Genetic factor increased identification efficiency of predictive models for dyslipidemia: a prospective cohort study

Miaomiao Niu
Zhengzhou University
Liying Zhang
Zhengzhou University
Yikang Wang
Zhengzhou University
Runqi Tu
Zhengzhou University
Xiaotian Liu
Zhengzhou University
Jian Hou
Zhengzhou University
Wenqian Huo
Zhengzhou University
Zhenxing Mao
Zhengzhou University
Zhenfei Wang
Zhengzhou University
Chongjian Wang (tjwcj2005@126.com)
Zhengzhou University https://orcid.org/0000-0001-5091-6621

Research

Keywords: dyslipidemia, genetic risk score, machine learning, risk model, predictive performance

Posted Date: November 18th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-109727/v1

Version of Record: A version of this preprint was published on February 12th, 2021. See the published version at https://doi.org/10.1186/s12944-021-01439-3.

Abstract

Background: Few studies have developed risk models for dyslipidemia, especially for rural population. Further, the performance of genetic factors in predicting dyslipidemia was not explored. The purpose is to develop and evaluate the prediction models with and without genetic factor for dyslipidemia in Chinese rural population.

Methods: A total of 3596 individuals from the Henan Rural Cohort study were included in this study. All subjects were divided into training set and testing set in a ratio of 7:3. The conventional models and conventional+GRS models were developed with COX regression, artificial neural network (ANN), random forest (RF), and gradient boosting machine (GBM) classifiers in training set. Area under the receiver operating characteristic curve (AUC), net reclassification index (NRI), and integrated discrimination index (IDI) were used to assess the discrimination ability of models and the calibration curve was used to show calibration ability in testing set.

Results: Compared to the lowest GRS quartile, *HR* (95%*CI*) of individuals in the highest GRS quartile was 1.23(1.07, 1.41) in total population. Age, family history of diabetes, physical activity, BMI, TG, HDL-C, and LDLC-C were included and developed the conventional models, and the AUC of COX, ANN, RF, and GBM classifiers were 0.702(0.673, 0.729), 0.736(0.708, 0.762), 0.787 (0.762, 0.811), and 0.816(0.792, 0.839), respectively. After adding GRS, the AUC increased by 0.005, 0.018, 0.023, and 0.015 with COX, ANN, RF, and GBM classifiers, respectively. The corresponding NRI and IDI were 25.6%, 7.8%, 14.1%, 18.1% and 2.3%, 1.0%, 2.5%, 1.8%, respectively.

Conclusion: Genetic factors could improve the predictive ability for dyslipidemia risk model, suggesting genetic information could be provided as a potential predictor to screen clinical dyslipidemia.

Trial Registration

The Henan Rural Cohort Study has been registered at Chinese Clinical Trial Register. (Trial registration: ChiCTR-OOC-15006699. Registered 6 July 2015 - Retrospectively registered)
http://www.chictr.org.cn/showproj.aspx?proj=11375

Contributions To The Literature

What is already known on this topic?

Previous studies have developed prediction model for dyslipidemia, but rare focused on the resource-limited area. Besides, the association between genetic factors and dyslipidemia have been widely reported, but none have explored whether the predictive ability could be enhanced when the conventional model further combined with genetic factors, especially in the rural area.

What does this study add?

The established conventional models showed better performance in predicting dyslipidemia, especially using gradient boosting machine (GBM) classifier. More importantly, the predictive ability of models was significantly improved when incorporating genetic factor in conventional models, suggesting the better potential utility of genetic factor in predicting dyslipidemia in rural population.

Background

Dyslipidemia is an important risk factor for the development of cardiovascular diseases (CVD)[1]. Studies have shown that about 20% of patients with atherosclerosis have either high triglyceride (TG) or low high-density lipoprotein cholesterol (HDL-C) lipid levels[2], and low HDL-C levels could reduce the incidence of heart disease and ischemic stroke[3]. Elevated serum levels of total cholesterol (TC), TG, and low-density lipoprotein cholesterol (LDL-C) could be used as independent predictors of CVD due to the close relationship[1, 4, 5]. The prevalence of dyslipidemia has declined in the developed countries such as the United States in nearly a decade[6], while the prevalence in China, the biggest developing country, remains at a high level and continue growth[7]. A total of 9.2 million cardiovascular events will occur due to the serum cholesterol levels in Chinese population between 2010 and 2030[8]. In the rural area, the age-standardized prevalence of dyslipidemia was 32.21% in adults with a relatively low rate of awareness, treatment, and control (15.07%, 7.23%, 3.25%, respectively)[9]. The prevention of dyslipidemia remains a huge public health problem in China, especially in the rural area. The establishment of disease risk prediction models has received extensive attention globally in preventing diseases. In previous studies, the effective disease prediction models for CVD and diabetes were built based on the Framingham study[10, 11]. In recent years, some researchers had established other effective risk models to diagnose and predict varieties of diseases[12–15]. Few studies have involved the prediction model for dyslipidemia[16–19], and most of them were limited to certain groups of people such as children and adolescents to some extent.

As reported, a genetic risk score consists of multiple single nucleotide polymorphisms (SNPs) conferred a strong prediction in risk but each SNP only donated little individually[20]. Although the role of SNPs in dyslipidemia are well known[21–23], no studies have been interpreted that how polygenetic genetic risk scores (GRS) affect dyslipidemia when the prediction of the risk of dyslipidemia is needed, especially in the resource-limited area. To that end, this study was constructed to set up the dyslipidemia prediction model and to reveal the prediction performance of the model incorporating genetic factors in predicting the occurrence of dyslipidemia in Chinese rural adults.

Methods

Study population

Participants were recruited from the Henan Rural Cohort study which was registered in Chinese Clinical Trial Register (Registration number: ChiCTR-OOC-15006699). The baseline examination and follow-up information have been previously described in detail[24]. In brief, the baseline investigation included a

questionnaire interview, anthropometry measurements, blood tests. The subjects were then asked about the occurrence of chronic diseases, including the type and duration of the disease, as well as the status of treatment and medication at the follow-up survey.

A total of 6930 individuals had completed follow-up survey, and 3596 individuals were finally analyzed after excluding participants who 1) had dyslipidemia at baseline; 2) were using lipid-lower drugs; 3) were missing important information about the key variables. According to a ratio of 7: 3, the subjects were then randomly divided into training set (n = 2517) for model construction and testing set (n = 1079) for performance evaluation of the models.

Definition of dyslipidemia

As reported by the Chinese guidelines on prevention and treatment of dyslipidemia in adults[7], dyslipidemia was defined as having one or more of the following conditions: TC ≥ 6.2 mmol/L (240 mg/dl); TG ≥ 2.3 mmol/L (200 mg/dl); HDL-C ≤ 1.0 mmol/L (40 mg/dl); LDL-C ≥ 4.1 mmol/L (160 mg/dl) or use of lipid-lower drugs in recent two weeks.

Calculation of weighted genetic risk score (GRS)

A weighted genetic risk score (GRS) was calculated using 21 SNPs to assess the predictive performance of genetic factors. GRS was calculated by multiplying the weight of each SNP by the number of risk alleles. As shown in Table S1, the weights of each SNP were calculated based on our own population. The mean value and standard deviation of GRS were 1.329 and 0.337, ranging from 0.195 to 2.451.

Classifier

COX regression model, also known as the "proportional hazards model", is a semiparametric regression model. The model takes survival outcome and survival time as dependent variables. This model has been widely used in medical follow-up studies and is by far the most used multi-factor analysis method in survival analysis.

Artificial neural network (ANN), which simulates neuron activity with a mathematical model, is an information processing system based on imitating the structure and function of brain neural networks. Compared with traditional data processing methods, neural network technology has obvious advantages in processing fuzzy data, random data, and non-linear data, and is particularly suitable for systems with large scale, complex structure, and ambiguous information.

Random forest (RF) mixes up the bagging ensemble learning theory and random subspace technique. In training set, numerous decision trees are produced to randomly separate data. Random feature selection is bringing in feature selection process in RF. After randomly selecting a subset containing K attributes from the attribute set of each base decision tree node, an optimal attribute is adopted from the sub-set for partitioning. At last, RF chooses the classification with the most votes which are voted by all trees.

Gradient boosting machine (GBM) trains a stronger classifier by combining different weak classifiers which are trained in the same training set. As an iterative algorithm, GBM can master inadequacy of the

combination of weak classifier by putting each weak classifier into iteration process, and these series of iteration will ameliorate the results of classification. At the time when training weak classifier, GBM strengthen models using the residual of training set which was adapted by preceding weak classifier.

Statistical analysis

Statistical significance was inferred at a two-tailed value of $P < 0.05$. T-test and chi-square test were used to compare differences in characteristics between training and testing set. All the subjects were divided into quartiles according to GRS. Taking Q1 as a reference, we calculated the hazard ratios (*HR*s) of the remaining three groups of subjects in total population, as well as training and testing set.

Previous studies revealed a dozen of variables as predictors of dyslipidemia: age, sex, educational level, smoking, high-fat diet, more vegetable and fruit intake, family history of hyperlipidemia, physical activity, waist circumference (WC), family history of diabetes, BMI, TG, HDL-C and LDL-C (Table S2) [16, 18]. In the training set, all the variables were analyzed using univariate COX regression. Then, those variables presenting a significant impact on dyslipidemia entered the conventional models. GRS mentioned above was then incorporated into the conventional models to constitute the conventional + GRS models. COX regression, artificial neural network (ANN), random forest (RF), and gradient boosting machine (GBM) were employed to construct the conventional models and conventional + GRS models. In COX classifiers, the conventional model and conventional + GRS model were constructed in training set, as for ANN, RF, and GBM, prediction models were trained and tested through 10-fold cross-validation during the iteration process, which repeated 100 times.

Model performance was calculated in the testing set. In COX classifier, the coefficients in training set was used to predict dyslipidemia risk in testing set. The parameters of each model were determined by grid search and ten cross-validations of the training set to ensure the best performance value with ANN, RF, and GBM. The discrimination of models was assessed using the area under the receiver operating characteristic curve (AUC). Net reclassification index (NRI) and integrated discrimination index (IDI) were used to evaluate the improvement of predictive ability of the conventional models when adding GRS. The calibration of models was assessed by calibration curves. Statistical analyses were performed with R 3.6.2, and Python 3.8.

Results

Baseline characteristics

The baseline characteristics for training and testing set were shown in Table 1. The average age of all subjects was 50.49 ± 12.16, and the proportion of men was 31.2%. No significant differences of demographic characteristics and lipid measurements were observed between training and testing set (all $P > 0.05$).

Table 1
Baseline characteristic between training and testing set

Characteristic	Total (n = 3596)	Training set(n = 2517)	Testing set(n = 1079)	*P*-value
Age	50.49 ± 12.16	50.38 ± 12.16	50.76 ± 12.15	0.387
Man, n (%)	1122(31.2)	774(30.8)	348(32.3)	0.373
Education, n (%)				0.410
No education	596(16.6)	429(17.0)	167(15.5)	
Primary school	1246(34.6)	848(33.7)	398(36.9)	
Middle school	1428(39.7)	1013(40.2)	415(38.5)	
High school	296(8.2)	206(8.2)	90(8.3)	
College and above	30(0.8)	21(0.8)	9(0.8)	
Smoking, n (%)	794(22.1)	553(22.0)	241(22.3)	0.809
High-fat diet, n (%)	70(2.0)	52(2.1)	18(1.7)	0.420
More vegetable and fruit intake, n (%)	1372(38.2)	997(38.8)	395(36.6)	0.205
Family history of hyperlipidemia, n (%)	354(9.8)	239(9.5)	115(10.7)	0.284
Family history of diabetes, n (%)	186(5.2)	132(5.2%)	54(5.0)	0.088
Physical activity, n (%)				0.967
Low	1656(46.1)	1156(45.9)	500(46.3)	
Moderate	810(22.5)	567(22.5)	243(22.5)	
High	1130(31.4)	794(31.5)	336(31.1)	
Waist circumference (WC), cm	80.42 ± 9.53	80.49 ± 9.52	80.26 ± 9.58	0.521
Body mass index (BMI), kg/m^2				
Total cholesterol (TC), mmol/L	4.45 ± 0.72	4.45 ± 0.73	4.45 ± 0.71	0.801
Triglyceride (TG), mmol/L	1.18 ± 0.44	1.18 ± 0.44	1.18 ± 0.44	0.960
Low density lipoprotein (LDL-C), mmol/L	2.61 ± 0.63	2.61 ± 0.63	2.60 ± 0.61	0.724
High density lipoprotein (HDL-C), mmol/L	1.31 ± 0.21	1.31 ± 0.21	1.31 ± 0.21	0.942

Association between GRS and dyslipidemia

The mean value of GRS in all participants was 1.33 (SD: 0.34). The overall association was significant between GRS and dyslipidemia, with crude *HR* (95% *CI*) 1.366 (1.187, 1.572); adjusted *HR* (95% *CI*) 1.353 (1.172, 1.561) (Table 2). The GRS was divided into quartiles. Compared with Q1, subjects in Q2, Q3, Q4 group had an *HR* (95% *CI*) = 1.043 (0.900, 1.210), 1.188 (1.028, 1.374), and 1.229 (1.069, 1.412), respectively, when adjusting for age, family history of diabetes, physical activity, BMI, TG, HDL-C, LDL-C, which suggested a steady increase in the risk of dyslipidemia occurrence with the rise of GRS. By the same token, adjusted and crude *HR*s also showed the same constant increment in training set and testing set.

Table 2
Association between GRS and incidence of dyslipidemia

	Subjects	Crude *HR*s (95%*CI*)	Adjusted *HR*s (95%*CI*)
Total population			
Q1	900	1.00 (reference)	1.00 (reference)
Q2	898	1.110 (0.958, 1.287)	1.043 (0.900, 1.210)
Q3	900	1.244 (1.077, 1.437)	1.188 (1.028, 1.374)
Q4	898	1.276 (1.111, 1.466)	1.229 (1.069, 1.412)
Continuous GRS	3596	1.366 (1.187, 1.572)	1.353 (1.172, 1.561)
$P_{\text{for trend}}$		< 0.001	0.001
Training set			
Q1	633	1.00 (reference)	1.00 (reference)
Q2	638	0.996 (0.834, 1.188)	1.023 (0.855, 1.223)
Q3	624	1.182 (0.995, 1.404)	1.166 (0.979, 1.388)
Q4	622	1.207 (1.023, 1.424)	1.213 (1.028, 1.433)
Continuous GRS	2517	1.337 (1.129, 1.584)	1.318 (1.110, 1.565)
$P_{\text{for trend}}$		0.006	0.008
Testing set			
Q1	267	1.00 (reference)	1.00 (reference)
Q2	260	1.456 (1.112, 1.907)	1.081 (0.820, 1.425)
Q3	276	1.405 (1.080, 1.827)	1.225 (0.940, 1.596)
Q4	276	1.454 (1.129, 1.874)	1.273 (0.986, 1.643)
Continuous GRS	1079	1.432 (1.113, 1.843)	1.466 (1.127, 1.907)
$P_{\text{for trend}}$		0.009	0.040

Note: GRS is divided into four groups. Q1, Q2, Q3, Q4 represent the first, second, third, fourth quartile of GRS, respectively. Adjusted *HR*s adjust for the following covariates: age, family history of diabetes, physical activity, BMI, TG, HDL-C, LDL-C.

Development and evaluation of the conventional models

In training set, the 14 predictors were analyzed using univariate COX regression, and 8 variables (age, family history of diabetes, physical activity, WC, BMI, TG, HDL-C, and LDL-C) showed statistically significant correlation with dyslipidemia. Eventually, the conventional models were composed of age, family history of diabetes, physical activity, BMI, TG, HDL-C, and LDL-C (Table 3, above), considering the collinearity between WC and BMI. It was worth noting that there was no collinearity between TC, HDL-C, and LDL-C. The AUC and differences of 4 conventional models with different classifiers were shown in Fig. 1 and Table 4. In testing set, the AUC of conventional models with COX, ANN, RF, and GBM classifiers were 0.702(0.673, 0.729), 0.736(0.708, 0.762), 0.787 (0.762, 0.811), and 0.816(0.792, 0.839), respectively, indicating the conventional models showed quite high predictive performance, especially the conventional model with GBM classifier.

Table 3
Multivariate COX regression analysis on significant factors of developing dyslipidemia in training set

Variables	*β*	S.E.	Wald	*P*	*HR* (95%*CI*)
Conventional model					
Age	0.005	0.003	3.017	0.082	1.005(0.999, 1.010)
Family history of diabetes	0.194	0.125	2.429	0.119	1.215(0.951, 1.551)
Physical activity					
Low	Reference				
Moderate	0.793	0.080	99.087	< 0.001	2.210(1.890, 2.583)
High	0.324	0.071	20.810	< 0.001	1.383(1.203, 1.590)
BMI	0.016	0.010	2.777	0.096	1.016(0.997, 1.036)
TG	0.292	0.074	15.609	< 0.001	1.339(1.158, 1.548)
HDL-C	-2.103	0.196	114.907	< 0.001	0.122(0.083, 0.179)
LDL-C	0.284	0.052	29.792	< 0.001	1.329(1.200, 1.472)
Conventional + GRS model					
Age	0.005	0.003	2.887	0.089	1.005(0.999, 1.010)
Family history of diabetes	0.198	0.125	2.517	0.113	1.219(0.954, 1.557)
Physical activity					
Low	Reference				
Moderate	0.802	0.080	101.097	< 0.001	2.230(1.907, 2.607)
High	0.328	0.071	21.347	< 0.001	1.389(1.208, 1.596)
BMI	0.017	0.010	2.998	0.083	1.017(0.998, 1.037)
TG	0.281	0.074	14.410	< 0.001	1.325(1.146, 1.532)
HDL-C	-2.095	0.195	114.889	< 0.001	0.123(0.084, 0.180
LDL-C	0.286	0.052	29.968	< 0.001	1.330(1.201, 1.474)
Weighted GRS	0.276	0.088	9.925	0.002	1.318(1.110, 1.565)

Note: The predictors of conventional model are variables that are significant associated with dyslipidemia in univariate COX regression analysis. GRS is added to conventional model to construct conventional + GRS model. **Abbreviations**: BMI: body mass index; TG: triglyceride; HDL-C: high density lipoprotein; LDL-C: low density lipoprotein.

Table 4
Performance of conventional and conventional + GRS model in predicting dyslipidemia

	AUC	△AUC	Continuous NRI, %	IDI, %
COX				
Conventional model	0.702(0.673, 0.729)			
Conventional + GRS model	0.707(0.679, 0.734)	0.00491(*P* = 0.0549)	**25.6 (13.8, 35.8)**	**2.3 (1.1, 3.7)**
ANN				
Conventional model	0.736(0.708, 0.762)			
Conventional + GRS model	0.754(0.727, 0.779)	**0.0183(P = 0.0031)**	7.8 (-2.7, 18.5)	1.0 (-0.3, 2.4)
RF				
Conventional model	0.787 (0.762, 0.811)			
Conventional + GRS model	0.810 (0.762, 0.811)	**0.0230(P = 0.023)**	**14.1 (1.1, 26.1)**	**2.5 (0.5, 4.2)**
GBM				
Conventional model	0.816(0.792, 0.839)			
Conventional + GRS model	0.831(0.808, 0.853)	**0.0151(P = 0.0135)**	**18.1 (4.4, 27.2)**	**1.8 (0.1, 3.5)**

Note: All bold fonts represent statistically significant values. **Abbreviations**: AUC: area under receiver operating characteristic curve; △AUC: difference between AUC of conventional model and conventional + GRS model; NRI: net reclassification improvement; IDI: integrated discrimination improvement; ANN: artificial neural network; RF: random forest; GBM: gradient boosting machine.

Development and evaluation of the conventional models with GRS

The conventional + GRS models combined conventional factors and GRS (Table 3, bellow). Table 4 showed the differences of discrimination between conventional model and conventional + GRS model. With the COX classifier, the addition of GRS improved the predictive ability of the conventional model in a limited way. The conventional model showed moderate discrimination, and the addition of GRS slightly increased AUC to 0.707(0.679, 0.734); the difference in AUC was 0.00491 but showed no statistical significance at a *P* = 0.0549. Notwithstanding, the addition of GRS resulted in a continuous NRI of 25.6% (13.8%, 35.8%) and an IDI of 2.3% (1.1%, 3.7%), which were statistically significant. As for ANN classifier, the addition of GRS increased AUC to 0.754 (0.727, 0.779); difference in AUC was 0.0183 (*P* = 0.0031).

Nevertheless, the continuous NRI and IDI were 7.8% (-2.7%, 18.5%) and 1.0% (-0.3%, 2.4%), presenting no statistically significant. Additionally, the conventional + GRS model with RF classifier resulted in significant improvements (NRI: 14.1% (1.1%, 26.1%); IDI: 2.5% (0.5%, 4.2%)), announcing a competent progress of GRS in predicting dyslipidemia. The discrimination of the prediction model showed significant improvements better than the GBM classifier when adding GRS into the conventional model. Figure 2 provided the receiver-operating characteristic curves for conventional and conventional + GRS models in different classifiers. Results suggested the addition of GRS could improve the prediction performance of the conventional models. Besides, the GBM classifier presented the best performance with an AUC of 0.831 (0.808, 0.853) of the conventional model.

Figure 3 demonstrated the calibration of conventional and conventional + GRS models. The calibration curves of the conventional + GRS models were closer to the reference line than the conventional models. The brier scores, which can be considered as a "calibration" measure of a set of probabilistic predictions, also declined with the addition of GRS (COX declined 0.048, ANN classifier slightly declined 0.005, and GBM declined 0.006), indicating models were provided with better calibration when incorporating GRS (The lower the brier score value, the better the prediction calibration). Other statistics for instance sensitivity, specificity, etc. were also provided in Table S3, which proved that the predictive ability of models was improved by adding GRS.

Discussion

As far as we know, this is the first study explored the utility of genetic factors in the prediction of dyslipidemia in resource-limited area based on a prospective study. Results of this study suggested those in higher GRS quartiles displayed increasing risk of dyslipidemia onset compared to participants with the lowest quartile of GRS. Then, the conventional models were constructed with COX, ANN, RF, and GBM classifiers, and the model with GBM classifier significantly outperformed the other classifiers. More importantly, the accession of GRS convincingly improved the capability of the conventional models to predict dyslipidemia, implying the genetic factors perform a meaningful role in predicting the occurrence of dyslipidemia.

This study elaborated the correlation between the genetic factor (GRS) and dyslipidemia by dividing GRS into quartiles. A previous study divided all participants into 3 groups according to GRSs of LDL-C, HDL-C, and TG, the highest GRS groups all presented higher lipids levels than the lowest GRS groups in HDL-C, LDL-C, and TG[23]. Similarly, in this study, we found that the higher GRS was associated with a higher risk of dyslipidemia onset regardless of age, family history of diabetes, physical activity, BMI, TG, HDL-C, and LDL-C. Although not every *HR* was statistically significant, dyslipidemia risk increased within each quartile of GRS, and a similar trend was observed in training set and testing set. The above announced a statistically significant enhanced occurrence of dyslipidemia risk with incremental GRS in rural area population.

Results showed that the conventional model with GBM classifier presented the best predictive performance. Yet, 7 variables demonstrated statistical significance in univariate COX regression analysis and finally were included in the conventional models. Based on the results, univariate COX regression tagged baseline lipoprotein including TG, HDL-C, and LDL-C as predictors, which was a reasonable result that currently plasma lipoproteins leading to abnormal future blood lipids. Besides, *HR*s of predictors in the conventional model were comparable to those reported in other reasearshes[9, 25–29]. Correspondingly, the *HR*s of these 7 variables were also consistent with those in early published studies[16, 18, 19]. What is noteworthy is that the three serum lipid parameters showed no collinearity. The findings pointed out that GBM classifier could predict the incidence of dyslipidemia better, which had been confirmed in the previous study[30]. This might be due to the GBM classifier could deal with the intricate relationship between predictors and dyslipidemia.

Considering the moderate but strong association between GRS and dyslipidemia, the performance of GRS to predict the occurrence of dyslipidemia was then figured out. All the 4 classifiers (COX, ANN, RF, and GBM) manifested that the discrimination and calibration of the prediction model were moderately improved by adding GRS into the conventional models. The NRI and IDI were not significantly corrected with the inclusion of GRS ($P > 0.05$) in the ANN classifier though the number of NRI and IDI were slightly increased. Still, a major improvement was observed in COX, RF, and GBM classifiers. As is shown in an earlier study[22], in the transition from childhood to adulthood, the predictive power of GRSs on HDL-C, LDL-C, and TG had been proved to be valuable in predicting adulthood lipids level. Any abnormal lipid index can be defined as dyslipidemia; thus, GRS might have a predictive effect on dyslipidemia, and our results confirm this. Further, the result also suggested the application of the machine learning technic might have a better effect on disease prediction than the statistical method, which was consistent with the results of other studies[31, 32]. By the same token, the elevation of other statistical (Table S3) value exhibited that GRS played a relatively important role in dyslipidemia prediction. Principally, the results of this study revealed that GRS could be a crucial predictor to the occurrence of dyslipidemia.

As was demonstrated in a former study[33], the disclosure of coronary heart disease risk estimates indicated that the inclusion of genetic risk information resulted in lower levels of LDL-C compared to the disclosure based on conventional risk factors only. Genetic risk information for common diseases could be incorporated into the conventional predictive model and used to guide treatment. Considering how lipids level impressed CVD[34, 35], it’s reasonable to infer that the addition of the GRS into the prediction model of dyslipidemia might help individuals prevent abnormal blood lipid levels and thus contribute to the prevention of cardiovascular events.

Strengths and limitations: this research clarified the crucial impact of genetic information in predicting dyslipidemia in rural area, signifying the certain guiding role of the gene in the prevention and treatment of clinical dyslipidemia. To some extent, the research indicated that the machine learning method might have certain advantages in the construction of the disease prediction model. As well, a cohort study was used to construct the conventional model and to analyze the relationship between genetic factors and dyslipidemia, making the results more convincing. Yet, several limitations need to be remarked. The

integration of the four lipid measurements (TC, TG, LDL-C, and HDL-C) into dyslipidemia might gloss over the ability of genetic information in each lipid indexes. But there was no denying that genetic information was impressive in blood lipids, providing a foundation for the follow-up studies about genetic factors and lipid levels. Another limitation concerns that the brier score failed to test statistically in assessing the calibration of models, though the value has declined. Thirdly, the extrapolation of the conclusions is restricted by the lack of external validation. However, 30% of subjects were randomly selected to conduct internal verification to increase the credibility of the study. Meanwhile, the representation might limit as a result of the recruited subjects only came from the rural area in China.

Conclusion

Based on the prospective cohort study, eight dyslipidemia prediction models were developed and evaluated with and without the genetic factor (GRS), respectively. The conventional models included age, family history of diabetes, physical activity, BMI, TG, HDL-C, and LDL-C, which showed better performance in predicting dyslipidemia, especially with GBM classifier. After adding genetic factor, the prediction performance of the conventional models was effectively enhanced. The results set the stage for future research to study the prediction ability of genetic factors in different lipid indexes.

List Of Abbreviations

GRS: weighted genetic risk score

AUC: area under ROC curve

NRI: net reclassification index

IDI: integrated discrimination index

CVD: cardiovascular diseases

TG: triglyceride

HDL-C: high-density lipoprotein cholesterol

TC: total cholesterol

LDL-C: low-density lipoprotein cholesterol

SNPs: single nucleotide polymorphisms

GRS: genetic risk scores

BMI: body mass index

WC: waist circumference

***HR*s**: hazard ratios

ANN: artificial neural network

RF: random forest

GMB: gradient boosting machine

SD: standard deviation

Declarations

Ethics approval and consent to participate

Ethics approval was obtained from the “Zhengzhou University Life Science Ethics Committee” (Ethic approval code: [2015] MEC (S128) and written informed consent was obtained from all participants before this study.

Consent for publication

All authors consent for publication.

Availability of data and materials

The data are available from the corresponding author on reasonable request.

Competing interests

All authors declare that they have no competing of interest.

Funding

This research was supported by the National Natural Science Foundation of China (Grant NO: 81573243, 81602925), Foundation of National Key Program of Research and Development of China (Grant NO: 2016YFC0900803, 2019YFC1710002), Foundation of Medical Science and Technology of Henan province (NO: 201702367, 2017T02098), Henan Natural Science Foundation of China (Grant NO: 182300410293). Discipline Key Research and Development Program of Zhengzhou University (Grant NO: XKZDQY202008, XKZDQY202002). The funders had no role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Authors’ contributions

Chongjian Wang, Zhenfei Wang: Conceived and designed the experiments. Miaomiao Niu, Liying Zhang, Yikang Wang, Ruqi Tu, and Xiaotian Liu: Gathered data. Miaomiao Niu, Liying Zhang: Analyzed the data.

Miaomiao Niu: Drafted the manuscript. Liying Zhang, Jian Hou, Wenqian Huo, and Zhenxing Mao: modified the manuscript. All the authors contributed to the revision of the manuscript and approved the final manuscript.

Acknowledgements

The authors thank all the participants, coordinators, and administrators for their support and help during the research.

References

1. Barter P, Gotto AM, LaRosa JC, Maroni J, Szarek M, Grundy SM, Kastelein JJP, Bittner V, Fruchart J-C: **HDL cholesterol, very low levels of LDL cholesterol, and cardiovascular events.** *The New England journal of medicine* 2007, **357:**1301-1310.
2. Halcox JP, Banegas JR, Roy C, Dallongeville J, De Backer G, Guallar E, Perk J, Hajage D, Henriksson KM, Borghi C: **Prevalence and treatment of atherogenic dyslipidemia in the primary prevention of cardiovascular disease in Europe: EURIKA, a cross-sectional observational study.** *BMC Cardiovasc Disord* 2017, **17:**160.
3. Cholesterol Treatment Trialists C, Baigent C, Blackwell L, Emberson J, Holland LE, Reith C, Bhala N, Peto R, Barnes EH, Keech A, et al: **Efficacy and safety of more intensive lowering of LDL cholesterol: a meta-analysis of data from 170,000 participants in 26 randomised trials.** *Lancet* 2010, **376:**1670-1681.
4. Mach F, Baigent C, Catapano AL, Koskinas KC, Casula M, Badimon L, Chapman MJ, De Backer GG, Delgado V, Ference BA, et al: **2019 ESC/EAS Guidelines for the management of dyslipidaemias: lipid modification to reduce cardiovascular risk.** *Eur Heart J* 2020, **41:**111-188.
5. Prospective Studies C, Lewington S, Whitlock G, Clarke R, Sherliker P, Emberson J, Halsey J, Qizilbash N, Peto R, Collins R: **Blood cholesterol and vascular mortality by age, sex, and blood pressure: a meta-analysis of individual data from 61 prospective studies with 55,000 vascular deaths.** *Lancet* 2007, **370:**1829-1839.
6. Peters SAE, Muntner P, Woodward M: **Sex Differences in the Prevalence of, and Trends in, Cardiovascular Risk Factors, Treatment, and Control in the United States, 2001 to 2016.** *Circulation* 2019, **139:**1025-1035.
7. Junren Z, Runlin G, Shuiping Z, Guoping L, Dong Z, Jianjun L: **Guidelines for prevention and treatment of dyslipidaemia in Chinese adults (revised in 2016).** *Chinese Circulation Journal* 2016, **31:**937-953.
8. Moran A, Gu D, Zhao D, Coxson P, Wang YC, Chen CS, Liu J, Cheng J, Bibbins-Domingo K, Shen YM, et al: **Future cardiovascular disease in china: markov model and risk factor scenario projections from the coronary heart disease policy model-china.** *Circ Cardiovasc Qual Outcomes* 2010, **3:**243-252.

9. Liu X, Yu S, Mao Z, Li Y, Zhang H, Yang K, Zhang H, Liu R, Qian X, Li L, et al: **Dyslipidemia prevalence, awareness, treatment, control, and risk factors in Chinese rural population: the Henan rural cohort study.** *Lipids Health Dis* 2018, **17:**119.
10. D'Agostino RB, Vasan RS, Pencina MJ, Wolf PA, Cobain M, Massaro JM, Kannel WB: **General cardiovascular risk profile for use in primary care: the Framingham Heart Study.** *Circulation* 2008, **117:**743-753.
11. Wilson PW, Meigs JB, Sullivan L, Fox CS, Nathan DM, D'Agostino RB, Sr.: **Prediction of incident diabetes mellitus in middle-aged adults: the Framingham Offspring Study.** *Arch Intern Med* 2007, **167:**1068-1074.
12. Viti A, Socci L, Congregado M, Ismail M, Nachira D, Munoz CG, Bolufer S, Ruckert JC, Margaritora S, Terzi A: **The everlasting issue of prolonged air leaks after lobectomy for non-small cell lung cancer: A data-driven prevention planning model in the era of minimally invasive approaches.** *J Surg Oncol* 2018, **118:**1285-1291.
13. Kermany DS, Goldbaum M, Cai W, Valentim CCS, Liang H, Baxter SL, McKeown A, Yang G, Wu X, Yan F, et al: **Identifying Medical Diagnoses and Treatable Diseases by Image-Based Deep Learning.** *Cell* 2018, **172:**1122-1131 e1129.
14. Viarasilpa T, Panyavachiraporn N, Marashi SM, Van Harn M, Kowalski RG, Mayer SA: **Prediction of Symptomatic Venous Thromboembolism in Critically Ill Patients: The ICU-Venous Thromboembolism Score.** *Crit Care Med* 2020.
15. Orozco-Beltran D, Quesada JA, Bertomeu-Gonzalez V, Lobos-Bejarano JM, Navarro-Perez J, Gil-Guillen VF, Garcia Ortiz L, Lopez-Pineda A, Castellanos-Rodriguez A, Lopez-Domenech A, et al: **A new risk score to assess atrial fibrillation risk in hypertensive patients (ESCARVAL-RISK Project.** *Sci Rep* 2020, **10:**4796.
16. Wang CJ, Li YQ, Wang L, Li LL, Guo YR, Zhang LY, Zhang MX, Bie RH: **Development and evaluation of a simple and effective prediction approach for identifying those at high risk of dyslipidemia in rural adult residents.** *PLoS One* 2012, **7:**e43834.
17. Marateb HR, Mohebian MR, Javanmard SH, Tavallaei AA, Tajadini MH, Heidari-Beni M, Mananas MA, Motlagh ME, Heshmat R, Mansourian M, Kelishadi R: **Prediction of dyslipidemia using gene mutations, family history of diseases and anthropometric indicators in children and adolescents: The CASPIAN-III study.** *Comput Struct Biotechnol J* 2018, **16:**121-130.
18. Yang X, Xu C, Wang Y, Cao C, Tao Q, Zhan S, Sun F: **Risk prediction model of dyslipidaemia over a 5-year period based on the Taiwan MJ health check-up longitudinal database.** *Lipids Health Dis* 2018, **17:**259.
19. Zhang X, Tang F, Ji J, Han W, Lu P: **Risk Prediction of Dyslipidemia for Chinese Han Adults Using Random Forest Survival Model.** *Clin Epidemiol* 2019, **11:**1047-1055.
20. Smith JA, Ware EB, Middha P, Beacher L, Kardia SL: **Current Applications of Genetic Risk Scores to Cardiovascular Outcomes and Subclinical Phenotypes.** *Curr Epidemiol Rep* 2015, **2:**180-190.

21. Piccolo SR, Abo RP, Allen-Brady K, Camp NJ, Knight S, Anderson JL, Horne BD: **Evaluation of genetic risk scores for lipid levels using genome-wide markers in the Framingham Heart Study.** *BMC proceedings* 2009, **3 Suppl 7:**S46.
22. Buscot M-j, Magnussen CG, Juonala M, Pitkänen N, Lehtimäki T, Viikari JSA, Kähönen M, Hutri-Kähönen N, Schork NJ, Raitakari OT, Thomson RJ: **The Combined Effect of Common Genetic Risk Variants on Circulating Lipoproteins Is Evident in Childhood: A Longitudinal Analysis of the Cardiovascular Risk in Young Finns Study.** *PloS one* 2016, **11:**e0146081.
23. Paquette M, Chong M, Theriault S, Dufour R, Pare G, Baass A: **Polygenic risk score predicts prevalence of cardiovascular disease in patients with familial hypercholesterolemia.** *J Clin Lipidol* 2017, **11:**725-732 e725.
24. Liu X, Mao Z, Li Y, Wu W, Zhang X, Huo W, Yu S, Shen L, Li L, Tu R, et al: **Cohort Profile: The Henan Rural Cohort: a prospective study of chronic non-communicable diseases.** *Int J Epidemiol* 2019, **48:**1756-1756j.
25. Kuwabara M, Kuwabara R, Niwa K, Hisatome I, Smits G, Roncal-Jimenez CA, MacLean PS, Yracheta JM, Ohno M, Lanaspa MA, et al: **Different Risk for Hypertension, Diabetes, Dyslipidemia, and Hyperuricemia According to Level of Body Mass Index in Japanese and American Subjects.** *Nutrients* 2018, **10**.
26. Shen Z, Munker S, Wang C, Xu L, Ye H, Chen H, Xu G, Zhang H, Chen L, Yu C, Li Y: **Association between alcohol intake, overweight, and serum lipid levels and the risk analysis associated with the development of dyslipidemia.** *J Clin Lipidol* 2014, **8:**273-278.
27. Lin HQ, Wu JY, Chen ML, Chen FQ, Liao YJ, Wu YT, Guo ZJ: **Prevalence of dyslipidemia and prediction of 10-year CVD risk among older adults living in southeast coastal regions in China: a cross-sectional study.** *Clin Interv Aging* 2019, **14:**1119-1129.
28. Zhang A, Yao Y, Xue Z, Guo X, Dou J, Lv Y, Shen L, Yu Y, Jin L: **A Study on the Factors Influencing Triglyceride Levels among Adults in Northeast China.** *Sci Rep* 2018, **8:**6388.
29. Liu HH, Li JJ: **Aging and dyslipidemia: a review of potential mechanisms.** *Ageing Res Rev* 2015, **19:**43-52.
30. Zhang L, Wang Y, Niu M, Wang C, Wang Z: **Machine learning for characterizing risk of type 2 diabetes mellitus in a rural Chinese population: the Henan Rural Cohort Study.** *Sci Rep* 2020, **10:**4406.
31. Ambale-Venkatesh B, Yang X, Wu CO, Liu K, Hundley WG, McClelland R, Gomes AS, Folsom AR, Shea S, Guallar E, et al: **Cardiovascular Event Prediction by Machine Learning: The Multi-Ethnic Study of Atherosclerosis.** *Circ Res* 2017, **121:**1092-1101.
32. Dalakleidi K, Zarkogianni K, Thanopoulou A, Nikita K: **Comparative assessment of statistical and machine learning techniques towards estimating the risk of developing type 2 diabetes and cardiovascular complications.** *Expert Systems* 2017, **34:**8.
33. Kullo IJ, Jouni H, Austin EE, Brown S-A, Kruisselbrink TM, Isseh IN, Haddad RA, Marroush TS, Shameer K, Olson JE, et al: **Incorporating a Genetic Risk Score Into Coronary Heart Disease Risk**

Estimates: Effect on Low-Density Lipoprotein Cholesterol Levels (the MI-GENES Clinical Trial). *Circulation* 2016, **133:**1181-1188.

34. Anderson KM, Castelli WP, Levy D: **Cholesterol and mortality. 30 years of follow-up from the Framingham study.** *JAMA* 1987, **257:**2176-2180.
35. Emerging Risk Factors C, Di Angelantonio E, Gao P, Pennells L, Kaptoge S, Caslake M, Thompson A, Butterworth AS, Sarwar N, Wormser D, et al: **Lipid-related markers and cardiovascular disease prediction.** *JAMA* 2012, **307:**2499-2506.

Figures

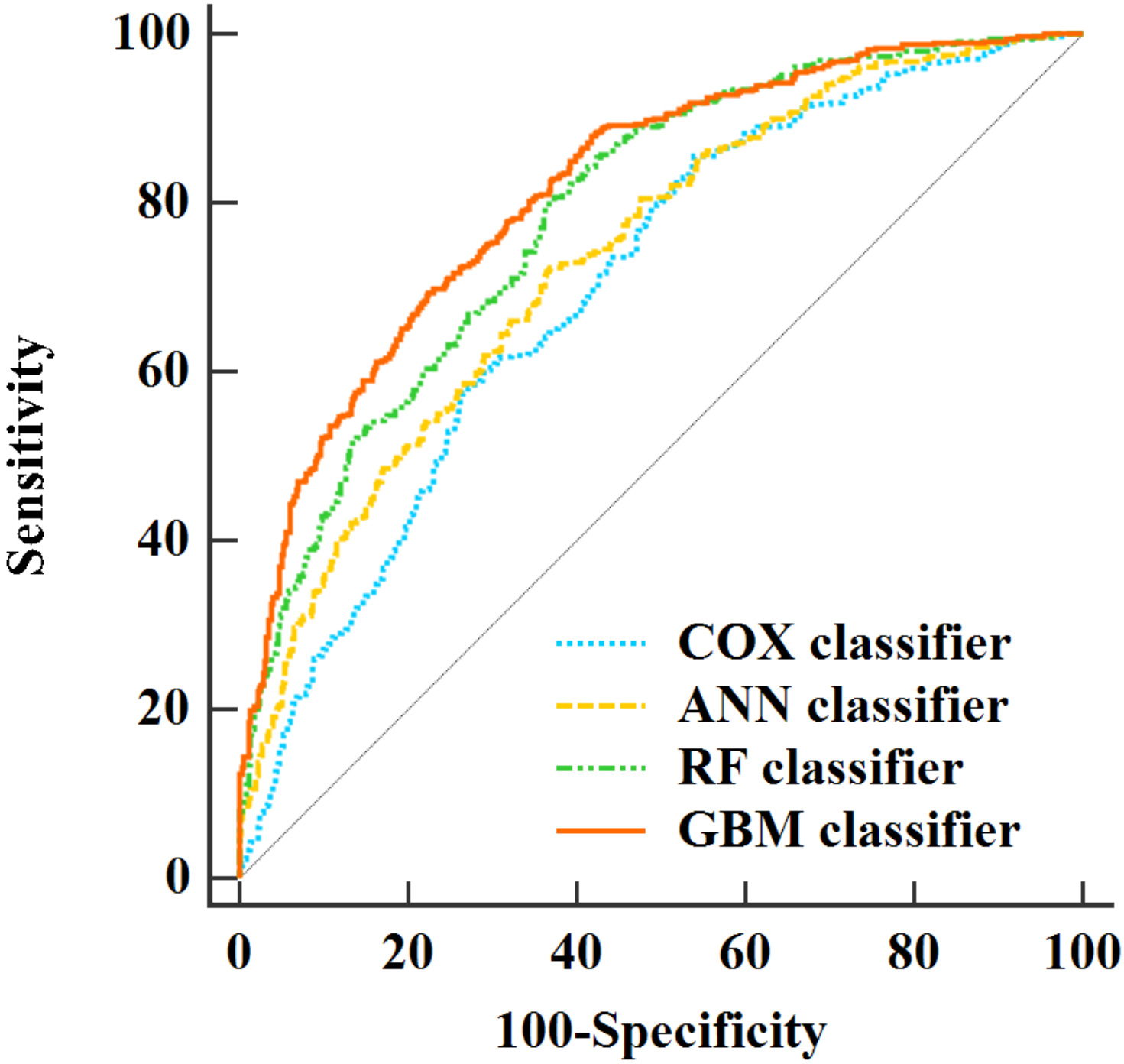

Figure 1

Receiver-operating characteristic curves of conventional models with four classifiers. Abbreviations: ANN: artificial neural network; RF: random forest; GBM: gradient boosting machine.

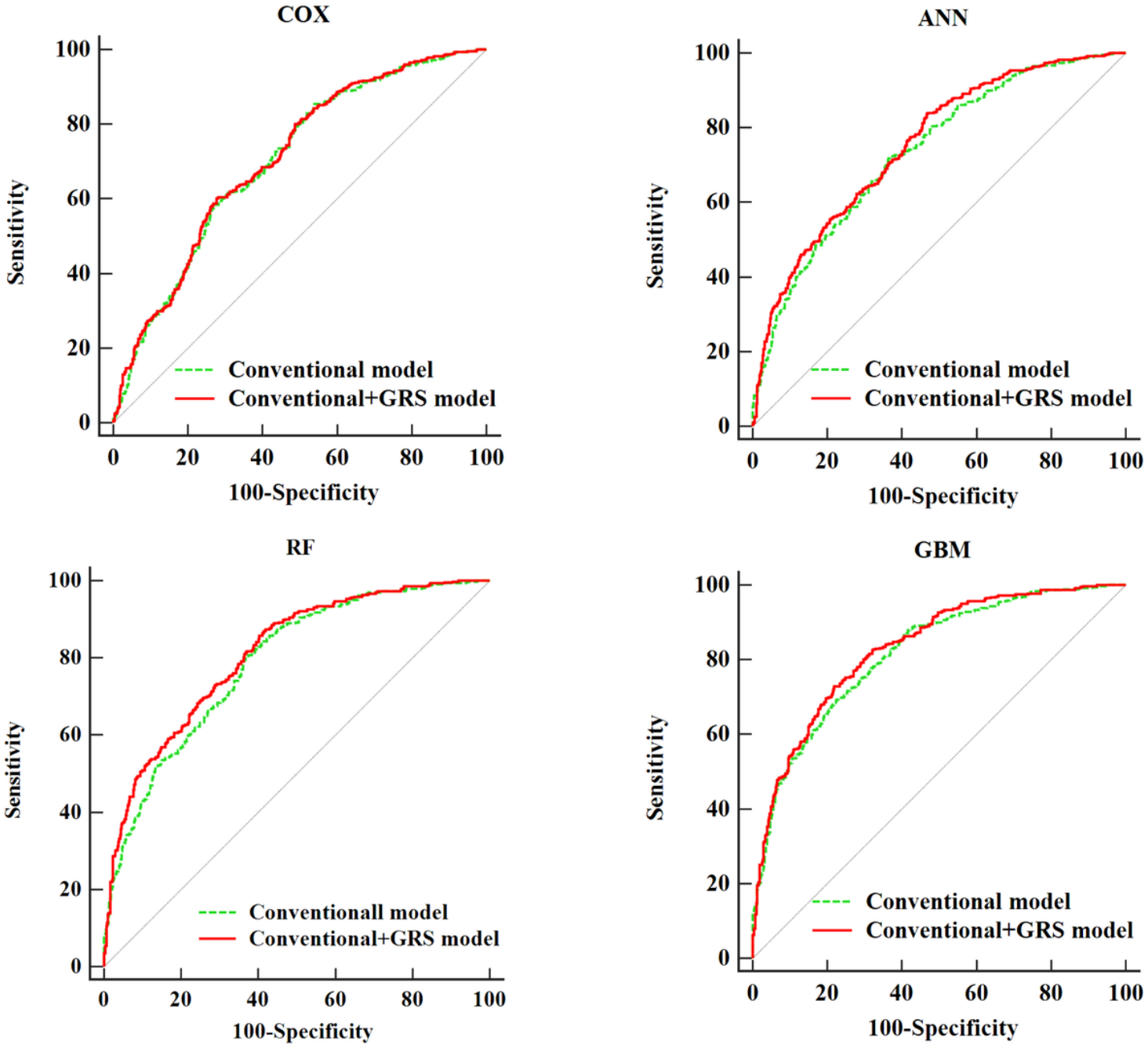

Figure 2

Receiver-operating characteristic curves of conventional model and conventional+GRS model with four classifiers. Note: Age, family history of diabetes, physical activity, BMI, TG, HDL-C, LDL-C are included in conventional model; conventional+GRS model includes age, family history of diabetes, physical activity, BMI, TG, HDL-C, LDL-C and GRS. Abbreviations:ANN: artificial neural network; RF: random forest; GBM: gradient boosting machine.

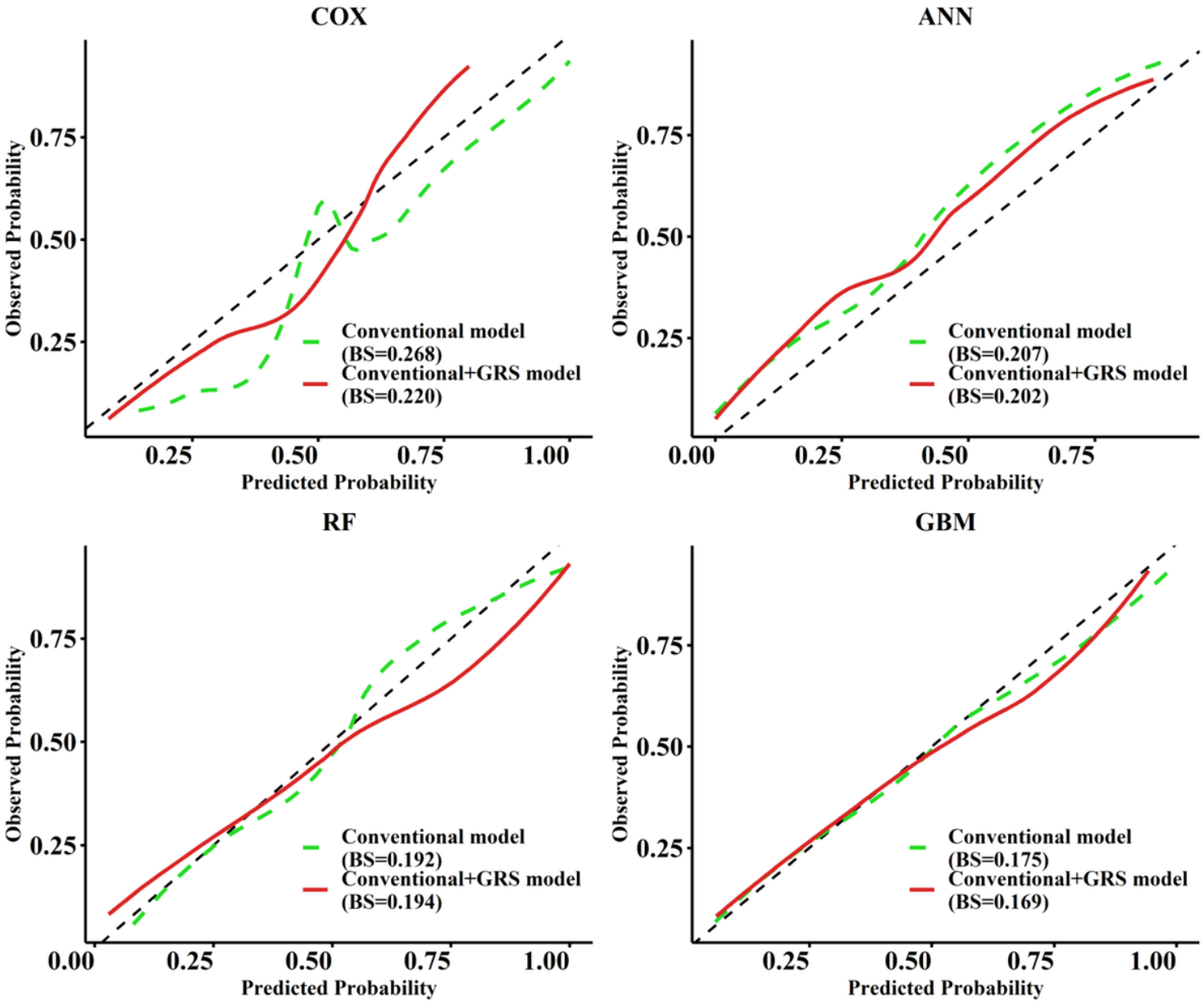

Figure 3

Calibration curves of conventional model and conventional+GRS model with four classifiers. Note: The dotted gray line is an ideal curve, indicating that the predicted probability is consistent with the observed one. The lower the brier score, the better the predictive calibration. Abbreviations:ANN: artificial neural network; RF: random forest; GBM: gradient boosting machine; BS: brier score.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Additionalfile1.docx