Iran J Public Health, Vol. 51, No.5, May 2022, pp.999-1009

Original Article

The Prediction Models for High-Risk Population of Stroke Based on Logistic Regressive Analysis and Lightgbm Algorithm Separately

Yicheng Xue [1], Silong Chen [1], Mengmeng Zhang [1], Xiaojuan Cai [1], Jialian Zheng [2], Shihua Wang [3], *Yan Chen [1]

1. *The Medical School of Jiaxing University, Jiahang Road, Jiaxing, China*
2. *The Medical Examination Center of the First Hospital of Jiaxing, Jiaxing, China*
3. *The Medical Examination Center of the Second Hospital of Jiaxing, Jiaxing, China*

***Corresponding Author:** Email: ychen88@sina.com

(Received 09 Feb 2021; accepted 15 Apr 2021)

Abstract

Background: We aimed to investigate the high-risk factors of stroke through logistic regressive analysis and using LightGBM algorithm separately. The results of the two models were compared for instructing the prevention of stroke.

Methods: Samples of residents older than 40 years of age were collected from two medical examination centers in Jiaxing, China from 2018 to 2019. Among the total 2124 subjects, 1059 subjects were middle-aged people (40-59 years old) and 1065 subjects were elder-aged people (≥60 years old). Their demographic characteristics, medical history, family history, eating habits etc. were recorded and separately input into logistic regressive analysis and LightGBM algorithm to build the prediction models of high-risk population of stroke. Four values including F1 score, accuracy, recall rate and AUROC were compared between the two models.

Results: The risk factors of stroke were positively correlated with age, while negatively correlated with the frequency of fruit consumption and taste preference. People with low-salt diet were associated with less risk of stroke than those with high-salt diet, and male had higher stroke risk than female. Meanwhile, the risk factors were positively correlated with the frequency of alcohol consumption in the middle-aged group, and negatively correlated with the education level in the elder-aged group. Furthermore, the four values from LightGBM were higher than those from logistic regression, except for the recall value of the middle-aged group.

Conclusion: Age, gender, family history of hypertension and diabetes, the frequency of fruit consumption, alcohol and dairy products, taste preference, and education level could as the risk predictive factors of stroke. The Model of using LightGBM algorithm is more accurate than that using logistic regressive analysis.

Keywords: Stroke; Logistic regressive analysis; LightGBM algorithm; Prediction

Introduction

Brain is one of the most important organs in human body. The irreversible brain damage and acute occurrence of stroke are associated with extremely high disability rate and death rate (1).

At present, China has an increasing incidence of stroke over years and an ever-growing population of stroke showing a trend of occurrence in younger age (2). Therefore, understanding the risk of stroke and performing primary prevention are critically important. The collected data in this study can be used as a reference to reduce the risk of stroke through facilitating changing into lower-risk lifestyle and improving the primary and emergency care (3). The risk factors in high-risk population of stroke can be divided into unpreventable factors and preventable factors. The former includes race, age, gender, and family history; the latter includes diabetes, hypertension, heart disease, hyperlipidemia, obesity, and smoking (4, 5).

Logistic regressive analysis, a linear regressive analysis model, is often applied in data analysis, automatic disease diagnosis, economical prediction and other fields. It can provide us with the importance of independent variables to roughly predict the risk factors of a disease. The LightGBM algorithm is a novel model based on a decision tree algorithm, which is more powerful and efficient than conventional algorithms.

We aimed to establish and evaluate these two mathematical models in predicting the high-risk population of stroke by analyzing demographic characteristics, medical history, and family history, lifestyle, behavior habits, etc. The objective of this study was to reduce and delay the occurrence of stroke by early identifying the high risk factors of stroke and intervening related preventable risk factors.

Materials and Methods

Research Subjects

The research subjects included residents aged 40 or elder who had undergone a medical examination at the Physical Examination Center of the First Hospital or the Second Hospital of Jiaxing, China from 2018 to 2019. The patients with severe organ dysfunction, malignant tumors, mental illness, or cognitive dysfunction were excluded. The included subjects were divided into two groups as middle-aged group (40-59 years old) and elder-aged group (60 years old and elder). Data were collected from a total of 2124 subjects including 1059 from the middle-aged group and 1065 from the elder-aged group.

The demographic characteristics, medical history, family history, lifestyle, dietary habits, physical examination, and laboratory analyses of the subjects are shown in Table 1.

Table 1: Summary of various types of variables

Variable category	*Variables*
Demographic characteristics	gender, age, marital status, current or pre-retirement occupation, education level
Medical history	history of hypertension, atrial fibrillation or heart valve disease, diabetes, previous stroke or TIA
Family history	family history of stroke, family history of coronary heart disease, family history of hypertension, family history of diabetes
Lifestyle	smoking, excessive drinking, exercise, wake-up time, bedtime, meal time
Eating habits	white meat, red meat, dairy products, beans and soy products, vegetables, fruits, nuts, preference for spicy, salty, sour and mild food
Physical examination	BMI, blood pressure
Laboratory examination	fasting blood glucose, triglycerides, total cholesterol, low-density lipoprotein cholesterol, high-density lipoprotein cholesterol

According to the *Technical Specification for Stroke Screening and Prevention* by National Health Commission of China (6), the following eight risk factors were considered:

(1) Hypertension: have a history of hypertension

or is taking antihypertensive drugs or blood pressure ≥ 140 / 90 mmHg;
(2) Atrial fibrillation or heart valve disease: definite diagnosis or has ECG or cardiac ultrasound to support the diagnosis;
(3) Smoking: more than 6 sticks per day in the past 6 months;
(4) Obesity: BMI ≥ 26 kg/m^2;
(5) Dyslipidemia: triglyceride ≥ 2.26 mmol / L, or total cholesterol ≥ 6.22 mmol / L, or low density lipoprotein cholesterol ≥ 4.14 mmol / L, or high density lipoprotein cholesterol ≤ 1.04 mmol / L;
(6) Diabetes: fasting blood glucose FPG ≥ 7.0 mmol / L or 2h postprandial blood glucose ≥ 11.1 mmol / L or has been diagnosed with diabetes and is taking blood glucose control drugs;
(7) Rare physical activity (criteria for physical exercise: ≥ 3 times / week, > 30 minutes each time and last more than 1 year);
(8) Have a family history of stroke.

Those with three or more risk factors or previous history of stroke or transient ischemic attack could be treated as high-risk population.

Research Methods

Logistic regressive analysis was performed by SPSS 17.0 (Chicago, IL, USA), and LightGBM algorithm analysis was performed by Python3.72.

Results

Prediction model based on logistic regressive analysis

The evaluation results (whether they were high-risk population of stroke) were used as outcome indicators, and the remaining variables were acted as factors. Detailed related-information of the variables are shown in Table 2. General distribution of two groups are shown in Table 3 and 4.

Table 2: Detailed related-information of the variables

Variables	*Assignment / data value range*
High-risk of stroke	Yes = 1, No = 0
Age (yr)	40-59 years, ≥ 60 years
Gender	Male = 1, Female = 2
Marital status	Married = 1, Unmarried or divorced = 2
Current or pre-retirement occupation	Mental worker = 1, Manual worker = 2
Educational level	Primary and below = 1, Secondary and above = 2
Family history of coronary heart disease	Yes = 1, No = 0
Family history of hypertension	Yes = 1, No = 0
Family history of diabetes	Yes = 1, No = 0
Excessive drinking	Yes = 1, No = 0
Wake-up time	6: 00 ~ 8: 00 = 1, Others = 2
Bedtime	21: 00 ~ 23: 00 = 1, Others = 2
Meal time	Regular = 1, Irregular = 0
White meat	Often = 1, Do not or less = 0
Red meat	Often = 1, Do not or less = 0
Dairy products	Often = 1, Do not or less = 0
Beans and soy products	Often = 1, Do not or less = 0
Vegetables	Often = 1, Do not or less = 0
Fruit	Often = 1, Do not or less = 0
Nut	Often = 1, Do not or less = 0
Spicy foods	Preference = 1, Dislike = 0
Salty foods	Preference = 1, Dislike = 0
Sour foods	Preference = 1, Dislike = 0
Mild foods	Preference = 1, Dislike = 0

Table 3: General distribution in the middle-aged group [$\bar{x} \pm s$, n (%)]

Variable	*High-risk of stroke*	*Non-high risk stroke*	*Total*	χ^2 *(t)*	*P*
Age (yr)	53.4 ± 4.4	51.3 ± 4.2		7.329	0.000
Gender					
Male	239 (79.9%)	211 (27.8%)	450	238.997	0.000
Female	60 (20.1%)	549 (72.2%)	609		
Total	299	760	1059		
Marital status					
Married	297 (99.3%)	741 (97.5%)	*1038*	*2.819*	0.093
Unmarried or divorced	2 (0.7%)	19 (2.5%)	*21*		
Total	299	760	1059		
Current or pre-retirement occupation					
Mental worker	155 (51.8%)	426 (56.1%)	581	1.538	0.215
Manual worker	144 (48.2%)	334 (43.9%)	478		
Total	299	760	1059		
Educational level					
Elementary and below	63 (21.1%)	130 (17.1%)	193	2.264	0.132
Secondary and above	236 (78.9%)	630 (82.9%)	866		
Total	299	760	1059		
Family history of coronary heart disease					
Yes	33 (11.0%)	67 (8.8%)	100	1.238	0.266
No	266 (89.0%)	693 (91.2%)	959		
Total	299	760	1059		
Family history of hypertension					
Yes	138 (46.2%)	241 (31.7%)	379	*19.479*	0.000
No	161 (53.8%)	519 (68.3%)	680		
Total	299	760	1059		
Family history of diabetes					
Yes	60 (20.1%)	94 (12.4%)	154	*10.233*	0.001
No	239 (79.9%)	666 (87.6%)	905		
total	299	760	1059		

Table 4: General situation in the elder-aged group [$\bar{x} \pm s$, n (%)]

Variable	*High-risk of stroke*	*Non-high risk stroke*	*Total*	χ^2 *(t)*	*P*
Age (yr)	69.1 ± 7.2	68.1 ± 6.5		2.519	0.012
Gender					
Male	259 (59.8%)	290 (45.9%)	549	19.961	0.000
Female	174 (40.2%)	342 (54.1%)	516		
Total	433	632	1065		
Marital status					
Married	412 (95.2%)	606 (95.9%)	1018	0.330	0.566
Unmarried or divorced	21 (4.8%)	26 (4.1%)	47		
Total	433	632	1065		
Current or pre-retirement occupation					
Mental worker	143 (33.0%)	349 (55.2%)	492	5.932	0.000
Manual worker	290 (67.0%)	283 (44.8%)	573		
Total	433	632	1065		
Educational level					
Elementary and below	269 (62.1%)	80 (12.7%)	349	285.390	0.000
Secondary and above	164 (37.9%)	552 (87.3%)	716		
Total	433	632	1065		
Family history of coronary heart disease					
Yes	34 (7.9%)	56 (8.9%)	90	0.338	0.561
No	399 (92.1%)	632 (91.1%)	975		
Total	433	632	1065		
Family history of hypertension					
Yes	131 (30.3%)	178 (28.2%)	309	*0.545*	0.460
No	302 (69.7%)	454 (71.8%)	756		
Total	433	632	1065		
Family history of diabetes					
Yes	70 (16.2%)	54 (8.5%)	124	*14.510*	0.000
No	363 (83.8%)	578 (91.2%)	941		
total	433	632	1065		

$\alpha = 0.05$ was set and each variable was included as a covariate in the logistic regressive model. We selected Forward LR as stepwise regression method and screened these variables until all variables had statistical significance in the model. The results of logistic regression analysis for the two groups are shown in table 5 and 6.

Table 5: The results of Logistic regression analysis in the middle-aged group

Variables	*B*	*SE*	*Wald* χ^2	P	OR *(95%*CI *)*
Age (yr)	0.176	0.023	56.435	0.000	1.193 (1.139 - 1.249)
Gender	-2.827	0.236	143.113	0.000	0.059 (0.037 - 0.094)
Family history of hypertension	1.045	0.219	22.688	0.000	2.845 (1.850 - 4.374)
Family history of diabetes	1.424	0.306	21.718	0.000	4.154 (2.282 - 7.562)
Excessive drinking	1.817	0.300	36.780	0.000	6.151 (3.420 - 11.065)
Dairy products	-1.378	0.209	43.571	0.000	0.252 (0.167 - 0.379)
Fruits	-1.233	0.242	25.937	0.000	0.292 (0.181 - 0.469)
Preference for salty foods	1.158	0.221	27.476	0.000	3.185 (2.065 - 4.911)
Preference for mild foods	-1.259	0.212	35.403	0.000	0.284 (0.187 - 0.430）
Constant	-4.947	1.251	15.650	0.000	0.007

Table 6: The results of Logistic regressive analysis in the elder-aged group

Variable	*B*	*SE*	*Wald* χ^2	P	OR *(95%*CI *)*
Gender	-1.588	0.208	58.042	0.000	0.204 (0.136 - 0.307)
Educational level	-3.890	0.253	236.789	0.000	0.020 (0.012 - 0.034)
Family history of hypertension	0.466	0.192	5.862	0.015	1.593 (1.093 - 2.322)
Family history of diabetes	1.888	0.272	48.346	0.000	6.609 (3.881 - 11.254)
Beans and soy products	-0.445	0.170	6.802	0.009	0.641 (0.459 - 0.895)
Fruits	-0.458	0.186	6.064	0.014	0.632 (0.439 - 0.911)
Preference for mild foods	-1.547	0.193	64.541	0.000	0.213 (0.146 - 0.310)
Constant	9.391	0.667	198.035	0.000	11977.152

Prediction model based on LightGBM algorithm

It was a challenge for machine to learn to select the appropriate feature variables. We described the correlation between features in the form of a heat map, as shown in Fig. 1 and 2. The value in the corresponding cross-lattice was the value of the correlation between two features. The absolute value of the correlation ranges between 0-1. The value was more lager, the more relevant between two features. A positive value indicated a positive correlation, and a negative value indicated a negative correlation.

The importance scores of features of the LightGBM model were outputted. Ten most important features were got finally in the middle-aged group, ranking by importance from the highest to the lowest were: age, dairy products, beans and soy products, gender, mild food, salty food, excessive drinking, nuts, family history of hypertension, and fruits. Seven most important characteristics were got finally in the elder-aged group, ranking by importance from highest to lowest were: age, gender, education level, mild food, wake-up time, fruits, and salty foods (Fig. 3 and 4).

Evaluation of the prediction performance of the two models

The performance of the two prediction models was evaluated by four values including F1 score, precision, recall and AUROC (area under the receiver operating characteristic curve) (Table 7). All the values of the LightGBM model were higher than the corresponding logistic regressive model, except for the recall value in the middle-aged group. The LightGBM model performed better than the logistic regressive model.

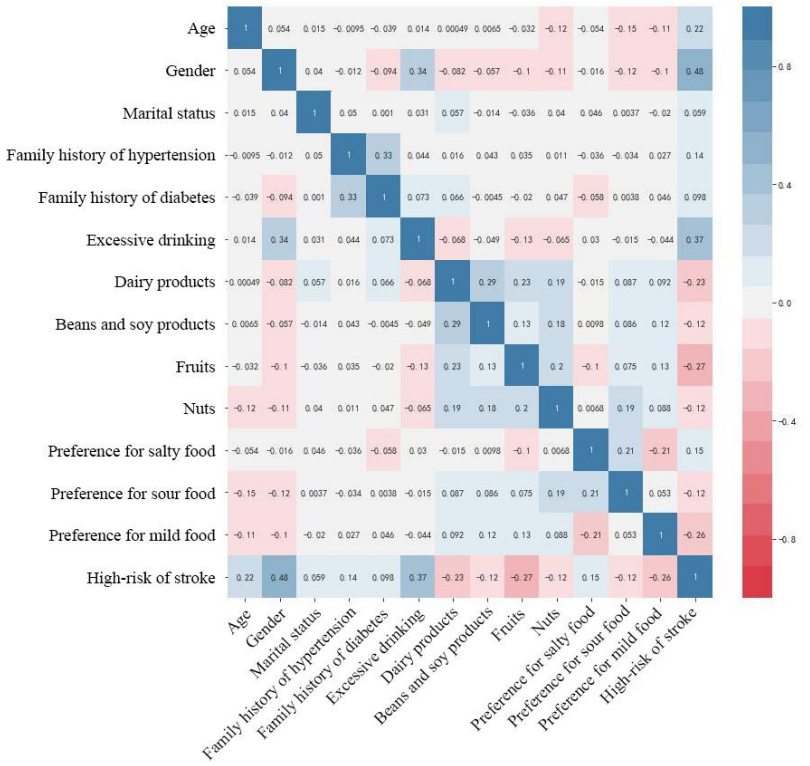

Fig. 1: Heat map of correlations between features in the middle-aged group

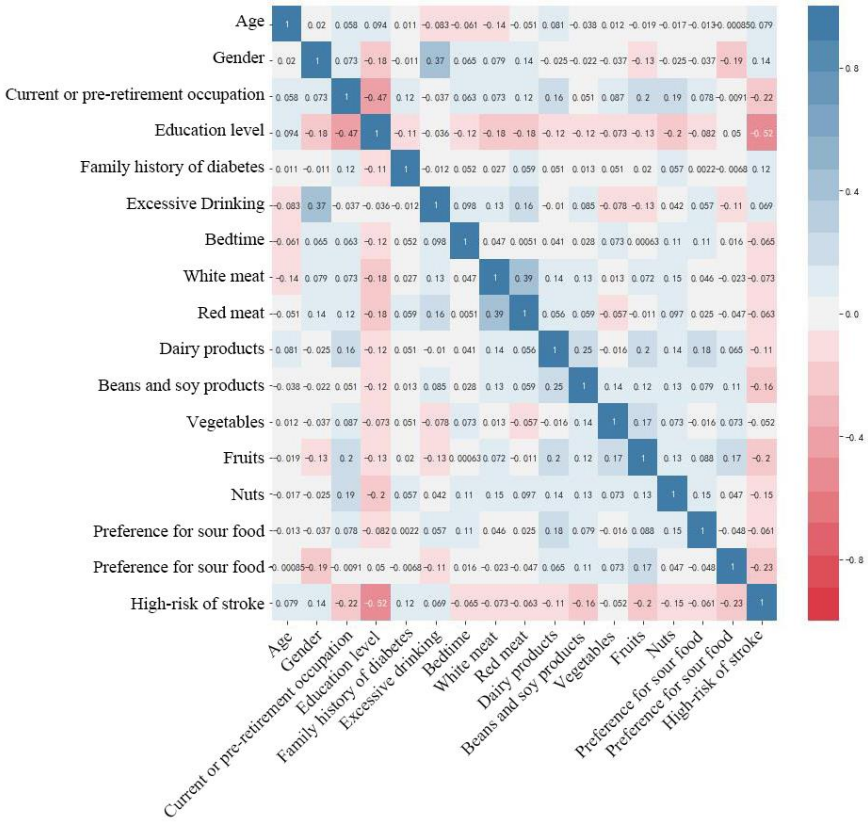

Fig. 2: Heat map of correlations between features in the elder-aged group

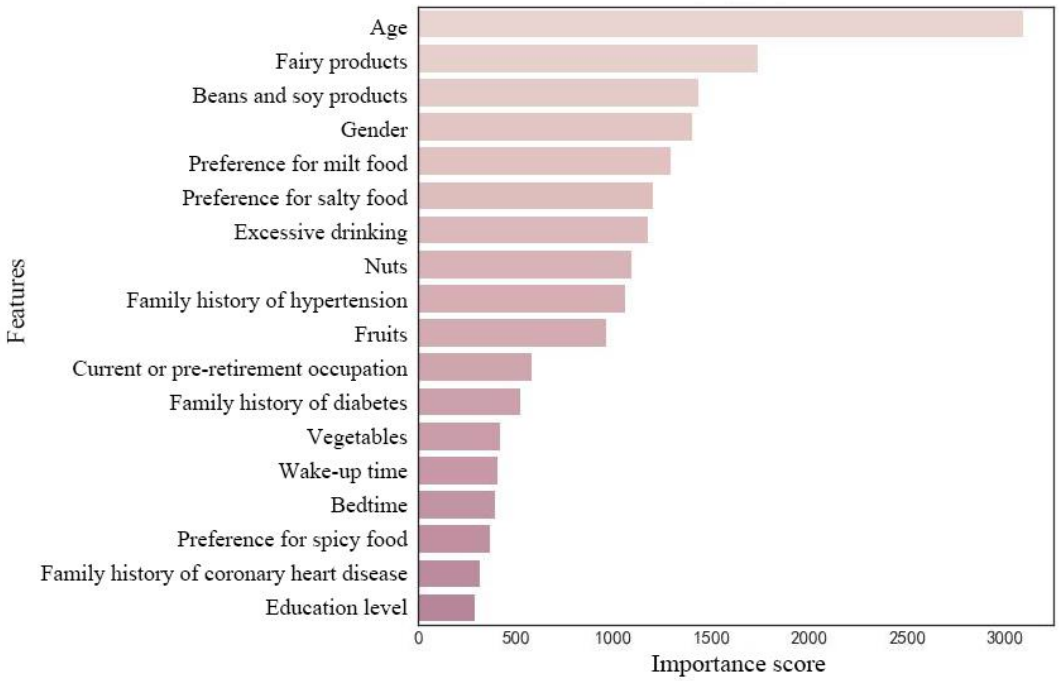

Fig. 3: Importance scores of features presented by LightGBM model in the middle-aged group

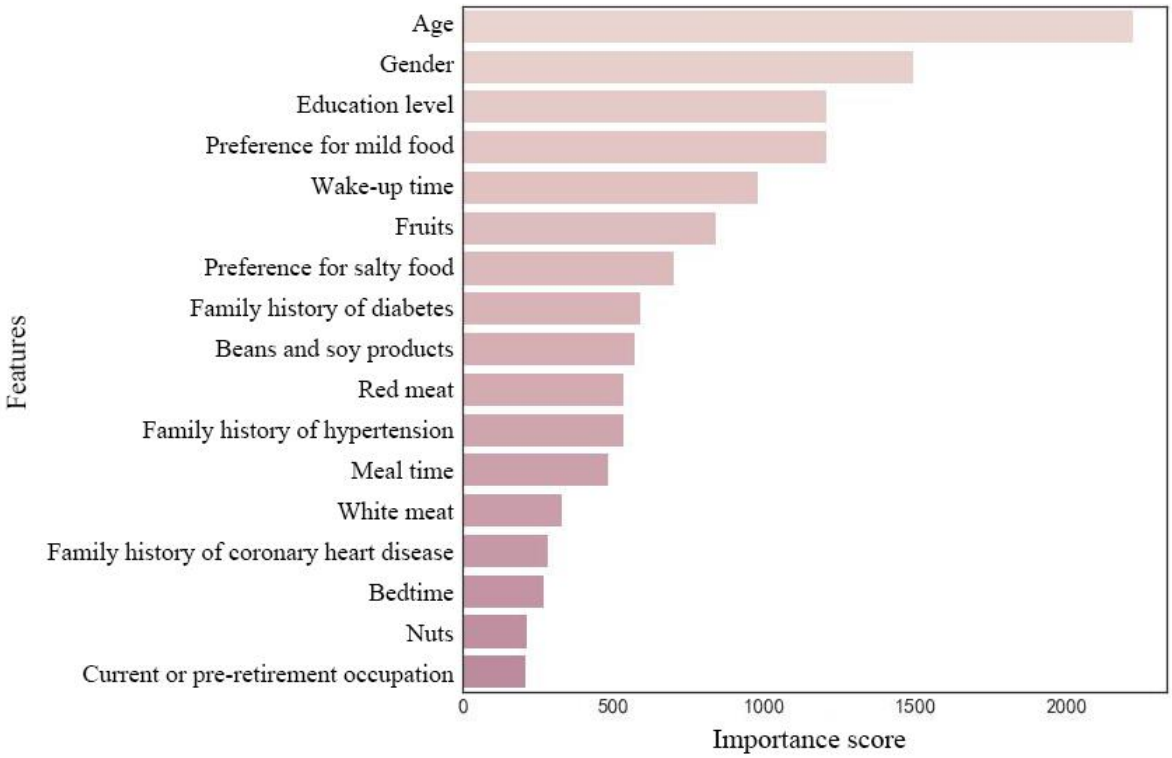

Fig. 4: Importance scores of features presented by LightGBM model in the elder-aged group

Table 7: The result of comparison between two prediction models

Model	*Group*	*F1 Score*	*Accuracy*	*Recall*	*AUROC*
Logistic regressive model	Middle-age group	0.734	0.735	0.732	0.911
	Elder-age group	0.728	0.747	0.709	0.875
LightGBM model	Middle-age group	0.743	0.802	0.692	0.913
	Elder-age group	0.751	0.784	0.721	0.883

Discussion

In this study, logistic regressive analysis and LightGBM algorithm were applied for screening out the relevant risk factors of stroke, and establishing the prediction models. Then, we evaluated the performance and compared the prediction between the two models.

Most of the previous related studies (7-9) only enrolled the middle-aged and elder-aged population together as one single group, while our study had separated them into two groups. Through comprehensively analyzing the results of the two models, we found that the risks of stroke were positively correlated with age, family history of hypertension, family history of diabetes, and preference for salty foods, while negatively correlated with the consumption frequency of fruits and preference for mild foods. Men were associated with a higher risk of stroke than women. In addition, the risk factors were positively correlated with excessive drinking and negatively correlated with the consumption frequency of dairy products in the middle-aged group. For the elder-aged group, the risk factors had a negative correlation with the education level.

Age was the most significant factor of stroke in both the middle-aged and elder-aged groups. Males were at higher risk of stroke than females due to their lifestyle, stress condition and hormone levels (10, 11). High blood pressure is an important risk factor of stroke (12), while the blood pressure can be controlled by taking calcium channel blocker drugs in preventing stroke (13).Vascular disease and hyperlipidemia caused by diabetes can increase the death rate and relapse rate of the patients with stroke, people with diabetes had several times more chances of having a stroke than normal people, type 2 diabetes can enhance platelets level to promote the occurrence of ischemic stroke (14). The diseases mentioned above are genetic predisposition, indicating that a family history of hypertension or diabetes is associated with higher risk of stroke, which is consistent with our research findings. Meanwhile, a previous study has found that moderately eating fruits could reduce the risk of stroke (15), attributed to the benefits of vitamins and trace elements contained in the fruits. However, a high-salt diet was recognized as a main risk factor of stroke. People on longstanding high-salt diet are associated with rising plasma sodium concentration, which inhibits vascular smooth muscle and myocardial enzymes, potentially increasing cardiac contractility and vascular tone.

Moderate alcohol consumption has a protective effect on the blood vessel, while excessive drinking may increase the risk of stroke (16). The people aged 45-59 years old usually drink much more than others because of their social activities (17), which might explain the reason that excessive drinking was positively correlated with the risk of stroke only in the middle-aged group. Some other studies have shown that the intake of dairy products could reduce the risk of stroke (18, 19). One of the explanation is that the animal proteins can reduce the tissue-type plasminogen activator (t-PA) antigen, which converts plasminogen into plasmin that reduces fibrin and blocks blood clots, eventually reducing the risk of thrombosis (20). Consistently, we also found that the consumption frequency of dairy products was a protective factor of stroke in the middle-aged group (21).

The education level showed a negative correlation with the risk of stroke only in the elder-aged group, which might be because that the propor-

tion of people with primary school education or below in the elder-aged group (32.77%) was much larger than that in the middle-aged group (18.22%). Less literacy lead to a lower cognition level of stroke, thus increasing the risk of stroke. In addition, we compared the prediction results from the logistic regressive model and the LightGBM model. The AUROCs of the middle-aged group and the elder-aged group were higher than 0.85 in both models. The AUROC of the middle-aged group reached more than 0.90, indicating reliability for prediction. Through comparing the F1 score, accuracy, recall rate and AUROC between the two models, we found that the LightGBM model had a slightly better prediction effect than the logistic regressive model. In Comparison with traditional methods, the LightGBM algorithm had an overall better performance.

Conclusion

Age, gender, family history of hypertension, family history of diabetes, consumption frequency of fruit, and preferences for salty and mild foods were the common risk factors of stroke in both middle-aged and elder-aged people. Excessive drinking and consumption frequency of dairy products were risk factors of stroke in the middle-aged people. Education level was a unique risk factor of stroke in the elder-aged group. The logistic regressive model and the LightGBM model established by this study could have an important significance in a way of discovering high-risk stroke people and taking preventive measures early. The prediction effect of the LightGBM algorithm for high-risk population of stroke was more accurate than that of logistic regressive analysis.

Journalism Ethics considerations

Ethical issues (Including plagiarism, informed consent, misconduct, data fabrication and/or falsification, double publication and/or submission, redundancy, etc.) have been completely observed by the authors.

Acknowledgements

We thank all participants for their time and contributions, and the research staff and students for their help in this study.

Funding

This study was supported by National Innovation and Entrepreneurship Training Program for College Students (201810354036).

Conflict of interest

All authors declare that they have no competing interest.

References

1. World Health Organization (2010). World Health Report 2010: Changing history.
2. Wang L, Liu J, Yang Y (2018). Essentials of report on the prevention for Chinese s troke 2017. *Chin J Cerebrovasc Dis*, 15(11):611-617.
3. Lee H, Nam YS, Lee KM (2015). Development-assistance Strategies for Stroke in Low- and Middle-income Countries. *J Korean Med Sci*, 30(Suppl 2):139-142.
4. Chhabra M, Sharma A, Ajay KR, et al (2019). Assessment of risk factors, cost of treatment and therapy outcome in stroke patients: evidence from cross-sectional study. *Expert Rev Pharmacoecon Outcomes Res*, 19(5):575-580.
5. Guo L, Hu R, Gong W, et al (2017). Research progress on risk factors of stroke. *Chin J Gerontol*, 37(17):4413-4416.
6. National Health Commission of the People's Republic of China. (2013). Technical Specification for Stroke Screening and Prevention. *Chinese Journal of the Frontiers of Medical Science (Electronic Version)*, 5(09):44-50.
7. Zhao D, Liu J, Wang W, et al (2008). Epidemiological Transition of Stroke in China Twenty-One-Year Observational Study from the Si-

no-MONICA-Beijing Project. *Stroke*, 39(6):1668-74.
8. Wieberdink R G, Ikram M A, Hofman A, et al (2012). Trends in stroke incidence rates and stroke risk factors in Rotterdam, the Netherlands from 1990 to 2008. *Eur J Epidemiol*, 27(4):287-95.
9. Yu W, Yan Y, Jiang C (2010). Logistic analysis of epidemiological characteristics and risk factors of stroke in Shijiazhuang City. *Hebei Medical Journal*, 32(017):2433-2434.
10. Wang S, Wang Z, Ma Z, et al (2016). The role of estrogen in ischemic stroke. *Prevent Treat Cardil Cerebr Vasc Dis*, 16(6):459-461.
11. Sohrabji F, Okoreeh A, Panta A (2019). Sex hormones and stroke: Beyond estrogens. *Horm Behav*, 111:87-95.
12. Feigin VL, Roth GA, Naghavi M, et al (2016). Global burden of stroke and risk factors in 188 countries, during 1990-2013: a systematic analysis for the Global Burden of Disease Study 2013. *Lancet Neurol*, 15(9):913-924.
13. Harada W, Miyauchi S, Higashihara T, et al (2016). Visit-to-visit blood pressure variability and classes of antihypertensive agents; associations with artery remodeling and the risk of stroke. *Curr Pharm Des*, 22(3):383-389.
14. Deng X L, Liu Z, Wang C, et al (2017). Insulin resistance in ischemic stroke. *Metab Brain Dis*, 32(5):1323-1334.
15. Zhang D, Han W (2018). Research progress on prevalence and risk factors of stroke in China. *World Latest Med Inf*, 18(80):128-129.
16. Iso H, Baba S, Mannami T, et al (2004). Alcohol Consumption and Risk of Stroke among Middle-Aged Men: The JPHC Study Cohort I. *Stroke*, 35(5):1124-9.
17. Xu X, Zhao L, Fang H, et al (2016). [Status of alcohol drinking among aged 15 and above in China in 2010-2012]. *Wei Sheng Yan Jiu*, 45(4):534-5٦٧.
18. Larsson S C, Virtamo J, Wolk A (2012). Dairy Consumption and Risk of Stroke in Swedish Women and Men. *Stroke*, 43(7):1775-80.
19. Dalmeijer GW, Struijk EA, Van der Schouw YT, et al (2013). Dairy intake and coronary heart disease or stroke-A population-based cohort study. *Int J Cardiol*, 167(3):925-9.
20. Abbott RD, Curb JD, Rodriguez BL, et al (1996). Effect of dietary calcium and milk consumption on risk of thromboembolic stroke in older middle-aged men. The Honolulu Heart Program. *Stroke*, 27 (5):813-8.
21. Yang Z, Lv X, Xiao X, et al (2019). Analysis on consumption structure and trend of dairy products in China-a dairy consumption survey based on 3000 consumers in 6 typical cities. *Chin Dairy*, (9):23-27.