Journal of Geriatric Cardiology | JGC
RESEARCH ARTICLE
J Geriatr Cardiol 2025; 22(11): 930–940

Machine learning based model for predicting cardiovascular disease using dynamic triglyceride-glucose index: a longitudinal study cohort CHARLS database

Yi YANG[1,*], Zen-Gao YANG[2,3,*], Hong-Hong ZHANG[2,4,*], Zheng-Feng WU[2,4,*], Hai-Jing ZHAO[2,4], Yue ZHU[2,4], Yu-Han MA[2], Yu-Qi LIU[5,6,7,✉]

1. Department of Faculty of Engineering and Information Technology of University of Technology Sydney, Syndey, Australia; 2. Department of Cardiology, the Sixth Medical Centre, Chinese PLA General Hospital, Beijing, China; 3. School of Medicine, South China University of Technology, Guangzhou, China; 4. Medical School of Chinese PLA, Chinese PLA General Hospital, Beijing, China; 5. National Key Laboratory of Kidney Diseases, Beijing, China; 6. Department of Cardiology & National Clinical Research Center of Geriatric Disease, Beijing, China; 7. Beijing Key Laboratory of Chronic Heart Failure Precision Medicine, Beijing, China

[*]The authors contributed equally to this manuscript

✉ Correspondence to: ametuofo980869@163.com

https://doi.org/10.26599/1671-5411.2025.11.006

ABSTRACT

Background Cardiovascular disease (CVD) remains a major health challenge globally, particularly in aging populations. Using data from the China Health and Retirement Longitudinal Study (CHARLS), this study examines the Triglyceride-glucose (TyG) index dynamics, a marker for insulin resistance, and its relationship with CVD in Chinese adults aged 45 and older.

Methods This reanalysis utilized five waves of CHARLS data with multistage sampling. From 17,705 participants, 5,625 with TyG index and subsequent CVD data were included, excluding those lacking 2011 and 2015 TyG data. TyG derived from glucose and triglyceride levels, CVD outcomes via self-reports and records. Participants divided into four groups based on TyG changes (2011–2015): low-low, low-high, high-low, high-high TyG groups.

Results Adjusting for covariates, stable high group showed a significantly higher risk of incident CVD compared to stable low group, with an HR of 1.18 (95% CI: 1.03–1.36). Similarly, for stroke risk, stable high group had a HR of 1.45 (95% CI: 1.11–1.89). Survival curves indicated that individuals with stable high TyG levels had a significantly increased CVD risk compared to controls. The dynamic TyG change showed a greater risk for CVD than abnormal glucose metabolism, notably for stroke. However, there was no statistical difference in single incidence risk of heart disease between stable low and stable high group. Subgroup analyses underscored demographic disparities, with stable high group consistently showing elevated risks, particularly among < 65 years individuals, females, and those with higher education, lower BMI, or higher depression scores. Machine learning models, including random forest, XGBoost, CoxBoost, Deepsurv and GBM, underscored the predictive superiority of dynamic TyG over abnormal glucose metabolism for CVD.

Conclusions Dynamic TyG change correlate with CVD risks. Monitoring these changes could predict and manage cardiovascular health in middle-aged and older adults. Targeted interventions based on TyG index trends are crucial for reducing CVD risks in this population.

Cardiovascular disease (CVD) is the most common cause of mortality and morbidity worldwide.[1] The number of CVD patients in China is currently as high as 330 million, including hypertension, coronary heart disease (CHD), heart failure (HF) and other chronic non-communicable diseases.[2] CVD accounts for 46.74% of the all causes of death. Two out of every five deaths occur due to CVD. In addition, China is facing the pressure of population aging and the continued prevalence of metabolic risk factors, and the burden of CVD will continue to increase.[3] The relationship between metabolic indices and CVD has long intrigued

© 2025 JGC All rights reserved; www.jgc301.com

the medical community, with the triglyceride-glucose (TyG) index recently identified as a promising surrogate marker for insulin resistance—a predictor to both type 2 diabetes and CVD.[4,5] Despite its potential, most of the current research has focused on static measurements of single TyG index,[6] disregarding dynamic changes over time might influence cardiovascular health. CVD continues to be the leading cause of global morbidity and mortality,[7] demanding improved methodologies for risk assessment and intervention. Therefore, a deeper understanding of the TyG index's longitudinal dynamics could improve cardiovascular risk stratification and preventative healthcare approaches.

However, current studies on the TyG index and CVD risk are limited by the cross-sectional or short longitudinal range,[8] which fail to evaluate the complexities of metabolic fluctuations over extended periods. To address these issues, our study utilized longitudinal data from the China Health and Retirement Longitudinal Study (CHARLS) and employed machine learning models to analyze the dynamic changes in TyG index and their association with CVD. This approach[9] allows for a detailed exploration of how dynamic changes in the TyG index over time correlate with cardiovascular risk, thereby offering a more accurate and dynamic method for predicting CVD. By incorporating a broad range of demographic, lifestyle, and clinical covariates, our study aimed to provide a risk stratification that could lead to earlier interventions and individualized treatment strategies. The innovative use of machine learning[10] in this context not only enhanced the predictive accuracy of our models but also established a rigorous framework for cardiovascular risk factor research methodology, advancing methodological precision in predictive healthcare.

Our study represents a significant advancement in CVD research by utilizing the power of machine learning to explore the dynamic change of the TyG index over time. Unlike traditional statistical methods, machine learning can manage and interpret the complexities and interactions within large datasets.[11] This innovative approach enables us to capture and analyze the fluctuations in the TyG index over time, which is a crucial factor in understanding and predicting CVD risk that previous research has often overlooked.[12] This could lead to the discovery of novel biomarkers for early detection and the development of personalized medicine strategies tailored to individual metabolic profiles. Ultimately, we utilized machine learning in this study to enhance prediction accuracy and improve the prognosis of cardiovascular healthcare patients.

METHODS

Study Design and Population

This study constitutes a reanalysis of data derived from the China Health and Retirement Longitudinal Study (CHARLS), which received ethical approval from the Biomedical Ethics Review Committee of Peking University (IRB00001052–11015). All participants provided informed consent prior to their inclusion in the study.[13] CHARLS is a nationally representative, longitudinal survey targeting the Chinese population aged 45 years and older. This survey is based on a multistage probability sampling design and has publicly released data across five waves of follow-up,[14] occurring in 2011 (wave 1), 2013 (wave 2), 2015 (wave 3), 2018 (wave 4), and 2020 (wave 5). Due to 2013 survey without blood test data, we used 2011 and 2015 TyG dynamic triglyceride-glucose index.

Figure 1 showed the selection process for the study cohort. Initially, the CHARLS database contained records for 17,705 participants. For the purposes of this analysis, we included 7947 participants who had a recorded TyG index and subsequent cardiovascular disease (CVD) data available from the 2011 wave. To assess the impact of dynamic changes in the TyG index on CVD outcomes, we excluded 2322 participants who lacked TyG index data in the 2015 wave.

Assessment of TyG Index, Abnormal Glucose Metabolism and Cardiovascular Disease

In CHARLS, TyG index and CVD were determined based on blood test and self-reported questionnaires respectively. The TyG-index was defined as follows: TyG-index = ln [(Tg[mg/dL] × fasting glucose[mg/dL])/2]. Based on the TyG values of 2011 and 2015, values above 8.4 are classified as the high TyG group. Based on two dynamic changes, they were divided into four groups: group 1: low-low TyG group; group 2: low-high TyG group; group 3: high-low TyG group; group 4: high-high TyG group.

The dynamic TyG index base for group assignment can be seen in statistical analysis. People who reported being diagnosed with diabetes or high blood sugar (including impaired glucose tolerance and elevated fasting

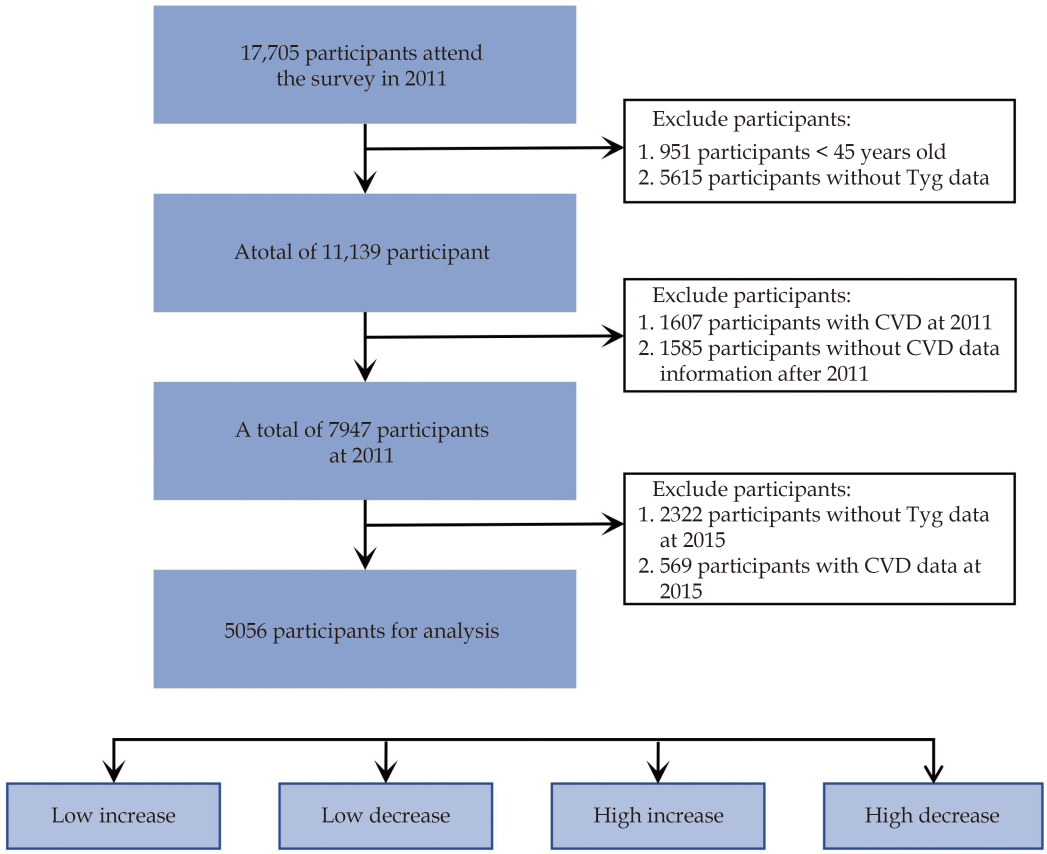

Figure 1 Flow chart.

blood glucose) were defined to have abnormal glucose metabolism (AGM). People who reported being diagnosed with heart disease (including angina, heart attack, congestive heart failure, and other heart problems) or/and stroke were defined to have CVD. The criteria for determining CVD amd AGM are detailed in the CHARLS questionnaire (https://charls.charlsdata.com/). In this study, CVD includes heart disease and/or stroke. The endpoints of follow-up were the initial incidence of CVD.

Ascertainment of Covariates

The covariates included age, sex, rural residence, marital status, education level, current smoke, current drinking, body mass index (BMI), depression score, sleep, hypertension, diabetes, total cholesterol (TC), HDL cholesterol (HDL-C), LDL cholesterol (LDL-C), uric acid (UA), creatinine, blood urea nitrogen (BUN), C-reactive protein (CRP). Marital status was divided into two groups: currently married (including married with spouse present, and married but temporarily not living with spouse due to work or other reasons) and other marital status (separated, divorced, unmarried, widowed, or cohabitated). Education was divided into two levels: primary school or lower and secondary school or higher. Smoking status was categorized as current smokers or not. Similarly, drinking status was categorized as current drinkers or not.

Statistical Analyses

The mean and standard deviation (SD) of continuous variables, as well as percentage of categorical variables are used to describe basic features. Continuous variables were compared by Kruskal-Wallis rank sum test, and categorical variables were compared by Chi-square test or Fisher exact probability test. Missing rates of covariates, including marital status (0.08%), education (0.14%), current drinker (0.42%), current smoker (1.68%), depression score (3.42%), sleep (3.78%), BMI (0.22%), LDL-C (0.22%), creatinine (0.04%), systolic blood pressure (11.06%), diastolic bloood pressure (11.06%) were summarized in Supplementary Table 1. The missing data of covariates were imputed using the multiple imputation method of randomforest (R package "Mice"). All continuous variables exhibit non-normal distributions. Reli-

ably, the baseline characteristics of participants using data not being imputed (see Supplementary Table 2) were similar with Table 1.

The X-tile software has been extensively employed to determine optimal cutpoints for continuous variables, proving particularly effective for the TyG index. Utilizing X-tile, we established the optimal cutpoint at 8.4 based on the 2011 TyG index data, categorizing it into low and high groups, which shows significant statistical difference with $P < 0.001$ (Figure 2A). Similarly, the 2015 TyG index was divided into low and high groups using the same cutpoint of 8.4. We further analyzed the dynamic TyG status, classifying it into four categories: low-low group, low-high group, high-low group and high-high group, based on TyG indices from 2011 and 2015.

The Cox-lasso method was used to screen variables, enhancing model performance while minimizing overfitting. Multivariable Cox proportional hazard models estimated the risk of CVD incidents. We assessed three models: Model I was crude; Model II adjusted for age, sex, smoke, depression, sleep, LDL-C, CRP, BMI, SBP, DBP; Model III adjusted for Model II + marital status, TC, HDL-C, UA and BUN (all these covariates were obtained by Cox-lasso method).

Table 1 Baseline characteristics.

Characteristics	Total	Group 1	Group 2	Group 3	Group 4	*P-value*
Number	5625	1183	882	732	2828	
Age, yrs	58.19 (8.49%)	58.79 (8.89%)	57.63 (8.70%)	59.15 (8.61%)	57.87 (8.19%)	<0.001
Females	2555 (45.42%)	647 (54.69%)	406 (46.03%)	367 (50.14%)	1135 (40.13%)	<0.001
Current married	5074 (90.20%)	1056 (89.26%)	799 (90.59%)	648 (88.52%)	2571 (90.91%)	0.153
Rural	3782 (67.24%)	889 (75.15%)	582 (65.99%)	524 (71.58%)	1787 (63.19%)	<0.001
Drink	1909 (33.94%)	464 (39.22%)	313 (35.49%)	272 (37.16%)	860 (30.41%)	<0.001
Smoke	1692 (30.08%)	418 (35.33%)	283 (32.09%)	255 (34.84%)	736 (26.03%)	<0.001
Education						
Primary school or lower	3913 (69.56%)	831 (70.25%)	610 (69.16%)	531 (72.54%)	1941 (68.64%)	
Secondary school or higher	1712 (30.44%)	352 (29.75%)	272 (30.84%)	201 (27.46%)	887 (31.36%)	0.208
BMI, kg/m^2	23.61 (3.82%)	21.97 (3.33%)	23.27 (3.74%)	22.77 (3.72%)	24.62 (3.75%)	<0.001
Depression, score	8.21 (6.19%)	8.67 (6.19%)	7.74 (6.21%)	9.18 (6.42%)	7.91 (6.09%)	<0.001
Sleep, h	6.38 (1.86%)	6.35 (1.88%)	6.36 (1.82%)	6.36 (1.88%)	6.41 (1.86%)	0.62
BUN	15.61 (4.36%)	15.89 (4.66%)	16.15 (4.52%)	15.39 (4.18%)	15.37 (4.20%)	<0.001
Creatine	0.76 (0.18%)	0.76 (0.19%)	0.76 (0.17%)	0.77 (0.18%)	0.77 (0.18%)	0.457
TC	193.12 (38.10%)	179.44 (32.21%)	186.54 (34.87%)	187.28 (35.47%)	202.42 (39.53%)	<0.001
HDL-C	51.20 (15.32%)	60.20 (15.39%)	56.99 (13.97%)	51.52 (14.50%)	45.56 (13.28%)	<0.001
LDL-C	115.77 (34.51%)	108.13 (28.28%)	117.84 (30.43%)	109.94 (31.93%)	119.84 (37.83%)	<0.001
CRP	2.44 (6.79%)	2.38 (8.16%)	2.60 (7.14%)	2.17 (7.87%)	2.49 (5.65%)	<0.001
UA	4.37 (1.23%)	4.23 (1.17%)	4.16 (1.15%)	4.33 (1.26%)	4.51 (1.25%)	<0.001
SBP	128.21 (20.75%)	124.48 (20.53%)	125.97 (20.30%)	127.26 (20.84%)	130.71 (20.63%)	<0.001
DBP	75.00 (12.06%)	72.53 (11.82%)	74.20 (11.87%)	73.80 (12.05%)	76.60 (11.98%)	<0.001
Abnormal glucose metabolism	3044 (54.12%)	430 (36.35%)	369 (41.84%)	405 (55.33%)	1840 (65.06%)	<0.001
Cardiovascular disease	1658 (29.48%)	290 (24.51%)	245 (27.78%)	207 (28.28%)	916 (32.39%)	<0.001
Heart disease	1268 (22.54%)	228 (19.27%)	195 (22.11%)	163 (22.27%)	682 (24.12%)	0.01
Stroke	467 (8.30%)	74 (6.26%)	62 (7.03%)	51 (6.97%)	280 (9.90%)	<0.001

Continuous variables were expressed as mean ± SD in case of normal distribution and compared between two groups by Kruskal Wallis rank sum test. If the count variable had a theoretical number < 10, Fisher's exact probability test was used. Categorical variables are presented as counts (percentages) and compared by Chi-square test. Group 1: low-low group; Group 2: low-high group; Group 3: high-low group; Group 4: high-high group. BMI: body mass index; BUN: blood urea nitrogen; CRP: C-reactive protein; CVD: cardiovascular disease; DBP: diastolic blood pressure; HDL-C: high-density lipoprotein cholesterol; LDL-C: low-density lipoprotein cholesterol; UA: uric acid; SBP: systolic blood pressure.

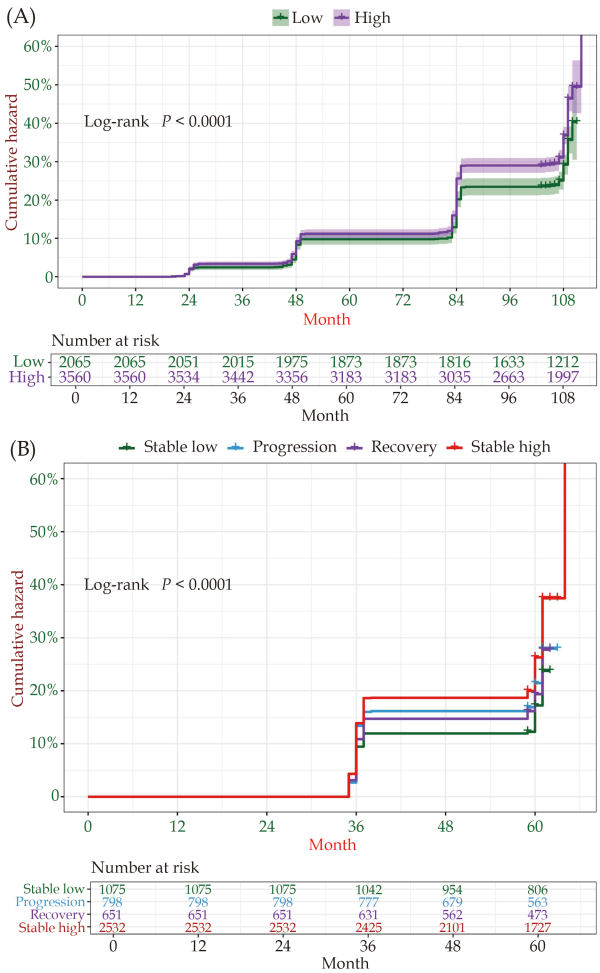

Figure 2 Nelson-Aalen cumulative risk curve for TyG index. (A): the survival curve for low and high TyG groups (2011 wave) according cutpoint at 8.4; (B): the survival curve for four groups based on TyG 2011 and TyG 2015 (group 1: low-low TyG group; group 2: low-high TyG group; group 3: high-low group; group 4: high-high group). TyG: triglyceride-glucose.

In addition, a machine-learning approach was used to develop a CVD model. Initially, the optimal model was selected based on the unadjusted parameters through three-fold cross-verification benchmark test involving Coxph, Kaplan-Meier, Random Forest, XGBoost, CoxBoost, DeepSurv, and Gradient Boosting Machine (GBM). Random Forest exhibited best performance in the concordance index (C-index). The data were then divided into a training set (70%) and a test set (30%) for model construction. We conducted hyperparameter tuning via three-fold cross-validation on the training set with a random search method to determine the optimal hyperparameter configuration evaluated by C-index. After applying the optimal hyperparameters (ntree = 581, mtry = 1, nodesize = 7) to the random forest model, we retrained it on the training dataset and subsequently evaluated its performance using the test dataset. The model's effectiveness was assessed using the C-index, Brier score, recall, and D-calibration. Machine-learning features included age, sex, smoking, depression, sleep, LDL-C, CRP, BMI, SBP, DBP, marital status, TC, HDL-C, UA, BUN and TyG or AGM.

Sensitivity analyses were performed to ensure the validity and consistency of the results. Firstly, to verify the consistency of the CVD, we examined the association of dynamic TyG status with heart disease and stroke separately. Secondly, to assess the robustness of our results, subgroup analyses were conducted with adjustments for covariates including age, sex, marital status, rural or urban living, education level, smoking status, alcohol consumption, BMI, sleep and depression score. Age groups were delineated as < 65 years and ≥ 65 years, BMI was categorized into < 24 kg/m^2 and ≥ 24 kg/m^2, and sleep duration was divided into < 6.5 h and ≥ 6.5 h. All statistical analyses were performed using R software (Version 4.3.0). *P* values were two-sided, and a value of $P < 0.05$ was considered statistically significant.

RESULTS

Baseline Characteristics Based on Dynamic Changes of TyG

This study involved 5,056 participants divided into four groups based on the dynamic change of TyG: Group 1 (low-low group, 1,075 people), Group 2 (low-high, 798 people), Group 3 (high-low group, 651 people), and Group 4 (high-high, 2532 people, Table 1). Baseline characteristics included age, sex, marital status, rural residency, lifestyle factors such as drinking and smoking, education levels, BMI, and various biochemical markers like cholesterol levels and C-reactive protein. Significant differences were observed among these groups. Notably, the prevalence of cardiovascular diseases was highest in Group 4 and lowest in Group 1. This showed the complexity of CVD incidence in different groups based on dynamic change of TyG, with significant variations in demographic, lifestyle, and metabolic factors impacting overall health outcomes. Furthermore, we also conducted abnormal glucose metabolism baseline feature analysis solely, and the baseline features remained consistent before and after data imputation (Supplemental Ta-

ble 3&4). Similarly, participants with abnormal glucose metabolism have higher CVD incidence (see Supplemental Table S3&4) with significant statistical difference ($P < 0.001$).

Increased CVD and Stroke Risk Associated with Stable High TyG Index

Table 2 showed the association between dynamic TyG status and the risk of incident CVD. The results across Model 1, Model 2, and Model 3 are consistent. After adjusting for covariates, the group 4 showed a significantly increased risk of incident CVD compared to the group 1, with a hazard ratio (HR) of 1.36 (95% CI 1.12–1.64). A similar trend was observed for heart disease and stroke risk, with the group 4 having an HR of 1.33 (95% CI: 1.06–1.67) and 1.48 (95% CI: 1.07–2.04) respectively (refer to Supplementary Table 5 & Supplementary Table 6). Similarly, the survival curves also show that the group with a high TyG has a significantly higher risk of CVD compared to the control group, with the dynamic TyG grouping curves indicating that the stable high TyG group has a significantly higher CVD risk than the stable low TyG group (Figure 2A and 2B). The risk of CVD associated with a stable high TyG index exceeds that of abnormal glucose metabolism and is even more pronounced for stroke risk.

Sensitivity Analyses of Dynamic TyG and CVD

Crude subgroup analysis revealed significant differences in CVD risk associated with dynamic TyG status (Supplementary Table 7–9). Among subgroups groups by features including age, sex, address, depression, education level, smoking, drinking, and sleep, higher risks were evident in Groups 4 ($P < 0.05$, respectively). BMI less than 24 kg/m^2 showed elevated risk in Group 4 (HR = 1.50, 95% CI: 1.21–1.84). Individuals with other marital status in Groups 4 had increased risk (HR = 1.68 95% CI: 1.40–2.01). In subgroup analyses without adjusting for confounding factors, Group 4 consistently showed robust increases in CVD risk, including heart disease and stroke.

In addition, subgroup analysis adjusted for covariates showed significant higher risk in group 4 for age < 65 (HR = 1.46, 95% CI: 1.16–1.83, Supplemental Table 10–12), rutal (HR = 1.36, 95% CI: 1.08–1.70). Other notable associations included higher risk in Group 4 for those with sleep time < 6.5 (HR = 1.40, 95% CI: 1.08–1.81), BMI < 24 kg/m^2 (HR = 1.32, 95% CI: 1.04–1.69), other marital status (HR = 1.46, 95% CI: 1.19–1.79) and depression score ≥10 (HR = 1.50, 95% CI: 1.11–2.02). Subgroups grouped by features of sex and education level also showed higher risk in Group 4. It is noteworthy that in the subgroup analyses adjusted for confounding factors, group 4 primarily led to an increased risk of stroke rather than heart disease. These results highlight differential impacts of demographic and lifestyle factors on CVD risk across dynamic TyG status subgroups.

COX-Lasso Regression Screening for Model Variables to Construct the Predictive Model

We screened out 16 variables through Cox-Lasso regression, including age, sex, smoking, depression, sleep, LDL-C, CRP, BMI, SBP, DBP, marital status, TC, HDL-C, UA, BUN, and dynamic TyG status or abnormal glucose metabolism (AGM) (Figure 3). seven variables showed no statistically significant differences after adjusting for confounding factors and were therefore excluded from the model, leaving a final set of 9 variables for plotting

Table 2 Cox regression analysis for relationship between TyG index, diabetes and CVD.

Group	CVD	Model 1	Model 2	Model 3
Stable Low	1183 (21.0%)	Ref	Ref	Ref
Progression	882 (15.7%)	1.15 (0.97–1.37), $P = 0.100$	1.13 (0.95–1.34), $P = 0.173$	1.11 (0.93–1.31), $P = 0.250$
Recovery	732 (13.0%)	1.18 (0.99–1.41), $P = 0.066$	1.07 (0.90–1.28), $P = 0.450$	1.06 (0.89–1.27), $P = 0.497$
Stable High	2828 (50.3%)	1.39 (1.22–1.58), $P < 0.001$	1.21 (1.06–1.39), $P = 0.006$	1.18 (1.03–1.36), $P = 0.018$
Abnormal glucose metabolism (No)	2581 (45.9%)	Ref	Ref	Ref
Abnormal glucose metabolism (Yes)	3044 (54.1%)	1.24 (1.13–1.37), $P < 0.001$	1.15 (1.04–1.27), $P = 0.005$	1.15 (1.04–1.27), $P = 0.008$

Model 1: crude; Model 2 adjusted for age, marital status, alcohol consumption, smoking, depression score, sleep, BMI, SBP, DBP, and education; Model 3 adjusted for age, marital status, alcohol consumption, smoking, depression score, sleep, BMI, SBP, DBP, education, LDL-C, CRP, and BUN. BMI: body mass index; BUN: blood urea nitrogen; CRP: C-reactive protein; CVD: cardiovascular disease; DBP: diastolic blood pressure; HDL-C: high-density lipoprotein cholesterol; LDL-C: low-density lipoprotein cholesterol; UA: uric acid; SBP: systolic blood pressure.

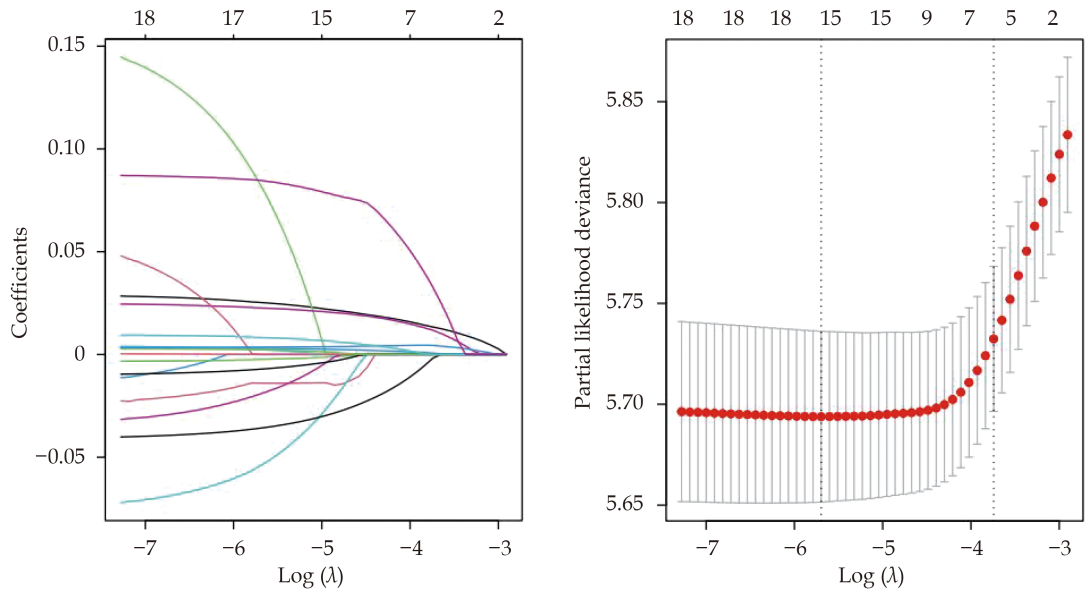

Figure 3 Cox-lasso method for variable screen. According to Cox-Lasso regression, the optimal set of variables includes 16 variables, including age, sex, smoking, depression, sleep, LDL-C, CRP, BMI, SBP, DBP, marital status, TC, HDL-C, UA and BUN, and dynamic TyG status or AGM. AGM: abnormal glucose metabolism; BMI: body mass index; BUN: blood urea nitrogen; CRP: C-reactive protein; DBP: diastolic blood pressure; HDL-C: high-density lipoprotein cholesterol; LDL-C: low-density lipoprotein cholesterol; UA: uric acid; SBP: systolic blood pressure.

the nomogram (Figure 4). The results showed that the CVD incidence risk score for TyG is higher than that for glucose metabolism according to the nomogram and exhibits a similar pattern for heart disease and stroke (refer to Supplementary Figure S1 and Supplementary Figure S2).

TyG Index and AGM in Predicting CVD by Machine Learning

To explore the role of TyG index in predicting CVD compared to abnormal glucose metabolism (AGM) through different models in machine learning. Initially, we compared the predictive performance of seven machine learning methods, including Coxph, Kaplan, Random Forest, XGBoost, CoxBoost, DeepSurv, and GBM. Based on the unadjusted parameters through three-fold cross-verification benchmark test, the results showed that Coxph, Random Forest, and CoxBoost exhibited relatively higher predictive performance, with Random Forest performing the best (Figure 5).

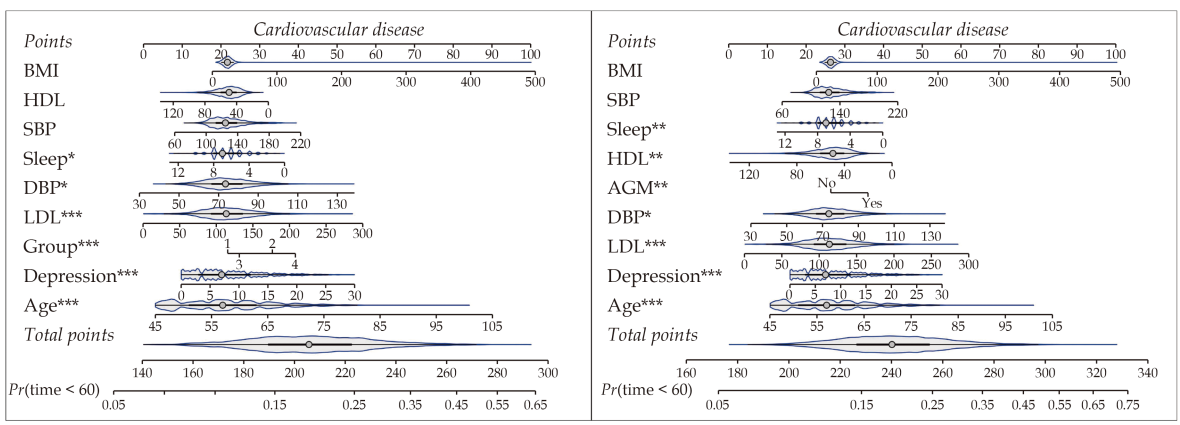

Figure 4 Nomogram for TyG index and abnormal glucose metabolism in CVD. (A): Dynamic TyG group; (B): AGM. AGM: abnormal glucose metabolism; CVD: cardiovascular disease; BMI: body mass index; CVD: cardiovascular disease; DBP: diastolic blood pressure; HDL-C: high-density lipoprotein cholesterol; LDL-C: low-density lipoprotein cholesterol; SBP: systolic blood pressure; TyG: triglyceride-glucose.

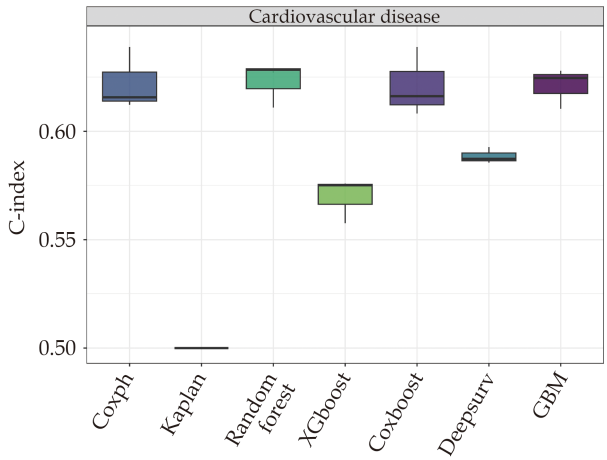

Figure 5 Benchmark test for machine learning.

To further compare the contributions of dynamic TyG group, TyG change value, TyG 2015 baseline, AGM, and other variables in the Random Forest model, we calculated the c-index values for each variable in the model. The data was split into training set (70%) and test set (30%) at random. The two data set are comparable, and there is no statistically significant differences observed (P = 0.910, see Supplementary Figure S3). The analysis showed that the dynamic TyG group exhibit higher importance than AGM, its difference value (Dta) and TyG 2015 baseline (Figure 6). Figure 7 demonstrates that dynamic TyG group and its difference value outperform abnormal glucose metabolism in predicting CVD, including higher C-index, lower brier score, and higher D-calibration which indicate excellent performance of the model. Specifically, the model accuracy for dynamic group is superior, characterized by a higher C-index of 0.862, recall rate of 0.662 and improved D-calibration of 485.535.

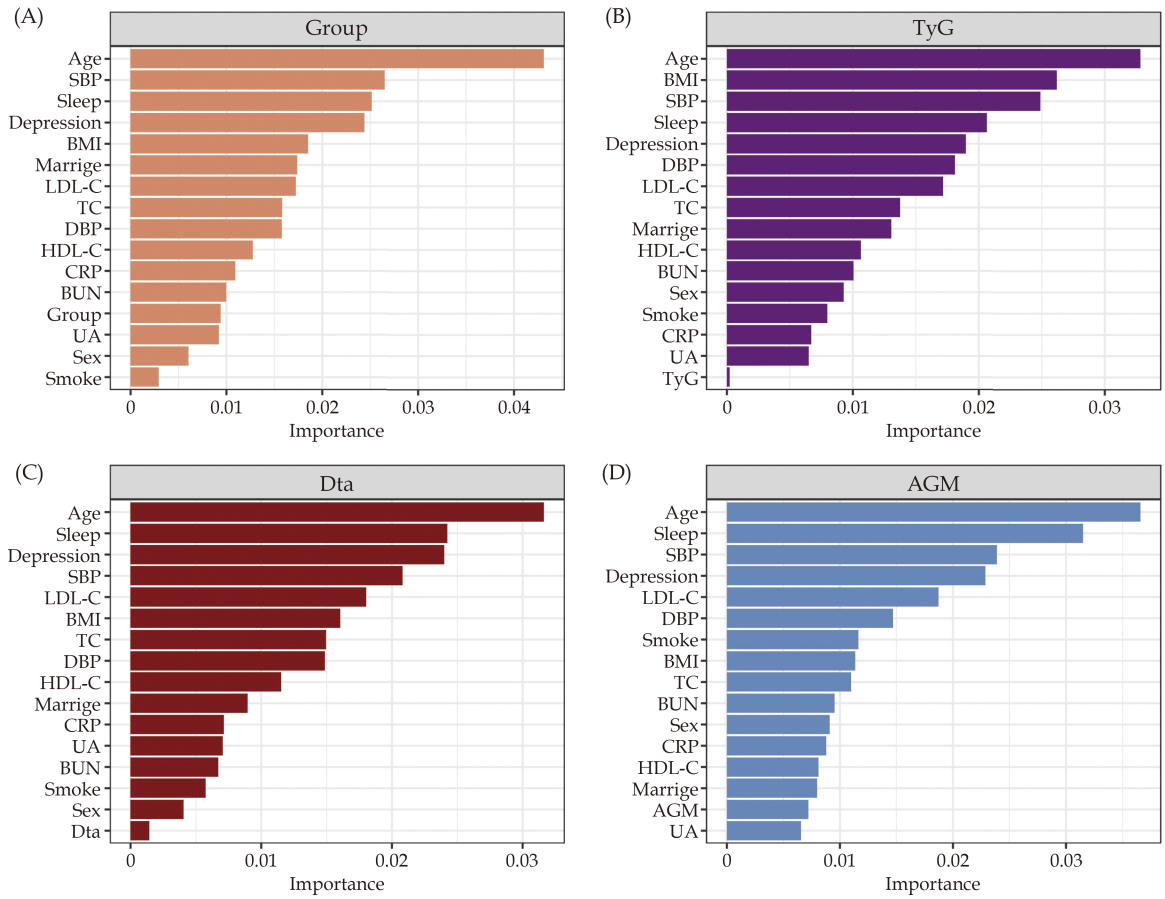

Figure 6 Variable importance rank in random forest model. (A): Dynamic TyG group; (B): group replaced by TyG index 2015; (C): group replaced by difference value (Dta= TyG 2015 - TyG 2011); (D): Group replaced by AGM (abnormal glucose metabolism). AGM: abnormal glucose metabolism; CVD: cardiovascular disease; BMI: body mass index; CVD: cardiovascular disease; DBP: diastolic blood pressure; HDL-C: high-density lipoprotein cholesterol; LDL-C: low-density lipoprotein cholesterol; SBP: systolic blood pressure; TyG: triglyceride-glucose.

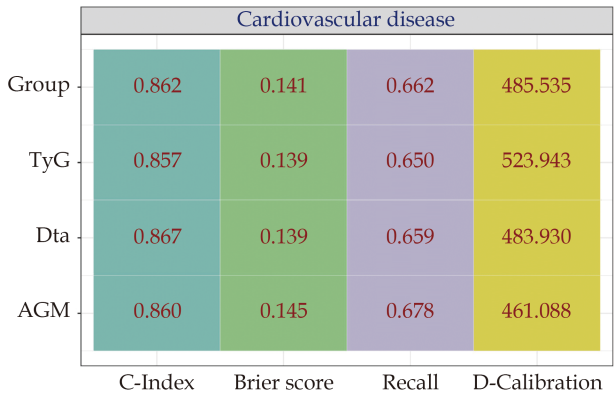

Figure 7 The CVD machine learning model evaluation. AGM: abnormal glucose metabolism; CVD: cardiovascular disease; TyG: triglyceride-glucose.

DISCUSSION

Previous research has established that the TyG index, as a biomarker of insulin resistance,[15] is closely associated with the risk of CVD.[16,17] Recent research has increasingly highlighted the TyG index as a pivotal marker for assessing the risk of multiple CVD.[18] The TyG index, a surrogate marker of insulin resistance, has shown a strong correlation with the development and progression of atherosclerosis,[19] particularly in populations at risk for or with established coronary artery disease (CAD).[20] One notable advancement in this area is the utilization of the TyG index for improving cardiovascular risk stratification models.[21] Incorporating the TyG index alongside traditional risk factors—like glycated hemoglobin (HbA1c) and cholesterol levels—enhances the predictive accuracy for major adverse cardiovascular events (MACEs).[22] This integrated approach helps in identifying high-risk individuals who may benefit from early and more aggressive therapeutic interventions.[23] Moreover, the clinical relevance of the TyG index extends beyond risk prediction. It also provides insights into potential therapeutic targets for reducing insulin resistance, a known contributor to CVD pathology.[24] Ongoing research continues to explore the molecular mechanisms linking the TyG index with cardiovascular health, offering promising directions for future therapeutic strategies.[25] Previous researches and analysis have shown that combining the TyG index with other predictive indicators, such as WhtR, BMI CVAI *et al*, can better predict the occurrence of CVD.[26,27]

This study innovatively investigates the dynamic changes of the TyG index and their association with CVD using the longitudinal CHARLS cohort. Unlike previous cross-sectional studies, this approach allows for an understanding of how TyG index trajectories impact CVD risk over time. By analyzing these temporal changes, our research provides a more detailed prediction model for CVD, enhancing earlier findings and offering new insights into the progression of cardiovascular risk factors. We utilized the Cox-Lasso regression and machine learning to analyze the association between the TyG index and the risk of CVD. By utilizing multiple models, we not only enhanced the predictive performance of the model and minimized overfitting but also innovatively applied machine learning in the field of cardiovascular research. This demonstrated the powerful potential and novelty of machine learning in handling large-scale biostatistical data. Studies have pointed out that the TyG index is an independent predictor of future cardiovascular events.[12,18] Consistent results across multiple models in our study further validate the reliability and effectiveness of the TyG index in predicting CVD risk. Notably, the significant association between the TyG index and CVD risk persisted even after adjusting for age, hypertension, lifestyle factors, and other biochemical markers, highlighting its significant role in the prediction of CVD. Previous studies have shown that dynamic changes in the TyG index are closely associated with hypertension,[28,29] and all-cause mortality.[30] This study employed COX-Lasso regression to construct a predictive model, identifying 9 key variables including age, blood pressure, lifestyle factors (BMI, sleep), sex, *et al* and dynamic TyG change. This model demonstrates that TyG index contributes significantly to predicting CVD and stroke risk, further highlighting its utility in risk assessment. With the explosive growth of data in medical research, traditional statistical methods face challenges in handling big data. Machine learning methods, particularly advanced statistical learning techniques like Cox-Lasso, Random forest, offer new perspectives and tools for cardiovascular disease research.[31]

Strengths and Limitations

This study had innovation and strengths in three main areas. Firstly, this study initially used CHARLS longitudinal data to explore dynamic TyG changes and CVD correlation. Secondly, it employs COX-Lasso regression to handle complex variables and multicollinearity effectively, identifying key factors like dynamic TyG, age, blood pressure, lifestyle, education, and BMI for CVD risk prediction. Thirdly, through subgroup analysis, the

study provides robust evidence supporting the correlation between dynamic TyG and CVD, while further exploring which populations are more susceptible to developing CVD. Finally, it enhances predictive accuracy using multiple machine learning models like Random Forest, CoxBoost, offering new mathod for CVD risk assessment and personalized medicine.

This study had several limitations. Firstly, due to its data source, there remains a possibility of potential confounding variables, despite efforts with multiple models and sensitivity analyses aimed at bolstering result robustness. Secondly, the reliance on data primarily sourced from a single population hampers the broad applicability of the findings. To address these limitations, future research should prioritize large-scale cohort studies spanning multiple centers and diverse populations to facilitate more generalized conclusions and broader insights into the relationship between dynamic TyG and cardiovascular health.

In conclusion, the application of the Cox-Lasso machine learning model in this study not only enhanced our understanding of cardiovascular disease risk factors but also provided new perspectives and methods for future disease prediction and management. This marks a deep application of machine learning technology in the field of cardiovascular diseases, enabling systematic improvements in the methodologies of medical research.

DISCLOSURE

Disclosure of Interest

None.

Research Funding

This work was supported by the National Natural Science Foundation of China (grant numbers 82070434, LYQ).

Ethical Approval

The China Health and Retirement Longitudinal Study was approved by the Ethics Review Committee of Peking University.

Data Availability

Data supporting the results of this study are available from official websites of China Health and Retirement Longitudinal Study (https://charls.pku.edu.cn/).

REFERENCES

[1] Timmis A, Vardas P, Townsend N, *et al*. European Society of Cardiology: cardiovascular disease statistics 2021. *Eur Heart J* 2022; 43: 716–799.
[2] Writing committee of the report on cardiovascular health and diseases in china. Report on Cardiovascular Health and Diseases in China 2021: An Updated Summary. *Biomed Environ Sci* 2022; 35: 573–603.
[3] Fang EF, Xie C, Schenkel JA, *et al*. A research agenda for ageing in China in the 21st century (2nd edition): Focusing on basic and translational research, long-term care, policy and social networks. *Ageing Res Rev* 2020: 64: 101174.
[4] Mirjalili SR, Soltani S, Heidari Meybodi Z, *et al*. An innovative model for predicting coronary heart disease using triglyceride-glucose index: a machine learning-based cohort study. *Cardiovasc Diabetol* 2023; 22: 200.
[5] Li H, Zuo Y, Qian F, *et al*. Triglyceride-glucose index variability and incident cardiovascular disease: a prospective cohort study. *Cardiovasc Diabetol* 2022; 21: 105.
[6] Dang K, Wang X, Hu J, *et al*. The association between triglyceride-glucose index and its combination with obesity indicators and cardiovascular disease: NHANES 2003–2018. *Cardiovasc Diabetol* 2024; 23: 8.
[7] Townsend N, Kazakiewicz D, Lucy Wright F, *et al*. Epidemiology of cardiovascular disease in Europe. *Nat Rev Cardiol* 2022; 19: 133–143.
[8] Guo D, Wu Z, Xue F, *et al*. Association between the triglyceride-glucose index and impaired cardiovascular fitness in non-diabetic young population. *Cardiovasc Diabetol* 2024; 23: 39.
[9] He D, Wang Z, Li J, *et al*. Changes in frailty and incident cardiovascular disease in three prospective cohorts. *Eur Heart J* 2024; 45: 1058–1068.
[10] Li C, Liu X, Shen P, *et al*. Improving cardiovascular risk prediction through machine learning modelling of irregularly repeated electronic health records. *Eur Heart J Digit Health* 2023; 5: 30–40.
[11] Barbieri S, Mehta S, Wu B, *et al*. Predicting cardiovascular risk from national administrative databases using a combined survival analysis and deep learning approach. *Int J Epidemiol* 2022; 51: 931–944.
[12] Tao LC, Xu JN, Wang TT, *et al*. Triglyceride-glucose index as a marker in cardiovascular diseases: landscape and limitations. *Cardiovasc Diabetol* 2022; 21: 68.
[13] Zhao Y, Hu Y, Smith JP, *et al*. Cohort profile: the China Health and Retirement Longitudinal Study (CHARLS). *Int J Epidemiol* 2014; 43: 61–68.
[14] Gong J, Wang G, Wang Y, *et al*. Nowcasting and forecasting the care needs of the older population in China: analysis of data from the China Health and Retirement Longitudinal Study (CHARLS). *Lancet Public Health* 2022; 7: e1005–e1013.
[15] Tahapary DL, Pratisthita LB, Fitri NA, *et al*. Challenges in the diagnosis of insulin resistance: Focusing on the role of HOMA-IR and Tryglyceride/glucose index. Diabetes Metab Syndr. 2022;16: 102581.
[16] Moon JH, Kim Y, Oh TJ, *et al*. Triglyceride-glucose index predicts future atherosclerotic cardiovascular diseases: a 16-year follow-up in a prospective, community-dwelling

cohort study. *Endocrinol Metab (Seoul)* 2023; 38: 406–417.
[17] Zhang Q, Xiao S, Jiao X, Shen Y. The triglyceride-glucose index is a predictor for cardiovascular and all-cause mortality in CVD patients with diabetes or pre-diabetes: evidence from NHANES 2001–2018. *Cardiovasc Diabetol* 2023; 22: 279.
[18] Sun C, Hu L, Li X, *et al*. Triglyceride-glucose index's link to cardiovascular outcomes post-percutaneous coronary intervention in China: a meta-analysis. *ESC Heart Fail* 2024; 11: 1317–1328.
[19] Zhuang Y, Wang Y, Sun P, *et al*. Association between triglyceride glucose-waist to height ratio and coronary heart disease: a population-based study. *Lipids Health Dis* 2024; 23: 162.
[20] Chen S, Li Z, Li H, *et al*. Novel lipid biomarkers and ratios as risk predictors for premature coronary artery disease: A retrospective analysis of 2952 patients. *J Clin Hypertens (Greenwich)* 2023; 25: 1172–1184.
[21] Li K, Hou Q, Li X, *et al*. Triglyceride-glucose index predicts major adverse cardiovascular events in patients with chronic kidney disease. *Int Urol Nephrol* 2024; 56: 2793–2802.
[22] Tai S, Fu L, Zhang N, *et al*. Impact of baseline and trajectory of triglyceride-glucose index on cardiovascular outcomes in patients with type 2 diabetes mellitus. *Front Endocrinol (Lausanne)* 2022: 13: 858209.
[23] Zhang Z, Zhao L, Lu Y, *et al*. Association between non-insulin-based insulin resistance indices and cardiovascular events in patients undergoing percutaneous coronary intervention: a retrospective study. *Cardiovasc Diabetol* 2023; 22: 161.
[24] Yin T, Chen S, Zhu Y, *et al*. Insulin resistance, combined with health-related lifestyles, psychological traits and adverse cardiometabolic profiles, is associated with cardiovascular diseases: findings from the BHMC study. *Food Funct* 2024; 15: 3864–3875.
[25] Liu D, Yang K, Gu H, *et al*. Predictive effect of triglyceride-glucose index on clinical events in patients with acute ischemic stroke and type 2 diabetes mellitus. *Cardiovasc Diabetol* 2022; 21: 280.
[26] Cui C, Qi Y, Song J, *et al*. Comparison of triglyceride glucose index and modified triglyceride glucose indices in prediction of cardiovascular diseases in middle aged and older Chinese adults. *Cardiovasc Diabetol* 2024; 23: 185.
[27] Zhang X, Wang Y, Li Y, *et al*. Optimal obesity- and lipid-related indices for predicting type 2 diabetes in middle-aged and elderly Chinese. *Sci Rep* 2024; 14: 10901.
[28] Zhao Y, Yang X, Wu Y, *et al*. Association of triglyceride-glucose index and its 6-year change with risk of hypertension: A prospective cohort study. *Nutr Metab Cardiovasc Dis* 2023; 33: 568–576.
[29] Liu R, Wang L, Zhong W, *et al*. Triglyceride glucose index combined with body mass index and its 4-year change with the risk of hypertension in middle-aged and older Chinese: A prospective cohort study. *Nutr Metab Cardiovasc Dis* 2024; 34: 1381–1388.
[30] Cheng L, Zhang F, Xue W, *et al*. Association of dynamic change of triglyceride-glucose index during hospital stay with all-cause mortality in critically ill patients: a retrospective cohort study from MIMIC IV2. *Cardiovasc Diabetol* 2023; 22: 142.
[31] Tang CX, Guo BJ, Schoepf JU, *et al*. Feasibility and prognostic role of machine learning-based FFRCT in patients with stent implantation. *Eur Radiol* 2021; 31: 6592–6604.

Please cite this article as: YANG Y, YANG ZG, ZHANG HH, WU ZF, ZHAO HJ, ZHU Y, MA YH, LIU YQ. Machine learning based model for predicting cardiovascular disease using dynamic triglyceride-glucose index: a longitudinal study cohort CHARLS database. *J Geriatr Cardiol* 2025; 22(11): 930–940. DOI: 10.26599/1671-5411.2025.11.006