance of a feature can be measured by how much the objective score decreases as a result of removing the feature. Specifically, the variable weights were calculated through the permutation mechanism provided by XGboost [28], which is a boosting tree model with the capacity to handle missing values. Ten-fold cross-validation tests on the training set were conducted to calculate the feature weights.

We filtered out the feature variables in turn, according to their weight order. A newly added feature must increase the overall score of the internal verification cohort by at least 0.005 in the cross-validation. Considering the features selected in the previous studies [16–21] as well as recommended by the doctors, we further added three new variables—gender, current smoking, and C-reactive protein. Finally, we ended up with 12 variables of dysphagia, Glasgow Coma Score (GCS), age, gender, fasting blood glucose, uric acid, COPD, National Institutes of Health Stroke Scale admission score (NIHSS score), mRS, current smoking, serum creatinine, and C-reactive protein. Detailed descriptions of these variables are provided in Additional file 1: Figure S3.

Baseline scores

In Additional file 1: Tables S1 and S2, we list the scoring scales that can be used in ICH as the predictions of SAP in recent years. The aged variable is used by all scores, so is NIHSS except for the ACDD4 score [19]. In the following experiments, we mainly considered clinical variables that are easy to obtain. Therefore, we screened ICH-APS-A (from now on referred to as ICH-APS), PASS, ISAN, and PNA. It is worth mentioning that we made a compromise for ICH-APS; that is, we used drinking history instead of excessive drinking in ICH-APS. And considering the acquisition of variables, we set a score of 0 for the three medical variables (hematoma volume, infratentorial location, and extension into ventricles) that are not included in our data cohorts.

ICH-LR2S2

We used classic machine learning models of logistic regression. Calculating the medical risk score by using the regression coefficient [29] and the prior medical consensus [16, 19, 21], we developed the ICH-LR2S2 risk score shown in Table 1. We excluded features with scores of less than one point. As such, ICH-LR2S2 used the nine patient features: age, mRS, fasting blood glucose, NIHSS score, GCS, C-reactive protein, dysphagia, COPD, and current smoking. Compared with previous risk scores, ICH-LR2S2 has two new variables—fasting blood glucose and C-reactive protein.

External validation cohort

The performance of our model was tested on an independent cohort from the China National Stroke Registry II (CNSR II) [30]. As a nationwide initiative, the CNSR II, launched in 2012 by the Ministry of Health of China, established a reliable national stroke database for evaluating the delivery of stroke care in clinical practice. The CNSR II cohort included patients recruited from all 219 urban hospitals that voluntarily participated in the General Administration of Stroke Registration of China from June 2012 to January 2013. The study had been approved by the Central Institutional Review Board of Beijing Tiantan Hospital. Each participant provided written informed consent before participating.

Table 1 ICH-LR2S2

Item	Range	Score
Age group	< 60	0
	60–69	1
	70–79	2
	80–89	3
	≥ 90	4
mRS	< 4	0
	4	2
	5	3
Fasting blood glucose	< 6	0
	6–8	1
	9–11	2
	≥ 12	3
NIHSS score	< 5	0
	5–13	1
	14–21	2
	22–29	3
	≥ 30	4
GCS	3–5	2
	6–8	1
	≥ 9	0
C-reactive protein	< 7	0
	7–16	1
	≥ 17	2
Dysphagia	Yes	4
	No	0
COPD	Yes	3
	No	0
Current smoking	Yes	2
	No	0

Statistical analysis

Continuous variables are described by means and standard deviations (SD), while categorical variables are described by counts and percentages. The prediction performances of the models are measured by the area under the receiver operating characteristic curve (AUC), with a 95% confidence interval (CI). The AUCs of these models were compared using the DeLong test [31]. The student's *t*-test was used for continuous variables and the chi-square test for categorical variables. Two-sided $p < 0.01$ was considered to be statistically significant.

Based on logistic regression, the risk score used weight coefficients. By taking ten years as the interval, the ratio of a feature weight to the age weight coefficient was calculated to obtain the corresponding feature score and numerical interval. For a binary variable, the presence or absence of a feature was used as a scoring criterion (gender features give scores to men). For a continuous variable, in addition to considering the weight coefficient from the model, the actual meaning of the medical feature (medical risk range for this feature) must also be considered. In particular, the minimum unit of the score was 1 point, and features with less than 1 point were not scored. Based on the predictive score of the model, we stratified the risk of the population cohort and specified the risk threshold. We analyzed different risk groups by calculating the number of patients, the pneumonia rate, accuracy, sensitivity, specificity, PPV, and NPV (for more detailed information, refer to Additional file 1: Tables S7–S14).

Results

The composition and characteristics of the population

From July 2015 to June 2019, CSCA recruited 83,063 ICH patients, of which 21,194 (25.52%) were pneumonia patients. Specifically, 62.55% of all patients were male, with an average age of 62.48 years. For more detailed information, refer to Table 2. It can be found that there is a significant difference between the proportion of patients with dysphagia and pneumonia (36.63%) and those with dysphagia who did not have pneumonia (9.01%). This can also explain why the weight coefficient of this variable is large.

The external validation cohort of CNSR II included 24,680 patients with similar demographic characteristics (male 63.75%, mean age 64.1 years, SD 12.0). Related surveys show that patients from the 2012 to 2013 dataset had insurance through new rural cooperative medical schemes [30]. These patients received better medical assistance and thus had a lower incidence of stroke and stroke-related complications. Additional file 1: Table S4 lists other characteristics of the derived cohort and external verification cohort. By comparing the data populations of the two cohorts, the proportion of pneumonia patients in CNSR II is only 8.44% and in CSCA it is 25.52%. We compared their representative characteristics. In medical judgment, a high NIHSS score, or a low GCS score indicates that a patient's condition is serious. Patients with high scores for NIHSS variables (>16)

Table 2 Risk factors and basic knowledge in CSCA cohort (N: number of people)

Basic information	Total patients (n = 70,540)	With pneumonia (n = 18,190)	Without pneumonia (n = 52,350)	P values
Male (n%)	44,123 (62.55%)	11,864 (65.22%)	32,259 (61.62%)	<0.01
Age (mean)	62.48	65.03	61.59	<0.01
<60 (N)	28,305	5908 (32.48%)	22,397 (42.783%)	<0.01
$60 \leq$ age <70 (N)	18,435	4636 (25.49%)	13,799 (26.36%)	0.01
$70 \leq$ age <80 (N)	12,839	4,013 (22.06%)	8826 (16.86%)	<0.01
$80 \leq$ age <90 (N)	5644	2,170 (11.93%)	3474 (6.64%)	<0.01
age ≥ 90 (N)	5317	1,463 (8.04%)	3854 (7.36%)	<0.01
mRS at hospital (mean)	2.11	2.45	2.00	<0.01
≤ 4 (N)	60,044	13,764 (75.67%)	46,280 (88.41%)	<0.01
≥ 5 (5,6) (N)	10,496	4,426 (24.33%)	6,070 (11.60%)	0.02
NIHSS score (mean)	8.17	13.51	6.69	<0.01
<10 (N)	15,348	2061 (11.33%)	13,287 (25.38%)	<0.01
10–16 (N)	2825	928 (5.10%)	1897 (3.62%)	0.03
>16 (N)	2983	1523 (8.37%)	1460 (2.79%)	0.05
GCS (mean)	11.42	9.68	12.17	<0.01
≥ 10 (N)	24,615	5527 (30.38%)	19,088 (36.46%)	<0.01
<10 (N)	10,847	5165 (28.40%)	5,682 (10.85%)	<0.01
Smoking	14,165	3780 (20.78%)	10,385 (19.83%)	<0.01
COPD	1,026	517 (2.84%)	509 (0.97%)	0.01
Dysphagia	11,379	6663 (36.63%)	4,716 (9.01%)	<0.01
CRP (>10 mg/l)	854	318 (1.75%)	536 (1.02%)	<0.01
Creatinine (μmol/l)	83.16	86.37	82.06	<0.01
Uric acid (μmol/l)	288.56	280.83	291.25	<0.01
Fasting blood glucose (mmol/l)	6.54	7.00	6.38	<0.01
<7.8	56,617	13,421 (73.78%)	43,196 (82.51%)	<0.01
7.8–11.1	8708	3059 (16.81%)	5649 (10.79%)	<0.01
≥ 11.1	3805	1349 (7.42%)	2456 (4.69%)	0.02

accounted for 4.93% in CNSR II, which was lower than in CSCA (13.18%). Patients with low scores for GCS variables (< 10) accounted for 1.45% in CNSR II, which was also lower than in CSCA (30.67%). For dysphagia, which has the greatest impact on the weight coefficient of pneumonia, its incidence rate in CNSR II was 8.29%, compared to 15.98% in CSCA. All these results demonstrate that the patients in the CNSR II data cohort had milder cases of pneumonia.

Classification performance

The performance of the model was examined using the test data. The results of risk scores were as follows: ICH-LR2S2 (AUC = 0.749) [95% CI 0.739–0.759], the existing scoring method ICH-APS (AUC = 0.704) [95% CI 0.694–0.714], PASS (AUC = 0.684) [95% CI 0.674–0.694], ISAN (AUC = 0.676) [95% CI 0.666–0.686], and PNA (AUC = 0.636) [95% CI 0.626–0.646]. As a white-box model, ICH-LR2S2 is highly explanatory and intuitive, performing significantly better than the baseline of ICH-APS, PASS, and so on. The DeLong test results (see Additional file 1: Table S17) show that ICH-LR2S2 performed significantly better than other risk scores for ICH.

Compared with the previous scoring scales, we used newly added fasting blood glucose variables and C-reactive protein biomarkers in ICH-LR2S2. To validate this choice, we compared the effects of ICH-LR2S2 by deleting these two variables, respectively (see Additional file 1: Fig. S5 and S6). For the CSCA cohort data, AUC was reduced by 0.005 after removing fasting blood glucose, by 0.012 after removing C-reactive protein, and by 0.017 after removing both. For the external verification queue CNSR II, the AUC decreases by 0.009 after removing these two variables.

Furthermore, the machine learning model that uses more variables (gender, uric acid, and serum creatinine) shows better predictive performance in terms of comparing the risk score. As shown in Fig. 2(a), the results were as follows: XGboost (AUC = 0.772) [95% CI 0.762–0.782] and logistic regression (AUC = 0.755) [95% CI 0.745–0.765].

External validation

We tested the risk score of ICH-LR2S2 on an independent external cohort of CNSR II. The CNSR II data statistics are reported in Additional file 1: Table S4. The results of external validation are consistent with those of test data. The overall performance of ICH-LR2S2 was AUC = 0.784 [95% CI 0.774–0.794] and the best performance of our baseline risk scoring models was ICH-APS (AUC = 0.737) [95% CI 0.727–0.747]. Also, our machine learning model can be further improved, as shown in Fig. 2b. The detailed results are reported in Additional file 1: Table S6, indicating that ICH-LR2S2 has high potential application values.

Risk stratification

We divided the validation data set into different risk cohorts according to the given risk threshold. For the risk scoring scale, for example, we regarded people with a score higher than 13 as a high-risk group, people with a score lower than 6 as a low-risk group, and the rest as a medium-risk group. For the CSCA cohort, the number of patients in three risk groups were as follows: 556 (4.44%) in the high-risk group, with a pneumonia rate of 61.51%; 5659 (45.19%) in the middle-risk group, with a pneumonia rate of 33.89%; and 6308 (50.37%) in the low-risk group with a pneumonia rate of 11.79%. For XGboost, we also performed a risk stratification and cohort analysis based on predicted scores. The detailed results of the statistical analysis are given in Additional file 1: Tables S7–S14.

By using ICH-LR2S2, we listed the pneumonia probabilities in the table in terms of different scores (see Additional file 1: Tables S15 and S16). The bar graphs in Fig. 2c, d illustrate the corresponding relationships between ICH-LR2S2 scores and their probabilities of pneumonia. These figures provide doctors with an easy way of estimating a risk score in their clinical practice. For example, for the CSCA cohort, the probability of having pneumonia is 0.86 when the score reaches 20. As such, the information in this table can assist doctors in making decisions on patients with different conditions.

Discussion

Previous studies have compared the different SAP scores in AIS by listing their pros and cons [26, 33]. Similarly, we compare the widely used SAP scores in ICH in Additional file 1: Table S1 from a different perspective. As demonstrated, the performance of our new risk score is superior to all existing scoring scales for predicting SAP of ICH. Further, the obvious advantage of ICH-LR2S2 lies in the use of widely available variables upon patients' admissions. The corresponding score and pneumonia risk can then be easily calculated. More importantly, ICH-LR2S2 for physicians is a simple, intuitive, and easy-to-use pneumonia assessment model. Surprisingly, we discovered new variables that are influential on SAP in ICH, and the risk stratification based on our risk score corresponds to the probabilities of pneumonia from this study.

From this study, we concluded that fasting blood glucose levels and C-reactive protein play an important role in predicting SAP based on our large-scale data analysis. Both blood glucose level and diabetes history were regarded as important predictors in previous studies on

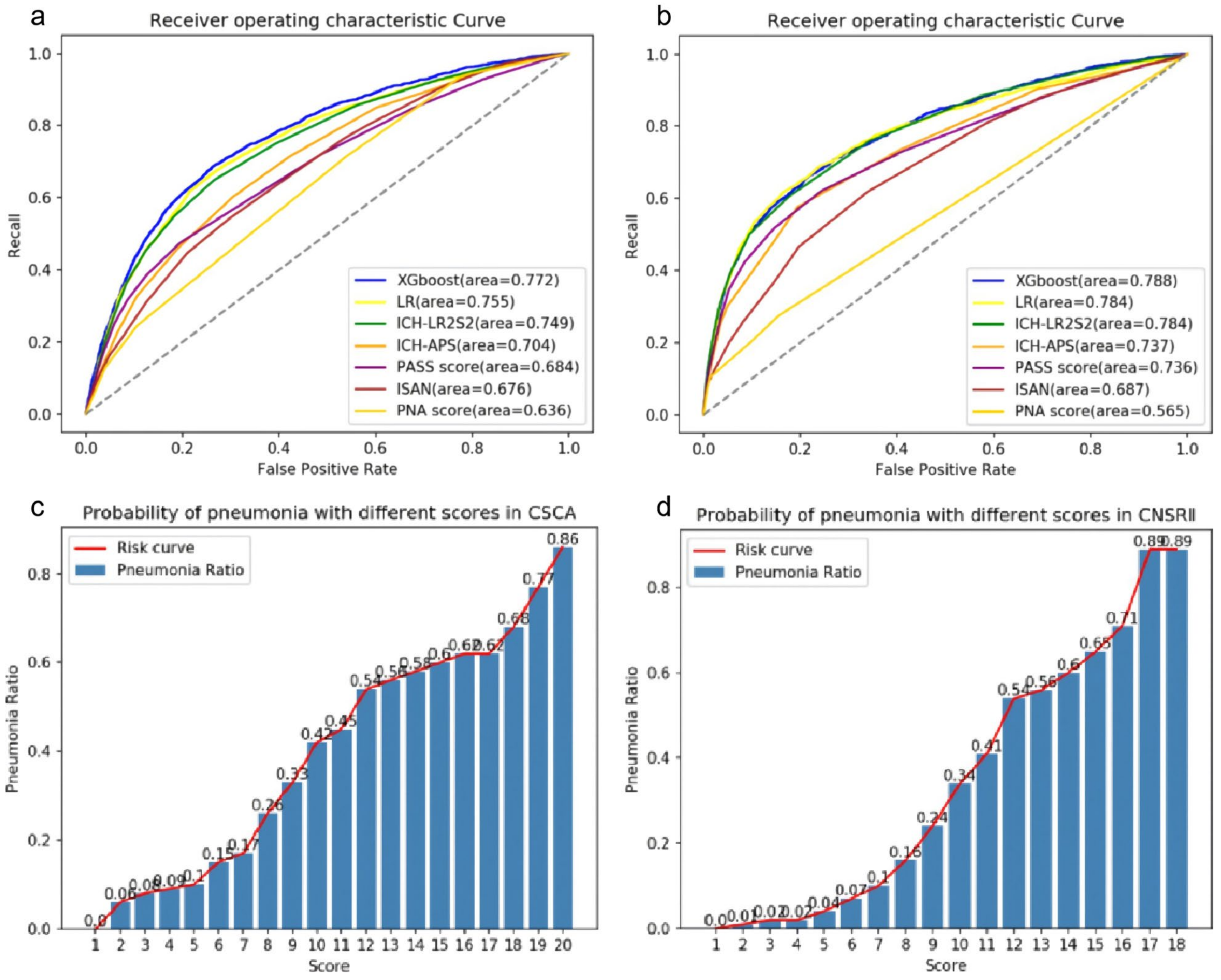

Fig. 2 a Receiver operating characteristic curves (ROC) for discriminatory abilities of the different scores for clinical diagnosis of stroke-associated pneumonia (SAP) of CSCA cohort. **b** Receiver operating characteristic curves (ROC) for discriminatory abilities of the different scores for clinical diagnosis of stroke-associated pneumonia (SAP) of CNSR II cohort. **c** Probability of pneumonia with different scores in the CSCA cohort. **d** Probability of pneumonia with different scores in the CNSR II cohort

SAP in AIS, but they were ignored in ICH [34]. Note that the history of diabetes does not represent abnormal blood glucose levels. Also, temporary hyperglycemia may indicate stress hyperglycemia rather than diabetes. As Hotter et al. [32] pointed out, diabetes history is not an independent risk factor for SAP; however, hyperglycemia reduces the bactericidal ability of white blood cells so that the possibility of lung infection increases, and patients with fasting hyperglycemia are likely to suffer from SAP [32, 33]. All these conclusions are consistent with our findings in this study. On the other hand, as an important biomarker, C-reactive protein (CRP) plays a significant role in predicting pneumonia [34]. Specifically, elevated CRP is an important sign of the poor prognosis of acute respiratory distress syndrome, reflecting the persistent state of inflammation [34, 35]. Adnet et al. [36] showed that high CRP levels help diagnose pneumonia patients with drug-induced coma and secondary inhalation. At the same time, they found that the sensitivity and specificity of other parameters (such as fever and white blood cell count) are poor indicators for diagnosing pneumonia. Based on our big data analysis, this study confirmed a strong correlation between CRP and SAP in ICH. Note that biomarkers and genetic data are needed to further investigate the mechanism of this correlation.

In terms of pneumonia in hemorrhagic stroke, it is clear from the data that patients with a history of hypertension are indeed more susceptible to pneumonia. Angiotensin can cause high blood pressure [37, 38], while high blood pressure is associated with a variety of diseases including stroke, diabetes, and so on [39]. As a drug, ARBs are one of the most commonly used first-line treatment drugs for hypertension [40]. By selectively blocking the angiotensin II receptor (AT1 type) and angiotensin II (Ang II), ARBs produce a pharmacological effect that is similar to that of angiotensin-converting enzyme inhibitors (ACEI) [41, 42]. They dilate blood vessels and lower blood pressure. Therefore, the use of ARB antihypertensive drugs can reduce the blood pressure of patients, which may reduce the probability of stroke and pneumonia in patients. In addition to ARB, it can be concluded from the data in this paper that the use of antihypertensive drugs, hypoglycemic drugs, and anticoagulants (see Additional file 1: Fig. S3) may be related to subsequent pneumonia infection in stroke patients. This warrants further research.

This study was conducted based on a large data registry collected from multi-centered hospitals in China. Compared to our CSCA study cohort, patients in the CNSR II cohort had fewer strokes and were less likely to have pneumonia. The statistical analysis of important characteristic indicators shows that the proportion of the population within the characteristic risk threshold is significantly smaller. This is due to the different admission times of patients and different periods of policies followed. Being validated by an external validation cohort, ICH-LR2S2 can still make accurate predictions on these datasets. This indicates that ICH-LR2S2 has good generalization capabilities. All validation results demonstrate that the variables we choose for ICH-LR2S2 can effectively distinguish those who do not have pneumonia from those who do.

For the stratification of patients at the greatest risk, we used machine learning and scoring methods to conduct risk stratification of the current patient population, and statistically analyzed various indicators of different risk population cohorts. For ICH-LR2S2, we plotted the histogram of the probabilities of pneumonia under different scores. We also tested the performance of risk stratification in an external verification cohort, and the results demonstrated its effectiveness in distinguishing different risk populations. With the same pneumonia risk rate (the error does not exceed 0.02) in the data, we compared our risk score ICH-LR2S2 with the two best-performing risk scores of ICH-APS and PASS in terms of population coverage in different risk cohorts. As shown in Additional file 1: Fig. S7 and S8, ICH-LR2S2 has a higher coverage in high-risk and low-risk populations, indicating that it can distinguish different risk population cohorts better than ICH-APS and PASS. Specifically, in the CSCA cohort, the population coverage rate for the high-risk cohort was 4.44% by ICH-LR2S2, which was higher than 3.08% by ICH-APS and 2.89% by PASS. ICH-LR2S2 has also demonstrated consistent performance on different datasets (see Additional file 1: Fig. S9 and S10). With such good performance, ICH-LR2S2 can be used to score the risk of a new patient in clinical practice. A patient can be classified into the corresponding risk population cohort by using ICH-LR2S2. As such, medical expenses can be saved. Further, different post-response measures can be adopted for targeted treatment and care. On the other hand, several factors should be considered before using our ICH-LR2S2. For example, the threshold in ICH-LR2S2 should be set based on the different incidences of SAP or a hospital's clinical priorities. Factors such as the acceptability of risks or the adoption of reasonable precautions may also affect the choice of a particular threshold.

By using the limited number of scoring variables, clinical risk scoring can be simplified at the expense of compromising the accuracy of prediction. For example, ICH-LR2S2 does not use the variables and factors of medical comorbidities, additional imaging data or additional medication history information such as anticoagulants or antibiotics. However, the prediction from our scores should be combined with the use of any other clinically relevant information. In other words, ICH-LR2S2 is intended for use as a tool to aid a clinical decision-making process. Moreover, future efforts will be geared toward investigating its best use. The experiment results also have indicated that ICH-LR2S2 is a valuable tool to help identify clinicians' triage patients with ICH by predicting their risks of pneumonia reliably and accurately. In other words, our score can identify high-risk ICH patients who require additional interventional treatment.

Limitations

First, stroke cohorts other than CSCA and CNSR can be selected for further validation. Second, the data lacks variables such as hematoma volume, submeningeal location, and intraventricular hemorrhage, making ICH-APS slightly less effective. Finally, ICH-LR2S2 is a model for predicting pneumonia risk. This model can be transferred to other types of stroke complication prediction. Its application and prognostic model in daily clinical practice remain to be studied.

Conclusion

In this paper, we have presented ICH-LR2S2, which can accurately predict pneumonia associated with spontaneous intracranial hemorrhage by using data-driven machine learning methods. ICH-LR2S2 can be used easily by physicians of varying specialties.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12967-022-03389-5.

Additional file 1.

Acknowledgements

The authors express their immense gratitude to all of the patients who participated in each of the individual studies as well as the many personnel who helped with recruitment, collection, curation, management, and processing of the samples and data.

Author contributions

ZL, XZ and SZ conceived and supevised the study. WZ, JY, ZL and SZ designed the study. WZ and JY performed the meta-analysis of the collected estimates and drafted the article. ZL, LD, JY, JZ, XY, CW, XM, YJ, XH, SW, YW, ZL, SZ, YW, XZ, and JF revised the draft paper, critical revision of the manuscript for important intellectual content. WZ conducted the statistical analysis and takes responsibility for the accuracy of the data analysis. All the authors have read and approved the final manuscript.

Funding

This work has been supported by the Chinese Academy of Medical Sciences Innovation Fund for Medical Sciences (2019-I2M-5–029); Beijing Natural Science Foundation (Z200016), National Natural Science Foundation of China (92046016). This work has been also supported by Shanghai Municipal Science and Technology Major Project (No.2018SHZDZX01), 111 Project (No. B18015), ZJ Lab, and Shanghai Center for Brain Science and Brain-Inspired Technology. SZ has been supported by the Shanghai Municipal Science and Technology Major Project (No. 2017SHZDZX01) and Information Technology Facility, CAS-MPG Partner Institute for Computational Biology, Shanghai Institute for Biological Sciences, Chinese Academy of Sciences.

Availability of data and materials

The datasets used and/or analyzed in the current study are available from the corresponding author upon reasonable request.

Code availability

Not applicable.

Declarations

Ethics approval and consent to participate

Participating hospitals received either healthcare quality assessment and research approval to collect data in the CSCA project without requiring individual patient informed consent under the common rule or a waiver of authorization and exemption from subsequent review by their Institutional Review Board. The CNSR is funded by the Ministry of Science and Technology and the Ministry of Health of the People's Republic of China. The Grant Numbers are 2006BA101A11 and 2009CB521905.

Consent for publication

Not applicable.

Competing interests

Not applicable.

Author details

[1]Vascular Neurology, Department of Neurology, Beijing Tiantan Hospital, Capital Medical University, Beijing 100070, China. [2]China National Clinical Research Center for Neurological Diseases, Beijing, China. [3]Institute of Science and Technology for Brain-Inspired Intelligence, Fudan University, Shanghai 200433, China. [4]Chinese Institute for Brain Research, Beijing, China. [5]Research Unit of Artificial Intelligence in Cerebrovascular Disease, Chinese Academy of Medical Sciences, Beijing, China. [6]School of Computing, Mathematics and Engineering, Charles Sturt University, Albury, NSW 2640, Australia. [7]Ministry of Education, Key Laboratory of Computational Neuroscience and Brain-Inspired Intelligence (Fudan University), Shanghai 200433, China. [8]MOE Frontiers Center for Brain Science and Shanghai Institute of Artificial Intelligence Algorithms, Fudan University, Shanghai 200433, China. [9]Zhangjiang Fudan International Innovation Center, Shanghai 200433, China.

Received: 1 February 2022 Accepted: 9 April 2022
Published online: 04 May 2022

References

1. Wang Y, Li Z, Gu H, On behalf of China Stroke Statistics 2019 Writing Committee, et al. China Stroke Statistics 2019: A Report from the National Center for Healthcare Quality Management in Neurological Diseases, China National Clinical Research Center for Neurological Diseases, the Chinese Stroke Association, National Center for Chronic and Non-communicable Disease Control and Prevention, Chinese Center for Disease Control and Prevention and Institute for Global Neuroscience and Stroke Collaborations. Stroke Vasc Neurol. 2020;5(3):211–39.
2. Westendorp WF, Nederkoorn PJ, Vermeij JD, Dijkgraaf MG, van de Beek D. Post-stroke infection: a systematic review and meta-analysis. BMC Neurol. 2011;11(1):110.
3. Bustamante A, Giralt D, García-Berrocoso T, Rubiera M, Álvarez-Sabín J, Molina C, et al. The impact of post-stroke complications on in-hospital mortality depends on stroke severity. Eur Stroke J. 2017;2(1):54–63.
4. Finlayson O, Kapral M, Hall R, Asllani E, Selchen D, Saposnik G, on behalf of the Investigators of the Registry of the Canadian Stroke Network. Risk factors, inpatient care, and outcomes of pneumonia after ischemic stroke. Neurology. 2011;77:1338–45.
5. Lou M, Ding J, Hu B, Zhang Y, Li H, Tan Z, et al. Chinese Stroke Association Stroke Council Guideline Writing Committee. Chinese Stroke Association guidelines for clinical management of cerebrovascular disorders: executive summary and 2019 update on organizational stroke management. Stroke Vasc Neurol. 2020;5(3):260–9.
6. Teh WH, Smith CJ, Barlas RS, Wood AD, Bettencourt-Silva JH, Clark AB, et al. Impact of stroke-associated pneumonia on mortality, length of hospitalization, and functional outcome. Acta Neurol Scand. 2018;138(4):293–300.
7. Meisel A. Preventive antibiotic therapy in stroke: PASSed away? Lancet. 2015;385(9977):1486–7.
8. Badve MS, Zhou Z, Anderson CS, Hackett ML. Effectiveness and safety of antibiotics for preventing pneumonia and improving outcome after acute stroke: systematic review and meta-analysis. J Stroke Cerebrovasc Dis. 2018;27(11):3137–47.
9. Faura J, Bustamante A, Miró-Mur F, Montaner J. Stroke-induced immunosuppression: implications for the prevention and prediction of post-stroke infections. J Neuroinflammation. 2021;18(1):127.
10. Kwan J, Pickering RM, Kunkel D, Fitton C, Jenkinson D, Perry VH, Stroke Association Rehabilitation Research Centre, et al. Impact of stroke-associated infection on long-term survival: a cohort study. J Neurol Neurosurg Psychiatry. 2013;84:297–304.
11. Smith CJ, Bray BD, Hoffman A, Meisel A, Heuschmann PU, Wolfe CD, Tyrrell PJ, Rudd AG, Intercollegiate Stroke Working Party Group. Can a novel clinical risk score improve pneumonia prediction in acute stroke care? A UK multicenter cohort study. J Am Heart Assoc. 2015;4(1): e001307.
12. Hoffmann S, Malzahn U, Harms H, Berlin Stroke Register and the Stroke Register of Northwest Germany. Development of a clinical score (A2DS2) to predict pneumonia in acute ischaemic stroke. Stroke. 2012;43:2617–23.
13. Gong S, Zhou Z, Zhou M, Lei Z, Guo J, Chen N, He L. Validation of risk scoring models for predicting stroke-associated pneumonia in patients with ischaemic stroke. Stroke Vasc Neurol. 2016;1:122–6.
14. Kwon HM, Jeong SW, Lee SH, Yoon BW. The pneumonia score: a simple grading scale for prediction of pneumonia after acute stroke. Am J Infect Control. 2006;34(2):64–8.
15. Chumbler NR, Williams LS, Wells CK, Lo AC, Nadeau S, Peixoto AJ, et al. Derivation and validation of a clinical system for predicting pneumonia in acute stroke. Neuroepidemiology. 2010;34(4):193–9.

16. Ji R, Shen H, Pan Y, Du W, Wang P, Liu G, China National Stroke Registry investigators, et al. Risk score to predict hospital-acquired pneumonia after spontaneous intracerebral hemorrhage. Stroke. 2014;45(9):2620–8.
17. Friedant AJ, Gouse BM, Boehme AK, Siegler JE, Albright KC, Monlezun DJ, et al. A simple prediction score for developing a hospital-acquired infection after acute ischemic stroke. J Stroke Cerebrovasc Dis. 2015;24(3):680–6.
18. Smith CJ, Bray BD, Hoffman A, Meisel A, Heuschmann PU, Wolfe CD, Intercollegiate Stroke Working Party Group, et al. Can a novel clinical risk score improve pneumonia prediction in acute stroke care? A UK multicenter cohort study. J Am Heart Assoc. 2015;4(1): e001307.
19. Kumar S, Marchina S, Massaro J, Feng W, Lahoti S, Selim M, Herzig SJ. ACDD4 score: a simple tool for assessing risk of pneumonia after stroke. J Neurol Sci. 2017;372:399–402.
20. Ji R, Shen H, Pan Y, Wang P, Liu G, Wang Y, Li H, Wang Y, China National Stroke Registry Investigators. Novel risk score to predict pneumonia after acute ischemic stroke. Stroke. 2013;44(5):1303–9.
21. Westendorp WF, Vermeij JD, Hilkens NA, Brouwer MC, Algra A, van der Worp HB, et al. Development and internal validation of a prediction rule for post-stroke infection and post-stroke pneumonia in acute stroke patients. Eur Stroke J. 2018;3(2):136–44.
22. Finlayson O, Kapral M, Hall R, et al. Risk factors, inpatient care, and outcomes of pneumonia after ischemic stroke. Neurology. 2011;77:1338–45.
23. Sellars C, Bowie L, Bagg J, et al. Risk factors for chest infection in acute stroke: a prospective cohort study. Stroke. 2007;38:2284–91.
24. Wang Y, Li Z, Wang Y, Zhao X, Liu L, Yang X, et al. Chinese stroke center alliance: a national effort to improve healthcare quality for acute stroke and transient ischaemic attack: Rationale, design and preliminary findings. Stroke Vasc Neurol. 2018;3:256–62.
25. Smith CJ, Kishore AK, Vail A, Chamorro A, Garau J, Hopkins SJ, Di Napoli M, Kalra L, Langhorne P, Montaner J, et al. Diagnosis of stroke-associated pneumonia: recommendations from the pneumonia in stroke consensus group. Stroke. 2015;46:2335–40.
26. Kishore AK, Vail A, Bray BD, Chamorro A, Napoli MD, Kalra L, Langhorne P, Montaner J, Roffe C, et al. Clinical risk scores for predicting stroke-associated pneumonia: a systematic review. Eur Stroke J. 2016;1(2):76–84.
27. Altmann A, Toloşi L, Sander O, Lengauer T. Permutation importance: a corrected feature importance measure. Bioinformatics. 2010;26(10):1340–7.
28. Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016.
29. Sullivan LM, Massaro JM, DAgostino RB. Presentation of multivariate data for clinical use: the Framingham Study risk score functions. Stat Med. 2004;23(10):1631–60.
30. Li Z, Wang C, Zhao X, Liu L, Wang C, Li H, China National Stroke Registries, et al. Substantial progress yet significant opportunity for improvement in stroke care in China. Stroke. 2016;47:2843–9.
31. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a non-parametric approach. Biometrics. 1988;44:837–45.
32. Hotter B, Hoffmann S, Ulm L, Meisel C, Bustamante A, Montaner J, Katan M, Smith CJ, Meisel A. External Validation of Five Scores to Predict Stroke-Associated Pneumonia and the Role of Selected Blood Biomarkers. Stroke. 2021;52(1):325–30.
33. Rayfield EJ, et al. Infection and diabetes: the case for glucose control. Am J Med. 1982;72(3):439–50.
34. Ridker PM, et al. Rosuvastatin to prevent vascular events in men and women with elevated C-reactive protein. N Engl J Med. 2008;359:2195–207.
35. Bajwa EK, et al. Plasma C-reactive protein levels are associated with improved outcome in ARDS. Chest. 2009;136:471–80.
36. Adnet F, Borron SW, Vicaut E, et al. Value of C-reactive protein in the detection of bacterial contamination at the time of presentation in drug-induced aspiration pneumonia. Chest. 1997;112:466–71.
37. Crowley SD, Gurley SB, Oliverio MI, Pazmino AK, Griffiths RI, Flannery P, Spurney RF, Kim H, Smithies O, Le TH, Coffman TM. Distinct roles for the kidney and systemic tissues in blood pressure regulation by the renin-angiotensin system. J Clin Investig. 2005;115(4):1092–9.
38. Oparil S, Haber E. The renin-angiotensin system. N Engl J Med. 1974;291(8):389–401.
39. Ahmadian E, Pennefather PS, Eftekhari A, Heidari R, Eghbal MA. Role of renin-angiotensin system in liver diseases: an outline on the potential therapeutic points of intervention. Expert Rev Gastroenterol Hepatol. 2016;10:1279–88.
40. Zhang Y, Ding X, Hua B, Liu Q, Chen H, Zhao XQ, Li W, Li H. Real-world use of ACEI/ARB in diabetic hypertensive patients before the initial diagnosis of obstructive coronary artery disease: patient characteristics and long-term follow-up outcome. J Transl Med. 2020;18(1):150.
41. Onozato ML, et al. Oxidative stress and nitric oxide synthase in rat diabetic nephropathy: effects of ACEI and ARB. Kidney Int. 2002;61(1):186–94.
42. Cheng J, Zhang W, Zhang XH, He Q, Tao XJ, Chen JH. ACEI/ARB therapy for IgA nephropathy: a meta analysis of randomised controlled trials. Int J Clin Pract. 2009;63(6):880–8.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.