MEDS Clinical Medicine (2024)
Clausius Scientific Press, Canada

DOI: 10.23977/medsc.2024.050206
ISSN 2616-1907 Vol. 5 Num. 2

Research on the Influencing Factors of Brain Stroke Based on Binary Logistic Regression and Random Forest

Qiuyue Lin*

School of Science, Jimei University, Xiamen, 361000, China
**Corresponding author: 202321392022@jmu.edu.cn*

Keywords: Brain stroke, logistic regression, random forest

Abstract: Stroke is characterized by high incidence, recurrence, disability, mortality and economic burden, which is also a serious threat to people's lives and health and quality of life. This study analyzes data from the Kaggle website to develop a risk assessment model to explore the risk factors affecting the occurrence of stroke. The dataset includes 4981 cases and 10 variables such as gender, age and so on. The target variable is whether the respondent has had a brain stroke or not. In this study, at first, the accuracy of the training set is about 100%, and then the fitted model is tested, and the accuracy rate is still in a good state at 99%. Therefore, the assessment results of this random forest model are acceptable. To better assess the risk factors that trigger stroke, more comprehensive data and better sampling methods are needed. This model provides an important ranking of risk factors for stroke occurrence and provides a reference for stroke prevention.

1. Introduction

Acute cerebrovascular disease, which includes hemorrhagic and ischemic strokes, is often referred to as "apoplexy" or stroke. These conditions cause damage to brain tissue when a blood vessel in the brain suddenly bursts or becomes blocked, preventing blood from reaching the brain [1]. Stroke is second globally in terms of cause of death and third globally in terms of disability-combined death, according to the 2019 Global Burden of Disease (GBD) research [2]. In addition to high lethality, stroke is also characterized by high rates of disability, with 15-30% of post-stroke patients having severe disability [3]. Patients' ability to live on their own is significantly reduced, which makes them unable to take on appropriate social responsibilities and increases the difficulty of caregiving for their families and the financial burden of maintaining treatment. Every year, China is burdened with up to about 40 billion yuan of economic losses caused by stroke, and this economic burden continues to grow linearly [4]. Stroke causes great harm to human health and economic society. Therefore, it is necessary to construct a preventive approach based on risk factors to predict stroke.

Strokes can be caused by many reasons: genetics, family history, fundus markers, thyroid hormone markers, etc. Among them, the phenomenon of causative genes causing stroke has received much attention recently. Based on the study that homocysteinemia (HHcy) may be related to the occurrence of stroke, Zhou et al. tested 110 normal participants and 80 stroke patients using polymerase chain reaction-restriction endonuclease fragment length polymorphism (PCR-RFLP),

and found that Compared to the normal group, stroke patients had a considerably greater prevalence of homozygous and heterozygous mutations in the MTHF gene [5]. It was suggested that MTHFR gene mutation and HHcy were related to the occurrence of stroke and that the MTHFR gene might be one of the susceptibility genes for stroke, and HHcy was separate risk factor for the occurrence of stroke. The development of stroke was independently associated with HHcy. However, this study did not rank the effect of blood homocysteine (Hcy) levels in blood corresponding to each of the three mutated genes. Based on Zhou et al.'s study, Zhang et al. further investigated the connection between the polypeptide nature of the MTHFR gene and plasma homocysteine levels and the incidence of stroke in the elderly and performed logistic analyses of the variables, which improved Zhou et al.'s experiments, study's findings demonstrated that the importance of the effects of the three genotypes on the plasma levels of Hcy was ranked as TT > CT type > CT type > CC type [6]. This indicated that the C667T mutation in the MTHFR gene was closely related to the increased concentration of Hcy in the blood. With the trend of younger strokes, the problem of stroke in young people has gained attention.

Wang et al. found that elevated Hcy levels were associated with the onset of cerebral infarction in young and middle-aged adults after adjusting for traditional risk factors [7]. This suggested that Hcy is an important independent risk factor for the development of cerebral infarction in young and middle-aged Chinese and that the mutation in the C677T locus of the MTHFR gene may be an important genetic factor for HHcy, which may interact with other risk factors. With the advancement of genetic screening technology, more and more disease-causing genes have been identified. Liu et al. found that the risk of acute coronary heart disease and ischaemic stroke was significantly higher in people carrying the ApoE4 allele after genotyping for ApoE [8]. Subsequently, Liu et al. further examined the role of the ApoE gene in acute stroke using magnetic resonance imaging and further concluded that ApoE4 has both harmful and protective effects on people who have already had a stroke [9]. The study of stroke genetics allows the identification of single-gene causative factors that influence the occurrence of stroke and facilitates early intervention in stroke by determining the risk genes in the family [10].

However, there are certain problems with genetic screening. Stroke is a multifactorial outcome of genes, lifestyle, and the environment, and unless there are strong clinical clues, such as the clinical features of Fabry disease, it is not advisable to perform a large number of genetic tests [11]. In addition, genetic screening has a high cost. With limited accuracy, it becomes less cost-effective. Therefore, it is important to be able to determine the influencing factors of the disease through routine indicators. This paper aims to explore the factors influencing the risk of stroke occurrence using common indicators and to rank the importance of the risk. 4982 cases of clinical variables were obtained from the Kaggle data site and included 11 independent predictor variables such as gender, age, and mean blood glucose level. This paper now uses logistic screening to identify the factors of interest and then uses random forests to model these factors.

2. Methods

2.1. Data Source

The data for this study is collected from the Kaggle website, which was compiled by Jillani Soft Tech from the UIC open-source data warehouse and published and updated in 2022 for 4981 individuals. The dataset offers crucial insights into the correlation between risk factors and the incidence of brain stroke, which categorizes individuals depended on the presence or absence of brain stroke. The original dataset remained in csv format.

2.2. Variable Selection

The data used in this paper was totalled 7942 cases, including 11 variables. An algorithm was created using the available characteristics (including gender, age, hypertension, heart disease, marriage, type of work, type of residence, mean blood glucose level, body mass index, and smoking status) to categorize stroke patients and non-stroke patients. Table 1 lists all 11 variables.

Table 1: Different types of variables.

Term	Type	Range
Age	Numeric	18 to 82
Gender	Categorical	0-Female, 1-male
Hypertension	Categorical	0-No, 1-Yes
Heart disease	Categorical	0-No, 1-Yes
Ever_married	Categorical	0-No, 1-Yes
Work type	Categorical	0-Private, 1-Govt_job, 2-Self-employed
Residence type	Categorical	0-Rural, 1-Urban
Avg_glucose level	Numeric	55.12 to 271.74 mg/dl
BMI	Numeric	14.1 to 48.9 kg/m2
Smoking status	Categorical	0-Never smoked, 1- smoked, 2-smokes
Stroke	Categorical	0-No, 1-Yes

2.3. Method Introduction

In this study, Pearson's correlation analysis as well as covariance diagnosis were first applied successively to the 11 variables to exclude multicollinearity among the variables. Then, binary logistic regression analyses were performed on the variables that did not have multicollinearity to de-rive the influencing factors related to stroke onset. Finally, a random forest model was constructed based on the results of binary logistic regression.

2.4. Data Preprocessing

The original dataset was hardly preprocessed and there were 1500 missing values in the column of smoking status. According to the missing values, the article utilized the “missForest” package in R software to fill in the missing values using the random forest method. Meanwhile, since the age of identification of young stroke is 18-45 years old, and there are 823 samples whose actual age is less than 18 years old in the original dataset, this article chooses to exclude these samples. In addition, the dataset has the problem of data imbalance, for which this paper adopts the method of over-sampling to process the data. Finally, the data used in this paper totalled 7942 cases, including 11 variables.

3. Results and Discussion

3.1. Descriptive Analysis

The following box-plot divides the variables on the x-axis by whether the respondent has a stroke or not. Figure 1 is the box-plot of the age of the respondent. 'Stroke' produces a box-plot with a median of 71, while 'No stroke' produces a box-plot with a median of 49. Therefore, the box-plot generated by 'Stroke' is in the upper position, which indicates that the overall age of stroke patients is higher. Furthermore, the box plot for 'Stroke' has an IQR of 19 while the box plot for 'No stroke'

has an IQR of 27, which suggests that stroke patients were concentrated in the 1970s.

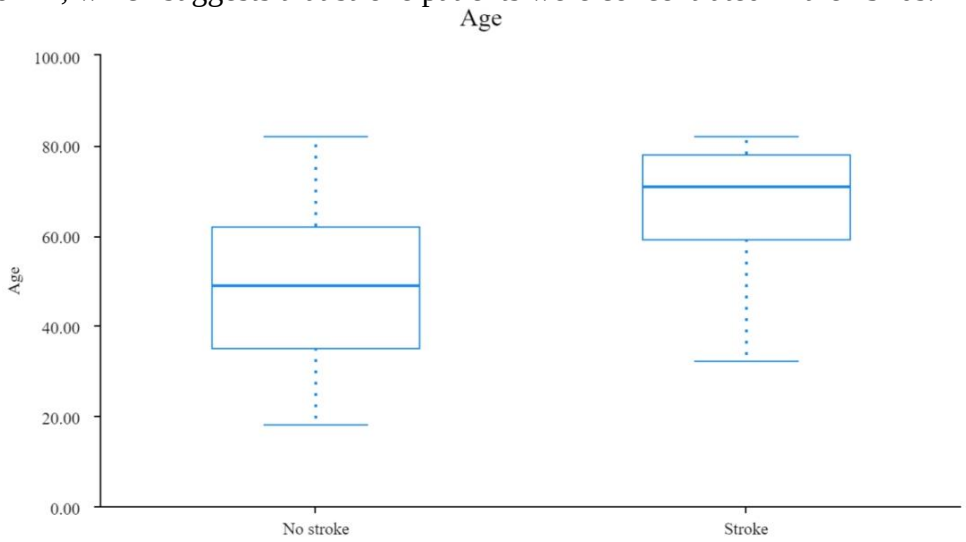

Figure 1: The box-plot of the age.

Figure 2 is the box plot of the respondent's BMI. The "Stroke" and "No stroke" generate 215 and 42 large outliers respectively, indicating that stroke patients are more likely to be overweight. Furthermore, the median of the box plot generated by "Stroke" is 29.6, and other medians of the box plot generated by "No stroke" is 29.2, and the IQR of the box plot generated by "Stroke" is 5.9, and the IQR of the box plot generated by "No stroke" is 8, indicating that stroke patients are more concentrated in the obese population.

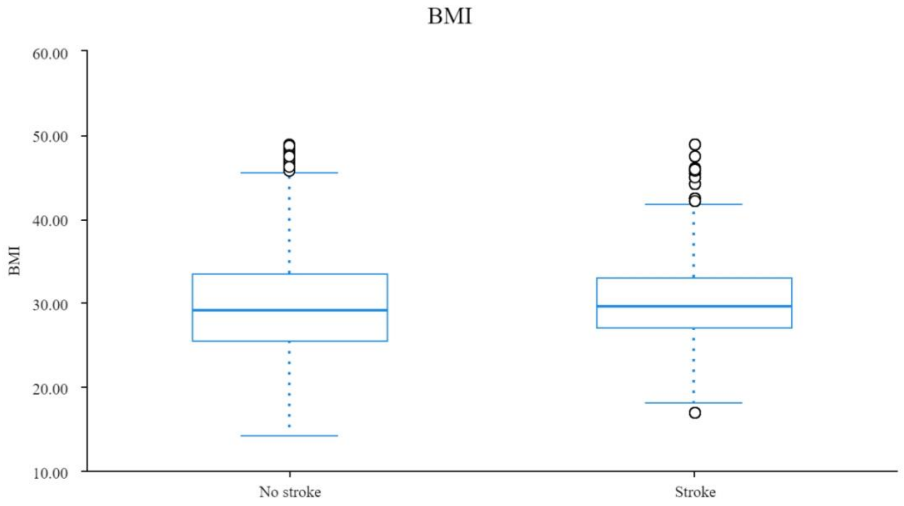

Figure 2: The box-plot of the BMI.

Figure 3 is the box plot of the respondent's Avg_glucose_level. The "No stroke" produces 502 outliers, which may be due to the sample size being too large. The median of the box plot generated by "Stroke" is 105.920 and the other medians of the box plot generated by "No stroke" are 91.885. Therefore, the box plot generated by "Stroke" is in the higher position, indicating that stroke patients often suffer from high blood sugar.

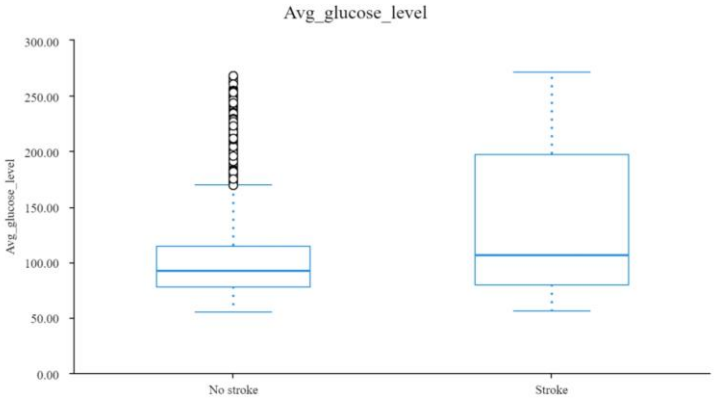

Figure 3: The box-plot of the Avg_glucose_level.

3.2. Random Forest Modelling

To exclude the multicollinearity between independent variables, this paper utilizes SPSS to perform Pearson correlation analysis, and the independent variables with more significant correlation coefficients are screened out by Pearson correlation coefficient. The covariance and standard quotient between two variables: X, Y are defined as Pearson's correlation coefficient between two variables. The overall correlation coefficient is given in the equation ρ. which is expressed as:

$$\rho_{X,Y} = \frac{cov(X,Y)}{\sigma_X \sigma_Y} = \frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y} \tag{1}$$

Estimating the covariance and standard deviation of the overall sample gives the Pearson's correlation coefficient, which is expressed as:

$$r = \frac{\sum_{i=1}^{n}(X_i-X_m)(Y_i-Y_m)}{\sqrt{\sum_{i=1}^{n}(X_i-X_m)^2 \cdot \sum_{i=1}^{n}(Y_i-Y_m)^2}} \tag{2}$$

Figure 4 shows the heat map of Pearson's correlation coefficient using SPSS correlation analysis results and Origin, in which the correlation coefficient between Age and Ever_married is 0.326, and the two have a low degree of correlation.

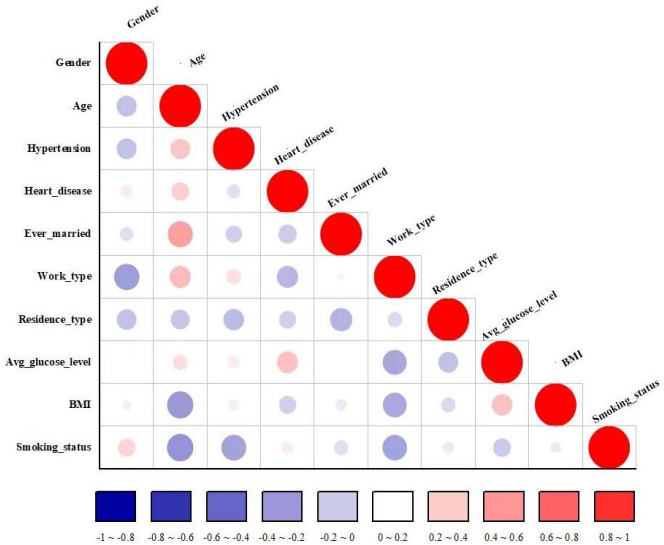

Figure 4: The heat map of Pearson's correlation.

To further exclude multicollinearity, this paper utilizes SPSS for covariance diagnosis. The analysis results show that none of the 10 independent variables have multicollinearity. (VIF) As shown in Table 2.

Table 2: Diagnosis of covariance of variables.

Variable	B	S.E.	Beta	t	P	VIF
Gender	-0.021	0.010	-0.021	-2.191	0.029	1.063
Age	0.015	0.000	0.520	47.467	<0.001	1.408
Hypertension	0.094	0.012	0.073	7.536	<0.001	1.104
Heart_disease	0.052	0.015	0.035	3.453	<0.001	1.184
Ever_married	-0.048	0.014	-0.036	-3.566	<0.001	1.166
Work_type	-0.025	0.006	-0.042	-4.325	<0.001	1.088
Residence_type	0.030	0.009	0.030	3.264	0.001	1.011
Avg_glucose_level	0.001	0.000	0.086	8.409	<0.001	1.221
Bmi	0.002	0.001	0.018	1.840	0.066	1.101
Smoking_status	0.051	0.006	0.080	8.315	<0.001	1.072

The presence of stroke (0-No, 1-Yes) is used as the dependent variable, and independent variables that don't have multicollinearity are included in the final regression equation. The results indicate that the variables gender, age, hypertension, heart disease, work_type, residence_type, avg_glucose, BMI and smoking_status are all statistically significantly associated with the occurrence of stroke. As shown in Table 3.

Table 3: Multivariate Logistic regression analysis of the occurrence of the stroke.

Variable	B	S.E.	Wals	P	Exp(B)	95%CI for EXP(B)	
						inf	sup
Gender	-0.137	0.057	5.678	0.017	0.872	0.780	0.976
Age	0.081	0.002	1335.227	<0.001	1.085	1.080	1.090
Hypertension	0.526	0.073	52.149	<0.001	1.692	1.467	1.952
Heart_disease	0.254	0.090	7.978	0.005	1.289	1.081	1.537
Ever_married	-0.049	0.091	0.293	0.589	0.952	0.797	1.137
Work_type	-0.143	0.034	17.985	<0.001	0.867	0.811	0.926
Residence_type	0.197	0.055	12.708	<0.001	1.218	1.093	1.357
Avg_glucose_level	0.004	0.001	52.318	<0.001	1.004	1.003	1.005
BMI	0.016	0.005	9.858	0.002	1.016	1.006	1.026
Smoking_status	0.347	0.038	85.224	<0.001	1.414	1.314	1.522

The items Gender, Age, Hypertension, Heart_disease, Work_type, Residence_type, Avg_glucose_level, BMI, Smoking_status are used as independent variables while Stroke is used as the dependent variable, the training set scale is set to 0.8, the number of decision trees is 100, the node splitting criterion is gini, and the maximum depth of the tree is not restricted for random forest modeling. The variables are categorized into two states "Stroke" and "No stroke". As shown in Table 4, the accuracy of the model for this training set is 100%.

The data from the test set is then tested. In Table 5, the accuracy is still in good condition. This shows that the model has better prediction results. The accuracy of the final model is 98.8%, the recall is 98.83%, and the f1-score on the test set is 0.99. The results of the model are acceptable.

Without screening the 11 variables with binary logistic regression, the recall of "No stroke" in the test set is 0.97, which is lower than the recall obtained after performing binary logistic regression of 0.98. The precision of "Stroke" in the test set is 0.97, which is lower than the precision

obtained after performing binary logistic regression of 0.98. Therefore, screening with binary logistic regression is more beneficial to the precision of the random forest model. As shown in Table 6.

Table 4: Training set of the random forest model evaluation results.

	Precision	Recall	f1-score	Samples
No stroke	1	1	1	3133
Stroke	1	1	1	3220
Accuracy	-	-	1	6353
Average	1	1	1	6353
Average(comprehensive)	1	1	1	6353

Table 5: Testing set of random forest model evaluation results.

	Precision	Recall	f1-score	Samples
No stroke	1.00	0.98	0.99	779.00
Stroke	0.98	1.00	0.99	810.00
Accuracy	-	-	0.99	1589.00
Average	0.99	0.99	0.99	1589.00
Average(comprehensive)	0.99	0.99	0.99	1589.00

Table 6: Testing set of random forest model evaluation results (without binary logistic regression).

	Precision	Recall	f1-score	Samples
No stroke	1	0.97	0.99	779
Stroke	0.97	1	0.99	810
Accuracy	-	-	0.99	1589
Average	0.99	0.99	0.99	1589
Average(comprehensive)	0.99	0.99	0.99	1589

3.3. Discussion

The figure 5 is the random forest plot of stroke occurrence. The results of the study indicate that the risk factors affecting the occurrence of stroke are age, avg_glucose_level, smoking_status, work_type, hypertension, residence_type, gender, and heart_disease in that order.

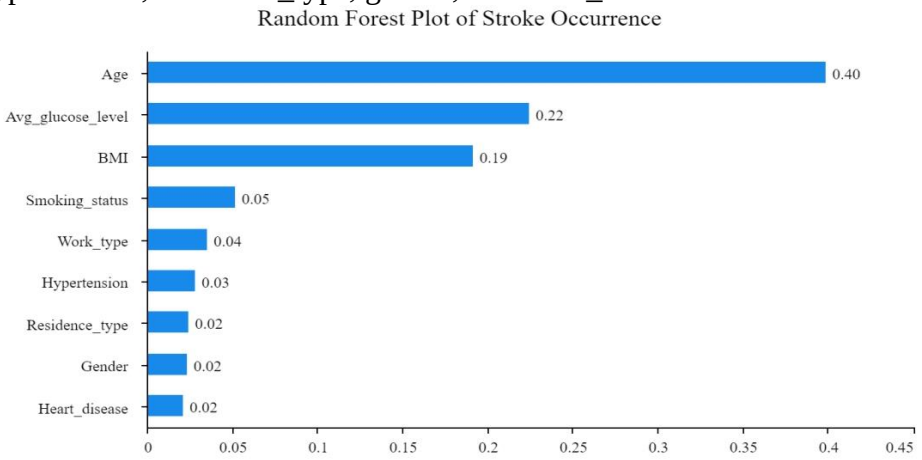

Figure 5: The random forest plot of stroke occurrence.

4. Conclusion

In this study, the variables: age, avg_glucose_level, smoking_status, work_type, hypertension, residence_type, gender and heart_disease are selected from 10 possible variables by multivariate logistic regression analysis as influencing factors for constructing a random forest model. The accuracy of both test and training sets of random forest is around 99%. The random forest model can present the degree of influence of each factor on the occurrence of stroke in the form of feature weights, which can provide a reference basis for the early detection of susceptible populations and the implementation of targeted stroke prevention. This study finds that the factors affecting the risk of stroke are age, blood glucose level, and BMI. Therefore, people with high age, high blood glucose, and obesity should pay more attention to the prevention of stroke and carry out regular medical checkups according to these three indexes to avoid the occurrence of stroke.

The stroke prediction dataset selected in this paper is characterized by sample imbalance. There are only about 4.5% of diseased samples in the dataset, and the number of diseased and non-diseased people to improve the prediction accuracy of the model. This study adopts the method of oversampling to process the dataset. However, oversampling may allow the noisy samples to be replicated multiple times, leading to model overfitting. In the future, studies need to collect more diseased clinical data. Besides, the study should apply stratified sampling to the negative samples of the data using a combination of integrated learning and under-sampled to better balance the number of samples to prevent overfitting and improve the accuracy of model predictions.

References

[1] Yang, Z. M. (2020) Progress of respiratory training in the rehabilitation of patients with stroke combined with pulmonary infection. Infectious Diseases International (electronic version), 9(02), 246-248.

[2] Tang, C.H., et al. (2023) Interpretation of the Global Stroke Data Report 2022. Diagnostics Theory and Practice, 22(03), 238-246.

[3] Zhao Y. L. (2022) A mixed study on the current status and influencing factors of stroke patients' quality of life based on ICF framework, Qingdao University.

[4] Huang, Y. (2024) Clinical Characteristics Analysis of Type H Hypertension and Ischaemic Cerebrovascular Disease, Jilin University.

[5] Zhou, X.L., et al. (1999) Relationship between MTHFR gene polymorphisms and plasma homocysteine levels and stroke. Chinese Journal of Cardiovascular Disease, 40-42.

[6] Zhang, C.L., Ye, J., Hui, R.T. (2004) Association of MTHFR gene polymorphisms and plasma homocysteine levels with the development of stroke in the elderly. Chinese Journal of Molecular Cardiology, 3-7.

[7] Wang, L.J., et al. (2006) Association of MTHFR gene polymorphisms and plasma homocysteine levels with cerebral infarction in young and middle-aged people. Journal of Stroke and Neurological Diseases, 4, 478-481.

[8] Liu, Q., Chen, L.J., Tan, X.R. (2009) A prospective study of ApoE genotype and cardiovascular disease. Chinese Journal of Molecular Cardiology, 9(03), 139-143.

[9] Liu, Y.W., Hartikainen, P.I., Vanninen. R.L. (2010) Effect of the ApoE gene on acute ischaemic stroke: a magnetic resonance imaging study. International Journal of Medical Radiology, 33(05), 456-458.

[10] Lan, T, Huzhiletmuer. (2015) Current status of stroke epidemiology and progress in genetics research. Journal of Difficult Diseases, 14(09), 986-989.

[11] Xu, W.H., Li, S., Cheng, A.Q. (2022) Diagnosis and treatment of stroke in young people: don't overdo it. Chinese Journal of Stroke, 17(09), 915-918.