Identification of Carotid Atherosclerosis in Populations at Moderate-to-High Risk for Cardiovascular Disease: A Machine Learning-Based Predictive Model and Validation Postprint

Authors: Liu Zhongdian, Xu Qi, Chen Yijing, Qin Lingqiao, Chen Shuping, Tang Weiting, Zhong Qiuan, Zhong Qiu' an

Date: 2024-05-20T00:00:00+00:00

Abstract

Background: Carotid atherosclerosis (CAS) is often regarded as an early warning sign of cardiovascular disease (CVD). Its diagnostic technique, carotid Doppler ultrasound examination, has not been included in public health service programs, while the Framingham Risk Score (FRS) exhibits insufficient accuracy in assessing CAS risk, which is not conducive for primary healthcare personnel to identify CAS. Currently, research on using machine learning methods to identify CAS in medium- and high-risk populations according to FRS remains lacking.

Objective: To construct CAS prediction models for medium- and high-risk FRS populations using machine learning methods, compare their discriminative performance, select the model with optimal performance, with the aim of assisting primary healthcare personnel in identifying CAS more conveniently and accurately.

Methods: A total of 674 local residents who met the inclusion and exclusion criteria were selected from two townships in Liuzhou City, Guangxi Zhuang Autonomous Region, during 2019–2021 and 2023 as study subjects. Relevant information was collected, and fasting blood and urine samples were obtained to detect biochemical indicators. The FRS was used to assess CVD risk, and carotid ultrasound was employed to diagnose CAS. The 517 study subjects from 2019–2021 were randomly divided into training and validation sets at an 8:2 ratio; the training set was used to construct Logistic regression, Random Forest (RF), Support Vector Machine (SVM), eXtreme Gradient Boosting (XGBoost), and Gradient Boosting Decision Tree (GBDT) models, while the validation set was used for internal validation. The 157 study subjects from 2023 served as a

test set for external validation. Feature variables were selected through Lasso regression analysis, and discriminative performance was evaluated using sensitivity, specificity, accuracy, F1 score, and area under the curve (AUC) values. External validation employed AUC values to assess the generalization ability of the optimal model, and the Shapley Additive exPlanation (SHAP) method was used to explore important variables influencing CAS identification by the optimal model.

Results: Lasso regression selected 15 non-zero feature variables: age, BMI, systolic blood pressure (SBP), smoking, alcohol consumption, hypertension, total cholesterol, high-density lipoprotein cholesterol, C-reactive protein (CRP), fasting blood glucose, apolipoprotein B (ApoB), lipoprotein(a) (LPA), aspartate aminotransferase (AST), AST/alanine aminotransferase ratio, and urinary albumin-to-creatinine ratio. The constructed Logistic regression, RF, SVM, XGBoost, and GBDT models all achieved high AUC values, among which the GBDT model demonstrated optimal discriminative performance with sensitivity, specificity, accuracy, F1 score, and AUC values of 0.7551, 0.8364, 0.7981, 0.7789, and 0.8349, respectively; the external validation AUC value was 0.7940. The SHAP method identified age, SBP, CRP, LPA, and ApoB as the top 5 factors influencing CAS identification by the GBDT model.

Conclusion: The Logistic regression, RF, SVM, XGBoost, and GBDT models based on machine learning for CAS identification all demonstrated high discriminative performance, with the GBDT model exhibiting the best comprehensive discriminative efficacy and strong generalization ability.

Full Text

Identification of Carotid Atherosclerosis in Medium-High Risk Population of Cardiovascular Disease: Prediction Model and Validation Based on Machine Learning

LIU Zhongdian, XU Qi, CHEN Yijing, QIN Lingqiao, CHEN Shuping, TANG Weiting, ZHONG Qiuan*

Department of Epidemiology, School of Public Health, Guangxi Medical University, Nanning 530021, China

Corresponding author: ZHONG Qiuan, Professor/Doctoral supervisor; E-mail: qazhong@gxmu.edu.cn

---

Abstract

Background: Carotid atherosclerosis (CAS) is often considered an early warning signal for cardiovascular diseases (CVD). However, carotid Doppler ultrasonography, the diagnostic technique for CAS, has not been included in public

health service programs, and the Framingham Risk Score (FRS) lacks sufficient accuracy in assessing CAS risk, hindering primary healthcare personnel from identifying CAS effectively. Currently, research on using machine learning methods to identify CAS in the FRS-defined medium-high risk population remains scarce.

Objective: To construct CAS prediction models for the FRS medium-high risk population using machine learning methods, compare their discriminative efficacy, and identify the optimal model to assist primary healthcare personnel in identifying CAS more conveniently and accurately.

Methods: A total of 674 local residents from two townships in Liuzhou City, Guangxi Zhuang Autonomous Region, who met the inclusion and exclusion criteria during 2019–2021 and 2023 were selected as study subjects. Relevant information was collected, and fasting blood and urine samples were obtained for biochemical testing. The FRS was used to assess CVD risk, and carotid ultrasound was employed to diagnose CAS. Among the 517 subjects from 2019–2021, an 8:2 random split was used to create a training set and a validation set. The training set was used to build Logistic regression, Random Forest (RF), Support Vector Machine (SVM), Extreme Gradient Boosting (XGBoost), and Gradient Boosting Decision Tree (GBDT) models, while the validation set was used for internal validation. The 157 subjects from 2023 served as a test set for external validation. Feature variables were selected using Lasso regression analysis. Discriminative efficacy was evaluated using sensitivity, specificity, accuracy, F1 score, and area under the curve (AUC). External validation assessed the generalization ability of the optimal model using AUC, and the Shapley Additive exPlanation (SHAP) method was used to explore important variables influencing the optimal model' s identification of CAS.

Results: Lasso regression identified 15 non-zero feature variables: age, BMI, systolic blood pressure (SBP), smoking, alcohol consumption, hypertension, total cholesterol, high-density lipoprotein cholesterol, C-reactive protein (CRP), fasting plasma glucose, apolipoprotein B (ApoB), lipoprotein(a) (LPA), aspartate aminotransferase (AST), AST/alanine aminotransferase ratio, and urinary microalbumin-to-creatinine ratio. All constructed models (Logistic regression, RF, SVM, XGBoost, and GBDT) achieved high AUC values. The GBDT model demonstrated the best discriminative performance, with sensitivity, specificity, accuracy, F1 score, and AUC values of 0.7551, 0.8364, 0.7981, 0.7789, and 0.8349, respectively. The external validation AUC was 0.7940. SHAP analysis revealed that age, SBP, CRP, LPA, and ApoB were the top five factors influencing the GBDT model' s identification of CAS.

Conclusion: All machine learning models (Logistic regression, RF, SVM, XGBoost, and GBDT) showed high discriminative performance for identifying CAS, with the GBDT model exhibiting the best comprehensive discriminative efficacy and strong generalization ability.

Keywords: cardiovascular diseases; carotid atherosclerosis; machine learning;

Framingham risk score; identification; forecasting

Introduction

Cardiovascular disease (CVD) is one of the leading causes of death among urban and rural residents in China, with its incidence and mortality continuing to rise, representing the primary health risk factor for Chinese residents [1]. Atherosclerosis is the main pathological basis of CVD, and the carotid artery is often the earliest affected site. Consequently, carotid atherosclerosis (CAS) is generally considered an early warning signal for CVD [2]. In terms of diagnosis, Doppler ultrasonography measuring carotid intima-media thickness (CIMT) is a reliable technique for assessing CAS severity [3]. Since 2009, the Basic Public Health Service Program has been continuously expanded, reaching 12 service categories by 2019 [4]; however, carotid Doppler ultrasonography has not been included, failing to meet the needs of early CVD prevention and control. The Framingham Risk Score (FRS) is a widely used cardiovascular risk assessment tool, but it lacks accuracy in evaluating CAS risk [5,6], potentially preventing primary healthcare personnel from accurately identifying CAS. Therefore, there is an urgent need to explore more convenient and effective methods for early CAS identification. In recent years, an increasing number of scholars have employed machine learning to identify diseases using easily obtainable factors, achieving promising results for both individual self-assessment and clinical applications [7].

Currently, research on using machine learning to identify CAS in the FRS-defined medium-high risk population remains relatively limited. To strengthen research in this area, this study selected Logistic regression, Random Forest (RF), Support Vector Machine (SVM), Extreme Gradient Boosting (XGBoost), and Gradient Boosting Decision Tree (GBDT) to construct CAS prediction models for the FRS medium-high risk population (FRS $>$ 6%), compare their performance, and identify the optimal model. Our goal is to assist primary healthcare personnel in identifying CAS more conveniently, accurately, and earlier, providing a scientific basis for clinical prevention and treatment.

Methods

1.1 Study Subjects

Using convenience sampling, 1,169 local residents from two townships in Liuzhou City, Guangxi Zhuang Autonomous Region, were selected as potential study subjects during 2019–2021 and 2023. Among them, 852 residents from 2019–2021 were intended for model construction and internal validation, while 317 residents from 2023 were designated for external validation. Inclusion criteria were: (1) aged 30–74 years; (2) FRS $>$ 6%; (3) willingness to undergo

carotid Doppler ultrasonography. Exclusion criteria included: (1) individuals with major diseases such as malignant tumors, severe infectious diseases, or psychiatric disorders; (2) those already diagnosed with coronary heart disease, stroke, or peripheral arterial disease; and (3) subjects with missing covariates. Based on these criteria, 674 subjects were ultimately enrolled (517 from 2019-2021 and 157 from 2023). This study was approved by the Ethics Committee of Guangxi Medical University (2019-SB-094), and all participants provided informed consent.

1.2 Research Methods

1.2.1 General Data Collection and Physical Examination General data were collected through a questionnaire designed by the research team, including information on gender, age, ethnicity, education level, physical activity, smoking history, alcohol consumption history, disease history, and medication use. Physical examinations primarily included measurements of BMI, waist circumference, heart rate, systolic blood pressure (SBP), and diastolic blood pressure (DBP). Laboratory test indicators comprised total cholesterol (TC), triglycerides (TG), low-density lipoprotein cholesterol (LDL-C), high-density lipoprotein cholesterol (HDL-C), fasting plasma glucose (FPG), urinary microalbumin (ALB), C-reactive protein (CRP), urinary creatinine (UCR), lipoprotein(a) (LPA), apolipoprotein A (ApoA), apolipoprotein B (ApoB), alanine aminotransferase (ALT), and aspartate aminotransferase (AST). The urinary microalbumin-to-creatinine ratio (ACR) was calculated as ALB/UCR. Physical activity was calculated using the International Physical Activity Questionnaire (Short Form) [8] and expressed as metabolic equivalent (MET-min/week).

1.2.2 FRS Criteria This study used the FRS to assess CVD risk in the population, with FRS $> 6\%$ defined as medium-high CVD risk [9].

1.2.3 CAS Diagnosis CAS was defined as increased CIMT ≥ 1 mm or presence of plaque [10]. The definition of CIMT and detailed measurement methods were described in a previous study [11]. Plaque was defined as a focal structure encroaching into the arterial lumen by at least 0.5 mm or 50% of the surrounding CIMT value, or CIMT > 1.5 mm [12]. Professional ultrasound physicians performed carotid Doppler ultrasonography, and trained investigators recorded the corresponding data. Based on CAS diagnosis results, the 517 residents were divided into two groups: normal group (272 cases) and CAS group (245 cases).

1.2.4 Definitions of Relevant Variables

(1) **Smoking**: Never smoking was defined as total cigarette consumption $<$ 100 cigarettes; former smoking as $>$ 100 cigarettes but no smoking in the 30 days before the survey; current smoking as $>$ 100 cigarettes and smoking in the 30 days before the survey [13]. (2) **Alcohol consumption**: Never drinking was defined as $<$ 12 standard drinking units; former drinking as $\geq$

12 standard drinking units previously but < 1 unit in the most recent year; current drinking as ≥ 12 standard drinking units previously and ≥ 1 unit in the most recent year [14]. (3) **Renal function**: Estimated glomerular filtration rate (eGFR) was calculated using the Chronic Kidney Disease Epidemiology Collaboration formula. eGFR ≥ 90 mL $\cdot$ min^{-1} $\cdot$ (1.73 m^2)$^{-1}$ was defined as normal renal function, and eGFR < 90 mL $\cdot$ min^{-1} $\cdot$ (1.73 m^2)$^{-1}$ as decreased renal function [15]. (4) **Hypertension**: According to the Chinese Guidelines for the Prevention and Treatment of Hypertension (2018 Revision), hypertension was defined as SBP ≥ 140 mmHg (1 mmHg = 0.133 kPa) and/or DBP ≥ 90 mmHg, previous diagnosis of hypertension, or current use of antihypertensive medication [16]. (5) **Diabetes**: Defined as FPG ≥ 7.0 mmol/L in this survey, or self-reported current use of hypoglycemic medication or diabetes diagnosis [17]. (6) **Dyslipidemia**: Defined as meeting any of the following criteria: TC ≥ 200 mg/dL, TG ≥ 150 mg/dL, LDL-C ≥ 130 mg/dL, HDL-C < 40 mg/dL, or current use of lipid-lowering medication [18,19]. (7) **Metabolic syndrome**: Defined according to the International Diabetes Federation criteria [20]. (8) **Family history of disease in first-degree relatives**: Defined as at least one first-degree relative (father, mother, sibling, son, or daughter) having the disease [11].

1.3 Model Construction and Evaluation

Models were constructed using Python 3.7.4 with the scikit-learn 2.2.2 library. Feature variables selected by Lasso regression (with continuous variables normalized) were used as input variables, and CAS was the outcome variable. The train_{{{test}}}_{{split}}} module in scikit-learn 2.2.2 was used to randomly divide all samples into training and validation sets at an 8:2 ratio while maintaining the same ratio of positive to negative cases as in the full dataset. In the training set, LogisticRegression, RandomForestClassifier, SVC, XGBClassifier, and GradientBoostingClassifier modules were used to build Logistic regression, RF, SVM, XGBoost, and GBDT models, respectively. The GridSearchCV module (grid search algorithm) was used for parameter optimization for each model, with the area under the curve (AUC) as the evaluation metric. In the validation set, sensitivity, specificity, accuracy, F1 score, and AUC were used to evaluate the discriminative performance of the five models and select the optimal model. The optimal model was externally validated in the test set using AUC to assess generalization ability. The Shapley Additive exPlanation (SHAP) method was used to explore the specific impact of each feature variable on the optimal prediction model.

1.4 Statistical Analysis

Statistical analysis was performed using R (4.1.3). Normally distributed continuous variables were expressed as mean $\pm$ standard deviation ($\bar{x} \pm s$) and compared between groups using independent samples t-tests. Non-normally

distributed continuous variables were expressed as median (P25, P75) and compared using Mann-Whitney U tests. Categorical variables were expressed as percentages and compared using ^{2} tests. Lasso regression analysis was performed with CAS as the dependent variable to select feature variables. Receiver operating characteristic (ROC) curves were plotted for each model in the validation set, and AUC values were calculated and compared. $P < 0.05$ was considered statistically significant.

Results

2.1 General Characteristics of Study Subjects

Among the 517 residents, 210 (40.6%) were male and 307 (59.4%) were female, with a mean age of 60.2 ± 7.9 years. CAS was diagnosed in 245 cases (47.4%), while 272 cases (52.6%) were normal. There were no statistically significant differences between the two groups in gender, ethnicity, education level, waist circumference, heart rate, DBP, smoking history, alcohol consumption history, diabetes, physical activity, FPG, TC, LDL-C, dyslipidemia, metabolic syndrome, ApoA, ApoB, ApoA/ApoB ratio, ALT, AST, UCR, ALB, or ACR ($P > 0.05$). However, significant differences were observed between the groups in age, BMI, SBP, hypertension, HDL-C, TG, renal function, CRP, LPA, and AST/ALT ratio ($P < 0.05$) .

2.2 Feature Variable Selection Using Lasso Regression

With CAS diagnosis as the dependent variable and 36 potential influencing factors as independent variables, Lasso regression was performed to select variables. The assignment table for categorical variables is shown in , while continuous variables including age, heart rate, waist circumference, BMI, physical activity, SBP, DBP, FPG, TC, HDL-C, LDL-C, TG, CRP, LPA, ApoA, ApoB, ApoA/ApoB ratio, UCR, ALB, ALT, AST, AST/ALT ratio, and ACR were included as measured values. The analysis identified 15 variables with non-zero coefficients: age, BMI, SBP, smoking, alcohol consumption, hypertension, TC, HDL-C, CRP, FPG, ApoB, LPA, AST, AST/ALT ratio, and ACR [Figure 1: see original paper] and .

2.3 Construction of Machine Learning Models

The 15 variables selected by Lasso regression were incorporated into Logistic regression, RF, SVM, XGBoost, and GBDT models. Using grid search with AUC as the evaluation metric, the optimal parameters for each model were determined in the training set as follows: Logistic regression: solver = "liblinear" , max_{iter} = 500, penalty = "l2" ; RF: bootstrap = True, max_{depth} = 20, max_{features} = "auto" , min_{{{samples}}*{{leaf}}} = 2, *min*{{{samples}}*{{split}}} = 2; SVM: *kernel* = "*rbf*" , *C* = 1, *gamma* =

0.01; XGBoost: learning{rate} = 0.007, n_{estimators} = 500, max_{depth} = 2, min_{{{child}}*{{weight}}} = 8, gamma = 0.8, subsample = 0.8, colsample*{bytree} = 0.8, objective = "binary:logistic" , nthread = 4; GBDT: n_{estimators} = 500, learning_{rate} = 0.008, max_{depth} = 2, subsample = 0.8, max_{features} = "sqrt" , min_{{{samples}}*{{split}}} = 5, min*{{{samples}}*{{leaf}}} = 2, random*{state} = 1117.

2.4 Comparison of Discriminative Performance Among Models

Internal validation in the validation set showed that all models achieved high AUC values. While SVM had the highest AUC value, the GBDT model achieved the highest sensitivity, specificity, accuracy, and F1 score. Overall, the GBDT model demonstrated the optimal discriminative performance [Figure 2: see original paper] and .

2.5 External Validation

The GBDT model, which performed best in internal validation, was externally validated to assess its generalization ability. The results showed that the AUC in the external validation set was 0.7940, slightly lower than that in the internal validation set (0.8349) but still >0.7, indicating that the GBDT model constructed in this study possesses strong external generalization capability.

2.6 Interpretation of the Optimal Model Using SHAP

[Figure 3: see original paper]A displays the factors influencing CAS identification by the model, sorted by mean absolute SHAP values, providing an intuitive understanding of each factor' s contribution to model identification. In [Figure 3: see original paper]B, the y-axis shows the importance of each variable, with the most important at the top and least important at the bottom. The x-axis represents SHAP values, which measure each variable' s contribution to model identification. Positive values increase the likelihood of identification, while negative values decrease it. This visualization clearly shows the relationship between original variable values (color-coded: red for high values, blue for low values) and their impact on identification. The results indicate that the top five variables affecting the GBDT model' s discriminative performance were age, SBP, CRP, LPA, and ApoB. [Figure 3: see original paper]B shows that higher values of these variables increase CAS risk.

Discussion

This study found that 52.6% of individuals in the FRS medium-high risk population were not identified as having CAS, consistent with previous research [5,6], suggesting that the FRS has insufficient accuracy for CAS identification. To improve early CAS identification accuracy in this population, we constructed CAS

risk prediction models and identified the optimal model to enable more accurate identification, optimize individual prevention and treatment strategies, reduce medical burden, and avoid waste of medical resources.

This study constructed five prediction models using Logistic regression, RF, SVM, XGBoost, and GBDT algorithms based on machine learning. All models achieved high AUC values, with the GBDT model demonstrating the best comprehensive discriminative performance (sensitivity = 0.7551, specificity = 0.8364, accuracy = 0.7981, F1 score = 0.7789, AUC = 0.8349). Compared with similar studies [21-23], this model is considered to have high predictive accuracy. The model also showed strong generalization ability in external validation (AUC = 0.7940). GBDT, a machine learning method also known as multiple additive regression trees, offers more accurate identification and a more sophisticated algorithm than Logistic regression, decision trees, and RF [24]. It features numerous nonlinear transformations and robust representation capabilities without requiring complex feature engineering and transformation [25]. GBDT models have been widely applied in disease identification and have demonstrated good discriminative performance. WU et al. [21] used four machine learning methods (XGBoost, GBDT, RF, and SVM) to construct a carotid plaque identification model in an asymptomatic population, with the GBDT model achieving an AUC of 0.8367 and high discriminative performance. YE et al. [26] used multiple indicators including vital signs and laboratory tests from the Medical Information Mart for Intensive Care (MIMIC) IV database to establish machine learning-based prediction models for in-hospital mortality in intensive care unit patients with chronic kidney disease and coronary artery disease, with the GBDT model as the optimal model achieving an AUC of 0.946. LIU et al. [27] constructed an artificial intelligence-based risk prediction model for myocardial infarction to warn of its occurrence in hospitalized patients, with the GBDT model as the optimal model achieving an AUC of 0.91. LIU et al. [28] used machine learning methods to construct a sepsis risk prediction model for acute pancreatitis patients and compared the optimal GBDT model with Logistic regression and scoring systems, showing superior discriminative performance. SU et al. [29] used machine learning methods combined with longitudinal data to predict the 2-year risk of chronic kidney disease development in Chinese elderly, with the GBDT model showing good discriminative performance.

The feature variables in this study' s CAS risk prediction model are detection indicators included in public health service programs, making them easily obtainable and enhancing the model' s practicality. This can improve the convenience and accuracy of CAS identification by primary healthcare personnel, facilitating early identification and effective preventive and treatment strategies before disease progression, thereby improving patients' quality of life. By reducing cardiovascular events caused by CAS, this approach is expected to yield significant social and economic benefits, alleviate medical burden, and improve health resource utilization efficiency.

This study used SHAP method for visual interpretation of the GBDT model.

The top five variables affecting model discriminative performance were age, SBP, CRP, LPA, and ApoB, indicating that younger age, lower SBP, lower CRP, lower LPA, and lower ApoB can reduce CAS risk. ZHANG et al. [30] demonstrated that with increasing age, the proportion of collagen and elastic fibers in arterial wall structure becomes imbalanced, leading to arterial wall thickening and reduced compliance, combined with vascular endothelial dysfunction and structural abnormalities caused by some diseases, promoting atherosclerosis development. TANG et al. [31] also found that age is a risk factor for carotid plaque formation, with plaque increasing significantly with age, and many studies consider it an independent risk factor. Research shows that CAS incidence is higher in hypertensive patients, with more pronounced SBP elevation [32]. Previous studies have demonstrated that inflammation alone can trigger CAS formation even without other CVD risk factors [33]. High inflammation levels may cause excessive increase in endothelial permeability, indicating compromised integrity of the endothelial barrier. Damaged endothelial cells further express adhesion molecules and chemokines, enabling leukocytes to roll, adhere, and ultimately enter the vascular wall, thereby promoting vascular wall inflammation development [34]. Studies have shown that LPA is closely related to carotid atherosclerotic plaque occurrence, with mechanisms primarily related to cholesterol metabolism and fibrinolytic activity. Patients with high LPA have higher incidence of myocardial infarction and coronary heart disease than healthy individuals, and LPA in patients with cerebral arteriosclerosis is not only significantly higher than in healthy individuals but also closely related to lesion severity [35,36]. A meta-analysis including eight cohort and four case-control studies concluded that elevated ApoB levels are a risk factor for first ischemic stroke [37]. These findings are consistent with our results and clinical practice, demonstrating the strong rationality of our GBDT model.

This study has several limitations. First, convenience sampling may introduce selection bias. Second, the proportion of females was relatively high, possibly because more males work away from home. Third, information on medication use was not available, which may affect the results. Finally, most study subjects were from township areas, which may limit generalizability of the findings.

In summary, this study identified feature variables related to CAS through Lasso regression and constructed prediction models for the FRS medium-high risk population based on Logistic regression, RF, SVM, XGBoost, and GBDT. Comprehensive evaluation using sensitivity, specificity, accuracy, F1 score, and AUC showed that the GBDT model achieved the best performance in identifying CAS and demonstrated strong generalization ability. SHAP interpretation revealed that age, SBP, CRP, LPA, and ApoB were the most important variables for model discriminative performance and are also CAS risk factors. These findings are expected to help primary healthcare personnel make more accurate assessments, improve CAS identification and treatment coverage, facilitate rational allocation of medical resources, and provide a scientific basis for early intervention in the FRS medium-high risk population, thereby improving cardiovascular health, healthcare service levels, and public health in community

settings. Future research should further validate and expand the model' s applicability to ensure its effectiveness across different populations.

Author Contributions: LIU Zhongdian, XU Qi, CHEN Yijing, QIN Lingqiao, CHEN Shuping, and TANG Weiting conducted the study implementation, data collection, and organization. LIU Zhongdian performed statistical analysis, result interpretation, and manuscript writing. LIU Zhongdian and ZHONG Qiuan revised the manuscript. ZHONG Qiuan conceived and designed the study, conducted feasibility analysis, and was responsible for quality control and final review.

Conflict of Interest Statement: The authors declare no conflict of interest.

ORCID: LIU Zhongdian: https://orcid.org/0009-0003-3135-6800

References

[1] HU Shengshou, WANG Zengwu. Overview of the "China Cardiovascular Health and Disease Report 2022"[J]. Chinese Journal of Cardiovascular Research, 2023, 21(7): 577-600.

[2] SAKELLARIOS A I, BIZOPOULOS P, PAPAFAKLIS M I, et al. Natural history of carotid atherosclerosis in relation to the hemodynamic environment [J]. Angiology, 2017, 68(2): 109-118. DOI: 10.1177/0003319716644138.

[3] JOHRI A M, NAMBI V, NAQVI T Z, et al. Recommendations for the assessment of carotid arterial plaque by ultrasound for the characterization of atherosclerosis and evaluation of cardiovascular risk: from the American Society of Echocardiography [J]. J Am Soc Echocardiogr, 2020, 33(8): 917-933. DOI: 10.1016/j.echo.2020.04.021.

[4] YOU Lili, CHEN Xinyue, YANG Linghe, et al. Ten-year evaluation of the National Basic Public Health Service Program (2009-2019) series report (3)—Implementation of the National Basic Public Health Service Program for ten years: challenges and recommendations [J]. Chinese General Practice, 2022, 25(26): 3221-3231. DOI: 10.12114/j.issn.1007-9572.2022.0406.

[5] PEN A, YAM Y, CHEN L, et al. Discordance between Framingham Risk Score and atherosclerotic plaque burden [J]. Eur Heart J, 2013, 34(14): 1075-1082. DOI: 10.1093/eurheartj/ehs473.

[6] YI Yanshan, NONG Qingjiao, MAO Baoyu, et al. Study on cardiovascular disease risk factors based on Framingham Risk Score and vascular endothelial function classification [J]. Chinese General Practice, 2018, 21(16): 1959-1964. DOI: 10.3969/j.issn.1007-9572.2018.16.011.

[7] RIDKER P M, BURING J E, RIFAI N, et al. Development and validation of improved algorithms for the assessment of global cardiovascular risk

in women: the Reynolds Risk Score [J]. JAMA, 2007, 297(6): 611-619. DOI: 10.1001/jama.297.6.611.

[8] FAN Mengyu, LÜ Yun, HE Pingping. Calculation method for physical activity level in the International Physical Activity Questionnaire [J]. Chinese Journal of Epidemiology, 2014, 35(8): 961-964. DOI: 10.3760/cma.j.issn.0254-6450.2014.08.019.

[9] D'AGOSTINO R B Sr, VASAN R S, PENCINA M J, et al. General cardiovascular risk profile for use in primary care: the Framingham Heart Study [J]. Circulation, 2008, 117(6): 743-753. DOI: 10.1161/CIRCULATIONAHA.107.699579.

[10] WANG X J, LI W Z, SONG F J, et al. Carotid atherosclerosis detected by ultrasonography: a national cross-sectional study [J]. J Am Heart Assoc, 2018, 7(8): e008701. DOI: 10.1161/JAHA.118.008701.

[11] CHEN Runlin, HE Tufeng, TAO Lijun, et al. Study on the effect of cardiovascular risk factors on carotid intima-media thickness progression [J]. Chinese General Practice, 2023, 26(14): 1709-1715. DOI: 10.12114/j.issn.1007-9572.2022.0750.

[12] TOUBOUL P J, HENNERICI M G, MEAIRS S, et al. Mannheim carotid intima-media thickness and plaque consensus (2004-2006-2011). An update on behalf of the advisory board of the 3rd, 4th and 5th watching the risk symposia, at the 13th, 15th and 20th European Stroke Conferences, Mannheim, Germany, 2004, Brussels, Belgium, 2006, and Hamburg, Germany, 2011 [J]. Cerebrovasc Dis, 2012, 34(4): 290-296. DOI: 10.1159/000343145.

[13] HORNE D J, CAMPO M, ORTIZ J R, et al. Association between smoking and latent tuberculosis in the U.S. population: an analysis of the National Health and Nutrition Examination Survey [J]. PLoS One, 2012, 7(11): e49050. DOI: 10.1371/journal.pone.0049050.

[14] KUO C C, WEAVER V, FADROWSKI J J, et al. Arsenic exposure, hyperuricemia, and gout in US adults [J]. Environ Int, 2015, 76: 32-40. DOI: 10.1016/j.envint.2014.11.015.

[15] LEVEY A S, STEVENS L A, SCHMID C H, et al. A new equation to estimate glomerular filtration rate [J]. Ann Intern Med, 2009, 150(9): 604-612. DOI: 10.7326/0003-4819-150-9-200905050-00006.

[16] DAI Ye. Investigation of antihypertensive drug application in outpatients based on the "Chinese Guidelines for the Prevention and Treatment of Hypertension (2018 Revision)" [J]. Chinese Community Doctor, 2022, 38(12): 13-16.

[17] Chinese Clinical Guideline for the Prevention and Treatment of Type 2 Diabetes in the Elderly Writing Group. Chinese clinical guideline for the prevention and treatment of type 2 diabetes in the elderly (2022 edition) [J]. Chinese Journal of Diabetes, 2022, 30(1): 2-51. DOI: 10.3969/j.issn.1006-6187.2022.01.002.

[18] EXPERT PANEL ON DETECTION E. Executive summary of the

third report of the national cholesterol education program (NCEP) expert panel on detection, evaluation, and treatment of high blood cholesterol in adults (adult treatment panel) [J]. JAMA, 2001, 285(19): 2486-2497. DOI: 10.1001/jama.285.19.2486.

[19] ZHU Junren, GAO Runlin, ZHAO Shuiping, et al. Chinese guidelines for the prevention and treatment of dyslipidemia in adults (2016 revision) [J]. Chinese Circulation Journal, 2016, 31(10): 937-953.

[20] JIN Wensheng, PAN Changyu. Global consensus on the definition of metabolic syndrome by the International Diabetes Federation [J]. Chinese Journal of Endocrinology and Metabolism, 2005, 21(4): Appendix 4b-1-Appendix 4b-2. DOI: 10.3760/j.issn:1000-6699.2005.04.054.

[21] WU D, CUI G S, HUANG X X, et al. An accurate and explainable ensemble learning method for carotid plaque prediction in an asymptomatic population [J]. Comput Methods Programs Biomed, 2022, 221: 106842. DOI: 10.1016/j.cmpb.2022.106842.

[22] YU J, ZHOU Y, YANG Q, et al. Machine learning models for screening carotid atherosclerosis in asymptomatic adults [J]. Sci Rep, 2021, 11(1): 22236. DOI: 10.1038/s41598-021-01603-8.

[23] GONG Jun, ZHONG Xiaogang, TAN Juntao, et al. Establishment of a pediatric septic shock prediction model using "grid search + XGBoost"algorithm [J]. Medical Journal of Chinese People' s Liberation Army, 2020, 45(12): 1285-1289.

[24] ZHOU Z H, FENG J. Deep forest [J]. Natl Sci Rev, 2019, 6(1): 74-86. DOI: 10.1093/nsr/nwy108.

[25] ZHANG Z D, JUNG C. GBDT-MO: gradient-boosted decision trees for multiple outputs [J]. IEEE Trans Neural Netw Learn Syst, 2021, 32(7): 3156-3167. DOI: 10.1109/TNNLS.2020.3009776.

[26] YE Z X, AN S Y, GAO Y X, et al. The prediction of in-hospital mortality in chronic kidney disease patients with coronary artery disease using machine learning models [J]. Eur J Med Res, 2023, 28(1): 33. DOI: 10.1186/s40001-023-00995-x.

[27] LIU R, WANG M Y, ZHENG T, et al. An artificial intelligence-based risk prediction model of myocardial infarction [J]. BMC Bioinformatics, 2022, 23(1): 217. DOI: 10.1186/s12859-022-04771-2.

[28] LIU F, YAO J, LIU C Y, et al. Construction and validation of machine learning models for sepsis prediction in patients with acute pancreatitis [J]. BMC Surg, 2023, 23(1): 267. DOI: 10.1186/s12893-023-02151-y.

[29] SU D, ZHANG X Y, HE K, et al. Individualized prediction of chronic kidney disease for the elderly in longevity areas in China: machine learning approaches [J]. Front Public Health, 2022, 10: 998549. DOI: 10.3389/fpubh.2022.998549.

[30] ZHANG Ping, GUO Xiuli, ZHANG Penghua. Correlation between carotid atherosclerosis and vascular risk factors [J]. Chinese Journal of Gerontology, 2017, 37(5): 1132-1134. DOI: 10.3969/j.issn.1005-9202.2017.05.041.

[31] TANG Yan, ZHOU Hong, LUO Guanghua, et al. Analysis of ultrasound, CT angiography, and clinical risk factors for CAS plaques in patients with ischemic stroke [J]. Chinese Journal of Arteriosclerosis, 2016, 24(4): 391-395.

[32] GAO Suying, YAN Yinglin, YU Kai, et al. Study on risk factors of carotid atherosclerosis in acute ischemic stroke [J]. Chinese General Practice, 2021, 24(3): 327-332. DOI: 10.12114/j.issn.1007-9572.2020.00.401.

[33] TALEB S. Inflammation in atherosclerosis [J]. Arch Cardiovasc Dis, 2016, 109(12): 708-715. DOI: 10.1016/j.acvd.2016.04.002.

[34] XU S W, ILYAS I, LITTLE P J, et al. Endothelial dysfunction in atherosclerotic cardiovascular diseases and beyond: from mechanism to pharmacotherapies [J]. Pharmacol Rev, 2021, 73(3): 924-967. DOI: 10.1124/pharmrev.120.000096.

[35] KONG Xiangfeng, WANG Ping, CHEN Ming. Relationship between lipoprotein(a) and carotid atherosclerosis, fibrinogen, and D-dimer in patients with cerebral infarction [J]. Journal of Chongqing Medical University, 2011, 36(9): 1101-1102. DOI: 10.13406/j.cnki.cyxb.2011.09.027.

[36] ZHANG Wei, XI Yan, SUN Huijun, et al. Relationship between lipoprotein A and thrombosis and atherosclerosis [J]. China Journal of Modern Medicine, 2007, 17(20): 2500-2502, 2505. DOI: 10.3969/j.issn.1005-8982.2007.20.019.

[37] DONG H L, CHEN W, WANG X Y, et al. Apolipoprotein A1, B levels, and their ratio and the risk of a first stroke: a meta-analysis and case-control study [J]. Metab Brain Dis, 2015, 30(6): 1319-1330. DOI: 10.1007/s11011-015-9732-7.

[38] ZHOU Zhongliang, FAN Xiaojing. Quality and improvement strategies of primary medical and health services in western China [J]. Journal of Xi' an Jiaotong University (Social Sciences Edition), 2023, 43(6): 188-200. DOI: 10.15896/j.xjtuskxb.202306016.

[39] ZHANG Yuhui, ZHAI Tiemin, CHAI Peipei, et al. Study on the accounting and prediction of treatment costs for cardiovascular and cerebrovascular diseases in China [J]. Chinese Health Economics, 2019, 38(5): 18-22. DOI: 10.7664/CHE20190505.

(Received: February 19, 2024; Revised: April 30, 2024)

(Editor: KANG Yanhui)

Note: Figure translations are in progress. See original paper for figures.