Epilepsy Research 205 (2024) 107397

Contents lists available at ScienceDirect

Epilepsy Research

journal homepage: www.elsevier.com/locate/epilepsyres

Development and validation of an interpretable machine learning model for predicting post-stroke epilepsy

Yue Yu [a,b], Zhibin Chen [c], Yong Yang [a], Jiajun Zhang [b], Yan Wang [a,*]

[a] Affiliated Hospital of Qingdao University, Qingdao, China
[b] Qingdao Municipal Hospital, Qingdao, China
[c] Department of Neuroscience, Central Clinical School, Monash University, Melbourne, Victoria, Australia

ARTICLE INFO

Keywords:
Ischemic stroke
Post-stroke epilepsy
Machine learning
SHapley Additive explanation

ABSTRACT

Background: Epilepsy is a serious complication after an ischemic stroke. Although two studies have developed prediction model for post-stroke epilepsy (PSE), their accuracy remains insufficient, and their applicability to different populations is uncertain. With the rapid advancement of computer technology, machine learning (ML) offers new opportunities for creating more accurate prediction models. However, the potential of ML in predicting PSE is still not well understood. The purpose of this study was to develop prediction models for PSE among ischemic stroke patients.

Methods: Patients with ischemic stroke from two stroke centers were included in this retrospective cohort study. At the baseline level, 33 input variables were considered candidate features. The 2-year PSE prediction models in the derivation cohort were built using six ML algorithms. The predictive performance of these machine learning models required further appraisal and comparison with the reference model using the conventional triage classification information. The Shapley additive explanation (SHAP), based on fair profit allocation among many stakeholders according to their contributions, is used to interpret the predicted outcomes of the naive Bayes (NB) model.

Results: A total of 1977 patients were included to build the predictive model for PSE. The Boruta method identified NIHSS score, hospital length of stay, D-dimer level, and cortical involvement as the optimal features, with the receiver operating characteristic curves ranging from 0.709 to 0.849. An additional 870 patients were used to validate the ML and reference models. The NB model achieved the best performance among the PSE prediction models with an area under the receiver operating curve of 0.757. At the 20 % absolute risk threshold, the NB model also provided a sensitivity of 0.739 and a specificity of 0.720. The reference model had poor sensitivities of only 0.15 despite achieving a helpful AUC of 0.732. Furthermore, the SHAP method analysis demonstrated that a higher NIHSS score, longer hospital length of stay, higher D-dimer level, and cortical involvement were positive predictors of epilepsy after ischemic stroke.

Conclusions: Our study confirmed the feasibility of applying the ML method to use easy-to-obtain variables for accurate prediction of PSE and provided improved strategies and effective resource allocation for high-risk patients. In addition, the SHAP method could improve model transparency and make it easier for clinicians to grasp the prediction model's reliability.

1. Introduction

Stroke remains a major cause of mortality and disability worldwide (ANON, 2021), and it is the main cause of acquired epilepsy among adults over the age of 60 (Xu, 2019). Previous studies have reported that the incidence of post-ischemic stroke epilepsy (PISE) ranges from 3.1 % to 15.2 % (Bladin et al., 2000; Lamy et al., 2003; Lossius et al., 2005; Roivainen et al., 2013; Bentes et al., 2017). Research indicates that the majority of patients (85 %) with post-stroke epilepsy (PSE) experienced their first unprovoked seizures within two years following stroke onset (Guo et al., 2015). In addition, epilepsy has been identified as a serious complication after acute stroke, resulting in a poorer neurological prognosis, lower quality of life, and a greater risk of death than in stroke patients without epilepsy (Zelano et al., 2016; Bryndziar et al., 2016;

* Correspondence to: Affiliated Hospital of Qingdao University, 16 Jiangsu Road, Qingdao, Shandong 266000, China.
E-mail address: drwangyanqd@126.com (Y. Wang).

https://doi.org/10.1016/j.eplepsyres.2024.107397
Received 19 April 2024; Received in revised form 19 June 2024; Accepted 20 June 2024
Available online 28 June 2024

Winter et al., 2018; De Reuck et al., 2006; Zelano, 2020).

Given the burden of disease from PSE, there is a pressing need to build an effective prediction model. Such a model, could assist in identifying potentially high-risk patients, facilitate randomized controlled trials, and guide clinical practices. Although previous studies have established the SeLECT score and PSEiCARe score as means to predict epilepsy after ischemic stroke, data on their clinical applicability are lacking, and interpretable models are needed to improve PSE prediction (Galovic et al., 2018; Chi et al., 2018).

Feature selection and supervised learning modeling methods in the field of machine learning (ML) have been applied to address disease prediction and diagnose challenges. However, only one study applied ML to the characteristic variables of the CAVE model to predict early seizures after acute intracerebral hemorrhage (Bunney et al., 2022; Haapaniemi et al., 2014). The prediction of post-stroke seizures using ML technology is still in the preliminary exploration stage. To the best of our knowledge, there is a scarcity of research comparing the predictive performance of models' based on ML and traditional statistical regression algorithms.

Our research aimed to construct an interpretable model to anticipate the 2-year risk of epilepsy in ischemic stroke patients. We employed various ML algorithms to develop PSE predictive models. The optimal model was selected based on the model performance evaluation, and then compared with the commonly regarded SeLECT score as the reference model. In addition, to promote the transparency and adoption of the ML model, SHapley additive exPlanations (SHAP) were used to ascertain the contributions of input variables in model prediction (Lundberg and Lee, 2017). This marks a significant effort to achieve accurate prediction of PSE using ML algorithms, providing a reference for future research, and helping clinicians make better decisions and develop effective prevention strategies.

2. Methods

2.1. Study design and population

This retrospective cohort study included individuals diagnosed with acute ischemic stroke from two stroke centers. The derivation cohort, used for developing the predictive model consisted of patients admitted to the Affiliated Hospital of Qingdao University from January 2018 to December 2019. The validation cohort, on the other hand, comprised individuals hospitalized at Qingdao Municipal Hospital from January to December 2019. The Ethical Committees of Affiliated Hospital of Qingdao University and Qingdao Municipal Hospital approved the study. Telephonic interviews were conducted to obtain verbal informed consent from the participants or their legally authorized representatives. Since this study was retrospective in nature and did not involve any intervention, the requirement for written informed consent was waived.

The same inclusion and exclusion criteria were applied both cohorts. Inclusion criteria for patient participation were as follows: aged 18 years or older with acute ischemic stroke confirmed by a CT or MRI scan during the hospital admission. Exclusion criteria were: 1) previous history of seizures; 2) history of hemorrhagic stroke and subarachnoid hemorrhage; 3) history of intracranial tumors, traumatic brain injury, and craniocerebral operations; 4) history of central nervous system infections; 5) Death within 2 years after stroke and without any occurrence of late seizures during this period.

2.2. Data collection

The data collected including the basic demographic characteristics, clinical features, laboratory test results, reperfusion therapy, and neuroimaging information of acute ischemic stroke patients. Stroke severity was assessed by using the National Institute of Health Stroke Scale (NIHSS) (Brott et al., 1989; Kasner et al., 1999). Etiology classification was based on the Trial of ORG 10172 in Acute Stroke Treatment (TOAST) into the following categories: large-artery atherosclerosis, cardioembolism, small-vessel occlusion, stroke of other determined etiology, and stroke of undetermined etiology. We collected laboratory indicators, including fasting blood glucose, lipid profiles, and D-dimer levels. Patients who received reperfusion after stroke (thrombolytic therapy and mechanical thrombectomy) were identified based on thorough examinations of the patients' medication lists, prescriptions and medical records (Adams et al., 1993). All patients underwent neuroimaging examination of the brain with computed tomography (CT) or magnetic resonance imaging (MRI), then obtained imaging information such as cortical involvement, the infarcted lesion in the middle cerebral artery (MCA) area, and multiple lobes involvement. In addition, the SeLECT score was calculated by collecting data on the NIHSS score at admission, large-artery atherosclerosis, early seizure, cortical involvement, and territory of the middle cerebral artery (Galovic et al., 2018). Detailed definitions of all variables included in this study were presented in Supplementary Table S1.

2.3. Clinical follow-up

To identify late seizures that occurred within 2 years after stroke among the enrolled patients, we conducted follow-ups and seizure assessments using a mix of methods, including telephone interview, the electronic medical record, and outpatient visits with epilepsy specialist. The follow-up and seizure assessment methods as follow: First, we retrospectively identified patients who experienced late seizures during their hospital stay by checking their electronic medical records. Then, for those patients who returned for regular check-ups within two years after being discharged, we assessed potential post-stroke epilepsy (PSE) by reviewing their outpatient medical records. As for patients who didn't come back for check-ups, we screened for PSE through telephone interviews and specific questionnaires (see details in Supplementary Questionnaires) (Tan et al., 2012) for epilepsy symptoms. Those who screened positive were invited for outpatient visits and underwent electroencephalogram examinations, assessed by two epilepsy specialists. These follow-up procedures were carried out between 3 June and 21 October 2022 for the derivation cohort and from 22 October to 28 December 2022 for the validation cohort.

2.4. Definitions

Acute ischemic stroke was defined as sudden neurologic dysfunction caused by focal brain ischemia with imaging evidence of acute infarction (Mendelson and Prabhakaran, 2021). We used previously published studies to classify patients as early seizure($\leq$7 days post-stroke) or late seizure (spontaneous unprovoked seizures $>$7 days post-stroke, congruent with post-stroke epilepsy) (Zelano et al., 2020). A seizure occurring 7 days after a stroke carries a recurrence risk of **>** 60 % within the next ten years based on previous epidemiological studies (Hesdorffer et al., 2009). Therefore, one late seizure can be diagnosed as PSE according to the latest International League Against Epilepsy (ILAE) definitions (Fisher et al., 2014).

2.5. Machine learning development process

2.5.1. Data preprocessing

2.5.1.1. Split database. Data from the derivation cohort was randomly partitioned into 70 % training and 30 % test sets. The training set was utilized for constructing the prediction model, while the test set was employed for model validation.

2.5.1.2. Missing value handling. Variables with missing data are an inevitable problem in retrospectively collected data. The presence of missing values could affect the performance of a model developed.

Within the training set data, missing values (details shown in Supplementary Table S2 and S3) were imputed using the "mice" package for R through a multiple imputation process performed five times (Version 4.1.0) (Austin et al., 2021).

2.5.1.3. Balancing dataset. The number of patients with PSE is far less than those without PSE in the dataset. In this situation, the ML algorithm may have difficulties learning the concept related to the minority class. To address this class imbalance problem, the Combination of Synthetic Minority Oversampling Technique and Tomek Link (SMOTE Tomek) technique was applied to balance the training dataset, aimed to improve the performance of predictive models. This is a hybridization between under-sampling and oversampling techniques. First, the original data in the training set was over-sample with SMOTE, and after that, Tomek Link was identified and deleted, generating a class-balanced data set (Batista et al., 2004; Swana et al., 2022).

2.5.1.4. Feature selection. Feature selection was taken to reduce the redundant features and find optimal features to improve ML efficiency, which is an important pre-processing step in ML. We adopted the Boruta feature filtering based on a random forest (RF) approach. In short, this algorithm first created "shadow" variables to extend the dataset by adding random disturbance terms in the original data set, and then built a model based on RF to calculate an importance score (Z score) for each variable in the original and the shadow ones. Variables that proved significantly more important than their shadow counterparts, were deemed critical and retained in the dataset, while the insignificant variables were removed. The significance ordering of the variables was determined after several iterations (Kursa and Rudnicki, 2010; Huang et al., 2022).

2.5.2. Algorithm training and validation

Six ML algorithms were used to construct prediction models for identifying high-risk PSE patients, including logistic regression (LR), naive Bayes (NB), support vector machine (SVM), multilayer perceptron (MLP), adaptive boost (AdaBoost), and gradient boosting decision tree (GBDT).

The predictive performance of the ML model requires further appraisal. The test set was used to verify the repeatability and the overfitting of the model (Bedogni, 2009). Models were validated against an independent validation cohort to evaluate whether it has general applicability. The area under the receiver operating characteristic (ROC) curve (AUC) was used to evaluate the discriminability of the model. In ROC analysis, the AUC value in the range of 0.50–0.59, 0.60–0.75, and more than 0.75 reflects poor discriminability, moderate discriminability, and high discriminability, respectively (Alba et al., 2017). The calibration of the model was assessed using the Brier score, which ranges from 0 for an ideal model to 0.25 for a non-informative model with a 50 % occurrence of the outcome (Van Calster et al., 2016). We calculated the accuracy, sensitivity, specificity, and F1 score of the ML model and the SeLECT score when the absolute risk threshold was fixed at 20 %. Patients who reached this threshold should receive extensive attention and could be considered for inclusion in clinical trials (Galovic et al., 2018). The F1 score, serving as the harmonic mean of precision and recall, proves particularly useful in evaluating models trained and tested on imbalanced datasets. F1 score ranges between 0 and 1, with a higher score indicating superior model performance (Tan et al., 2023). The predictive performance of the ML model is compared with the reference SeLECT score model.

2.5.3. Machine learning interpretation

ML models are often seen as "black boxes", making it hard to interpret how they make accurate disease predictions. To address this, we adopted the SHAP method to interpret the optimum ML model, showing the degree of contribution and the direction of the effect on the model by calculating the Shapley value (Lundberg and Lee, 2017). The Shapley value is a fair way to distribute profit among many stakeholders based on their contributions, SHAP represents the outcome as the sum of each feature contribution calculated as the Shapley value. Additionally, we employed the SHAP summary plot to visualize how features impact the optimum ML model.

2.6. Statistical analysis

Descriptive statistics of the dataset and evaluation indices of the ML models were calculated using R version 4.1.0 (R Core Team). The data preprocessing, algorithm training, and interpretability analysis were conducted using Python version 3.8.0 (Python Software Foundation). Categorical variables were presented as frequencies and percentages. Continuous variables were presented as mean and standard deviation (SD) if approximately normally distributed, and as median and interquartile range (IQR) if not. DeLong's test was used to compare the AUCs for various models. Because the study which develop SeLECT model showed the risk of PSE within the 2 years after stroke under the different SeLECT value by graphs, the software GetData Graph Digitizer 2.24 was applied to digitize and extract the data (See Supplementary Table S4). P values < 0.05 were considered to be statistically significant.

3. Results

3.1. Patient characteristics

A total of 1977 and 870 patients were included in the derivation and validation cohorts, respectively (Fig. 1). The median age of the patients was 64.33±11.84 years, with 65.1 % being men, and a median follow-up duration of 3.6 years (IQR 3.2–4.1). During the follow-up period, 133 patients developed PSE, with 87 (4.4 %) from the derivation cohort and 46 (5.3 %) from the validation cohort. Table 1 presents baseline information of patients with and without PSE in the two cohorts. Supplementary Table S5 displays the baseline characteristics of patients who dropped out versus those who did not.

3.2. Model building and evaluation

We used 1384 patients to train six ML models for predicting the occurrence of epilepsy within two years after an ischemic stroke. A total of 33 features were initially extracted and four features were finally selected by the Boruta algorithm, namely the NIHSS score, hospital length of stay, D-dimer level, and cortical involvement. VIFs of the NIHSS score, hospital length of stay, D-dimer level, and cortical involvement in the ML models were found to be within acceptable limits, as 1.14, 1.12, 1.05, and 1.06, respectively.

In the training set, LR, NB, SVM, MLP, AdaBoost, and GBDT models were constructed using the four selected variables. In the test set, they yielded AUC values of 0.849, 0.831, 0.811, 0.751, 0.753, and 0.709, respectively (shown in Table 2 and Fig. 2(A)). The LR, NB, SVM, and reference SeLECT models demonstrated clear discrimination capabilities. However, DeLong's test indicated that there were no statistically significant differences between the ROC curves of each other (See Supplementary Table S6). The Brier scores for all models, except the AdaBoost and GBDT models, were less than 0.25. In predicting PSE, the LR and NB models showed the highest sensitivity, reaching 0.900, where the SeLECT model demonstrated the lowest sensitivity at just 0.150. F1 score for our ML models ranged between 0.113 and 0.222. The F1 scores of the SeLECT model is 0.158, and lower than NB, LR, SVM, and MLP models.

In the validation set, the LR, NB, and SVM achieved high AUCs (0.749, 0.757 and 0.754) (shown in Table 2 and Fig. 2(B)). The differences in AUC values between these three models and the SeLECT model were not statistically significant, as indicated by pairwise comparisons (See Supplementary Table S6). With the exception of the AdaBoost and

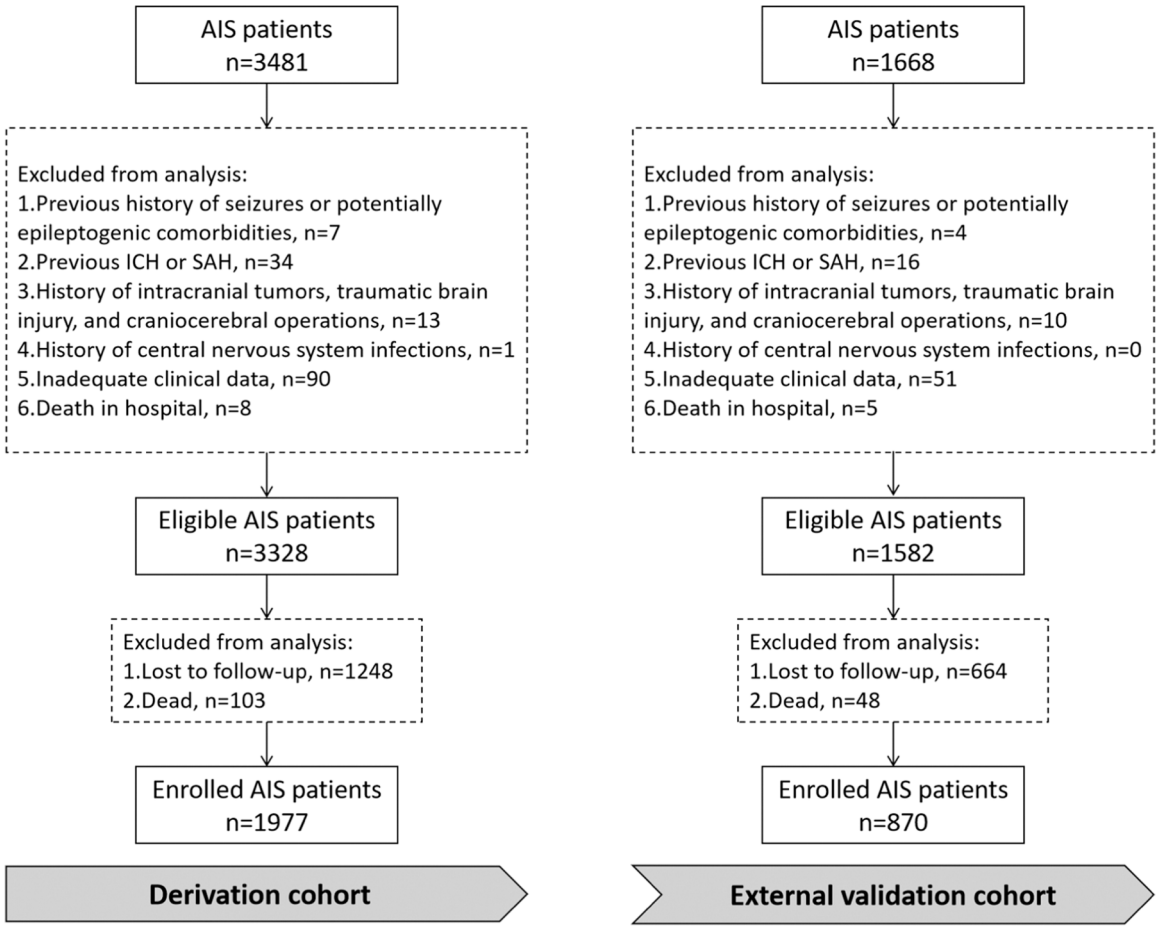

Fig. 1. Flow chart of this study. Abbreviations: AIS=acute ischemic stroke; ICH= intracerebral hemorrhage; SAH=spontaneous subarachnoid hemorrhage.

GBDT models, the Brier score for the other ML models ranged from 0.126 to 0.195. Moreover, the LR, NB, and SVM models showed higher sensitivities of 0.783, 0.739, and 0.783, respectively. However, the SeLECT model exhibited sensitivity and F1 score of 0.043 and 0.059, respectively, during for external verification, both markerly lower than ML models. Although the discrimination of the SeLECT model was slightly lower than ML models, its sensitivity and F1 score for predicting the PSE were the lowest compared with all ML models. This suggests that the SeLECT model may struggle to accurately identify patients with a higher risk of developing PSE. Therefore, the NB model obtained better evaluation scores than other PSE prediction models and was considered the optimal model.

3.3. Explanation of NB model with the SHAP method

The SHAP values were estimated to assess the importance of each predictor variable in relation to the outcome predicted by the NB model. In Fig. 3(A), the feature importance plot showed the predictors ranked in descending order of importance. The NIHSS score exhibited the highest predictive value for all prediction horizons, followed by the hospital length of stay, D-dimer, and cortical involvement.

The SHAP summary plot was utilized to identify both positive and negative associations between predictors and the target outcome (shown in Fig. 3(B)). The interpretation of this plot is based on three factors. First, a dot represents each feature attribution value for each participant in the model, meaning one participant is assigned one dot on the line for each feature. Second, the x-axis represents the Shapley value, indicating the prediction effect. A positive Shapley value means the feature is more likely to predict PSE, with a higher positive Shapley value indicating a greater PSE risk for baseline stroke patients. Third, dots are colored according to the feature values for each patient and stack vertically to show density. Red dots represent higher feature values, while blue dots represent lower values. For example, NIHSS score had a long tail to the right with red dots, indicating that higher NIHSS scores at baseline increased the risk of PSE. In summary, the plot showed the following effect directions: higher NIHSS scores, longer hospital stays, higher D-dimer levels, and the presence of cortical involvement predicted a higher risk of PSE.

4. Discussion

This study developed and validated six ML methods utilizing a set of 33 variables to predict the 2-year risk of epilepsy after ischemic stroke. The NB model demonstrated the most robust performance in PSE prediction, with an AUC of 0.757, accuracy of 0.722, sensitivity of 0.739, and specificity of 0.720. Our ML models achieved high predictive performance using only routinely available data to predict PSE.

Compared to the reference SeLECT model, we found similarly good discrimination capability with the NB model. However, the AUC is insensitive to changes in absolute risk estimates and cannot evaluate the model fit. Therefore, we further used the Brier score as a calibration measure to reflect the extent to which a model correctly estimates the absolute risk. We observed a good degree of correspondence between the predicted and observed risk of PSE for both the NB model and the SeLECT model. Moreover, it is necessary to use sensitivity, specificity, and F1 score to reflect the clinical application, especially when there was not much difference between the AUC of the two models. The NB model achieved the best sensitivity and F1 score, whereas the SeLECT model

Table 1
Characteristics of patients from derivation cohort and validation cohort with ischemic stroke stratified by seizure occurrence.

Patient characteristics		Derivation cohort		Validation cohort	
		No epilepsy (n = 1890)	Epilepsy (n = 87)	No epilepsy (n = 824)	Epilepsy (n = 46)
Sex (male), n (%)		1213 (64.2 %)	53 (60.9 %)	558 (67.7 %)	30 (65.2 %)
Age(years), median (IQR)		64 (55,72)	68.26±12.43[a]	66 (59,75)	66.22±10.75[a]
Length of stay(days), median (IQR)		9 (7,11)	12 (9,15)	11 (9,13)	13 (8,17)
Vascular risk factors, n (%)					
Hypertension		1365 (72.2 %)	64 (73.6 %)	644 (78.2 %)	34 (73.9 %)
Diabetes		632 (33.4 %)	29 (33.3 %)	312 (37.9 %)	16 (34.8 %)
Hyperlipidemia		189 (10.0 %)	8 (9.2 %)	150 (18.2 %)	10 (21.7 %)
Atrial fibrillation		140 (7.4 %)	28 (32.2 %)	78 (9.5 %)	8 (17.4 %)
Coronary heart disease		324 (17.1 %)	29 (33.3 %)	284 (34.5 %)	22 (47.8 %)
Cancer		68 (3.6 %)	3 (3.4 %)	36 (4.4 %)	6 (13.0 %)
History of ischemic stroke		277 (14.7 %)	19 (21.8 %)	172 (20.9 %)	4 (8.7 %)
Smoking		670 (35.4 %)	28 (32.2 %)	340 (41.3 %)	20 (43.5 %)
Alcohol-drinking		563 (29.8 %)	23 (26.4 %)	278 (33.7 %)	10 (21.8 %)
Vital signs, median (IQR)					
SBP, mmHg		149 (136,164)	152.92±23.55[a]	147 (135,162)	146 (126,163)
DBP, mmHg		83 (75,92.25)	83 (76,92)	81 (74,92)	79 (70,89)
NIHSS at admission, median (IQR)		2 (1,5)	5 (2,12)	3 (1,5)	4 (2,6)
Stroke cause, n(%)					
Large-artery atherosclerosis		1062 (56.2 %)	40 (46.0 %)	542 (65.8 %)	26 (56.5 %)
Cardioembolism		116 (6.1 %)	27 (31.0 %)	62 (7.5 %)	8 (17.4 %)
Small-vessel occlusion		536 (28.4 %)	11 (12.6 %)	186 (21.4 %)	6 (13.0 %)
Other determined cause		60 (3.2 %)	4 (4.6 %)	14 (1.7 %)	6 (13.0 %)
Undetermined cause		116 (6.1 %)	5 (5.7 %)	20 (2.4 %)	-
Laboratory variables, median (IQR)					
Fasting blood glucose, mmol/L		5.38 (4.71,6.99)	5.81 (4.86,7.65)	5.3 (4.62,7.31)	6.07 (5.61,7.90)
Total cholesterol, mmol/L		4.26 (3.59,5.01)	4.44±1.12[a]	4.51 (3.79,5.23)	4.72 (3.76,4.92)
Triglycerides, mmol/L		1.27 (0.94,1.77)	1.2 (0.98,1.71)	1.24 (0.95,1.65)	1.25 (0.88,1.82)
Cholesterol LDL, mmol/L		2.53 (2,3.12)	2.66±0.94[a]	2.65 (2.17,3.13)	2.82 (2.25,3.00)
D-dimer, ng/ML		230 (150,370)	450 (30,900)	400 (290,600)	620 (410,900)
Reperfusion treatment, n(%)					
Thrombectomy		5 (0.3 %)	2 (2.3 %)	14 (1.7 %)	2 (4.3 %)
Thrombolysis		147 (7.8 %)	8 (9.2 %)	38 (4.6 %)	2 (4.3 %)
Lung infection, n(%)		124 (6.6 %)	31 (35.6 %)	102 (12.4 %)	14 (30.4 %)
Neuroimaging markers, n(%)					
Multiple lobes involvement		444 (23.5 %)	52 (59.8 %)	188 (22.8 %)	28 (60.9 %)
Cortical involvement	483 (25.6 %)		65 (74.7 %)	272 (33.0 %)	32 (69.6 %)
Territory of MAC	986 (52.2 %)		65 (74.7 %)	452 (54.9 %)	36 (78.3 %)
Hemorrhagic transformation	47 (2.5 %)		11 (12.8 %)	30 (3.6 %)	10 (21.7 %)
Early seizure, n(%)	11 (0.8 %)		13 (14.9 %)	6 (0.7 %)	4 (8.7 %)

Abbreviation: DBP, diastolic blood pressure; HDL, high-density lipoprotein; IQR, interquartile range; LDL, low-density lipoprotein; MAC, middle cerebral artery; NIHSS National Institutes of Health Stroke Scale; SBP, systolic blood pressure.

[a] Continuous variables that approximated the normal distribution were expressed as means ± standard deviation (SD).

exhibited poorer sensitivity and F1 score. Due to the F1 score takes into accounts both precision and recall, we consider a more suitable metric for accessing the accuracy of the ML and SeLECT models in predicting patients at high risk of developing PSE. Our results showed that the SeLECT model misclassified a significant number of patients at high risk of PSE as low risk when the absolute risk threshold was fixed at 20 %. A possible explanation for this might be that the low proportion of high-risk patients leads to the low sensitivity and F1 score of the model. Sampling methods, including over-sampling combined with under-sampling, were taken to compensate for the imbalanced data in our study. The developers for the SeLECT model updated it to capture high-risk patients of PSE, however, they did not validate the model externally (Sinka et al., 2023).

We applied the SHAP method to enhance the interpretability, transparency and acceptability of our NB model. The result indicated that the NIHSS score was the most influential characteristic variable, followed by hospital length of stay, D-dimer level, and cortical involvement. Our study suggests that D-dimer levels and length of hospital stay have a stronger predictive role for PSE compared to cortical involvement. One possible explanation is that D-dimer levels and length of hospital stay, serving as a composite indicator of disease severity, demonstrated higher predictive value than individual feature indicators. The field of explainable ML can help counter opaque feature representation learning, but the black-box nature still requires careful consideration and justification in the light of clinical experience and results of previous studies (Hunter and Holmes, 2023).

As a widely used scoring system to assess neurological deficits in patients with ischemic stroke, the NIHSS was the most important variable for PSE prediction. Our finding is consistent with a prospective study that predicted epilepsy after ischemic stroke based on a classical regression algorithm (Galovic et al., 2018). An explanation for this might be that anterior circulation infarct or MCA territory involvement is a major factor in the development of PSE (Pitkänen et al., 2016). These patients may have a higher NIHSS score than patients with posterior circulation lesions because the NIHSS is heavily weighted toward neurological impairment caused by anterior circulation lesions, such as cortical signs and motor function (Alemseged et al., 2022). Additionally, hospital length of stay was also important factor in our ML models. This finding was broadly in line with previous research (Chi et al., 2018; Do et al., 2022; Hardtstock et al., 2021). Hospital length of stay may reflect the clinical condition of these patients who had suffered a severe stroke or early-onset seizure events that could prolong the duration of hospitalization (Alet et al., 2022). These events may potentially increase the risk of PSE.

One unanticipated finding was that D-dimer was the third important factor for predicting PSE. Stroke patients with elevated D-dimer levels showed a higher likelihood of developing PSE than patients with low D-dimer levels. There is compelling evidence showing a correlation

Table 2
Validation of different ML models and SeLECT score for predicting epilepsy after ischemic stroke.

	AUC	Accuracy	Sensitivity	Specificity	F1	Brier score
LR						
Internal validation	0.849	0.838	0.900	0.775	0.215	0.172
External validation	0.749	0.722	0.783	0.718	0.218	0.195
NB						
Internal validation	0.831	0.786	0.900	0.672	0.159	0.108
External validation	0.757	0.722	0.739	0.720	0.220	0.126
SVM						
Internal validation	0.811	0.793	0.850	0.737	0.181	0.154
External validation	0.754	0.706	0.783	0.701	0.207	0.176
MLP						
Internal validation	0.751	0.721	0.600	0.841	0.222	0.126
External validation	0.576	0.556	0.652	0.551	0.137	0.172
AdaBoost						
Internal validation	0.753	0.729	0.900	0.557	0.113	0.251
External validation	0.614	0.611	0.609	0.612	0.142	0.252
GBDT						
Internal validation	0.709	0.710	0.650	0.770	0.113	0.404
External validation	0.524	0.317	0.913	0.284	0.119	0.567
SeLECT score						
Internal validation	0.828	0.562	0.150	0.974	0.158	0.032
External validation	0.732	0.509	0.043	0.976	0.059	0.048

Abbreviation: AdaBoost, adaptive boost; GBDT, gradient boosting decision tree; LR, logistic regression; MLP, multilayer perceptron; NB, naive Bayes; SVM, support vector machine.

between D-dimer levels and PSE. First, previous research has established D-dimer as a biomarker of poor prognosis of stroke (Sato et al., 2020; Wang et al., 2020; Nam et al., 2023). A recent national study indicated that higher D-dimer level at baseline and 90 days was associated with unfavorable functional outcomes after stroke, as indicated by a modified Rankin Scale of 3–6) (Hou et al., 2021). Additionally, studies have identified high disability levels at discharge (Winter et al., 2018), delayed neurological recovery (Freiman et al., 2023), and high modified Rankin Scale scores (Alet et al., 2022)as risk factors for PSE. Thus, we hypothesize that D-dimer, as a biomarker of stroke prognosis, may also predict the risk of PSE. Furthermore, the coagulation cascade induced permanent structural changes after the ischemic stroke that contribute to PSE development (Altman et al., 2019). Although a correlation between hypercoagulable states and ES has been reported (Turaga et al., 2021), its correlation with PSE remains unclear. Given that D-dimer serves as a marker of the hypercoagulable stage and the activity of blood coagulation pathways (Tomimaru et al., 2006; Mu et al., 2022), investigating the association between D-dimer and PSE warrants further exploration in the future. However, our finding contrasts with a previous study aimed at identifying biomarkers associated with epilepsy (Abraira et al., 2020). One possible explanation for the difference could be that the previous study did not perform separate analyses for ischemic and hemorrhagic strokes, resulting in a diminished effect of higher D-dimer levels in ischemic strokes. Currently, researchers have hardly focused on associations between biomarkers and post-ischemic stroke epilepsy. In this regard, our study contributed to gaining an understanding of a new mechanism for PSE and brought forth prospects for further research in this aspect. Future studies with larger sample sizes are needed to confirm these findings.

5. Limitations

This study has several limitations warrants mentioning. First, the use of telephone follow-up to identify patients with PSE may reduce the accuracy of epilepsy diagnosis. To address this problem, we employed a combination of valid epilepsy screening questionnaires, the electronic medical record, and epilepsy specialist outpatient visits to improve the accuracy of identification. Second, data on the use of secondary prevention drugs for stroke were not collected during the follow-up period. Although some studies suggest that aspirin (Zhao et al., 2020), statin (Guo et al., 2015; Zhu et al., 2021; Vitturi and Gagliardi, 2020; Xu et al., 2020), and antihypertensive drugs (Sarfo et al., 2020) may reduce the risk of seizures, there is currently no direct evidence to support this conclusion. Third, the high dropout rate observed in our study, despite our efforts to minimize it, may introduce bias and uncertainty into the study results. Finally, our data were drawn from a relatively small sample for ML model building. Fourthly, our study excluding patients who died within two years without developing PSE may lead to potential bias in our model, as these patients may have had different risk profiles compared to those who survived the two-year period. This means that our model is more applicable to predicting PSE among stroke survivors

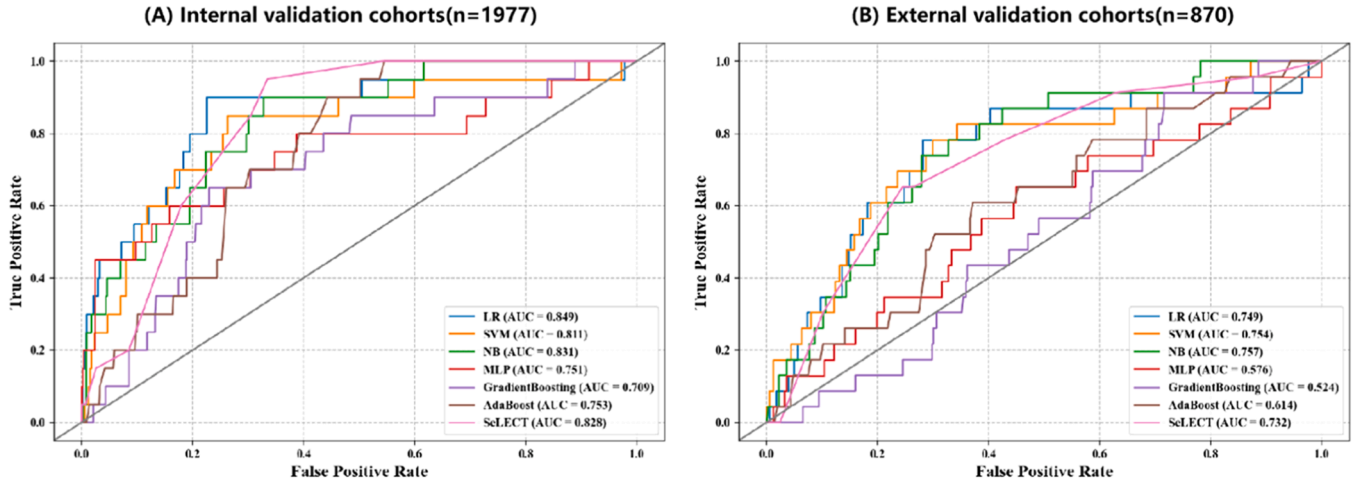

Fig. 2. This figure shows the classification performance of the machine learning model against SeLECT score for predicting PSE. (A) ROC analysis of risk score on internal validation. (B) ROC analysis of risk score on external validation.

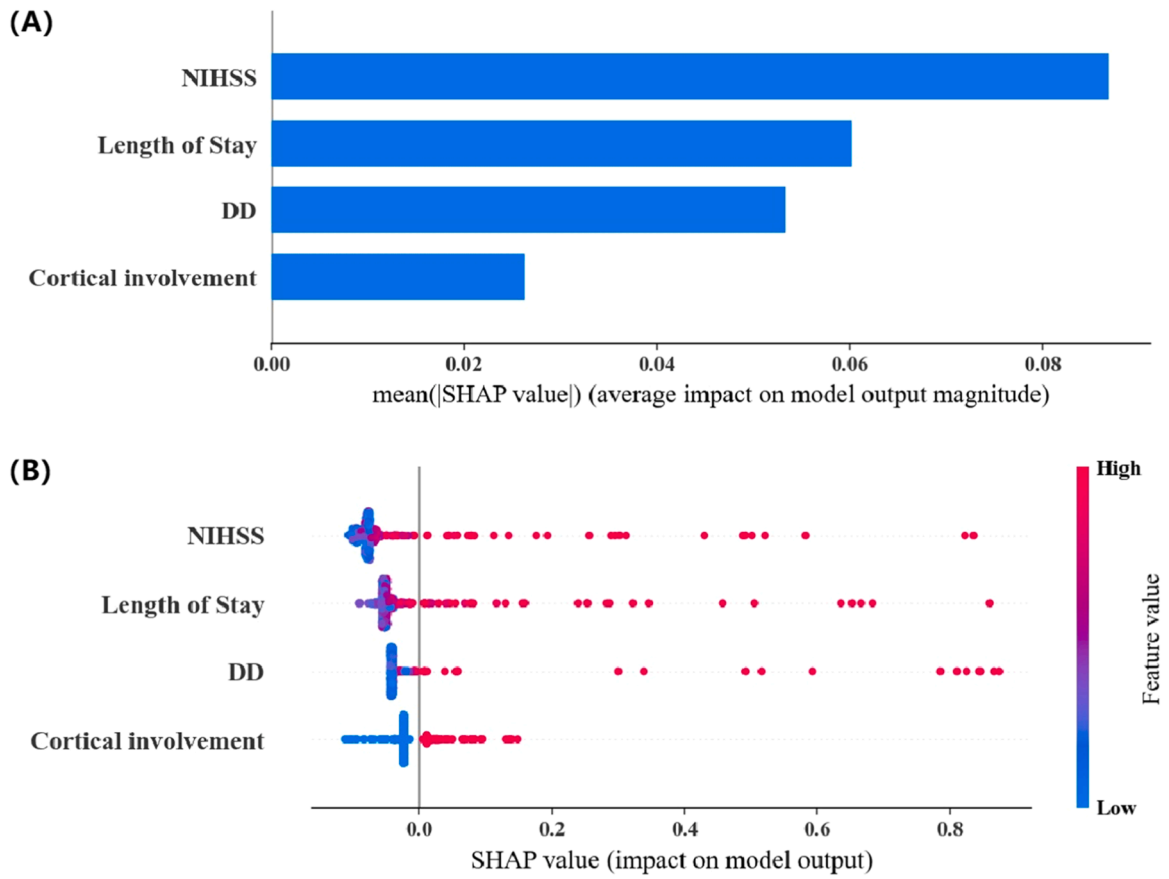

Fig. 3. Model explainability for post-ischemic stroke epilepsy. (A) The weights of variables importance. (B) The SHapley Additive exPlanation (SHAP) values summary plot of the Naive Bayes model. DD: D-dimer.

who live at least two years post-stroke, and it may not fully generalise to the broader stroke population. The future studies should aim to include all possible post-stroke patients and consider competing risks to provide a more comprehensive risk prediction. To mitigate concerns about overfitting, we validated our ML model in an external validation cohort. However, the model's performance in different regions and ethnic groups needs to be further examined.

6. Conclusions

Our study successfully explored the use of ML algorithms trained with clinical characteristics, laboratory examination results, and neuroimaging manifestations to achieve accurate predictions of PSE. This endeavor demonstrated ML has potential advantages and applications to support clinical decision-making. In the future, further validation and optimization of our ML model will be conducted in a large cohort.

Disclosure

None of the authors has any conflict of interest to disclose.

Funding

This research was supported by the National Natural Science Foundation of China (NSFC) [81901323].

CRediT authorship contribution statement

Zhibin Chen: Writing – review & editing, Validation, Methodology. **Yong Yang:** Writing – review & editing, Funding acquisition. **Yue Yu:** Writing – original draft, Visualization, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Jiajun Zhang:** Writing – review & editing. **Yan Wang:** Writing – review & editing, Validation, Supervision, Resources, Project administration, Investigation, Funding acquisition.

Data availability

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

Acknowledgments

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.eplepsyres.2024.107397.

References

Abraira, L., et al., 2020. Blood biomarkers predictive of epilepsy after an acute stroke event. Epilepsia 61 (10), 2244–2253.
Adams Jr., H.P., et al., 1993. Classification of subtype of acute ischemic stroke. Definitions for use in a multicenter clinical trial. TOAST. Trial of Org 10172 in Acute Stroke Treatment. Stroke 24 (1), 35–41.
Alba, A.C., et al., 2017. Discrimination and calibration of clinical prediction models: users' guides to the medical literature. Jama 318 (14), 1377–1384.
Alemseged, F., et al., 2022. Posterior national institutes of health stroke scale improves prognostic accuracy in posterior circulation stroke. Stroke 53 (4), 1247–1255.

Alet, M., et al., 2022. Predictive factors for the development of epilepsy after ischemic stroke. J. Stroke Cereb. Dis. 31 (12), 106858.
Altman, K., Shavit-Stein, E., Maggio, N., 2019. Post stroke seizures and epilepsy: from proteases to maladaptive plasticity. Front Cell Neurosci. 13, 397.
ANON, 2021. Global, regional, and national burden of stroke and its risk factors, 1990-2019: a systematic analysis for the Global Burden of Disease Study 2019. Lancet Neurol. 20 (10), 795–820.
Austin, P.C., et al., 2021. Missing Data in Clinical Research: A Tutorial on Multiple Imputation. Can. J. Cardiol. 37 (9), 1322–1331.
Batista, G.E.A.P.A., Prati, R.C., Monard, M.C., 2004. A study of the behavior of several methods for balancing machine learning training data. SIGKDD Explor. Newsl. 6 (1), 20–29. https://doi.org/10.1145/1007730.1007735.
Bedogni, G., 2009. Clinical Prediction Models-a Practical Approach to Development, Validation and Updating. J. R. Stat. Soc. Ser. A-Stat. Soc. 172, 944-944.
Bentes, C., et al., 2017. Post-stroke seizures are clinically underestimated. J. Neurol. 264 (9), 1978–1985.
Bladin, C.F., et al., 2000. Seizures after stroke: a prospective multicenter study. Arch. Neurol. 57 (11), 1617–1622.
Brott, T., et al., 1989. Measurements of acute cerebral infarction: a clinical examination scale. Stroke 20 (7), 864–870.
Bryndziar, T., et al., 2016. Seizures Following Ischemic Stroke: Frequency of Occurrence and Impact on Outcome in a Long-Term Population-Based Study. J. Stroke Cereb. Dis. 25 (1), 150–156.
Bunney, G., et al., 2022. Predicting Early Seizures After Intracerebral Hemorrhage with Machine Learning. Neurocrit Care 37 (Suppl 2), 322–327.
Chi, N.F., et al., 2018. Development and validation of risk score to estimate 1-year late poststroke epilepsy risk in ischemic stroke patients. Clin. Epidemiol. 10, 1001–1011.
De Reuck, J., et al., 2006. Computed tomographic changes of the brain and clinical outcome of patients with seizures and epilepsy after an ischaemic hemispheric stroke. Eur. J. Neurol. 13 (4), 402–407.
Do, P.T., et al., 2022. Risk Factors for Postischemic Stroke Epilepsy in Young Adults: A Nationwide Population-Based Study in Taiwan. Front Neurol. 13, 880661.
Fisher, R.S., et al., 2014. ILAE official report: a practical clinical definition of epilepsy. Epilepsia 55 (4), 475–482.
Freiman, S., et al., 2023. Post-stroke seizures, epilepsy, and mortality in a prospective hospital-based study. Front Neurol. 14, 1273270.
Galovic, M., et al., 2018. Prediction of late seizures after ischaemic stroke with a novel prognostic model (the SeLECT score): a multivariable prediction model development and validation study. Lancet Neurol. 17 (2), 143–152.
Guo, J., et al., 2015. Statin treatment reduces the risk of poststroke seizures. Neurology 85 (8), 701–707.
Haapaniemi, E., et al., 2014. The CAVE score for predicting late seizures after intracerebral hemorrhage. Stroke 45 (7), 1971–1976.
Hardtstock, F., et al., 2021. Poststroke epilepsy incidence, risk factors and treatment: German claims analysis. Acta Neurol. Scand. 143 (6), 614–623.
Hesdorffer, D.C., et al., 2009. Is a first acute symptomatic seizure epilepsy? Mortality and risk for recurrent seizure. Epilepsia 50 (5), 1102–1108.
Hou, H., et al., 2021. Association of Level and Increase in D-Dimer With All-Cause Death and Poor Functional Outcome After Ischemic Stroke or Transient Ischemic Attack. J. Am. Heart Assoc. 10 (3), e018600.
Huang, F., et al., 2022. Identifying COVID-19 Severity-Related SARS-CoV-2 Mutation Using a Machine Learning Method. Life (Basel) 12 (6).
Hunter, D.J., Holmes, C., 2023. Where Medical Statistics Meets Artificial Intelligence. N. Engl. J. Med 389 (13), 1211–1219.
Kasner, S.E., et al., 1999. Reliability and validity of estimating the NIH stroke scale score from medical records. Stroke 30 (8), 1534–1537.
Kursa, M.B., Rudnicki, W.R., 2010. Feature Selection with the Boruta Package. J. Stat. Softw. 36 (11), 1–13.
Lamy, C., et al., 2003. Early and late seizures after cryptogenic ischemic stroke in young adults. Neurology 60 (3), 400–404.
Lossius, M.I., et al., 2005. Poststroke epilepsy: occurrence and predictors–a long-term prospective controlled study (Akershus Stroke Study). Epilepsia 46 (8), 1246–1251.
Lundberg, S.M. and S.I. Lee. A Unified Approach to Interpreting Model Predictions. in 31st Annual Conference on Neural Information Processing Systems (NIPS). 2017. Long Beach, CA.
Mendelson, S.J., Prabhakaran, S., 2021. Diagnosis and management of transient ischemic attack and acute ischemic stroke: a review. Jama 325 (11), 1088–1098.
Mu, S., et al., 2022. Predictive factors for early-onset seizures in patients with cerebral venous sinus thrombosis. Front Neurol. 13, 842807.
Nam, K.W., Kwon, H.M., Lee, Y.S., 2023. Clinical significance of D-dimer levels during acute period in ischemic stroke. Thromb. J. 21 (1), 55.
Pitkänen, A., Roivainen, R., Lukasiuk, K., 2016. Development of epilepsy after ischaemic stroke. Lancet Neurol. 15 (2), 185–197.
Roivainen, R., et al., 2013. Young adult ischaemic stroke related acute symptomatic and late seizures: risk factors. Eur. J. Neurol. 20 (9), 1247–1255.
Sarfo, F.S., et al., 2020. Prevalence and predictors of post-stroke epilepsy among Ghanaian stroke survivors. J. Neurol. Sci. 418, 117138.
Sato, T., et al., 2020. D-dimer level and outcome of minor ischemic stroke with large vessel occlusion. J. Neurol. Sci. 413, 116814.
Sinka, L., et al., 2023. Association of mortality and risk of epilepsy with type of acute symptomatic seizure after ischemic stroke and an updated prognostic model. JAMA Neurol. 80 (6), 605–613.
Swana, E.F., Doorsamy, W., Bokoro, P., 2022. Tomek Link and SMOTE approaches for machine fault classification with an imbalanced dataset. Sensors 22 (9).
Tan, M.L., et al., 2012. Tissue plasminogen activator does not alter development of acquired epilepsy. Epilepsia 53 (11), 1998–2004.
Tan, W.Y., et al., 2023. A machine learning approach for early diagnosis of cognitive impairment using population-based data. J. Alzheimers Dis. 91 (1), 449–461.
Tomimaru, Y., et al., 2006. Plasma D-dimer levels show correlation with number of lymph node metastases in patients with esophageal cancer. J. Am. Coll. Surg. 202 (1), 139–145.
Turaga, S.P., et al., 2021. Frequency and predictors of early seizures following first acute stroke: data from a university hospital in South India. Neurol. India 69 (4), 847–855.
Van Calster, B., et al., 2016. A calibration hierarchy for risk models was defined: from utopia to empirical data. J. Clin. Epidemiol. 74, 167–176.
Vitturi, B.K., Gagliardi, R.J., 2020. The influence of statins on the risk of post-stroke epilepsy. Neurol. Sci. 41 (7), 1851–1857.
Wang, J., et al., 2020. D-dimer and its combination with blood lipid on prognosis of patients with acute ischemic stroke. J. Stroke Cereb. Dis. 29 (12), 105394.
Winter, Y., et al., 2018. Health-related quality of life in patients with poststroke epilepsy. Epilepsy Behav. 80, 303–306.
Xu, M.Y., 2019. Poststroke seizure: optimising its management. Stroke Vasc. Neurol. 4 (1), 48–56.
Xu, T., et al., 2020. Statin use and the risk of post-stroke seizures: a meta-analysis. Seizure 83, 63–69.
Zelano, J., et al., 2016. Association between poststroke epilepsy and death: A nationwide cohort study. Eur. Stroke J. 1 (4), 272–278.
Zelano, J., 2020. Prognosis of poststroke epilepsy. Epilepsy Behav. 104 (Pt B), 106273.
Zelano, J., et al., 2020. How to diagnose and treat post-stroke seizures and epilepsy. Epileptic Disord. 22 (3), 252–263.
Zhao, T., et al., 2020. Effects of atorvastatin and aspirin on post-stroke epilepsy and usage of levetiracetam. Medicine (Baltim.) 99 (50), e23577.
Zhu, Y., et al., 2021. Effects of double-dose statin therapy for the prevention of post-stroke epilepsy: a prospective clinical study. Seizure 88, 138–142.