

RESEARCH

Open Access



Development of a flexible self-calculation scoring model to determine stroke occurrence

Sunghyon Kyeong¹ and Dae Hyun Kim^{2*}

*Correspondence:

Dae Hyun Kim

dh23.kim@samsung.com

¹Institute of Behavioral Science in Medicine, Yonsei University College of Medicine, Seoul, Republic of Korea

²Department of Physical and Rehabilitation Medicine, Center for Prevention and Rehabilitation, Heart Vascular Stroke Institute, Samsung Medical Center, Sungkyunkwan University School of Medicine, Irwon-ro 115, Gangnam-gu, Seoul 06355, South Korea

Abstract

Stroke has become a significant threat to global public health, the ideal solution to which is primary prevention. Identification and management of determinants of stroke among various variables in different datasets are essential steps for its primary prevention. This study aimed to develop a flexible scoring model, which can easily modify different datasets. The public dataset containing 41,931 cases with 643 occurrences of stroke was randomly divided into training, validation, and test datasets comprising 25,158 (60%), 8,386 (20%), and 8,387 (20%) cases, respectively. Three continuous variables (age, body mass index, and average glucose level) and seven categorical variables (heart disease, hypertension, sex, married/smoking/work/residence status) in the dataset were converted using the weight of evidence method. The significant variables among 10 transformed variables were selected using multivariable logistic regression analyses. The scoring model for stroke occurrence was developed in the training and validation datasets, and performance was evaluated in the test dataset. Age, average glucose level, heart disease, and hypertension were significant variables of stroke occurrence. The scoring model was easily calculated using four determinants and indicates that the stroke occurrence ranged from 0.04 to 12.50%. The performance of the scoring model on the test dataset was similar to that on the validation dataset. This novel point scoring model is flexible enough to modify various datasets and can be used for determinant identification. Furthermore, its simplicity allows individuals to manage determinants by self-calculating stroke occurrence. Our model contributes to primary prevention using determinant identification and management.

Keywords Scoring model, Stroke, Prediction, Risk, Prevention

Introduction

Approximately 12.2 million incidents of stroke occur worldwide each year [1, 2]. Stroke is the leading cause of chronic disability, with an increasing burden on patients and society [1, 3]. Management of stroke leads to an improved prognosis due to recanalization and restoration [4–6]. Nevertheless, primary prevention remains the most effective means of reducing the stroke burden.

Identifying and managing risk factors is important to prevent strokes. Because the pathophysiology differs between ischemic and hemorrhagic stroke, risk factors can differ

between the two types [7, 8]. Nevertheless, risk factors, including old age, hypertension, diabetes mellitus (DM), smoking, and obesity, are common to ischemic and hemorrhagic stroke [1].

Risk factors can be classified into modifiable and unmodifiable. The management and identification of modifiable risk factors such as hypertension, DM, dyslipidemia, and atrial fibrillation (AF) are a major focus. Stratification of stroke occurrence aids in primary prevention by selecting candidates for the appropriate management of risk factors.

The Framingham Stroke Risk Score (FSRS), CHADS₂, ABCD, and ABCD2 scores are widely used in clinical practice to stratify stroke occurrence [9–13]. The FSRS is based on cohort data in subjects aged 55–84 years and predicts the 10-year stroke risk [13, 14]. The model is calculated using variables, including age, sex, systolic blood pressure, use of antihypertensive drug, presence/absence of left ventricular hypertrophy on electrocardiogram, prevalent cardiovascular disease, current smoking status, current/previous AF, and DM [14]. The CHADS₂ investigates data from the National Registry of AF consisting of Medicare beneficiaries aged 65–95 years and is used to identify stroke risk and determine the eligibility of patients with AF for antithrombotic therapies [10]. The CHADS₂ assigns one point for the presence of congestive heart failure, hypertension, and DM and age ≥ 75 years and two points for stroke history or transient ischemic attack [10]. The ABCD score, proposed for risk stratification after transient ischemic attack [12], combines variables of age, blood pressure, clinical features, and duration of systems scored using a six-point scale. Moreover, ABCD2 is a modified model that includes history of DM as the clinical variable [9].

Previous models of stroke occurrence stratification have limitations for primary prevention: each model stratifies different risk categories in one person; some models are applicable to patients with specific conditions, and each model requires different fixed variables. These variables are required, but it is impossible to modify the variables to fit each dataset. Medical datasets include the potential risk factors of stroke and require a variable modification model. These models do not fit the current clinical situation, and modification by increasing the number of datasets remains difficult.

Prevention models require flexibility and easy calculation. A model that can be transformed according to the available variables in datasets, rather than fixed variables, can be widely used. Furthermore, if individuals can calculate their own stroke probability by themselves, motivation for the management of risk factors will further increase.

We aimed to develop a scoring model for stroke occurrence using a public dataset. The established model exhibits characteristics of flexibility for the included variables, which allows modification of other datasets and simplicity for self-calculation.

Methods

Dataset

In this study, we used the benchmark dataset of the Kaggle competition for stroke prediction (<https://www.kaggle.com/asaumya/healthcare-dataset-stroke-data>. Accessed December 1, 2018). This dataset is a subset of the original stroke data collected from healthdata.gov and accounts for 1.18% of the whole original dataset [15]. The dataset comprises a total of 43,400 records, of which 783 correspond to patients with stroke and the others to non-stroke participants. Overall, the occurrence ratio of stroke was 1.8%.

The dataset contained three continuous variables and seven categorical variables. Continuous variables include age, body mass index (BMI), and average glucose level, while categorical variables include heart disease, hypertension, sex, ever-married status, smoking status, work type, and residence type (Table 1). The detailed lists of variables in the dataset are also described in a previous study [15].

Notably, the dataset lacked some information. Specifically, the smoking status of 13,292 cases (approximately 30%) and BMI of 1,462 cases (approximately 3%) were missing. In the development datasets, we dropped cases with missing BMI information. Additionally, we dropped seven cases that selected “other” in the sex question to remove ambiguous data.

Finally, we randomly divided the development datasets containing 41,931 cases with 643 occurrences of stroke into three subsets: 60% (N=25,158) were used as the training dataset, 20% (N=8,386) were used as the validation dataset, and 20% (N=8,387) were used as the test dataset. We then preprocessed the three datasets, fit the model to the training dataset and validated it with the validation dataset, and evaluated the model performance on the test dataset (Table 1).

Preprocessing

We transformed 10 variables using the weight of evidence (WoE) method. For each variable, the WoE is computed by calculating the logarithm of the ratio of the proportion of non-strokes over the proportion of strokes as follows:

$$WoE_i = \ln \left(\frac{NS_i / \sum NS_i}{S_i / \sum S_i} \right),$$

where NS_i is the number of non-strokes in bin i , $\sum NS_i$ is the total number of non-strokes, S_i is the number of strokes in bin i , and $\sum S_i$ is the total number of strokes.

WoE transformation converts categorical variables into numerical values and has a linear relationship with the logistic function. Thus, WoE-transformed variables are well-suited for input features in the logistic regression model. High positive WoE values

Table 1 Descriptive statistics of numerical variables and number of cases of categorical variables for the whole datasets

Variable	Mean	SD	Minimum	Maximum
Continuous variable				
Age (years)	41.84	22.48	0.08	82.00
Body mass index (kg/m ²)	28.61	7.77	10.10	97.60
Average glucose level (mg/dL)	103.63	42.23	55.00	291.05
Categorical variable				
Heart disease, No. (%)	Yes:No		1,808 (4.4):40,123 (95.6)	
Hypertension, No. (%)	Yes:No		3,670 (8.8):38,261 (91.2)	
Sex, No. (%)	Female:Male		24,945 (59.5):16,986 (40.5)	
Ever-married status, No. (%)	Yes:No		26,781 (63.9):15,150 (36.1)	
Smoking status, No. (%)			15,746 (37.5):7,093 (16.9):6,226 (14.9):12,866 (30.7)	
Work type, No. (%)			23,980 (57.2):6,474 (15.4):5,243 (12.5):176 (0.4):6058 (14.5)	
Residence type, No. (%)	Urban:Rural		21,001 (50.1):20,930 (49.9)	

Abbreviations: SD, standard deviation; NS, never smoked; FS, formerly smoked; SE, self-employed; GJ, government job; NW, never worked

indicate a low risk, whereas high negative WoE values indicate a high risk. WoE is widely used in models for risk management, such as credit risk models [16–18].

Logistic regression

We used a logistic regression model for the assessment of stroke occurrence because the non-strokes/strokes odds ratio in the logistic regression is easy to calculate and interpret, and logistic regression has been widely used in building prediction models for various diseases [19, 20]. We first performed univariate logistic regression to evaluate the significance of 10 transformed variables. Finally, we performed a multivariable logistic regression with input transformed variables at $P < 0.05$ in the univariate logistic regression. We used a backward selection approach to build the final model. The significance of each transformed variable in the logistic regression model was determined at a threshold of $P < 0.05$.

Model performance

The developed model was used to predict stroke occurrence from the validation and test datasets. To evaluate the model performance, we computed the area under the receiver operating curve (AUROC) and Kolmogorov–Smirnov (KS) statistics for the validation and test datasets. The AUROC measured the discriminatory power of a stroke prediction model that can be interpreted as the probability that strokes receive better scores than non-strokes [21]. The KS statistic computed the maximum differences between the cumulative distributions of two discriminations, such as strokes and non-strokes, where each discrimination score had a value between 0 and 1 [21]. We have also computed the accuracy of the stroke-scoring model. Accuracy represents the ratio of correct predictions made by the model on both the test and validation datasets, which includes the number of true positives and true negatives divided by the total number of predictions.

Scorecard model

Borrowing the concept of the credit scoring model, [16, 17] we developed a scorecard model for strokes. In the clinic, the use of a disease score rather than the probability of disease ranging from 0 to 1 is usually beneficial to determine a patient's health status. According to the scorecard model, the odds ratios used in the logistic regression could be converted into a disease score as follows:

$$P_0 = A + B \times \ln(\text{odds ratio}), \quad (1)$$

where A and B are constraints that need to be determined through specific disease scorecard model settings and P_0 indicates the user-defined baseline score. The point of double odds (POD) representing the score that doubles the odds was used to determine these constraints. The sum of POD and P_0 yields the double odds ratio as follows:

$$P_0 + \text{POD} = A + B \times \ln(2 \times \text{odds ratio}). \quad (2)$$

By solving Eqs. (1) and (2), we calculated the two constraints as follows:

$$\begin{cases} A = P_0 - B \times \ln(\text{odds ratio}) \\ B = \text{POD} / \ln(2). \end{cases}$$

A detailed description of the scoring method with POD has been provided previously [22].

Results

WoE Transformation

To illustrate the effect of WoE transformation for each variable, we plotted both WoE values and stroke occurrences as a function of variable bins. For each variable bin, we additionally computed stroke occurrences as a ratio of the number of strokes to the number of strokes and non-strokes. Trends between WoE values and stroke occurrences show an inverse relationship (Fig. 1). The WoE values of the continuous variables, such as age, BMI, and average glucose level, showed decreasing trends as the bin number increased. Additionally, the categorical variables, including heart disease, hypertension, ever-married status, smoking status, and work type, showed a good WoE distribution, with a difference between the maximum and minimum WoE values of more than 1. In contrast, sex and residence type showed similar WoE values within each category.

Stroke occurrence Model using logistic regression

We summarized the univariate and multivariable logistic regression analysis (Table 2). Except for residence type, all variables were significant in the univariate logistic regression analysis. However, in the multivariable logistic regression, only four variables, namely age, average glucose level, heart disease, and hypertension, were selected by a backward stepwise selection method. All the included variables, such as age, average glucose level, heart disease, and hypertension, were significant in predicting stroke occurrences ($P < 0.05$). Our multivariable logistic regression model showed that the AUROC and KS values of the validation (test) datasets were 85.3% (83.4%) and 57.6% (53.3%), respectively.

Stroke-scoring model

The distribution of stroke-scoring for the validation and test datasets shows that our model comprised four variables and scores ranging from 361 to 965 (Fig. 2; Table 3). Using the score cards (Fig. 3), we can easily evaluate an individual's stroke probability. As

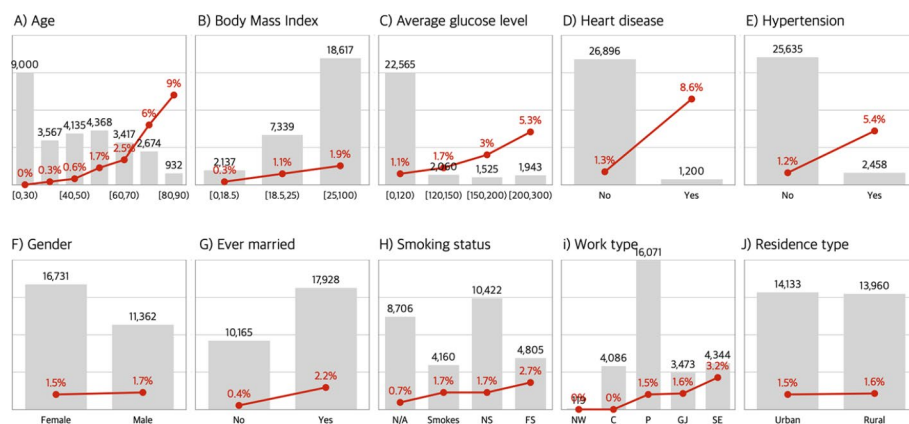


Fig. 1 WoE values, stroke risk, and the number of samples for each bin of 10 variables. The vertical bar shows the number of samples, while the line depicts the WoE value or the percentage of stroke incidence for each bin within the training dataset. Abbreviations: C, children; FS, formerly smoked; GJ, government job; N/A, not available; NW, never worked; NS, never smoked; P, private job; SE, self-employed; WoE, weight of evidence

Table 2 Logistic regression and fit results

Variable	Z value	P value	Coefficients	CI [0.025, 0.975]
Univariate logistic regression				
Age	-76.1	< 0.001	-0.73	[-0.75, -0.71]
Average glucose level	-41.5	< 0.001	-1.42	[-1.49, -1.36]
Body mass index	-25.7	< 0.001	-0.76	[-0.82, -0.70]
Ever-married status	-48.8	< 0.001	-0.77	[-0.80, -0.74]
Heart disease	-45.8	< 0.001	-2.99	[-3.12, -2.86]
Hypertension	-36.8	< 0.001	-1.52	[-1.60, -1.44]
Residence type	-1.5	0.137	-1.00	[-2.32, 0.32]
Sex	-1.0	0.305	-1.00	[-2.92, 0.91]
Smoking status	-31.1	< 0.001	-0.93	[-0.98, -0.87]
Work type	-43.0	< 0.001	-0.59	[-0.62, -0.57]
Multivariable logistic regression				
Constant	-59.2	< 0.001	-4.19	[-4.32, -4.05]
Age	-16.9	< 0.001	-0.94	[-1.04, -0.83]
Average glucose level	-4.3	< 0.001	-0.37	[-0.54, -0.20]
Heart disease	-4.8	< 0.001	-0.34	[-0.48, -0.20]
Hypertension	-2.7	0.007	-0.26	[-0.44, -0.07]

Abbreviation: CI, confidence interval

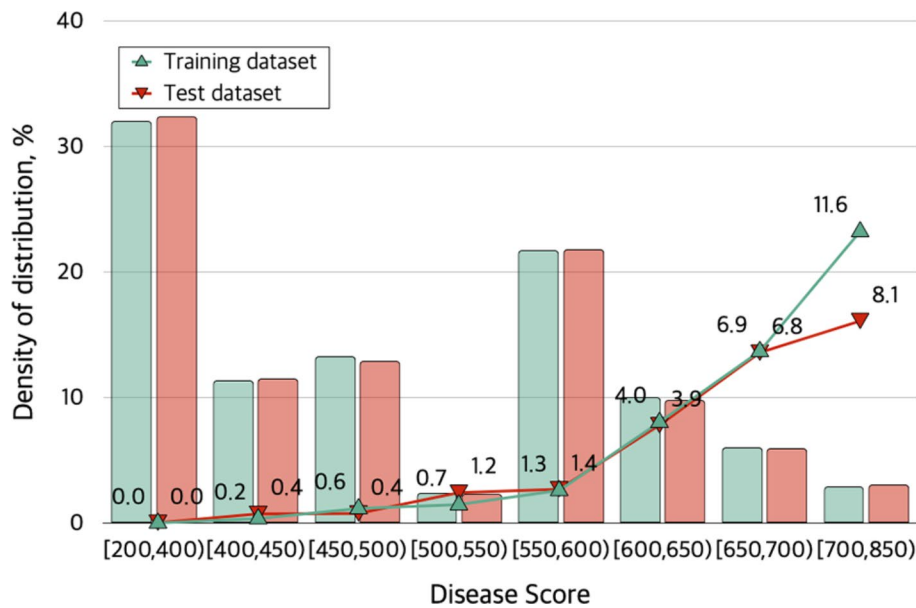


Fig. 2 Score distribution for the validation and test datasets and stroke occurrences for each score bin. The ticker values on the left y-axis represent the density of the score distribution. The numbers above the markers on the left (right) represent stroke risk for the validation (test) dataset

Table 3 Number of strokes and non-strokes in each bin of disease score

Score range	Validation dataset		Test dataset	
	Strokes	Non-Strokes	Strokes	Non-Strokes
0~499	0	2,731	1	2,685
500~599	8	2,092	15	2,127
600~699	10	1,086	10	1,048
700~799	55	1,635	38	1,672
800~899	46	647	54	657
900~100	9	67	10	70

Base points		Age		Heart Disease	
568		Intervals	Points	Intervals	Points
		0 - 29	-313	No	-6
		30 - 39	-109	Yes	50
Avg. Glucose Level		40 - 49	-63		
Intervals	Points	50 - 59	7	Hypertension	
0 - 119	-9	60 - 69	30	Intervals	Points
120 - 150	1	70 - 79	88	No	-8
150 - 200	18	80 - Inf	115	Yes	35
200 - Inf	34				

Fig. 3 Score cards for average glucose level, age, heart disease, and hypertension. Our stroke-scoring model allows to easily calculate an individual's stroke score using base points and four variables. The stroke risk is determined by the stroke score and score distribution in Fig. 2

an example, let's consider the case of a 63-year-old patient with heart disease and hypertension, and an average glucose level of 130. Based on the stroke-scoring model, the patient's score would be calculated as follows: Starting with a base score of 703 points, we add 42 points for their age, 50 points for having heart disease, 24 points for having hypertension, and 1 point for an average glucose level of 130, resulting in a total score of 820 points. According to Fig. 2, if the stroke score falls between 800 and 899 points, the probability of stroke is very high at 7.59%. Our stroke-scoring model was found to be 90.6% accurate for both the test and validation dataset when the disease score threshold was set at 800.

Discussion

This study developed a flexible and simple scoring model for stroke occurrences using a public dataset. The first step included the transformation of each variable using the WoE method. In this process, changes in stroke occurrences were observed according to the increase or decrease in continuous variables and the presence or absence of categorical variables. The next step was the selection of significant variables that affect stroke occurrences. Significant variables reflect risk factors of stroke. Finally, the scoring model was developed with the base point and scorecards using the significant variables.

Our model was based on the credit evaluation model of banks. The evaluation of credit is one of the most significant data for loans in the banking system [16, 23]. Banks use the credit evaluation model to predict the probability of default by loan applicants [16, 23]. In this study, loan applicants are cases in the dataset and default means stroke occurrences. The advantage of the credit evaluation model of banks is the flexibility of included variables. Because potential variables that can affect individual credit scoring have increased exponentially in recent years, it is necessary to determine whether each variable should be included in the credit evaluation model. Similar to that of the credit evaluation model, the flexibility of our model enables us to add new potential determinants for stroke occurrences in the process. Furthermore, our model is not limited to stroke; it can also be used for other disease models fitted to other datasets containing different variables.

Our model offers several advantages over previous models. First, it employs coefficients for individual variables, facilitating easy comprehension and interpretation of the

effects of each variable on an individual's stroke probability. In contrast, earlier models such as FSRS, CHADS2, ABCD, and ABCD2 have fixed variables, making it challenging to incorporate new variables beyond the fixed ones [9–13]. Second, the model uses a simple calculation method involving the addition of scores assigned to each variable. However, it should be noted that the model must be re-fitted to include new variables, which may result in changes to variable scores during the re-fitting process.

The simplicity of our scoring model allows individuals to easily self-calculate their probability of stroke. The WoE transformation is critical for the development of our scoring model [22]. The WoE values correspond to the stroke occurrences according to changes in each variable. The WoE values of significant variables are expressed as a monotonic relationship with stroke occurrences. The monotonic relationship easily enables each variable to transform our scoring model. The model provides a common base point and scorecards for variables included in the final model. The patients can intuitively understand that the management of included variables reduces the scores and stroke occurrences. This self-calculation of our model can motivate individuals to manage determinants to reduce stroke occurrences for primary prevention. In addition, the significant variables in our model are not essential for model development in other datasets. Each institution can easily replace other significant variables with their own medical data instead of unavailable significant variables in other datasets.

The selection of significant variables reveals risk factors of stroke. All variables except residence type were significant in the univariate regression; however, only well-known predisposing factors were selected as significant variables after multivariable logistic regression. The four significant variables in our model are consistent with previous studies on risk factors predisposing individuals to stroke. Hypertension is the single most important modifiable risk factor for both ischemic stroke and intracerebral hemorrhage [24–26]. Furthermore, the incidence of stroke increases with age [27], and increasing degrees of glucose intolerance lead to an increasing risk of thromboembolic stroke [28, 29]. The process of variable selection is used to find potential risk factors for specific diseases in medical datasets. However, we can estimate the limitations of our model based on the exclusion of smoking status, which is a well-known predisposing factor [30, 31]. The exclusion of smoking status may have two effects. First, stroke incidence shows a dose-dependent relationship with smoking with a more than twofold increase in heavy smokers compared with light smokers [32]. In addition, cessation of smoking decreases the risk of stroke in former smokers to the level of non-smokers after 5 years of abstinence [33]. Our dataset included smoking status (never smokers, former smokers, and current smokers), but the dose of smoking and the period of cessation of smoking were not included in the dataset. Second, the missing data corresponding to the smoking status in the dataset represented 30.6% (8,706/41,931) of included cases. The large percentage of missing data on smoking status may affect the exclusion of smoking status in our final model. Nonetheless, the selected variables in the scoring model are consistent with those reported in previous studies on risk factors of stroke [25, 28, 31]. Although the smoking status was excluded, the most well-known risk factors of stroke were included in our model. Therefore, the selection of variables in this study could be used as an example to identify new risk factors using various datasets and confirm risk factors in longitudinal studies.

This study had several limitations. First, the dataset contains limited information on stroke type and variables. Further information regarding diseases and variables could help improve the performance of the disease-scoring model. Second, the distribution of disease scores was not smooth. We could develop a smoother distribution of stroke scores if we had the numerical information of systolic and diastolic blood pressure in addition to a yes or no categorical variable for hypertension.

Conclusions

The disease-scoring model, based on the scorecard concept, is flexible and simple, making it suitable for individual use or adoption by government agencies responsible for public health. Although the current model is based on a limited number of variables extracted from publicly available datasets, it has the potential to expand into a model that can better reflect the risk factors for stroke by utilizing large-scale data collection at institutional and national levels. Even with the addition of more variables, the results will be presented in a scorecard format that can be easily evaluated by the public, making it highly applicable for individual use. If the stroke score is high due to factors such as high body weight or glucose levels, patients may be motivated to manage their health by adjusting their weight or managing their diet to reduce the risk factors of stroke. In conclusion, this scorecard model offers several advantages for identifying and managing risk factors, thereby promoting primary prevention of stroke.

Acknowledgements

None.

Authors' contributions

S.K. contributed to design the study concept and analysis. D.H.K. contributed to preparation of the manuscript. All authors read and approved the final manuscript.

Funding

This study was supported by a Chyung Ki Lee research grant of Korean Academy of Rehabilitation Medicine for 2023.

Data availability

The data used the Kaggle competition for stroke prediction (<https://www.kaggle.com/asaumya/healthcare-dataset-stroke-data>. Accessed December 1, 2018) and are available from the corresponding author upon reasonable request.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

None.

Received: 3 November 2022 / Accepted: 17 May 2023

Published online: 25 May 2023

References

1. Global regional. National burden of stroke and its risk factors, 1990–2019: a systematic analysis for the global burden of Disease Study 2019. *Lancet Neurol.* 2021;20:795–820.
2. Krishnamurthi RV, Ikeda T, Feigin VL. Global, Regional and Country-Specific Burden of Ischaemic Stroke, Intracerebral Haemorrhage and Subarachnoid Haemorrhage: a systematic analysis of the global burden of Disease Study 2017. *Neuro-epidemiology.* 2020;54:171–9.
3. Tervonen T, Prawitz T, Chua GN, Hyacinthe J, Pinto CA. Net clinical benefit of antiplatelet therapy was affected by patient preferences: a personalized benefit-risk assessment. *J Clin Epidemiol.* 2022;144:84–92.
4. Lansberg MG, Bluhmki E, Thijs VN. Efficacy and safety of tissue plasminogen activator 3 to 4.5 hours after acute ischemic stroke: a metaanalysis. *Stroke.* 2009;40:2438–41.

5. Saver JL, Jahan R, Levy EJ, Jovin TG, Baxter B, Nogueira RG, et al. Solitaire flow restoration device versus the Merci Retriever in patients with acute ischaemic stroke (SWIFT): a randomised, parallel-group, non-inferiority trial. *Lancet*. 2012;380:1241–9.
6. Lohse KR, Lang CE, Boyd LA. Is more better? Using metadata to explore dose-response relationships in stroke rehabilitation. *Stroke*. 2014;45:2053–8.
7. Vemuri P, Decarli C, Duering M. Imaging markers of vascular Brain Health: quantification, clinical implications, and future directions. *Stroke*. 2022;53:416–26.
8. O'Donnell MJ, Chin SL, Rangarajan S, Xavier D, Liu L, Zhang H, et al. Global and regional effects of potentially modifiable risk factors associated with acute stroke in 32 countries (INTERSTROKE): a case-control study. *Lancet*. 2016;388:761–75.
9. Fothergill A, Christianson TJ, Brown RD Jr, Rabinstein AA. Validation and refinement of the ABCD2 score: a population-based analysis. *Stroke*. 2009;40:2669–73.
10. Gage BF, Waterman AD, Shannon W, Boehler M, Rich MW, Radford MJ. Validation of clinical classification schemes for predicting stroke: results from the National Registry of Atrial Fibrillation. *JAMA*. 2001;285:2864–70.
11. Lip GY, Nieuwlaar R, Pisters R, Lane DA, Crijns HJ. Refining clinical risk stratification for predicting stroke and thromboembolism in atrial fibrillation using a novel risk factor-based approach: the euro heart survey on atrial fibrillation. *Chest*. 2010;137:263–72.
12. Rothwell PM, Giles MF, Flossmann E, Lovelock CE, Redgrave JN, Warlow CP, et al. A simple score (ABCD) to identify individuals at high early risk of stroke after transient ischaemic attack. *Lancet*. 2005;366:29–36.
13. Wolf PA, D'Agostino RB, Belanger AJ, Kannel WB. Probability of stroke: a risk profile from the Framingham Study. *Stroke*. 1991;22:312–8.
14. Flueckiger P, Longstreth W, Herrington D, Yeboah J. Revised Framingham Stroke Risk score, nontraditional risk markers, and Incident Stroke in a multiethnic cohort. *Stroke*. 2018;49:363–9.
15. Liu T, Fan W, Wu C. A hybrid machine learning approach to cerebral stroke prediction based on imbalanced medical dataset. *Artif Intell Med*. 2019;101:101723.
16. Chen K, Zhu K, Meng Y, Yadav A, Khan A. Mixed credit Scoring Model of Logistic regression and evidence weight in the background of Big Data. In: Abraham A, Cherukuri AK, Melin P, Gandhi N, editors. *Intelligent Systems Design and Applications*. Cham: Springer International Publishing; 2020. pp. 435–43.
17. Zeng G. A necessary Condition for a good Binning Algorithm in Credit Scoring. *Appl Math Sci*. 2014;8:3229–42.
18. Kyeong S, Kim D, Shin J. Can System Log Data enhance the performance of Credit Scoring? — evidence from an Internet Bank in Korea. *Sustainability*. 2022;14:130.
19. Kyeong S, Shin JE, Yang KH, Lee WS, Chung T-S, Kim J-J. Neural predisposing factors of postoperative delirium in elderly patients with femoral neck fracture. *Sci Rep*. 2018;8:7602.
20. Shigemizu D, Akiyama S, Asanomi Y, Borojevich KA, Sharma A, Tsunoda T, et al. Risk prediction models for dementia constructed by supervised principal component analysis using miRNA expression data. *Commun Biology*. 2019;2:77.
21. Chi B-W, Hsu C-C. A hybrid approach to integrate genetic algorithm into dual scoring model in enhancing the performance of credit scoring model. *Expert Syst Appl*. 2012;39:2650–61.
22. Wang C, Fang K, Zheng C, Xu H, Li Z. Credit scoring of micro and small entrepreneurial firms in China. *Int Entrepreneurship Manage J*. 2021;17:29–43.
23. Djeundje VB, Crook J. Incorporating heterogeneity and macroeconomic variables into multi-state delinquency models for credit cards. *Eur J Oper Res*. 2018;271:697–709.
24. Mohan KM, Wolfe CD, Rudd AG, Heuschmann PU, Kolominsky-Rabas PL, Grieve AP. Risk and cumulative risk of stroke recurrence: a systematic review and meta-analysis. *Stroke*. 2011;42:1489–94.
25. Lewington S, Clarke R, Qizilbash N, Peto R, Collins R. Age-specific relevance of usual blood pressure to vascular mortality: a meta-analysis of individual data for one million adults in 61 prospective studies. *Lancet*. 2002;360:1903–13.
26. Beckett NS, Peters R, Fletcher AE, Staessen JA, Liu L, Dumitrascu D, et al. Treatment of hypertension in patients 80 years of age or older. *N Engl J Med*. 2008;358:1887–98.
27. Carandang R, Seshadri S, Beiser A, Kelly-Hayes M, Kase CS, Kannel WB, et al. Trends in incidence, lifetime risk, severity, and 30-day mortality of stroke over the past 50 years. *JAMA*. 2006;296:2939–46.
28. Burchfiel CM, Curb JD, Rodriguez BL, Abbott RD, Chiu D, Yano K. Glucose intolerance and 22-year stroke incidence. The Honolulu Heart Program Stroke. 1994;25:951–7.
29. Najarian RM, Sullivan LM, Kannel WB, Wilson PW, D'Agostino RB, Wolf PA. Metabolic syndrome compared with type 2 diabetes mellitus as a risk factor for stroke: the Framingham offspring study. *Arch Intern Med*. 2006;166:106–11.
30. Etminan N, Chang HS, Hackenberg K, de Rooij NK, Vergouwen MDI, Rinkel GJE, et al. Worldwide Incidence of Aneurysmal Subarachnoid Hemorrhage according to Region, Time Period, blood pressure, and Smoking Prevalence in the Population: a systematic review and Meta-analysis. *JAMA Neurol*. 2019;76:588–97.
31. Shinton R, Beevers G. Meta-analysis of relation between cigarette smoking and stroke. *BMJ*. 1989;298:789–94.
32. Kawachi I, Colditz GA, Stampfer MJ, Willett WC, Manson JE, Rosner B, et al. Smoking cessation and decreased risk of stroke in women. *JAMA*. 1993;269:232–6.
33. Meschia JF, Bushnell C, Boden-Albala B, Braun LT, Bravata DM, Chaturvedi S, et al. Guidelines for the primary prevention of stroke: a statement for healthcare professionals from the American Heart Association/American Stroke Association. *Stroke*. 2014;45:3754–832.

Publisher's Note

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