Journal of Inflammation Research

Dovepress
open access to scientific and medical research

Open Access Full Text Article

ORIGINAL RESEARCH

Construction and Validation of a Predictive Model for Coronary Artery Disease Using Extreme Gradient Boosting

Zheng Zhang[1,2,*], Binbin Shao[3,*], Hongzhou Liu[2,4], Ben Huang[2,5], Xuechen Gao[1], Jun Qiu[1],
Chen Wang[1,2]

[1]Center of Clinical Laboratory, The First Affiliated Hospital of Soochow University, Suzhou, Jiangsu Province, People's Republic of China; [2]Center for Gene Diagnosis, Department of Laboratory Medicine, Zhongnan Hospital of Wuhan University, Wuhan, Hubei Province, People's Republic of China; [3]Department of Prenatal Diagnosis, Women's Hospital of Nanjing Medical University, Nanjing Women and Children's Healthcare Hospital, Nanjing, Jiangsu Province, People's Republic of China; [4]School of Clinical Medicine, The First Affiliated Hospital of Chengdu Medical College, Chengdu, Sichuang Province, People's Republic of China; [5]Department of Laboratory Medicine, The First Affiliated Hospital of Nanjing Medical University, Nanjing, Jiangsu Province, People's Republic of China

*These authors contributed equally to this work

Correspondence: Chen Wang; Jun Qiu, Center of Clinical Laboratory, The First Affiliated Hospital of Soochow University, No. 899 Pinghai Road, Suzhou, Jiangsu Province, 215006, People's Republic of China, Email 1617657224@qq.com; 13606136542@126.com

Purpose: Early recognition of coronary artery disease (CAD) could delay its progress and significantly reduce mortality. Sensitive, specific, cost-efficient and non-invasive indicators for assessing individual CAD risk in community population screening are urgently needed.

Patients and Methods: 3112 patients with CAD and 3182 controls were recruited from three clinical centers in China, and differences in baseline and clinical characteristics were compared. For the discovery cohort, the least absolute shrinkage and selection operator (LASSO) regression was used to identify significant features and four machine learning algorithms (logistic regression, support vector machine (SVM), random forest (RF) and extreme gradient boosting (XGBoost)) were applied to construct models for CAD risk assessment, the receiver operating characteristics (ROC) curve and precision-recall (PR) curve were conducted to evaluate their predictive accuracy. The optimal model was interpreted by Shapley additive explanations (SHAP) analysis and assessed by the ROC curve, calibration curve, and decision curve analysis (DCA) and validated by two external cohorts.

Results: Using LASSO filtration, all included variables were considered to be statistically significant. Four machine learning models were constructed based on these features and the results of ROC and PR curve implied that the XGBoost model exhibited the highest predictive performance, which yielded a high area of ROC curve (AUC) of 0.988 (95% CI: 0.986–0.991) to distinguish CAD patients from controls with a sensitivity of 94.6% and a specificity of 94.6%. The calibration curve showed that the predicted results were in good agreement with actual observations, and DCA exhibited a better net benefit across a wide range of threshold probabilities. External validation of the model also exhibited favorable discriminatory performance, with an AUC, sensitivity, and specificity of 0.953 (95% CI: 0.945–0.960), 89.9%, and 87.1% in the validation cohort, and 0.935 (95% CI: 0.915–0.955), 82.0%, and 90.3% in the replication cohort.

Conclusion: Our model is highly informative for clinical practice and will be conducive to primary prevention and tailoring the precise management for CAD patients.

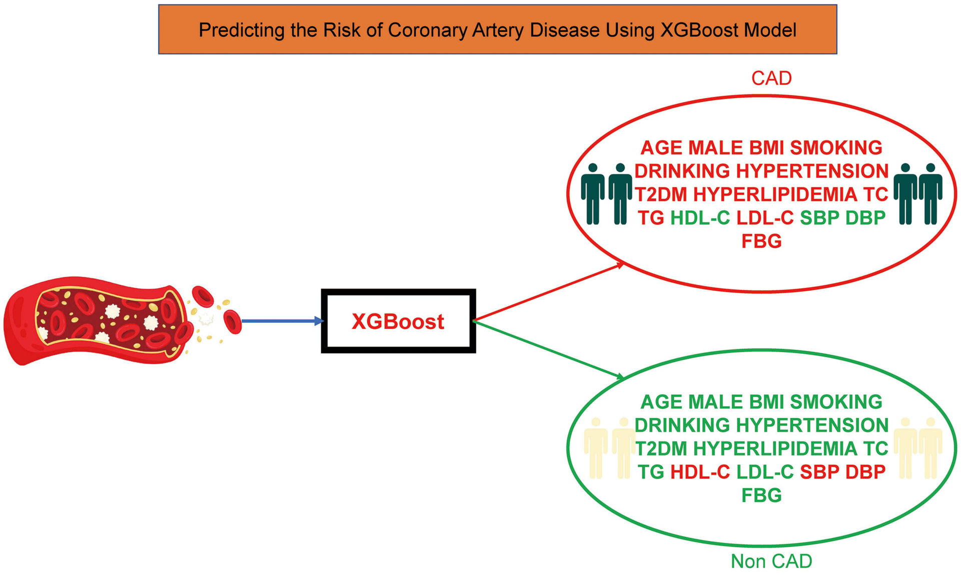
Keywords: coronary artery disease, predictive model, machine learning, XGBoost, primary prevention

Introduction

Coronary artery disease (CAD) and its complications are still the primary causes of morbidity and mortality.[1] The main etiology of CAD is atherosclerosis, a chronic vascular inflammation manifested by the accumulation of lipoprotein droplets, infiltration of various immune cells, and proliferation of smooth muscle cells, ultimately contributing to

Received: 17 February 2024
Accepted: 25 June 2024
Published: 1 July 2024

Graphical Abstract


Predicting the Risk of Coronary Artery Disease Using XGBoost Model
CAD
AGE MALE BMI SMOKING DRINKING HYPERTENSION T2DM HYPERLIPIDEMIA TC TG HDL-C LDL-C SBP DBP FBG
XGBoost
AGE MALE BMI SMOKING DRINKING HYPERTENSION T2DM HYPERLIPIDEMIA TC TG HDL-C LDL-C SBP DBP FBG
Non CAD

vascular lumen stenosis and the obstruction of blood supply to the heart.[2,3] The gold standard for CAD diagnosis is invasive coronary angiography, which can accurately evaluate the location and degree of coronary stenosis.[4] However, its routine use for population screening has been restricted by the requirement for specialized catheterization centers, experienced cardiologists, and underlying radiation exposure.[5,6] Accordingly, sensitive, specific, cost-efficient and noninvasive indicators for assessing individual CAD risk in community population screening are urgently needed.

The development and evolution of coronary atherosclerosis are regulated by various interactions between genetic and lifestyle factors.[7] Substantial evidence has verified a causal correlation between blood lipids and cardiovascular disease prevalence.[8,9] Higher circulating low-density lipoprotein cholesterol (LDL-C) and triglyceride-rich lipoproteins are correlated with increased CAD risk, whereas elevated high-density lipoprotein cholesterol (HDL-C) is associated with reduced CAD risk. In addition, genetic pedigree studies have implied that LDL-C, triglyceride, and HDL-C levels are largely determined by the individual genetic architecture. For example, rare variants in the LDL receptor (*LDLR*) and apolipoprotein B (*APOB*) genes, and common mutations in the apolipoprotein E (*APOE*) gene can increase LDL-C levels and are associated with higher CAD susceptibility. In addition to blood lipid levels, epidemiological studies have also confirmed other risk factors, including age, sex, smoking, alcohol consumption, obesity, diabetes, and hypertension.

In recent years, the rapid advancement of artificial intelligence (AI) has penetrated into various fields, particularly in medical applications.[10] Machine learning (ML) involves algorithms that are specifically geared towards identifying associations between data beyond the one-dimensional conventional statistical approaches and confers the perfect opportunity to utilize the increasingly complicated data that is accessible while improving predictions in an era of precision medicine. Furthermore, ML models have shown favorable performance in cardiovascular diagnosis, decision-making and risk prediction.[11,12] To devise and ameliorate preventive tactics for CAD, it is indispensable to recognize and accurately quantify the etiological contribution of these traditional risk factors. In this study, we aimed to evaluate the predictive value of these canonical risk factors in CAD by machine learning algorithms.

Materials and Methods

Study Population and Data Collection

This three-stage case-control design, including 3112 CAD patients and 3182 controls after removal of individuals with missing values, was retrospectively recruited from three clinical centers: a discovery cohort with 1727 CAD cases and

1756 controls from Wuhan Asia Heart Hospital between March 2016 and October 2019; a validation cohort with 1124 CAD cases and 1168 controls from Zhongnan Hospital of Wuhan University between May 2016 and December 2018; and a replication cohort with 261 CAD patients and 258 controls from Shandong Provincial Hospital between January 2017 and February 2018 (Figure 1). The first diagnosis of CAD was confirmed using coronary angiography, which showed > 50% stenosis in at least one main coronary artery or its important branches. Patients with other cardiac diseases, such as myocardial bridge, coronary artery spasm, congenital or valvular heart defects, and systemic diseases, such as severe liver or renal disease, autoimmune disease, and malignancy, were excluded. Controls were healthy individuals without cardiovascular or systemic diseases, based on physical examination and medical history. Conventional CAD risk factors, including age, sex, cigarette smoking, alcohol consumption, history of type 2 diabetes mellitus (T2DM), hyperlipidemia, and hypertension, and clinical indicators, including body mass index (BMI), fasting plasma glucose (FPG), systolic blood pressure (SBP), diastolic blood pressure (DBP), total triglycerides (TG), total cholesterol (TC), LDL-C, and HDL-C, were retrospectively collected from electronic medical records and laboratory test reports.

Machine Learning Algorithms

Logistic regression is a type of probabilistic statistical classification model, which can be used to predict the classifications of nominal variable based on some significant features. The classification is conducted by the logit function to calculate the outcome probability.

Support vector machine (SVM) is a kind of generalized linear classifiers that perform binary classification in a supervised learning manner, and its decision boundary is the maximum margin hyperplane after mapping the data into a multidimensional feature space. SVM utilizes the hinge loss function to assess the empirical venture and adds a regularization term to the solution system for structural risk optimization.

Random forest (RF) is a modified bagged tree that samples the original dataset with replacement and randomly picks out features to split at each node, the final classifications are determined by the voting results of multiple decision trees. It preserves many strengths of the decision tree and displays high predictive accuracy in disease diagnosis and risk assessment.

Extreme gradient boosting (XGBoost) is a robust ensemble machine learning algorithm based on gradient boosting decision trees. Its intention is to achieve precise classification by the iterative calculation of a weak classifier. K trees are built after the training process, and the model's variance and deviation collectively determine the prediction accuracy.

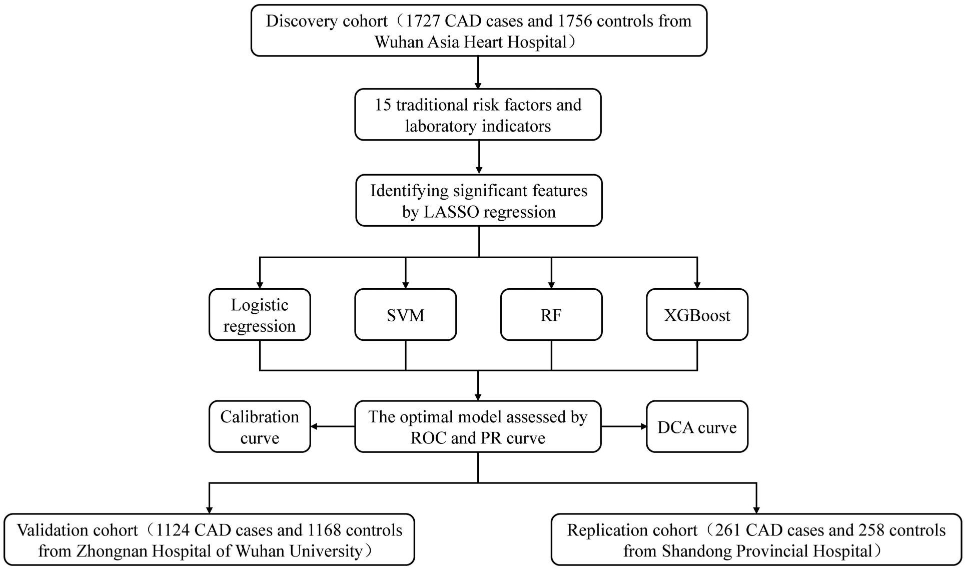

Discovery cohort（1727 CAD cases and 1756 controls from Wuhan Asia Heart Hospital）
15 traditional risk factors and laboratory indicators
Identifying significant features by LASSO regression
Logistic regression
SVM
RF
XGBoost
Calibration curve
The optimal model assessed by ROC and PR curve
DCA curve
Validation cohort（1124 CAD cases and 1168 controls from Zhongnan Hospital of Wuhan University）
Replication cohort（261 CAD cases and 258 controls from Shandong Provincial Hospital）

Figure 1 The flow chart of this study.

The variation of the model is embodied as the loss function, which is calculated by a second-order Taylor expansion. The XGBoost algorithm constantly splits the features to build trees that iteratively fit the residuals of the previous model. Each feature of a specific sample corresponding to a specific branch node of each tree was assigned a score and the cumulative score of each tree was the final predicted value. Assembling many weak learners together to form a strong ensemble learner, XGBoost exhibits the following advantages: (1) building different branches of each tree in parallel and reducing the running time, (2) dealing with missing data elastically and effectively, and (3) employing regularization to prevent overfitting.

Four aforementioned machine learning models were constructed and validated in accordance with the transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD) statement.[13]

Statistical Analysis

Qualitative parameters are shown as frequencies with percentages, and the chi-square test was used to compare differences between cases and controls. Continuous indicators are presented as mean with standard deviation, and normality distribution was assessed using the Kolmogorov–Smirnov test. Independent *t*-test and Mann–Whitney *U*-test were conducted to examine the differences between two groups with or without normal distribution, respectively. Statistical significance was set at $p < 0.05$.

For the discovery cohort, the least absolute shrinkage and selection operator (LASSO) regression analyses were performed to identify the most important features. The logistic regression could be fitted to the dataset by glm function with the family argument set to binomial and summary function was performed to check the coefficients and p-values. We utilize the e1071 package to construct linear SVM model since it has the tune.svm function which can optimize the tuning parameter and kernel function through 10-fold cross-validation. Randomforest function from randomForest package was applied to construct random forest model. The specific and optimal tree was dependent on the minimum mean of squared residuals and the number of trees constructed in this model was 498, and three variables were selected randomly to split at each node. The xgboost package was used to construct the XGBoost model and the caret package was used to tune the hyperparameters through 5-fold cross-validation. Considering both accuracy and overfitting, the hyperparameters of XGBoost were set with a maximum depth of 3, learning rate of 0.3, and iterations of 100. The predictive models were constructed by four aforementioned machine learning algorithms and the predictive performances were assessed by the receiver operating characteristic (ROC) curve and the precision-recall (PR) curve using the ROCR package.

Shapley additive explanations (SHAP) analysis is a flexible method that can provide insights into the relationship between feature observations and clinical outcomes.[14,15] When the SHAP value of some features is greater than zero, its influence on the model output is positive. This method can be used to explain the decision-making process of the machine learning model, including the relationship between observations and risk, determination of cutoff values for each feature, and ranking the features by importance. In this study, we used the SHAPforxgboost package to calculate the SHAP values for each feature and drew an SHAP summary plot to visualize the contribution of the feature to the optimal model's output. Calibration of the optimal model was evaluated by the calibration curve and the Brier scores were calculated to quantify how well the predicted results were in agreement with actual observations, with values closer to 0 indicating better calibration. The clinical usefulness was examined using decision curve analysis (DCA) by calculating the net benefit at different threshold probabilities. Finally, the discriminative abilities of the optimal model in the three cohorts were assessed by the ROC curve and the area under the ROC curve (AUC). R software (version 4.3.0) was used for all statistical analyses and figures drawings.

Results

Baseline and Clinical Characteristics of Participants in Three Study Cohorts

In the discovery cohort, CAD patients had considerably higher age; higher BMI; higher percentages of cigarette smoking and alcohol consumption; history of hypertension, hyperlipidemia, and T2DM; higher levels of TC, TG, LDL-C, and FPG; and lower HDL-C concentration than controls (Table 1). Hypertension is a universally recognized risk factor for cardiovascular disease, and the cardiovascular risk is positively correlated with blood pressure. Intriguingly, patients with CAD had lower SBP and DBP levels than the controls (Table 1), probably owing to the use of antihypertensive medications before their diagnosis. In the validation cohort, the baseline and clinical characteristics of the participants were analogous to those of the discovery cohort, except that the TC level and proportion of alcohol drinkers were similar between the two groups (Table 1). In the replication cohort, the baseline and clinical characteristics of the study population were also comparable to those of the discovery cohort, except that CAD patients had higher SBP, a higher proportion of male and lower TC levels than controls, and there were no obvious differences in DBP and LDL-C levels between the two groups (Table 1).

Model Construction and Validation

After feature selection using LASSO regression analysis, all fifteen indicators were considered to be statistically significant (Figure 2A and B). Therefore, we used four machine learning algorithms (logistic regression, SVM, RF and XGBoost) to construct a risk prediction model with all these indicators and the ROC curve (Figure 2C) and PR curve (Figure 2D) were applied to evaluate their performances. Ultimately, the XGBoost model was chosen for further analysis by virtue of its highest predictive capability.

Then, SHAP analysis was conducted to weigh the importance of each feature to the CAD risk predicted by the XGBoost model, suggesting that HDL-C had the strongest predictive value for all prediction horizons, followed by BMI, LDL-C, and TC (Figure 3A). Furthermore, SHAP values were calculated to assess the positive and negative relationships between feature observations and clinical outcomes, and the SHAP summary plot displayed the overall distribution of the influence of each feature on the model output. Male sex, advanced age, cigarette smoking, alcohol consumption, history of hypertension, hyperlipidemia, T2DM, BMI, TC, TG, LDL-C, and FPG had a positive influence on CAD risk, whereas HDL-C, SBP, and DBP had a negative influence (Figure 3B). The Brier score of the calibration curve was 0.0443, suggesting that the predicted Results were in good agreement with actual observations (Figure 4A), and DCA further indicated that the XGBoost model exhibited a better net benefit across a wide range of threshold probabilities (Figure 4B).

In the discovery cohort, this model yielded a high AUC of 0.988 (95% CI: 0.986–0.991) for distinguishing CAD patients from controls with a sensitivity of 94.6%, specificity of 94.6%, positive predictive value (PPV) of 0.946, and negative predictive value (NPV) of 0.946 (Figure 5A and B). In concordance with the discovery cohort, favorable

Table 1 Baseline and Clinical Characteristics of Participants in Three Study Cohorts

Variables	Discovery Cohort			Validation Cohort			Replication Cohort		
	Control (N = 1756)	CAD (N = 1727)	P	Control (N = 1168)	CAD (N = 1124)	P	Control (N = 258)	CAD (N = 261)	P
Age, year	59.60 ± 11.56	62.52 ± 9.66	<0.001	60.40 ± 12.08	62.19 ± 9.64	<0.001	57.26 ± 9.81	60.68 ± 9.87	<0.001
Male, n (%)	1120 (63.78)	1067 (61.78)	0.223	684 (58.56)	661 (58.81)	0.905	163 (63.18)	200 (76.63)	0.001
BMI, kg/m^2	23.79 ± 2.41	25.04 ± 3.72	<0.001	23.74 ± 2.48	25.16 ± 3.66	<0.001	24.11 ± 2.15	25.61 ± 4.37	0.001
Cigarette smoking, n (%)	501 (28.53)	690 (39.95)	<0.001	312 (26.71)	475 (42.26)	<0.001	58 (22.48)	137 (52.49)	<0.001
Alcohol drinking, n (%)	446 (25.40)	506 (29.30)	0.01	280 (23.97)	299 (26.60)	0.148	56 (21.71)	123 (47.13)	<0.001
Hypertension, n (%)	683 (38.90)	1065 (61.67)	<0.001	468 (40.07)	649 (57.74)	<0.001	86 (33.33)	161 (61.69)	<0.001
T2DM, n (%)	451 (25.68)	536 (31.04)	<0.001	290 (24.83)	362 (32.21)	<0.001	78 (30.23)	102 (39.08)	0.034
Hyperlipidemia, n (%)	411 (23.41)	509 (29.47)	<0.001	265 (22.69)	324 (28.83)	0.001	56 (21.71)	84 (32.18)	0.007
TC, mmol/L	4.58 ± 0.85	4.62 ± 1.14	0.001	4.61 ± 0.81	4.60 ± 1.05	0.135	4.81 ± 0.85	4.47 ± 1.40	<0.001
TG, mmol/L	1.26 ± 0.64	1.68 ± 1.04	<0.001	1.24 ± 0.58	1.77 ± 1.10	<0.001	1.14 ± 0.48	1.96 ± 1.29	<0.001
HDL-C, mmol/L	1.30 ± 0.25	1.01 ± 0.19	<0.001	1.31 ± 0.26	1.01 ± 0.19	<0.001	1.33 ± 0.28	1.00 ± 0.19	<0.001
LDL-C, mmol/L	2.70 ± 0.64	2.74 ± 0.88	<0.001	2.69 ± 0.61	2.76 ± 0.85	<0.001	2.89 ± 0.74	2.86 ± 1.00	0.143
SBP, mmHg	139.38 ± 25.12	131.69 ± 19.36	<0.001	140.28 ± 24.94	133.11 ± 19.72	<0.001	128.76 ± 16.67	132.22 ± 17.25	0.02
DBP, mmHg	84.79 ± 17.13	81.27 ± 11.95	<0.001	85.79 ± 17.33	80.97 ± 11.40	<0.001	80.97 ± 11.83	79.62 ± 11.48	0.348
FBG, mmol/L	5.32 ± 1.36	5.82 ± 1.99	<0.001	5.44 ± 1.65	5.91 ± 2.23	<0.001	5.41 ± 0.64	6.48 ± 1.95	<0.001

Abbreviations: BMI, body mass index; T2DM, type 2 diabetes mellitus; TC, total cholesterol; TG, total triglyceride; HDL-C, high-density lipoprotein cholesterol; LDL-C, low-density lipoprotein cholesterol; SBP, systolic blood pressure; DBP, diastolic blood pressure; FBG, fasting plasma glucose.

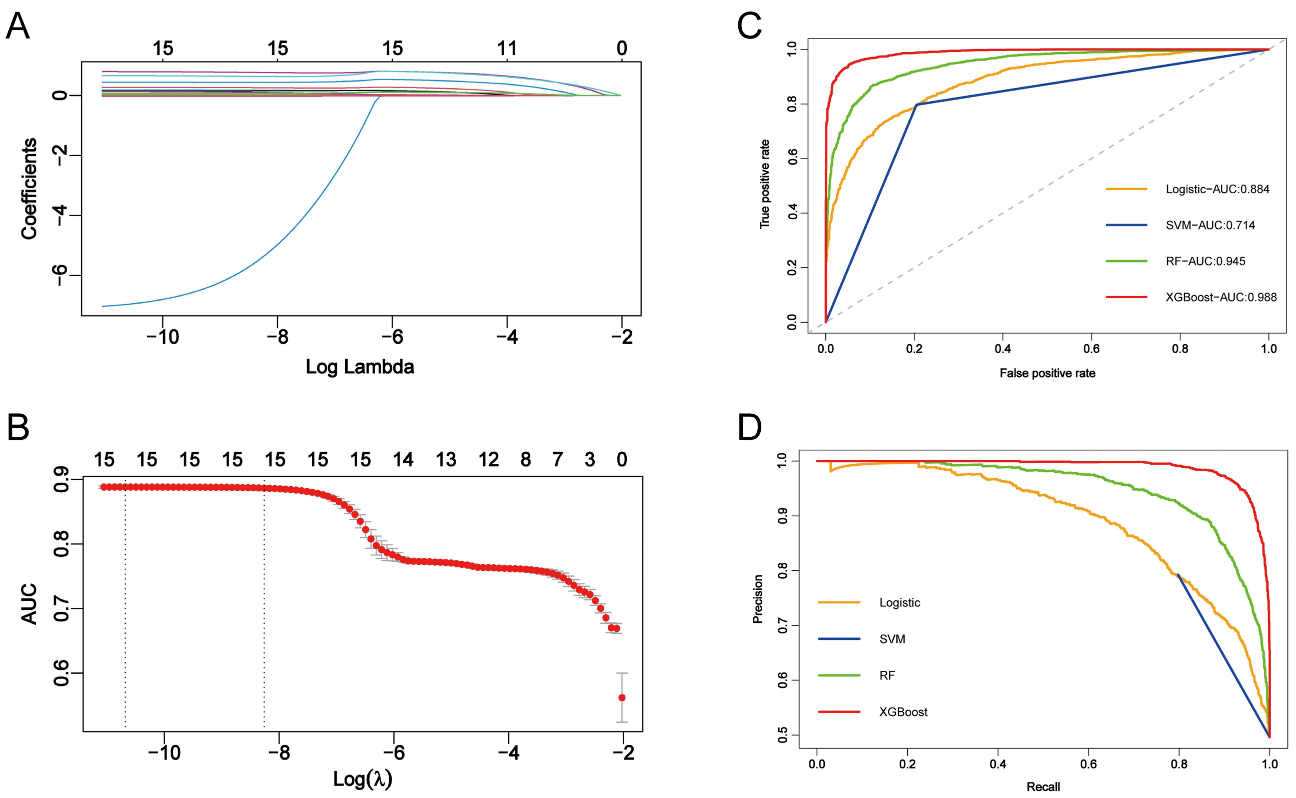


Figure 2 Features selection was conducted by the least absolute shrinkage and selection operator (LASSO) regression. (**A**) LASSO coefficients profiles (y-axis) of the fifteen high-dimensional features, the lower x-axis was the log(λ) and the upper x-axis was the average numbers of features, where λ was the tuning parameter. (**B**) Five-fold cross validation for tuning parameter optimization. The lower x-axis was the log(λ) and the upper x-axis was the average numbers of features, the area under the receiver operating characteristics with error bar was plotted against log(λ). The dotted vertical lines were drawn at the optimal values of minimum criteria and the one standard error of the minimum criteria (1se criteria). To avoid overfitting, 1se criteria (λ = 0.000259) was selected. (**C**) Four machine learning models were evaluated by the ROC curve, which plots a curve according to its true positive rate (y-axis) against its false positive rate (x-axis). The bigger area under the curve, the higher predictive capability of the model. (**D**) Four machine learning models were assessed by the PR curve, which presents the trade-off between precision (y-axis) and recall (x-axis). The larger area under the curve, the better predictive accuracy of the model.

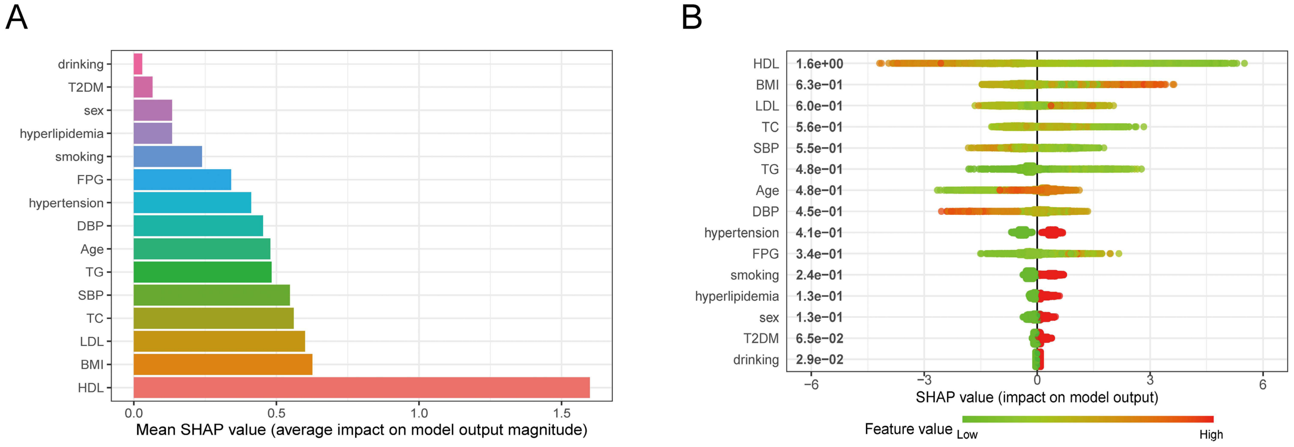


Figure 3 Interpretation of the XGBoost model by Shapley additive explanations (SHAP) analysis. (**A**) The variable important score of fifteen features in the XGBoost model, x-axis denoted the mean absolute SHAP values and y-axis represented the fifteen weighted features. (**B**) The features of each patient were described as different dots, which were colored according to the SHAP values of features for the respective individual. Red denoted a high or positive feature value, whereas green indicated a low or negative feature value. The further away a point is from the SHAP value of zero, the stronger its effects to the output.

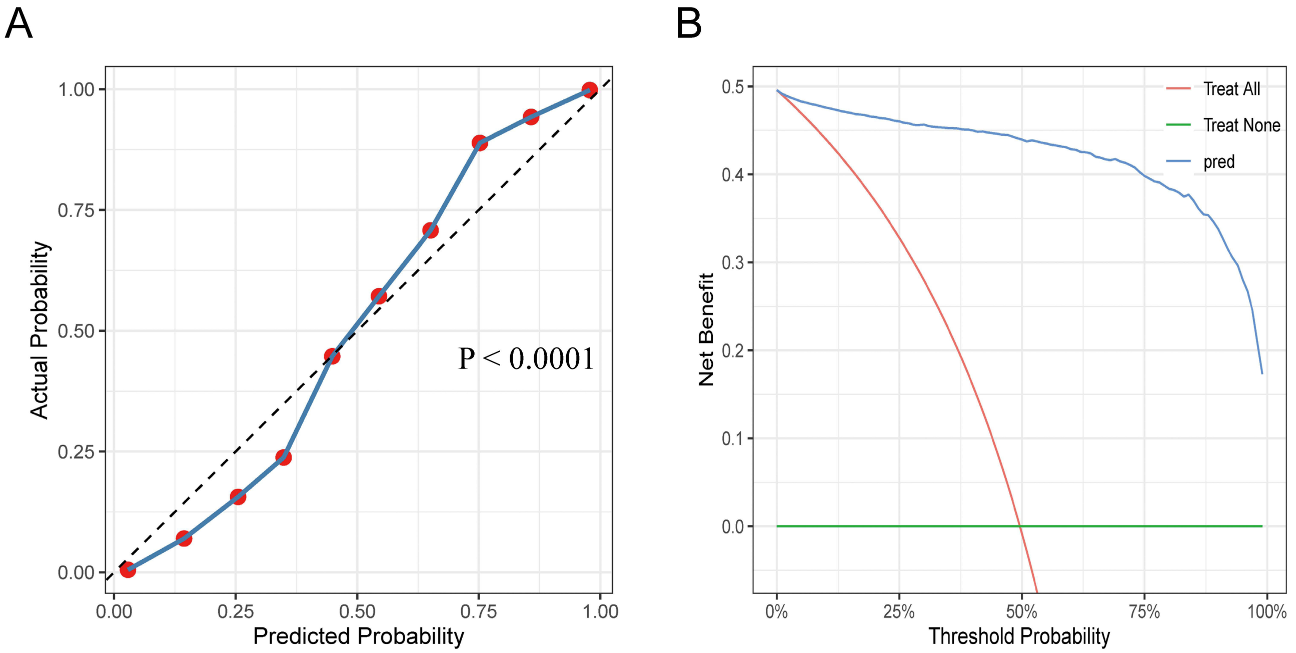


Figure 4 Evaluation of the XGBoost model in the discovery cohort. (**A**) The calibration curve of the XGBoost model. The black dotted line indicated that the observed and predicted values were perfectly matched and the blue curve exhibited that the distance between the observed and predicted values. (**B**) The decision curve analysis of the XGBoost model. The x-axis indicated the threshold probability of the outcome, while the y-axis denoted the net benefit of making a decision based on the model's prediction. The red curve indicated the net benefit of treating all patients, while the green curve represented the net benefit of not treating any patients.

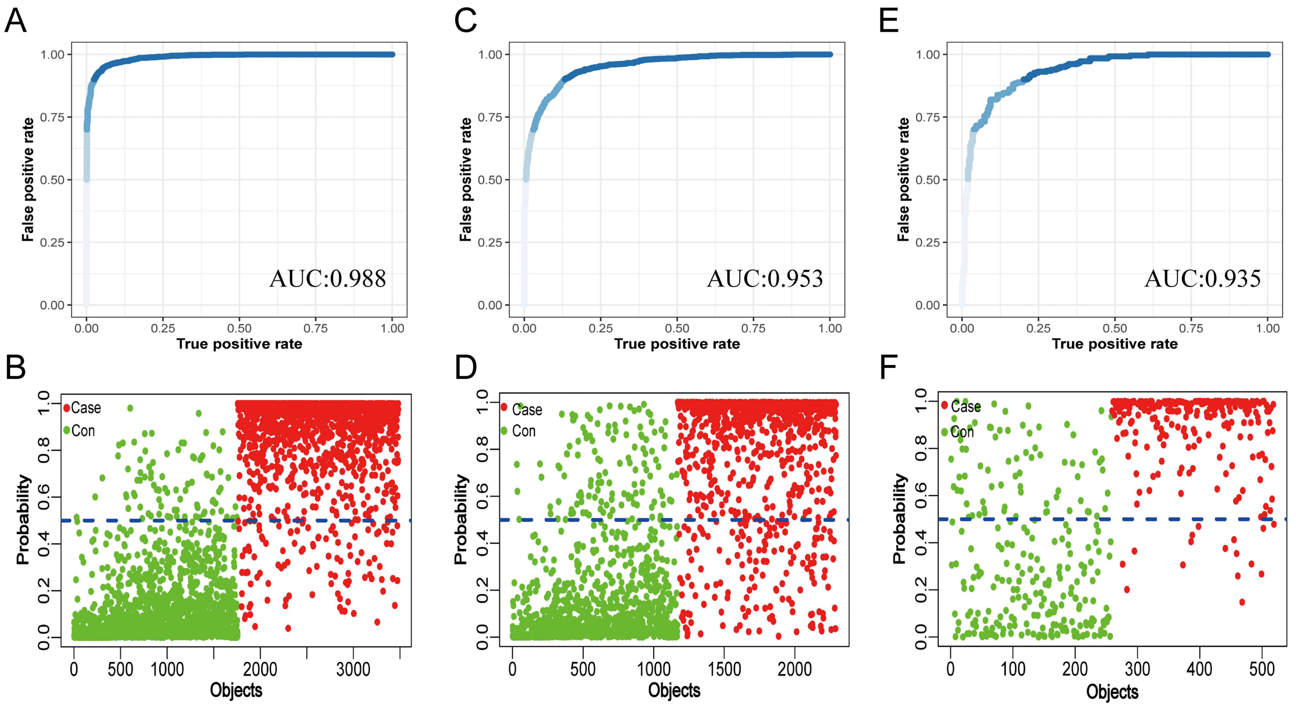


Figure 5 The discriminative performances of XGBoost model in three cohorts. (**A**) The receiver operating characteristics (ROC) curve of the XGBoost model in the discovery cohort, the area under the curve (AUC) was 0.988. (**B**) Risk probability plot of the XGBoost model in the discovery cohort. The red dots represented CAD patients, the green dots denoted controls and the blue dotted line indicated a probability of 0.5. (**C**) The ROC curve of the XGBoost model in the validation cohort, the AUC was 0.953. (**D**) Risk probability plot of the XGBoost model in the validation cohort. The red dots represented CAD patients, the green dots denoted controls and the blue dotted line indicated a probability of 0.5. (**E**) The ROC curve of the XGBoost model in the replication cohort, the AUC was 0.935. (**F**) Risk probability plot of the XGBoost model in the replication cohort. The red dots represented CAD patients, the green dots denoted controls and the blue dotted line indicated a probability of 0.5.

discriminatory performance was also externally validated by two other cohorts, with an AUC, sensitivity, specificity, PPV, and NPV of 0.953 (95% CI: 0.945–0.960), 89.9%, 87.1%, 0.870, and 0.899, respectively, in the validation cohort (Figure 5C and D), and 0.935 (95% CI: 0.915–0.955), 82.0%, 90.3%, 0.895, and 0.832, respectively, in the replication cohort (Figure 5E and F).

Discussion

In the present study, we illustrated the important contributions of sex, age, cigarette smoking, alcohol consumption, BMI, hypertension, T2DM, hyperlipidemia, SBP, DBP, FBG, TC, TG, LDL-C, and HDL-C to CAD risk. Subsequently, we developed and validated an XGBoost model that integrates these features with a favorable discrimination capability that can be conducive to improving risk stratification of CAD patients and helping guide downstream management.

The most significant feature identified by XGBoost was the HDL-C level. HDL-C had been considered as "good cholesterol" because there was a consistently negative relationship between HDL-C concentrations and the risk of atherosclerosis events in observational epidemiological evidences.[16,17] HDL is a highly heterogeneous population and transports various proteins, hormones, lipids, vitamins and miRNAs that confer HDL particles with many cardioprotective functions, including the promotion of macrophage reverse cholesterol efflux, inhibition of oxidation and inflammation, anti-apoptotic and anti-thrombotic properties.[18] Nevertheless, the Framingham study indicated that near 44% CAD clinical cases in men with HDL-C > 40 mg/dL and about 43% in women with HDL-C > 50 mg/dL. In addition, Mendelian randomization analyses implied that genetically elevated HDL-C levels were not associated with decreased CAD risk compared with rare variants related to lowering LDL-C levels.[19,20] Furthermore, HDL-C treatment failed to reduce cardiovascular events in large-scale clinical randomized controlled trials, partly because HDL-C did not reflect HDL function.[18] These results accentuated that HDL-C levels were not causally correlated with CAD and higher serum HDL-C levels did not ensure freedom from CAD events.[21,22] Intriguingly, our XGBoost model elucidated that HDL-C levels made a major contribution to CAD risk prediction.

Obesity is very prevalent worldwide, closely associated with many health risks, and directly contributes to the pathogenesis and progression of CAD.[23] In 2021, 1.95 million (95% CI: 1.12–2.91 million) cardiovascular deaths and 3.7 million (95% CI: 1.97–5.49 million) deaths overall were in consequence of higher BMI.[1] Obesity accelerated the atherosclerotic process through several mechanisms, including inflammation and insulin resistance.[24] Obesity-related inflammation increased the chance of low-density lipoprotein oxidation, which in turn worsened atherogenesis.[25] Insulin resistance was correlated with metabolic syndrome and dyslipidemia, which were both associated with atherosclerosis.[26] Endothelial dysfunction in obesity, mainly caused by decreased bioavailability of nitric oxide in the circumstance of oxidative stress and inflammation, is also fundamental to atherosclerosis progression.[27] Lifestyle intervention and weight loss could ameliorate both metabolic syndrome and associated endothelial dysfunction.[28] Nevertheless, clinical trials of medical weight loss have not authenticated an obvious reduction in CAD rates.[29,30] In contrast, prospective studies have suggested that obese patients undergoing bariatric surgery had significantly lower rates of fatal and nonfatal cardiovascular events than those nonsurgical patients with obesity.[31,32] In concordance with these reports, our model hinted that BMI was beneficial for CAD risk assessment.

Elevated LDL-C level continues to be the foremost modifiable risk factor and is one of the most closely related CAD markers. In 2021, 3.81 million (95% CI: 2.17–5.42 million) cardiovascular deaths and 3.81 million (95% CI: 2.17–5.42 million) deaths overall were ascribed to increased LDL-C levels.[1] Patients with familial hypercholesterolaemia had cumulative LDL-C burden at early ages and evolved into premature atherosclerotic CVD, which substantiated a causal role of LDL in atherosclerosis.[33] On the other hand, individuals with proprotein convertase subtilisin/kexin type 9 (*PCSK9*) loss-of-function variants had lifelong low LDL-C levels owing to decreasing degradation of the LDL receptors, and exhibited a greater reduction of cardiovascular diseases than stain treatment alone.[34,35] Furthermore, lowering serum LDL-C concentrations can significantly reduce mortality and morbidity of coronary events in both primary and secondary prevention.[36,37] In accord with these results, our model also indicated that LDL-C levels were the important feature for CAD risk estimation.

Large TG-rich lipoproteins, including chylomicrons (CM) and very low-density lipoprotein (VLDL), pass through the arterial wall by transcytosis of arterial macrophages, with excessive cholesterol accumulation and foam cell formation in the coronary arteries.[38] Nonetheless, the association between elevated plasma TG levels and cardiovascular disease is debatable in epidemiological studies. A meta-analysis involving 10,158 CAD patients among 262,525 participants in 29 western prospective studies indicated a modest correlation between TG levels and CAD risk, but it was abolished after multivariable adjustments for other conventional risk factors.[39] Meanwhile, a randomized controlled trial suggested that even slightly elevated TG concentrations were associated with increased recurrence risk of cardiovascular events in patients taking statin drugs and should be considered as a valuable risk marker.[40] In addition, genetic association studies had confirmed that elevated plasma TG levels were causally linked with higher CAD risk.[41] Furthermore, a Mendelian randomization study presented that genetically lowered non-fasting plasma TG levels were correlated with reduced all-cause mortality.[42] Consistent with these results, our model also implied that blood TG concentrations were conducive to CAD risk evaluation.

In addition to HDL-C, BMI, LDL-C, and TG levels, other indicators, such as age, sex, cigarette smoking, alcohol consumption, history of hypertension, hyperlipidemia, T2DM, TC, SBP, DBP, and FPG, are of relative importance for CAD risk stratification. Age is a major risk factor for the progression of atherosclerosis and mortality when coronary atherosclerosis manifests.[43] Prior studies have suggested that there is an obvious sex discrepancy in CAD prevalence and death. Generally, young women had a slight tendency to develop CAD and lower rates of myocardial infarction compared to men but that women drew near in the seventh decade of life and exceeded men by the ninth decade.[44] Cigarette smoking could affect all stages of atherosclerosis from endothelial dysfunction to acute clinical events by decreasing nitric oxide bioavailability, increasing adhesion molecules expression and adherence of platelets and macrophages.[45] The cardiovascular system is more susceptible to the toxic effects of alcohol and high dose of alcohol drinking could lead to extensive coronary arterial injuries and high CAD risk.[46] Hypertension was pathologically associated with atherosclerosis and high SBP was the leading risk factor for attributable premature cardiovascular deaths.[47] In addition, hypertension was frequently associated with other risk indicators, such as dyslipidemia and insulin resistance.[48,49] Excessive fat intake could elevate serum TC levels, which was correlated with increased CAD risk.[50] High fasting glucose is closely related with heavy burden of prediabetes, diabetes and obesity globally and both types of diabetes can contribute to atherosclerosis development or further accelerate its process.[51]

The Framingham risk score is a well-known predictive model that has been commonly used to assess individual CAD risk.[52] Nonetheless, considering that the risk equation was established in 1976 and the overwhelming majority of participants were of European descent, it is necessary to revalidate the predictive values of conventional risk factors for Chinese individuals owing to the inherent differences in diet and lifestyle, genetic predisposition, and social environment. Additionally, fast growing average income, westernized lifestyle, an ageing population and longer life expectancy in China gave rise to intricate alterations in CAD epidemics and risk factors pattern over the past decades.[53] Consequently, an evolutionary CAD risk evaluation tool based on recent data of Chinese population would be more applicable.[54] Many predictive models for cardiovascular diseases have been constructed and their predictive capacity (AUC: 0.61–1.0) varies considerably.[55] In addition, several models have been validated in more than one external population, with a trend towards lower discriminative performance over the past few years.[56] Indeed, differences in derivation (enrollment criteria of study population, different cut-off values for the definition of CAD and utilization of imputation methodologies for missing values), model selection as well as inconsistent external validation often exist, which limit their generalizable application in routine practice. XGBoost is an ensemble classifier which employs both categorical and continuous inputs without need for scaling or other pre-processing modifications. It has high degree of internal optimization and relatively modest computational cost to improve predictive accuracy. Our study constructed and validated a XGBoost model based on larger multicenter population, which exhibits favorable predictive capability and clinical application value.

Our study has some potential limitations. First, this was a retrospective study; some possible intrinsic biases cannot be disregarded, and causal effect argumentation is lacking. Second, we only considered 15 conventional risk factors for CAD, and future studies with more indicators, including deeper lipoprotein subtypes and individual genetic information, are needed to further validate our results. Finally, the recruited participants were all Chinese from the two provinces, which may limit their generalizability. Accordingly, more prospective multicenter studies from other regions of China are required to confirm our findings.

Conclusion

In summary, we constructed and validated an XGBoost model that integrated 15 conventional risk indicators to evaluate individual CAD risk. Our model is exceedingly informative for clinical practice, which will be conducive to primary prevention and tailoring the precise management of patients with CAD.

Abbreviations

CAD, Coronary artery disease; LDL-C, low-density lipoprotein cholesterol; HDL-C, high-density lipoprotein cholesterol; LDLR, LDL receptor; APOB, apolipoprotein B; APOE, apolipoprotein E; XGBoost, extreme gradient boosting; T2DM, type 2 diabetes mellitus; BMI, body mass index; FPG, fasting plasma glucose; SBP, systolic blood pressure; DBP, diastolic blood pressure; TG, total triglyceride; TC, total cholesterol; LASSO, least absolute shrinkage and selection operator; SHAP, Shapley additive explanations; DCA, decision curve analysis; ROC, receiver operating characteristic; AUC, area under the ROC curve; PPV, positive predictive value; NPV, negative predictive value; PCSK9, proprotein convertase subtilisin/kexin type 9; CM, chylomicrons; VLDL, very low-density lipoprotein.

Ethics Approval and Informed Consent

This study was approved by the Ethics Committees of Wuhan Asia Heart Hospital (No.2015076), Zhongnan Hospital of Wuhan University (No.2015183), and Shandong Provincial Hospital (No.2016094) and conformed to the tenets of the Declaration of Helsinki. All participants were Chinese Han, and they provided written informed consent.

Author Contributions

All authors made a significant contribution to the work reported, whether in the conception, study design, execution, acquisition of data, analysis, and interpretation, or in all these areas, took part in drafting, revising, or critically reviewing the article; gave final approval of the version to be published; have agreed on the journal to which the article has been submitted; and agree to be accountable for all aspects of the work.

Funding

This research was supported by the Natural Science Foundation of the First Affiliated Hospital of Soochow University (BXQN202229).

Disclosure

The authors report no conflicts of interest in this work.

References

1. Vaduganathan M, Mensah GA, Turco JV, et al. The global burden of cardiovascular diseases and risk: a compass for future health. *J Am Coll Cardiol*. 2022;80:2361–2371. doi:10.1016/j.jacc.2022.11.005
2. Wolf D, Ley K. Immunity and Inflammation in Atherosclerosis. *Circ Res*. 2019;124:315–327. doi:10.1161/CIRCRESAHA.118.313591
3. Kong P, Cui ZY, Huang XF, et al. Inflammation and atherosclerosis: signaling pathways and therapeutic intervention. *Signal Transduct Target Ther*. 2022;7:131. doi:10.1038/s41392-022-00955-7
4. Peper J, Becker LM, Van Den Berg H, et al. Diagnostic Performance of CCTA and CT-FFR for the Detection of CAD in TAVR work-up. *JACC: Cardiovasc Interv*. 2022;15(11):1140–1149.
5. De Gonzalo-Calvo D, Vilades D, Martínez-Camblor P, et al. Circulating microRNAs in suspected stable coronary artery disease: a coronary computed tomography angiography study. *J Intern Med*. 2019;286(3):341–355.
6. Messerli M, Panadero AL, Giannopoulos AA, et al. Enhanced radiation exposure associated with anterior-posterior x-ray tube position in young women undergoing cardiac computed tomography. *Am Heart J*. 2019;215:91–94.
7. Fahed AC, Wang M, Patel AP, et al. Association of the interaction between familial hypercholesterolemia variants and adherence to a healthy lifestyle with risk of coronary artery disease. *JAMA Network Open*. 2022;5:e222687.
8. Ference BA, Graham I, Tokgozoglu L, et al. Impact of lipids on cardiovascular health. *JACC Health Promotion Series. J Am Coll Cardiol*. 2018;72:1141–1156.
9. Yang X, Li Q, Liu D, et al. Joint effect of physical activity and blood lipid levels on all-cause and cardiovascular disease mortality: the Rural Chinese Cohort Study. *Nutr, Metab Cardiovasc Dis*. 2022;32:1445–1453.
10. Haug CJ, Drazen JM. Artificial intelligence and machine learning in clinical medicine, 2023. *N Engl J Med*. 2023;388:1201–1208. doi:10.1056/NEJMra2302038

11. Forrest IS, Petrazzini BO, Duffy Á, et al. Machine learning-based marker for coronary artery disease: derivation and validation in two longitudinal cohorts. *Lancet*. 2023;401:215–225. doi:10.1016/S0140-6736(22)02079-7
12. Al'aref SJ, Maliakal G, Singh G, et al. Machine learning of clinical variables and coronary artery calcium scoring for the prediction of obstructive coronary artery disease on coronary computed tomography angiography: analysis from the CONFIRM registry. *Eur Heart J*. 2020;41:359–367. doi:10.1093/eurheartj/ehz565
13. Collins GS, Reitsma JB, Altman DG, et al. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *BMJ*. 2015;350:g7594. doi:10.1136/bmj.g7594
14. Xue B, Li D, Lu C, et al. Use of machine learning to develop and evaluate models using preoperative and intraoperative data to identify risks of postoperative complications. *JAMA Network Open*. 2021;4:e212240. doi:10.1001/jamanetworkopen.2021.2240
15. Tideman LEM, Migas LG, Djambazova KV, et al. Automated biomarker candidate discovery in imaging mass spectrometry data through spatially localized Shapley additive explanations. *Anal Chim Acta*. 2021;1177:338522. doi:10.1016/j.aca.2021.338522
16. Hamer M, O'donovan G, Stamatakis E. High-density lipoprotein cholesterol and mortality: too much of a good thing? *Arterioscler Thromb Vasc Biol*. 2018;38(3):669–672. doi:10.1161/ATVBAHA.117.310587
17. Chen Y, Zhang F, Sun J, et al. Identifying the natural products in the treatment of atherosclerosis by increasing HDL-C level based on bioinformatics analysis, molecular docking, and in vitro experiment. *J Transl Med*. 2023;21:920. doi:10.1186/s12967-023-04755-7
18. Rohatgi A, Westerterp M, Von Eckardstein A, et al. HDL in the 21st century: a multifunctional roadmap for future HDL research. *Circulation*. 2021;143:2293–2309. doi:10.1161/CIRCULATIONAHA.120.044221
19. Cupido AJ, Reeskamp LF, Hingorani AD, et al. Joint genetic inhibition of PCSK9 and CETP and the association with coronary artery disease: a factorial Mendelian randomization study. *JAMA Cardiol*. 2022;7:955–964. doi:10.1001/jamacardio.2022.2333
20. Smith ML, Bull CJ, Holmes MV, et al. Distinct metabolic features of genetic liability to type 2 diabetes and coronary artery disease: a reverse Mendelian randomization study. *EBioMedicine*. 2023;90:104503. doi:10.1016/j.ebiom.2023.104503
21. Soppert J, Lehrke M, Marx N, et al. Lipoproteins and lipids in cardiovascular disease: from mechanistic insights to therapeutic targeting. *Adv Drug Deliv Rev*. 2020;159:4–33. doi:10.1016/j.addr.2020.07.019
22. Ouimet M, Barrett TJ, Fisher EA. HDL and Reverse cholesterol transport. *Circ Res*. 2019;124:1505–1518. doi:10.1161/CIRCRESAHA.119.312617
23. Lavie CJ, Laddu D, Arena R, et al. Healthy weight and obesity prevention: JACC health promotion series. *J Am Coll Cardiol*. 2018;72:1506–1531. doi:10.1016/j.jacc.2018.08.1037
24. Powell-Wiley TM, Poirier P, Burke LE, et al. Obesity and cardiovascular disease: a scientific statement from the American heart association. *Circulation*. 2021;143:e984–e1010. doi:10.1161/CIR.0000000000000973
25. Rotariu D, Babes EE, Tit DM, et al. Oxidative stress - Complex pathological issues concerning the hallmark of cardiovascular and metabolic disorders. *Biomed Pharmacother*. 2022;152:113238. doi:10.1016/j.biopha.2022.113238
26. Eley VA, Thuzar M, Navarro S, et al. Obesity, metabolic syndrome, and inflammation: an update for anaesthetists caring for patients with obesity. *Anaesth Crit Care Pain Med*. 2021;40(6):100947. doi:10.1016/j.accpm.2021.100947
27. Ait-Aissa K, Nguyen QM, Gabani M, et al. MicroRNAs and obesity-induced endothelial dysfunction: key paradigms in molecular therapy. *Cardiovasc Diabetol*. 2020;19:136. doi:10.1186/s12933-020-01107-3
28. Zheng X, Yu H, Qiu X, et al. The effects of a nurse-led lifestyle intervention program on cardiovascular risk, self-efficacy and health promoting behaviours among patients with metabolic syndrome: randomized controlled trial. *Int J Nurs Stud*. 2020;109:103638. doi:10.1016/j.ijnurstu.2020.103638
29. Ma C, Avenell A, Bolland M, et al. Effects of weight loss interventions for adults who are obese on mortality, cardiovascular disease, and cancer: systematic review and meta-analysis. *BMJ*. 2017;359:j4849. doi:10.1136/bmj.j4849
30. Jamshed H, Steger FL, Bryan DR, et al. Effectiveness of early time-restricted eating for weight loss, fat loss, and cardiometabolic health in adults with obesity: a randomized clinical trial. *JAMA Intern Med*. 2022;182:953–962.
31. Aminian A, Al-Kurd A, Wilson R, et al. Association of bariatric surgery with major adverse liver and cardiovascular outcomes in patients with biopsy-proven nonalcoholic steatohepatitis. *JAMA*. 2021;326:2031–2042. doi:10.1001/jama.2021.19569
32. Courcoulas AP, Daigle CR, Arterburn DE. Long term outcomes of metabolic/bariatric surgery in adults. *BMJ*. 2023;383:e071027. doi:10.1136/bmj-2022-071027
33. Mundal LJ, Igland J, Veierød MB, et al. Impact of age on excess risk of coronary heart disease in patients with familial hypercholesterolaemia. *Heart*. 2018;104:1600–1607. doi:10.1136/heartjnl-2017-312706
34. Peng J, Xing CY, Zhao K, et al. Associations of pro-protein convertase subtilisin-like kexin type 9, soluble low-density lipoprotein receptor and coronary artery disease: a case-control study. *Int J Cardiol*. 2022;350:9–15. doi:10.1016/j.ijcard.2022.01.014
35. Matyas C, Trojnar E, Zhao S, et al. PCSK9, A promising novel target for age-related cardiovascular dysfunction. *JACC Basic Transl Sci*. 2023;8:1334–1353. doi:10.1016/j.jacbts.2023.06.005
36. Mohammad S, Nguyen H, Nguyen M, et al. Pleiotropic effects of statins: untapped potential for statin pharmacotherapy. *Curr Vasc Pharmacol*. 2019;17:239–261. doi:10.2174/1570161116666180723120608
37. Fici F, Faikoglu G, Tarim BA, et al. Pitavastatin: coronary atherosclerotic plaques changes and cardiovascular prevention. *High Blood Press Cardiovasc Prev*. 2022;29:137–144.
38. Ginsberg HN, Packard CJ, Chapman MJ, et al. Triglyceride-rich lipoproteins and their remnants: metabolic insights, role in atherosclerotic cardiovascular disease, and emerging therapeutic strategies-A consensus statement from the European Atherosclerosis Society. *Eur Heart J*. 2021;42:4791–4806.
39. Sarwar N, Danesh J, Eiriksdottir G, et al. Triglycerides and the risk of coronary heart disease: 10,158 incident cases among 262,525 participants in 29 Western prospective studies. *Circulation*. 2007;115:450–458.
40. Faergeman O, Holme I, Fayyad R, et al. Plasma triglycerides and cardiovascular events in the treating to new targets and incremental decrease in end-points through aggressive lipid lowering trials of statins in patients with coronary artery disease. *Am J Cardiol*. 2009;104:459–463.
41. Do R, Willer CJ, Schmidt EM, et al. Common variants associated with plasma triglycerides and risk for coronary artery disease. *Nat Genet*. 2013;45:1345–1352.
42. Thomsen M, Varbo A, Tybjærg-Hansen A, et al. Low nonfasting triglycerides and reduced all-cause mortality: a Mendelian randomization study. *Clin Chem*. 2014;60:737–746.

43. Tyrrell DJ, Goldstein DR. Ageing and atherosclerosis: vascular intrinsic and extrinsic factors and potential role of IL-6. *Nat Rev Cardiol*. 2021;18:58–68.
44. Man JJ, Beckman JA, Jaffe IZ. Sex as a Biological Variable in Atherosclerosis. *Circ Res*. 2020;126:1297–1319.
45. Higashi Y. Smoking cessation and vascular endothelial function. *Hypertens Res*. 2023;46:2670–2678.
46. Hu C, Huang C, Li J, et al. Causal associations of alcohol consumption with cardiovascular diseases and all-cause mortality among Chinese males. *Am J Clin Nutr*. 2022;116:771–779.
47. Razo C, Welgan CA, Johnson CO, et al. Effects of elevated systolic blood pressure on ischemic heart disease: a burden of proof study. *Nat Med*. 2022;28:2056–2065.
48. Dąbrowska E, Narkiewicz K. Hypertension and dyslipidemia: the two partners in endothelium-related crime. *Curr Atheroscler Rep*. 2023;25:605–612.
49. Tagi VM, Mainieri F, Chiarelli F. Hypertension in patients with Insulin Resistance: etiopathogenesis and management in children. *Int J Mol Sci*. 2022;23:5814.
50. Berger S, Raman G, Vishwanathan R, et al. Dietary cholesterol and cardiovascular disease: a systematic review and meta-analysis. *Am J Clin Nutr*. 2015;102:276–294.
51. Poznyak A, Grechko AV, Poggio P, et al. The diabetes mellitus-atherosclerosis connection: the role of lipid and glucose metabolism and chronic inflammation. *Int J Mol Sci*. 2020;21:1835.
52. Chen G, Levy D. Contributions of the Framingham heart study to the epidemiology of coronary heart disease. *JAMA Cardiol*. 2016;1:825–830.
53. Cheng J, Zhao D, Zeng Z, et al. The impact of demographic and risk factor changes on coronary heart disease deaths in Beijing, 1999–2010. *BMC Public Health*. 2009;9:30.
54. Wang C, Zhao Y, Jin B, et al. Development and validation of a predictive model for coronary artery disease using machine learning. *Front Cardiovasc Med*. 2021;8:614204.
55. Damen JA, Hooft L, Schuit E, et al. Prediction models for cardiovascular disease risk in the general population: systematic review. *BMJ*. 2016;353:i2416.
56. He T, Liu X, Xu N, et al. Diagnostic models of the pre-test probability of stable coronary artery disease: a systematic review. *Clinics*. 2017;72:188–196.