Development and internal validation of risk prediction model of metabolic syndrome in oil workers

Jie Wang
North China University of Science and Technology
Chao Li
North China University of Science and Technology
Jing Li
North China University of Science and Technology
Sheng Qin
North China University of Science and Technology
Chunlei Liu
North China University of Science and Technology
Jiaojiao Wang
North China University of Science and Technology
Zhe Chen
North China University of Science and Technology
Jianhui Wu (wujianhui555@163.com)
North China University of Science and Technology
Guoli Wang
North China University of Science and Technology

Research article

Keywords: Data mining, oil workers, Metabolic syndrome, Risk prediction

Posted Date: July 2nd, 2020

DOI: https://doi.org/10.21203/rs.3.rs-31038/v2

Version of Record: A version of this preprint was published on November 30th, 2020. See the published version at https://doi.org/10.1186/s12889-020-09921-w.

Abstract

Background. The prevalence of metabolic syndrome continues to rise sharply worldwide, seriously threatening people's health. The optimal model can be used to identify people at high risk of metabolic syndrome as early as possible, to predict their risk, and to persuade them to change their adverse lifestyle so as to slow down and reduce the incidence of metabolic syndrome.

Objective. To develop and internally verify three risk prediction models for the metabolic syndrome of petroleum workers, compare the prediction performance of the three models, and find the optimal model.

Methods. Design existing circumstances research. A total of 1,468 workers from an oil company who participated in occupational health physical examination from April 2017 to October 2018 were included in this study. We established the Logistic regression model, the random forest model and the convolutional neural network model, and compared the prediction performance of the models according to the F1 score, sensitivity, accuracy and other indicators of the three models.

Results. The results showed that the accuracy of the three models in the training set was 83.45%, 94.21% and 86.34%, the sensitivity was 78.47%, 94.62% and 81.30%, the F1 score was 0.79, 0.93 and 0.83, the area under the ROC curve was 0.894, 0.987 and 0.935, and the Integrated Calibration Index was 0.074, 0.071 and 0.078, respectively. In the test set, the accuracy was 76.72%, 80.66% and 78.69%, the sensitivity was 70.00%, 77.50% and 68.33%, the F1 score was 0.70, 0.76 and 0.71, the area under the ROC curve was 0.797, 0.861 and 0.855, and the Integrated Calibration Index was 0.064, 0.051 and 0.096, respectively.

Conclusions. The study showed that the prediction performance of random forest model is better than other models, and the model has higher application value, which can better predict the risk of metabolic syndrome in oil workers, and provide corresponding theoretical basis for the health management of oil workers.

1. Introduction

Metabolic syndrome (MetS) refers to the accumulation of multiple metabolic risk factors in the body including obesity, impaired glucose regulation, dyslipidemia and hypertension. MetS is a group of complex clinical syndromes based on insulin resistance. Relevant literatures have shown that metabolic syndrome increases the risk of cardiovascular disease, type 2 diabetes and chronic kidney disease **[1-3]**.With the social and economic development and changes in people's lifestyles, the prevalence of metabolic syndrome has increased year by year and brought a heavy economic burden, which has become an important health issue of common concern to people worldwide.

At present, the definition and diagnostic criteria of metabolic syndrome have not been completely unified. In 1998, WHO officially named the "metabolic syndrome" and and proposed corresponding diagnostic criteria for the first time **[4]**.Over the course of the next decade, the diagnostic criteria for metabolic syndrome have undergone many changes and revisions, including 2001 national cholesterol education

program adult treatment group report for the third time (NCEP ATP □). Chinese diabetes association (CDS) diagnostic criteria in 2004. International diabetes federation (IDF) diagnostic criteria 2005. In 2009, the American heart association (AHA), the international diabetes federation, the national heart, lung and blood institute and other institutions jointly proposed a tentative unified standard **[5-8]**.According to a large number of epidemiological data, the global prevalence of MetS is about 30%**[9]**. DoosupShin based on 2007-2014 national health and nutrition survey data on MetS prevalence statistics found that American adults MetS prevalence rate has reached 34.3% (according to the revised NCEP-ATP □ diagnostic criteria) **[10]**.In South Korea, according to the same diagnostic criteria, the prevalence rate of metabolic syndrome in adults from 2009 to 2013 was as high as 30.52%**[11]**.In China, in 2010, Jieli Lu**[12]** and others conducted a data report analysis of 97,098 adults in China, and estimated the prevalence of MetS was 33.9% (according to the NCEP-ATP □ diagnostic criteria).In 2015, Ting Liu analyzed the prevalence of MetS among 34,025 residents in Jilin province and found that the prevalence of MetS was 32.5% (according to IDF diagnostic criteria) **[13]**.In 2016, Ri Li**[14]**and others conducted a meta-analysis showing that the prevalence of MetS in subjects over 15 years old was 24.5% (according to IDF diagnostic criteria).Although the diagnostic criteria are not uniform, it is undeniable that metabolic syndrome has become one of the chronic diseases with high incidence in China and even in the world.

Data mining refers to extracting hidden information and knowledge with potential research value from large data, which is often used in the medical field with large amounts of data and fast update speed. Among them, the classification algorithm has been widely concerned and applied in recent years. This algorithm takes a variety of risk factors affecting the occurrence of disease as a prerequisite, and uses statistical methods and computer algorithms to build a predictive model of disease risk. The constructed model is used to predict the probability of a certain population or individual developing a certain disease, and then provides a theoretical basis for personal health management and corresponding preventive measures**[15]**.At present, Logistic regression, Cox regression, BP neural network, decision tree, support vector machine and other models are mostly used to construct metabolic syndrome risk models at home and abroad **[16-18]**. These models can be used to identify high-risk groups of MetS, persuade them to change their unhealthy lifestyles, reduce and slow down the occurrence and development of the disease. Among them**[19-21]**, Logistic regression and Cox regression, as traditional statistical modeling methods, are widely used and have strong explanatory power. However, Cox regression is often used for survival analysis data, which requires two dependent variables at the same time and has relatively strict requirements on data. The decision-making tree model has strong visibility, but is prone to overfitting and poor generalization effect. The random forest model is a classifier composed of multiple decision-making trees, which improves the weak generalization ability of a single decision-making tree and balances the error of unbalanced data. As a kind of artificial neural network model, BP neural network is fault-tolerant to some extent, but local minimization problems often occur, and the learning speed is slow, and the phenomenon of overfitting is easy to occur. In the convolutional neural network model, the local receptive field and weight sharing of convolutional kernel reduce the computational complexity and have high accuracy and good generalization ability. Due to regional and cultural differences, the effects of existing models vary, and mature and accurate metabolic syndrome risk prediction systems have not been

established at home and abroad. Moreover, most of these models established at present are aimed at the assessment of the risk of disease in the general population, ignoring the special group of occupational population.

As an important part of China's non-renewable energy industry, the petroleum industry still accounts for a large proportion in the national economy. Oil workers are also the main laborers in the production of the secondary industry in China. Their health will affect the development of China's economy to a certain extent and should be paid more attention. Oil workers are affected by high temperature, noise, shift work and other harmful occupational factors, as well as a variety of adverse lifestyles caused by occupational stress, which can greatly increase the incidence of metabolic syndrome to some extent. For special occupational group, the risk prediction model of ordinary people is no longer suitable for them, so it is necessary to establish a risk prediction model of metabolic syndrome for them, so as to achieve early detection, diagnosis and treatment, and protect the health of oil workers. In this study, a certain oil industry worker was selected as the research object, and the traditional Logistic regression model, random forest model and the recent thermal convolutional neural network model were developed and internally verified. The prediction performance of each model is compared to find the optimal model, which provides a theoretical basis for the health management of this special occupation group of oil workers.

2. Methods

2.1 Data sources and Research objects

This study adopted the existing circumstances research method. The sample data was partitioned in the modeling process , with 60% as training set, 20% as verification set and 20% as test set. A total of 1,468 workers from an oil company who attended occupational examination and physical examination from April 2017 to October 2018 were selected as the research objects. Inclusion criteria: length of service 1 year or above. Aged between 18 and 60. Complete questionnaire and physical examination data. All subjects gave their informed consent for inclusion before they participated in the study. The study was conducted in accordance with the Declaration of Helsinki, and the protocol was approved by the Ethics Committee of North China University of Science and Technology(NO.16040).

2.2 Outcomes and Predictor variables

One-to-one questionnaire survey was conducted on oil workers by uniformly trained personnel to collect the following information: General situation: gender, age, education, income status, marital status, etc. Lifestyle: smoking, drinking, diet and physical exercise. history of personal and family diseases: hyperglycemia, hypertension, hyperlipidemia, etc. Working conditions: shifts, exposure to high temperature, noise and other harmful factors. Physical examination: height, weight, blood pressure and waist circumference, etc.

The study subjects took venous blood in the early morning after fasting for 12 hours, and tested the biochemical indicators such as fasting blood glucose, high-density lipoprotein, and triglyceride using the Dirion CS-1200 automatic biochemical analyzer (China Changchun Dirion Medical Technology Company). The diagnostic criteria of metabolic syndrome **[8]** can be diagnosed if it meets three or more of the following five indicators:

. Central obesity: Chinese people have a waist circumference ≥85cm (male). waist circumference ≥80cm (female).

. Elevated blood glucose: FBG ≥5.6mmol/L or those who have been diagnosed with diabetes and receive treatment.

. TG≥1.7mmol/L or those who have been diagnosed with hypertriglyceridemia and received treatment.

. HDL-C <1.04mmol / L (male). HDL-C <1.30mmol/L (female) or those who have been diagnosed with low-density lipoproteinemia and received treatment.

. Systolic / diastolic blood pressure ≥130/85mmHg or those diagnosed with hypertension and receiving treatment.

2.3 Quality control

The investigators can only take up their posts after unified training. The collected questionnaire data are collected on the spot for double and double check and input, and the questionnaires with incorrect input are checked for the third time to ensure the accuracy of the collected data. The same instrument was used for physical examination and laboratory test, and the biochemical indicators were tested by the same kit in North China Petroleum Underground Hospital.

2.4 Sample size

Through consulting a large number of relevant literatures, it was found that there were about 15 predictive factors related to metabolic syndrome.General neural network and random forest model require that the sample content is more than 2 times of explanatory variables.The newly developed Logistic regression model $R^2_{CS\_adj}$ (the estimated measure after adjusting the overfitting of the model) is at least 0.1, so to achieve the expected contraction coefficient of 0.9**[22]**, we finally need a sample size of at least 1274.

2.5 Statistical methods

CscrMainUI system developed by a scientific research company was used to scan and input questionnaires and establish a database. IBM SPSS19.0 was used for statistical analysis. The measurement data obeying the normal distribution were expressed as ±s, and the t test was used for comparison between groups. The non-normally distributed measurement data were represented by [M

(P25,P75)], and the rank sum test was used for comparison between groups. The count data were used as the ratio, and Pearson x^2 test was used for comparison between groups. Unconditional binary classification logistic regression was used for multivariate analysis. The independent variable introduction criterion was $P \leq 0.05$, and the test level $\alpha=0.05$(both sides).

2.6 Establishment and validation of the models

Input variables of the three models: predictors of metabolic syndrome of oil workers determined by multivariate logistic regression analysis and results of a large number of relevant literature reviews. The output variable was whether metabolic syndrome occurred. Logistic regression model (using forced entry method) and random forest model were constructed by SPSS Modeler18.0 (set the number of base classifiers as 100, set the sample number of data set used by each base classifier as 100, the maximum node number as 10000, the maximum tree depth as 10, and the minimum size as 5). The convolutional neural network model is constructed by using Pytorch (input 4*4 matrix, convolution kernel 3*3, step length =1, padding=1, maximum pooling, size 2*2, step length =1, input 144 in full connection layer, output 2).ROC curve was drawn with Medcalc and the area under the curve was compared.Cross-validation was used for internal validation of the model, and the predictive performance of the three models was compared with F1 score, sensitivity, specificity and area under ROC curve.

3. Results

1. 1 General situation

Of the 1,468 oil workers, 1,105 were male, with an average age of 43(38,48),363 were women, with an average age of 44(42,47). The prevalence rate of metabolic syndrome in petroleum workers was 40.67%, among which, the rate of central obesity was 56.81%, the rate of abnormal blood glucose was 49.39%, the rate of abnormal triglyceride was 32.90%, the rate of abnormal HDL was 19.28%, and the rate of abnormal blood pressure was 55.99%. As shown in Figure 1.

3.2 Independent variable screening

Single factor analyses were performed on the basic conditions, diet and lifestyle, occupational exposure factors and laboratory tests of 1,468 oil workers. The results showed statistically significant differences in age, gender, Body Mass Index(BMI), marital status, family history of hypertension, family history of diabetes mellitus, salt, meat intake, smoking status, drinking status, shift work situation, Occupational heat, noise, hemoglobin, uric acid(UA), alanine transaminase(ALT), etc ($P < 0.05$), are shown in table 1 to table 4.

The significant factors of univariate analysis were included in the multivariate nonconditional Logistic regression analysis. The results showed that the risk of metabolic syndrome increased with age, BMI, UA and ALT. People with a family history of diabetes, a strong salt taste, occasional consumption of dairy products, daily consumption of carbonated beverages, smoking, shift work, and exposure to high

temperatures are more likely to develop metabolic syndrome. The protective factors of metabolic syndrome include family income of 2000-3000 yuan per capita, daily consumption of dairy products and physical exercise. Combined with the results of relevant literature review, 13 significant factors in the multivariate analysis were taken as independent variables for the establishment of the model, as shown in table 5-6.

3.3 Collinearity diagnosis

The diagnosis of collinearity was made by using the binary correlation coefficient r, tolerance and variance inflation factor(VIF).The results showed that the correlation coefficient |r| was 0.31 at most and |r|<0.5, as shown in table 7.The minimum tolerance was 0.844, much higher than 0.1, and the maximum variance inflation factor was 1.185, less than 5, as shown in table 8.The above results indicate that there is no serious multicollinearity among the screened independent variables.

3.4 Logistic regression model

Logistic regression model in the training set, validation set and test set accuracy of 83.45%, 80.60% and 76.72% respectively, the sensitivity of 78.47%, 69.35% and 70.00% respectively, the specificity of 86.89%, 88.57% and 81.08% ,respectively, F1 score was 0.79, 0.75, 0.70,Youden's index was 0.65, 0.58, 0.51, positive likelihood ratio was 5.98, 6.07, 3.70, and negative likelihood ratio was 0.25, 0.35, 0.37, Kappa value was 0.66, 0.59, and 0.51, respectively, and the area under the ROC curve (AUC) was 0.894, 0.875, and 0797, respectively, and ICI (Integrated Calibration Index) was 0.074, 0.074, and 0.064, respectively. As shown in table 9-10.

3.5 Random forest model

Random forest model in the training set, validation set and test set accuracy of 94.21%, 81.27%, 80.66% respectively, the sensitivity of 94.62%, 77.42% and 77.50% respectively, the specificity of 93.93%, 84.00% and 82.70%, respectively, F1 score was 0.93, 0.77, 0.76, Youden's index was 0.89, 0.61, 0.60, positive likelihood ratio was 15.60, 4.84, 4.48, and negative likelihood ratio was 0.06, 0.27, 0.27, Kappa value was 0.88, 0.61, 0.60, and AUC values was 0.987, 0.878, and 0.861, respectively, and ICI was 0.071, 0.072, and 0.051, respectively. As shown in table 9-10.

3.6 Convolutional neural network model

Convolution neural network (CNN) model in the training set, validation set and test set accuracy of 86.34%, 82.61%, 78.69% respectively, the sensitivity of 81.30%, 73.39% and 68.33% respectively, the specificity of 89.82%, 89.14% and 85.41%, respectively, F1 score was 0.83, 0.78, 0.71, Youden's index was 0.71, 0.63, 0.54, positive likelihood ratio was 7.99, 6.76, 4.68, and negative likelihood ratio was 0.21, 0.30, 0.37, Kappa value was 0.72, 0.64, 0.55, and AUC values was 0.935, 0.872, and 0.855, respectively, and ICI was 0.078, 0.082, and 0.096, respectively. As shown in table 9-10.

3.7 Comparison of predictive performance of metabolic syndrome risk prediction models

In the training set, the accuracy, sensitivity, specificity, F1 score, Youden's index, positive likelihood ratio, Kappa index, positive predictive value and negative predictive value of the random forest model were all higher than those of the Logistic regression model and the convolutional neural network model. The area under ROC curve (AUC) of the random forest model was larger than that of the Logistic regression model and the convolutional neural network model, and the difference was statistically significant (P < 0.001). The calibration diagrams of Logistic regression model, Random forest model and CNN risk prediction models in the training set were all close to the diagonal, and there was no serious deviation from the calibration results.See table 11 and figure 2.

In the validation set, The accuracy, sensitivity, specificity, F1 score and other indexes of the three models were all higher. In order to further reflect the relationship between sensitivity and specificity, it is necessary to judge whether the models are overfitting and have good robustness. By plotting ROC curve and calculating AUC value, it was found that the three curves of Logistic regression model, random forest model and convolutional neural network model were basically identical, with no statistically significant difference (P>0.05). The area under the curve (AUC) was 0.875, 0.878 and 0.872 respectively.The calibration diagrams of Logistic regression model, Random forest model and CNN risk prediction models in the validation set were all close to the diagonal. See table 10,11 and figure 3.

In the test set, the accuracy, sensitivity, F1 score, Youden's index, Kappa index and negative predictive value of the random forest model were the highest, while the specificity, positive likelihood ratio and positive predictive value of the convolutional neural network model were the highest, but the sensitivity and negative predictive value were the lowest. The area under ROC curve (AUC) of the random forest model was larger than that of the Logistic regression model and the convolutional neural network model. Comparing the AUC of the three models in pairs, the difference between Logistic regression model and random forest model was statistically significant (Z=2.806, P=0.005), the difference between Logistic regression model and convolutional neural network model was statistically significant (Z=2.352, P=0.019), and the difference between random forest model and convolutional neural network model was not statistically significant (Z=0.320, P=0.749).The calibration diagrams of Logistic regression model, Random forest model and CNN risk prediction models in the test set were all close to the diagonal. See table 11 and figure 4.

Discussion

At present, all countries in the world have recognized that the establishment of disease risk prediction model has a greater role in preventing and controlling the occurrence of metabolic syndrome, and established the corresponding MetS model based on the epidemiological data. In 2008, Fabien Szabo DE Edelenyi et al. in France conducted a large case-control study and found that the prediction accuracy of metabolic syndrome status using random forest classification technique was 71.70%□72.10% in the control group and 70.70% in the case group)**[23]**.In 2010, Lin CC in Taiwan established an artificial neural network model and a Logistic regression model to identify metabolic syndrome in 383 patients with schizophrenia, and the results showed that the accuracy was 88.30% and 83.60%, the sensitivity was

93.10% and 86.20%, and the specificity was 86.90% and 83.80%, respectively **[24]**. In 2015, Worachartcheewan**[25]** et al. used the random forest model to establish a prediction model of metabolic syndrome for 5,646 adults living in Bangkok, and the accuracy was 98.11%.In 2016, karimi-alavijeh et al. used 2107 participants in the Iranian cohort study to establish the decision-making tree model and support vector machine model, and found that the accuracy was 73.90% and 75.70%, the sensitivity was 75.80% and 77.40%, and the specificity was 72.00% and 74.00%**[26]**.The established models have local applicability advantages due to the differences in region, population and input variables.

The results of this study showed that the prevalence of MetS in workers of an oil company was 40.67%, higher than the average level of Chinese adults **[12-14]**.At the same time, the prevalence rate of the five diagnostic criteria of metabolic syndrome ranged from high to low, which were: central obesity, abnormal blood pressure, abnormal blood glucose, abnormal triglyceride, and abnormal high-density lipoprotein. The occurrence of this phenomenon was related to the generally good living conditions, dietary habits, irregular life and rest oil workers. Independent variable screening found that age, income, BMI, family history of diabetes, salt intake and physical exercise were all influencing factors of metabolic syndrome, which was consistent with previous research results **[27-28]**.Different from the general population, oil workers have been in a special occupational environment for a long time. High temperature environment causes the body's circulatory system to be in a long-term stress state, resulting in decreased elasticity of blood vessel wall, increased blood viscosity, and increased blood pressure. In addition, studies have shown that high temperature contact can affect insulin hemodynamics, resulting in insulin resistance in the body**[29-30]**.Harmony between biological rhythm and natural rhythm is the basis of normal physiological activities. Irregular shift work will affect the biological rhythm of human body due to irregular circadian rhythm, resulting in the disturbance of nutrients and related hormones in the body, thus resulting in glucose and lipid metabolism disorder and energy imbalance **[31]**.On the other hand, the workers of night shift work lack of sleep time, and the incidence of unhealthy lifestyle such as smoking, drinking and irregular diet increases greatly, which are the driving forces for the occurrence of metabolic syndrome **[32]**.In this study, UA and ALT were found to be risk factors for MetS, and related studies showed that UA increased the risk of MetS by increasing insulin resistance, and increased ALT in the blood might cause fat accumulation in the liver. Through investigation, Mandana Khalili et al. found that patients with MetS had higher hepatic steatosis level, and there was a correlation between the elevation of ALT and MetS**[33-34]**.

In this study, Logistic regression model, random forest model and convolutional neural network model were established to compare their prediction performance. It was found that the random forest model was suitable for prediction model of MetS risk of oil workers. The prediction performance of the training set of the random forest model was higher than that of the Logistic regression model and the convolutional neural network model, and the difference was statistically significant. However, the specificity of the random forest model in the test set was slightly weaker than that of the convolutional neural network model, and the difference was not statistically significant. In general, the training ability of the model is directly proportional to the testing ability. On one hand, the above reasons may be due to the limitation of the sample size, which is not large enough, leading to poor model effect, on the other hand,

the instability of the network, the setting of parameters and the selection of input variables may affect the prediction performance of the model. In addition, although the specificity of the convolutional neural network model was high in the test set, its sensitivity was too low. As a prediction model for the risk of metabolic syndrome in petroleum workers, the model with higher sensitivity is more suitable for the early detection of patients, so as to play a real role in early detection, early diagnosis and early treatment of the disease, namely secondary prevention of the disease. As an emerging machine learning algorithm in recent years, random forest model**[35-36]** is a highly flexible classifier containing multiple decision trees. The random forest model solves the shortcoming of the decision tree algorithm, and adopts the random sampling method to enhance the generalization ability. Proposed by Yann Lecun of New York university in 1988, the convolutional neural network model is the first truly successful deep learning method using multi-layer hierarchical network, including input layer, hidden layer (convolutional layer, pooling layer, full connection layer) and output layer, which effectively reduces the number of network parameters and significantly reduces the computational complexity. Previously, convolutional neural network was mainly used for image, language and medical imaging processing. In recent years, it has also been used as a neural network model to predict the risk of various diseases **[37-39]**.However, the prediction effect of CNN for different diseases is uneven, which may be because the model construction needs to be further improved and there is no unified standard yet. At the same time, a certain amount of data is required for model training. Logistic regression model is a traditional statistical modeling method, which is widely used in the field of risk factor screening and disease prediction. It is convenient to use and the meaning of the parameters is clear, but it cannot solve the nonlinear problems and the application conditions are strict. The sample size increases with the increase of input variables, and the predictive power decreases when the data do not meet the requirements **[40]**.

Due to the limitation of research conditions, this study has certain limitations. This paper only developed and internally validated the metabolic syndrome risk prediction model for oil workers, and did not conduct external validation of the model. The choice of model input variables will directly affect the prediction effect of the model, which needs to be further explored.This study was based on a cross-sectional study.Only the prevalence data of metabolic syndrome of oil workers were available, and the causal relationship between the prevalence and predictive factors could not be determined.

Conclusions

Three risk prediction models (Logistic regression model, random forest model and convolutional neural network model) for the occurrence of metabolic syndrome in petroleum workers were established and compared. It is found that the random forest model has better discriminant degree and calibration degree in both training set and test set, and shows higher robustness. It shows that the random forest model can predict the risk of metabolic syndrome in oil workers more accurately, and can provide health education for high-risk employees with metabolic syndrome and put forward corresponding prevention strategies, so as to improve the allocation of national medical and health resources and the distribution of health services.

Abbreviations

MetS: Metabolic Syndrome

WHO:World Health Organization

NCEP ATP □:National Cholesterol Education Program Adult Treatment group report for the third time

CDS:Chinese Diabetes association

IDF:International Diabetes Federation

AHA:American Heart Association

BMI:Body Mass Index

RBC:Red Blood Cell

MCV:erythrocyte Mean Corpuscular Volume

BPC:Blood Platelet Count

MPV:Mean Platelet Volume

UA:Uric Acid

ALT:Alanine transaminase

OR:Odds ratio

95%CI:95% Confident limit

SE:Standard Error

VIF:Variance Inflation Factor

CNN:Convolution Neural Network

AUC:Area Under the Curve

ICI:Integrated Calibration Index

Declarations

Ethics approval and consent to participate

All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards.The study was approved by the Ethics Committee of North China University of Science and Technology(NO.16040).Informed consent was obtained from all individual participants included in the study.

Consent for publication

Not applicable.

Availability of data and materials

The data that support the findings of this study are available from [Institute of basic medicine, Chinese academy of medical sciences] but restrictions apply to the availability of these data, which were used under license for the current study, and so are not publicly available. Data are however available from the authors upon reasonable request and with permission of [Institute of basic medicine, Chinese academy of medical sciences].

Competing interests

The authors declare that they have no competing interests.

Funding

This work was supported by the National Key R&D Program of China (No.2016YFC0900605)

Authors' contributions

Design research, J.W. and J.H.W.; Methodology, C.L.L., Z.C. and G.L.W.; Project administration, C.L., J.L., S.Q. and J.J.W.; Software, J.W. and C.L.L.; Validation, J.H.W. and G.L.W.; Writing original draft, J.W.; Writing review, J.W. and J.H.W..All authors responded to the modification of the study protocol and approved the final manuscript.

Acknowledgements

We would like to thank for the support from the National Key R&D Program of China.We would also like to thank north China university of science and technology for providing a software and hardware platform and financial support to ensure the smooth progress of this research.We would also like to thank our teachers and classmates for their help and warmth in the research process.

References

[1]Li W, Song F, Wang X, et al. Relationship between metabolic syndrome and its components and cardiovascular disease in middle-aged and elderly Chinese population: a national cross-sectional survey.

BMJ Open. 2019; 9(8): 1-8.

[2]Low S, Khoo K C J, Wang J, et al. Development of a metabolic syndrome severity score and its association with incident diabetes in an Asian population—results from a longitudinal cohort in Singapore. Endocrine. 2019; 65(1): 73-80.

[3]Chen J, Kong X, Jia X, et al. Association between metabolic syndrome and chronic kidney disease in a Chinese urban population. Clinica Chimica Acta. 2017; 470: 103-108.

[4]Alberti K G M M, Zimmet P Z. Definition, diagnosis and classification of diabetes mellitus and its complications. Part 1: diagnosis and classification of diabetes mellitus. Provisional report of a WHO consultation. Diabetic medicine. 1998; 15(7): 539-553.

[5]Kuhar M B. Executive Summary of The Third Report of The National Cholesterol Education Program (NCEP) Expert Panel on Detection, Evaluation, and Treatment of High Blood Cholesterol In Adults (Adult Treatment Panel III). Circulation. 2001; 106(25): 2486-2497.

[6]Metabolic Syndrome Research Group of Diabetes Branch of Chinese Medical Association. Recommendations of the Chinese Medical Association Diabetes Branch on Metabolic Syndrome. Chinese Journal of Diabetes. 2004; 12(3): 156-161.

[7]Alberti K G, Zimmet P, Shaw J. Metabolic syndrome-a new world-wide definition.A Consensus Statement from the International Diabetes Federation. Diabetic medicine: a journal of the British Diabetic Association. 2006; 23(5): 469-480.

[8]Alberti K, Eckel R H, Grundy S M, et al. Harmonizing the metabolic syndrome: a joint interim statement of the international diabetes federation task force on epidemiology and prevention; national heart, lung, and blood institute; American heart association; world heart federation; international atherosclerosis society; and international association for the study of obesity. Circulation. 2009; 120(16): 1640-1645.

[9]Al-Thani M H, Al-Thani A A M, Cheema S, et al. Prevalence and determinants of metabolic syndrome in Qatar: Results from a National Health Survey. BMJ Open. 2016; 6(9): 1-10.

[10]Shin D, Kongpakpaisarn K, Bohra C. Trends in the prevalence of metabolic syndrome and its components in the United States 2007–2014. International journal of cardiology. 2018; 259: 216-219.

[11]Lee S E, Han K, Kang Y M, et al. Trends in the prevalence of metabolic syndrome and its components in South Korea: Findings from the Korean National Health Insurance Service Database (2009–2013). PloS one. 2018; 13(3): 1-12.

[12]Lu J, Wang L, Li M, et al. Metabolic syndrome among adults in China: the 2010 China noncommunicable disease surveillance. The Journal of Clinical Endocrinology & Metabolism. 2017; 102(2): 507-515.

[13]Ting Liu. Prevalence and Risk Factors of Metabolic Syndrome among Residents in Jilin Province. Jilin University. 2017.

[14]Li R, Li W, Lun Z, et al. Prevalence of metabolic syndrome in mainland china: a meta-analysis of published studies. BMC Public Health. 2016; 16(1): 296-306.

[15]Li-Qiang Zheng, Rui Zhang. Research progress on evaluation methods of fit degree of disease risk prediction model. China Health Statistics. 2015; 32(3): 544-546.

[16]Choe E K, Rhee H, Lee S, et al. Metabolic Syndrome Prediction Using Machine Learning Models with Genetic and Clinical Information from a Nonobese Healthy Population. Genomics & informatics. 2018; 16(4): 1-7.

[17]Worachartcheewan A, Schaduangrat N, Prachayasittikul V, et al. Data mining for the identification of metabolic syndrome status. EXCLI Journal. 2018; 17: 72-88.

[18]Dong-Yu Mu, Ya Ma, Xiao-Fan J, et al. Influencing factors and risk forecast model of metabolic syndrome among college faculties, Chengdu. Modern Preventive Medicine. 2019; 46(1):36-42.

[19]Fatekurohman M, Nurmala N, Anggraeni D. Comparison of exact, efron and breslow parameter approach method on hazard ratio and stratified cox regression model. Journal of Physics Conference Series. 2018; 1008(1): e012007.

[20]Sohrabi S, Atashi A, Dadashi A, et al. A Comparative Study of Multilayer Neural Network and C4. 5 Decision Tree Models for Predicting the Risk of Breast Cancer. Archives of Breast Cancer. 2018; 5(1): 11-14.

[21]Tran D P, Hoang V D. Adaptive Learning Based on Tracking and ReIdentifying Objects Using Convolutional Neural Network. Neural Processing Letters. 2019; 50(1): 263-282.

[22]Riley R D, Snell K I, Ensor J, et al. Minimum sample size for developing a multivariable prediction model: PART II - binary and time-to-event outcomes. Statistics in Medicine. 2019; 38: 1276-1296.

[23]de Edelenyi F S, Goumidi L, Bertrais S, et al. Prediction of the metabolic syndrome status based on dietary and genetic parameters, using Random Forest. Genes & nutrition. 2008; 3(3): 173-176.

[24]Lin C C, Bai Y M, Chen J Y, et al. Easy and low-cost identification of metabolic syndrome in patients treated with second-generation antipsychotics: artificial neural network and logistic regression models. Journal of Clinical Psychiatry. 2010; 71(3): 225-234.

[25]Worachartcheewan A, Shoombuatong W, Pidetcha P, et al. Predicting metabolic syndrome using the random forest method. The Scientific World Journal. 2015; 2015:1-10.

[26]Karimi-Alavijeh F, Jalili S, Sadeghi M. Predicting metabolic syndrome using decision tree and support vector machine methods. ARYA atherosclerosis. 2016; 12(3): 146-152.

[27]Soltani S, Moslehi N, Hosseini-Esfahani F, et al. The association between empirical dietary inflammatory pattern and metabolic phenotypes in overweight/obese adults. International journal of endocrinology and metabolism. 2018; 16(2): 1-7.

[28]Antonella A, Andrea M, Sarka K, et al. Association of Dietary Patterns with Metabolic Syndrome: Results from the Kardiovize Brno 2030 Study. Nutrients. 2018; 10(7):898-914.

[29]James S M, Honn K A, Gaddameedhi S, et al. Shift Work: Disrupted Circadian Rhythms and Sleep—Implications for Health and Well-being. Current Sleep Medicine Reports. 2017; 3(2): 104-112.

[30]Vinogradova I, Anisimov V. Melatonin prevents the development of the metabolic syndrome in male rats exposed to different light/dark regimens. Biogerontology. 2013; 14(4): 401-409.

[31]Schwartsburd P M. Catabolic and anabolic faces of insulin resistance and their disorders: a new insight into circadian control of metabolic disorders leading to diabetes. Future science OA. 2017; 3(3): 1-10.

[32]Kar D, Gillies C, Nath M, et al. Association of smoking and cardiometabolic parameters with albuminuria in people with type 2 diabetes mellitus: a systematic review and meta-analysis. Acta diabetologica. 2019; 56(8): 839-850.

[33]Rashidi H, Shahbazian H, Nokhostin F, et al. The comparison of insulin and uric acid levels in adolescents with and without metabolic syndrome. Frontiers in Biology. 2018; 13(6): 452-457.

[34]Khalili M, Shuhart M C, Lombardero M, et al. Relationship between metabolic syndrome, alanine aminotransferase levels, and liver disease severity in a multiethnic north American cohort with chronic hepatitis B. Diabetes care. 2018; 41(6): 1251-1259.

[35]Al-Quraishi T, Abawajy J H, Chowdhury M U, et al. Breast Cancer Recurrence Prediction Using Random Forest Model. Recent Advances on Soft Computing and Data Mining. 2018; 700:318-329.

[36]Dagliati A, Marini S, Sacchi L, et al. Machine learning methods to predict diabetes complications. Journal of diabetes science and technology. 2018; 12(2): 295-302.

[37]Wu J H, Li J, Wang J, et al. Risk prediction of type 2 diabetes in steel workers based on convolutional neural network. Neural Computing and Applications. 2019; 3: 1-16.

[38]Ševo I, Avramović A. Convolutional neural network based automatic object detection on aerial images. IEEE geoscience and remote sensing letters. 2016; 13(5): 740-744.

[39]Mu-han D. Prediction of epileptic seizures based on convolution neural network. Shandong Normal University. 2018.

[40]Zhang M, Wang L M, Chen Z H, et al. Multilevel logistic regression analysis on hypercholesterolemia related risk factors among adults in China. Chinese journal of preventive medicine. 2018; 52(2): 151-157.

Tables

Table 1 Comparison of the basic conditions of oil workers with and without metabolic syndrome

Basic conditions	Category(Unit)	MetS n(%)/M(P_{25},P_{75})		χ^2/Z	P
		No	Yes		
Age	Year	43(38,47)	44(40,49)	-5.79	<0.001
Gender	Male	601(69.00)	504(84.42)	45.26	<0.001
	Female	270(31.00)	93(15.58)		
BMI	Kg/m^2	23.9(21.90,25.90)	26.80(24.90,28.80)	-16.35	<0.001
Marital status	Unmarried	56(6.43)	15(2.51)	11.82	0.003
	Married	782(89.78)	559(93.63)		
	Others	33(3.79)	23(3.85)		
Education level	Junior high school and below	133(15.27)	104(17.42)	9.07	0.011
	High school/technical secondary school	374(42.94)	290(48.58)		
	College and above	364(41.79)	203(34.00)		
Per capita monthly household income(Yuan)	<2000	619(71.07)	454(76.05)	8.05	0.018
	2000~	212(24.34)	109(18.26)		
	3000~	40(4.59)	34(5.70)		
Family history of hypertension	No	489(56.14)	288(48.24)	8.88	0.003
	Yes	382(43.86)	309(51.76)		
Family history of hyperlipidemia	No	801(91.96)	538(90.12)	1.51	0.22
	Yes	70(8.04)	59(9.88)		
Family history of diabetes mellitus	No	725(83.24)	454(76.05)	11.58	0.001
	Yes	146(16.76)	143(23.95)		

Table 2 Comparison of diet and lifestyle of oil workers with and without metabolic syndrome

Factors	Category	MetS n(%)/M(P_{25},P_{75})		χ^2	P
		No	Yes		
Salt	Light	221(25.37)	88(14.74)	26.39	<0.001
	Moderate	381(43.74)	276(46.23)		
	Salty	269(30.88)	233(39.03)		
Meat intake	Never	23(2.64)	13(2.18)	9.38	0.025
	Occasionally	198(22.73)	101(16.92)		
	Regularly	335(38.46)	232(38.86)		
	Every day	315(36.17)	251(42.04)		
Fruit intake	Never	37(4.25)	27(4.52)	6.59	0.086
	Occasionally	278(31.92)	223(37.35)		
	Regularly	258(29.62)	146(24.46)		
	Every day	298(34.21)	201(33.67)		
Dairy intake	Never	127(14.58)	103(17.25)	119.81	<0.001
	Occasionally	230(26.41)	297(49.75)		
	Regularly	199(22.85)	111(18.59)		
	Every day	315(36.17)	86(14.41)		
Carbonated beverage intake	Never	370(42.48)	270(45.23)	10.52	0.015
	Occasionally	384(44.09)	258(43.22)		
	Regularly	79(9.07)	31(5.19)		
	Every day	38(4.36)	38(6.37)		
Physical exercise	No	307(35.25)	259(43.38)	9.90	0.002
	Yes	564(64.75)	338(56.62)		
Smoking status	No smoking	524(60.16)	262(43.89)	39.30	<0.001
	Quit smoking	51(5.86)	61(10.22)		
	Smoking	296(33.98)	274(45.90)		
Drinking status	No drinking	585(67.16)	309(51.76)	37.02	<0.001
	Alcohol withdrawal	16(1.84)	24(4.02)		
	Drinking	270(31.00)	264(44.22)		

Table 3. Comparison of occupational exposure factors of oil workers with and without metabolic syndrome

Factors	Category	MetS n(%)/M(P_{25},P_{75})		χ^2	P
		No	Yes		
Shift work situation	Never	535(61.42)	254(42.55)	51.44	<0.001
	Once	208(23.88)	202(33.84)		
	Now	128(14.70)	141(23.62)		
Labour intensity	Mild	93(10.68)	44(7.37)	5.36	0.069
	Moderate	434(49.83)	295(49.41)		
	Severe	344(39.49)	258(43.22)		
Occupational heat	No	548(62.92)	266(44.56)	48.34	<0.001
	Yes	323(37.08)	331(55.44)		
Noise	No	429(49.25)	206(34.51)	31.39	<0.001
	Yes	442(50.75)	391(65.49)		

Table 4. Comparison of laboratory tests in oil workers with and without metabolic syndrome

Biochemical Indicators	MetS n(%)/M(P_{25},P_{75})		Z	P
	No	Yes		
RBC（×10^{12}/L）	5.01(4.65,5.33)	5.29(4.99,5.54)	-6.94	<0.001
MCV（fl）	88.80(85.10,92.00)	88.20(84.80,91.80)	-0.85	0.397
BPC（×10^{12}/L）	256.00(219.50,290.75)	251.00(211.00,284.00)	-0.55	0.59
MPV（fl）	8.20(7.70,8.80)	8.20(7.70,8.80)	-0.83	0.405
Hemoglobin（g/L）	155(141,165)	160(151,169)	-6.44	<0.001
TBIL（mmol/L）	13.50(10.50,17.70)	13.45（10.30,17.10）	-0.81	0.421
UA（mmol/L）	307(242,373)	367（304,426）	-11.13	<0.001
ALT（U/L）	20.00(14.00,24.00)	35.00（21.00,45.00）	-17.07	<0.001

Table 5. Multivariate nonconditional Logistic regression analysis of influencing factors in oil workers with metabolic syndrome

Factors	*B*	*S.E*	*Wald*χ^2	*P*	*OR*	*95%CI*
Age	0.088	0.012	55.251	0.000	1.092	1.067, 1.118
Per capita monthly household income（2000~）	-0.77	0.22	12.244	0.000	0.463	0.301, 0.713
Per capita monthly household income(3000~)	0.166	0.388	0.184	0.668	1.181	0.552, 2.525
BMI	0.273	0.026	114.091	0.000	1.313	1.249, 1.381
Family history of diabetes mellitus	0.373	0.183	4.129	0.042	1.452	1.013, 2.080
Salt(Moderate)	0.86	0.206	17.429	0.000	2.362	1.578, 3.536
Salt(Salty)	0.555	0.214	6.759	0.009	1.742	1.146, 2.648
Dairy intake(Occasionally)	0.676	0.216	9.771	0.002	1.966	1.287, 3.003
Dairy intake（Every day）	-1.149	0.261	19.317	0.000	0.317	0.190, 0.529
Carbonated beverage intake（Every day）	1.102	0.365	9.148	0.002	3.012	1.474, 6.153
Physical exercise	-0.398	0.152	6.86	0.009	0.672	0.499, 0.905
Smoking status(Smoking)	0.431	0.181	5.675	0.017	1.539	1.079, 2.194
Shift work situation(Once)	0.974	0.172	32.184	0.000	2.648	1.892, 3.707
Shift work situation（Now ）	1.509	0.237	40.489	0.000	4.522	2.841, 7.198
Occupational heat	0.656	0.224	8.548	0.003	1.926	1.241, 2.989
UA	0.004	0.001	27.244	0.000	1.004	1.003, 1.006
ALT	0.029	0.005	40.946	0.000	1.030	1.020, 1.039

Table 6. Assignment of influencing factor variables

Variable name	Variable meaning	Assignment method
Y	MetS	0=No,1=Yes
X_1	Age	Continuous variable (year)
X_2	Per capita monthly household income	1=<2000,2=2000-3000,3=≥3000
X_3	BMI	Continuous variable（Kg/m2）
X_4	Family history of diabetes mellitus	1=No,2=Yes
X_5	Salt	1=Light,2=Moderate,3=Salty
X_6	Dairy intake	1=Never,2=Occasionally ,3=Regularly,4=Every day
X_7	Carbonated beverage intake	1=Never,2=Occasionally ,3=Regularly,4=Every day
X_8	Physical exercise	1=No,2=Yes
X_9	Smoking status	1=No smoking,2=Quit smoking,3=Smoking
X_{10}	Shift work situation	1=Never,2=Once,3=Now
X_{11}	Occupational heat	1=No,2=Yes
X_{12}	UA	Continuous variable（mmol/L）
X_{13}	ALT	Continuous variable（U/L）

Table 7 coefficient of correlation

Variable name	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	X_{11}	X_{12}	X_{13}
X_1	1												
X_2	0.062*	1											
X_3	-0.008	-0.110**	1										
X_4	0.068**	0.022	0.014	1									
X_5	-0.021	-0.015	0.141**	-0.008	1								
X_6	-0.063*	-0.004	-0.124**	-0.009	-0.070**	1							
X_7	-0.147**	-0.001	0.010	-0.010	0.065*	0.288**	1						
X_8	-0.019	-0.016	-0.043	-0.034	-0.027	-0.034	-0.045	1					
X_9	0.012	-0.137**	0.108**	-0.012	0.165**	-0.093**	0.034	-0.130**	1				
X_{10}	0.018	0.310**	0.081**	0.054*	0.004	-0.052*	0.011	0.039	-0.012	1			
X_{11}	0.028	-0.044	0.091**	0.028	0.000	-0.047	0.005	0.040	0.028	0.047	1		
X_{12}	-0.055*	-0.021	0.169**	0.012	0.043	-0.092**	0.035	-0.035	0.109**	-0.041	0.066*	1	
X_{13}	0.084**	0.015	0.168**	0.058*	0.026	-0.110**	-0.042	-0.078**	0.049	0.006	0.090**	0.226**	1

** P<0.05 ** P<0.01*

Table 8 Results of tolerance and variance inflation factor

Model	Collinearity statistics	
	Tolerance	VIF
(constant)	-	-
Age	0.966	1.036
Per capita monthly household income	0.881	1.135
BMI	0.897	1.115
Family history of diabetes mellitus	0.985	1.015
Salt	0.952	1.051
Dairy intake	0.844	1.185
Carbonated beverage intake	0.872	1.147
Physical exercise	0.963	1.038
Smoking status	0.922	1.085
Shift work situation	0.897	1.115
Occupational heat	0.975	1.026
UA	0.907	1.102
ALT	0.905	1.105

Table 9. Sample classification results of Logistic regression model, random forest model, convolutional neural network model training set, verification set and test set[n (%)]

Model	Data set	Model predictive value	Actual value		
			Yes	No	Total
Logistic regression model	Training set	Yes	277（78.47）	67（13.11）	344
		No	76（21.53）	444（86.89）	520
		Total	353	511	864
	Validation set	Yes	86（69.35）	20（11.43）	106
		No	38（30.65）	155（88.57）	193
		Total	124	175	299
	Test set	Yes	84（70.00）	35（18.92）	119
		No	36（30.00）	150（81.08）	186
		Total	120	185	305
Random forest model	Training set	Yes	334（94.62）	31（6.07）	365
		No	19（5.38）	480（93.93）	499
		Total	353	511	864
	Validation set	Yes	96（77.42）	28（16.00）	124
		No	28（22.58）	147（84.00）	175
		Total	124	175	299
	Test set	Yes	93（77.50）	32（17.30）	125
		No	27（22.50）	153（82.70）	180
		Total	120	185	305
CNN	Training set	Yes	287（81.30）	52（10.18）	339
		No	66（18.70）	459（89.82）	525
		Total	353	511	864
	Validation set	Yes	91（73.39）	19（10.86）	110
		No	33（26.61）	156（89.14）	189
		Total	124	175	299
	Test set	Yes	82（68.33）	27（14.59）	109
		No	38（31.67）	158（85.41）	196
		Total	120	185	305

Table 10. Comparison of predictive performance of the three models in training set, validation set and test set

Evaluation index	Training set			Validation set			Test set		
	Logistic regression model	Random forest model	CNN	Logistic regression model	Random forest model	CNN	Logistic regression model	Random forest model	CNN
Accuracy rate （%）	83.45	94.21	86.34	80.60	81.27	82.61	76.72	80.66	78.69
Sensitivity（%）	78.47	94.62	81.30	69.35	77.42	73.39	70.00	77.50	68.33
Specificity（%）	86.89	93.93	89.82	88.57	84.00	89.14	81.08	82.70	85.41
F1 Score	0.79	0.93	0.83	0.75	0.77	0.78	0.70	0.76	0.71
Youden’s index	0.65	0.89	0.71	0.58	0.61	0.63	0.51	0.60	0.54
Positive likelihood ratio	5.98	15.60	7.99	6.07	4.84	6.76	3.70	4.48	4.68
Negative likelihood ratio	0.25	0.06	0.21	0.35	0.27	0.30	0.37	0.27	0.37
Kappa value	0.66	0.88	0.72	0.59	0.61	0.64	0.51	0.60	0.55
Positive predictive value （%）	80.52	91.51	84.66	81.13	77.42	82.73	70.59	74.40	75.23
Negative predictive value （%）	85.38	96.19	87.43	80.31	84.00	82.54	80.65	85.00	80.61
AUC	0.894	0.987	0.935	0.875	0.878	0.872	0.797	0.861	0.855
AUC 95%CI									
lower	0.871	0.977	0.917	0.833	0.835	0.829	0.748	0.818	0.810
upper	0.913	0.994	0.951	0.911	0.913	0.908	0.841	0.898	0.892
ICI	0.074	0.071	0.078	0.074	0.072	0.082	0.064	0.051	0.096

Table 11. Comparison of training set, validation set and test set AUC of three models

Model	AUC difference	SE	95%CI		Z	P
			lower	upper		
Training set						
gistic regression VS Random forest	0.094	0.010	0.074	0.113	9.419	<0.001
Logistic regression VS CNN	0.042	0.008	0.027	0.057	5.371	<0.001
Random forest VS CNN	0.052	0.007	0.038	0.066	7.062	<0.001
Validation set						
gistic regression VS Random forest	0.002	0.018	-0.034	0.038	0.125	0.900
Logistic regression VS CNN	0.003	0.014	-0.024	0.031	0.248	0.804
Random forest VS CNN	0.006	0.016	-0.026	0.037	0.361	0.718
Test set						
gistic regression VS Random forest	0.064	0.023	0.019	0.109	2.806	0.005
Logistic regression VS CNN	0.058	0.025	0.010	0.106	2.352	0.019
Random forest VS CNN	0.007	0.020	-0.034	0.047	0.320	0.749

Figures

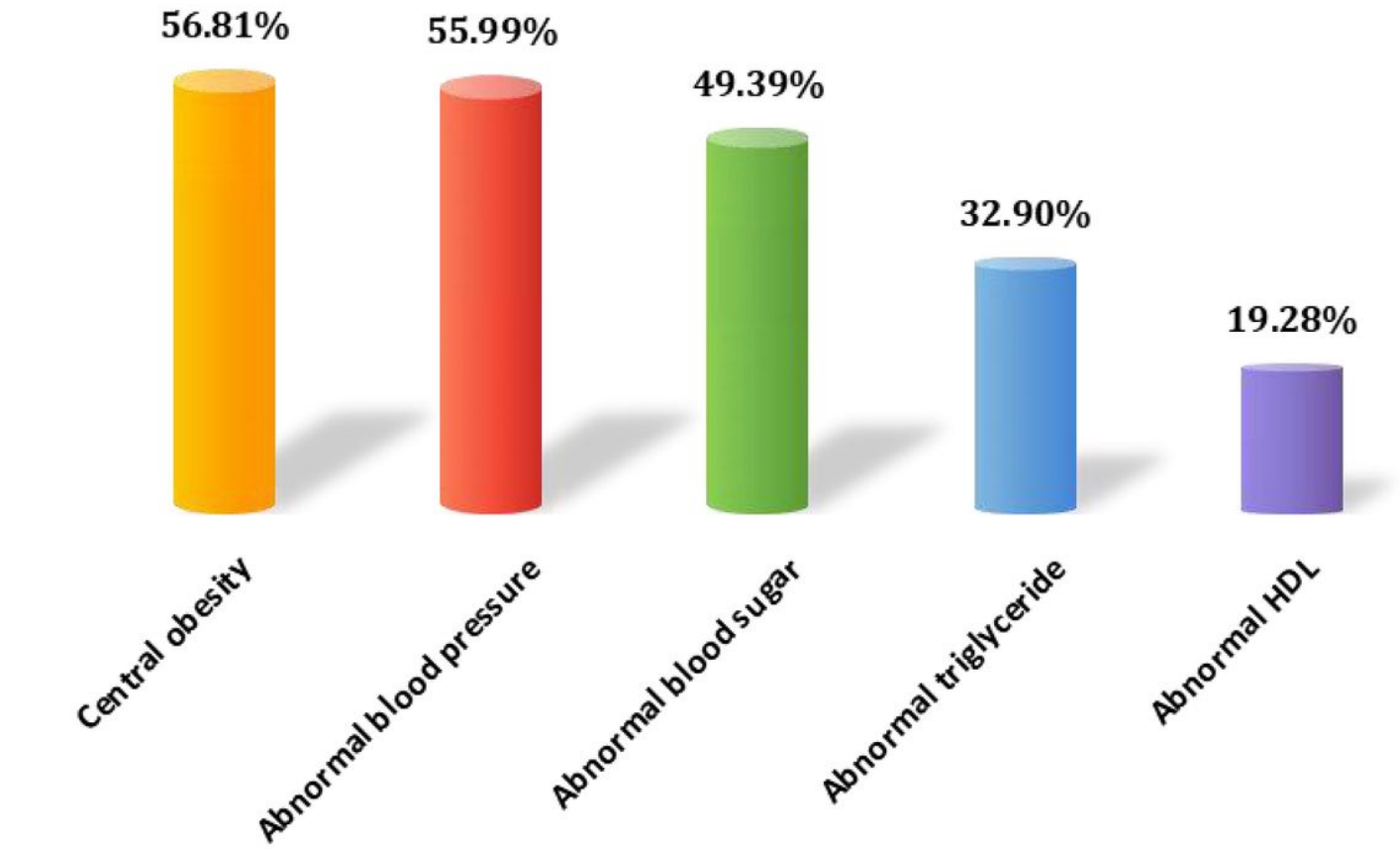

Figure 1

Comparison of abnormal rates among components of metabolic syndrome

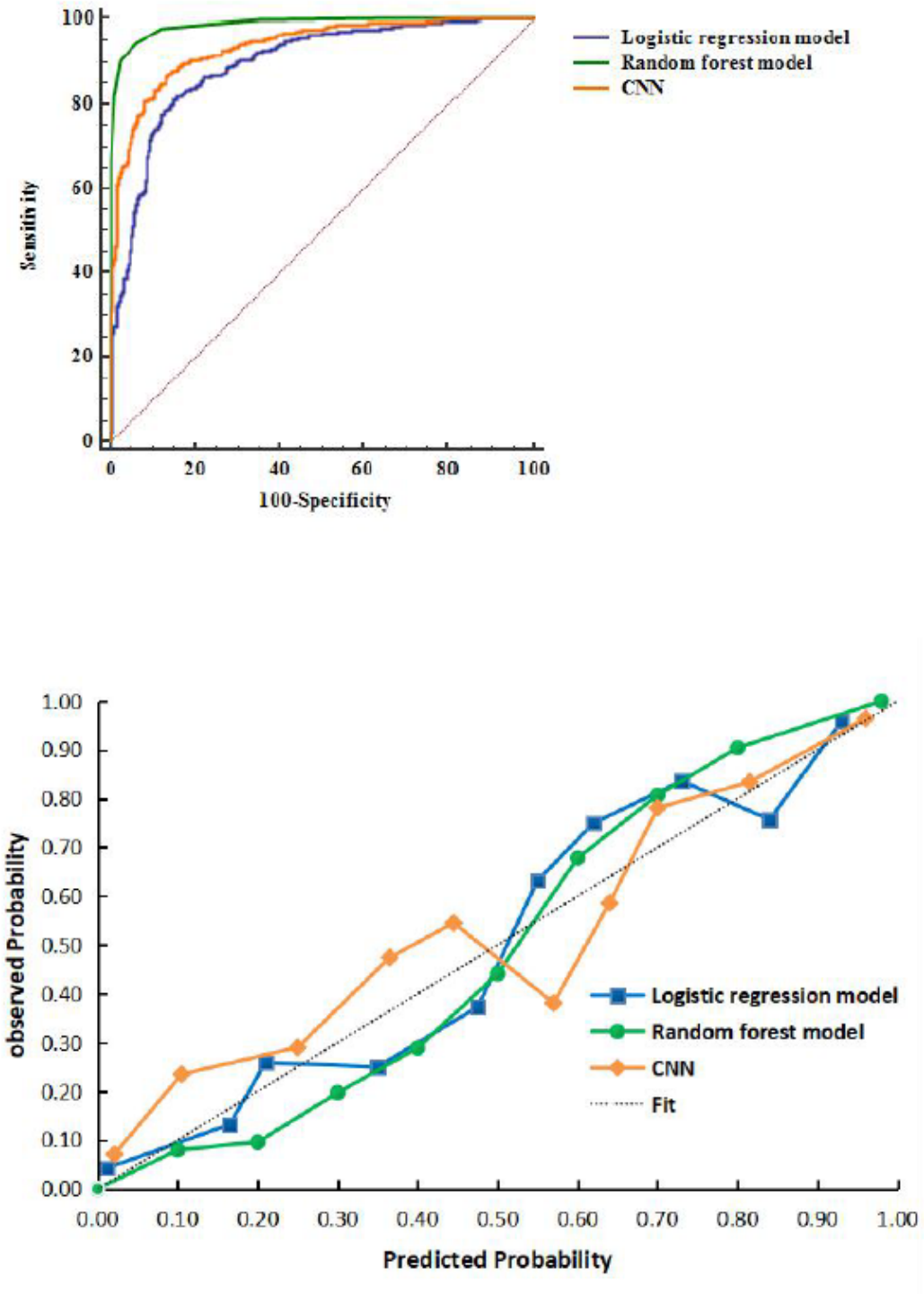

Figure 2

ROC curves and calibration curves of three predictive models in the training set

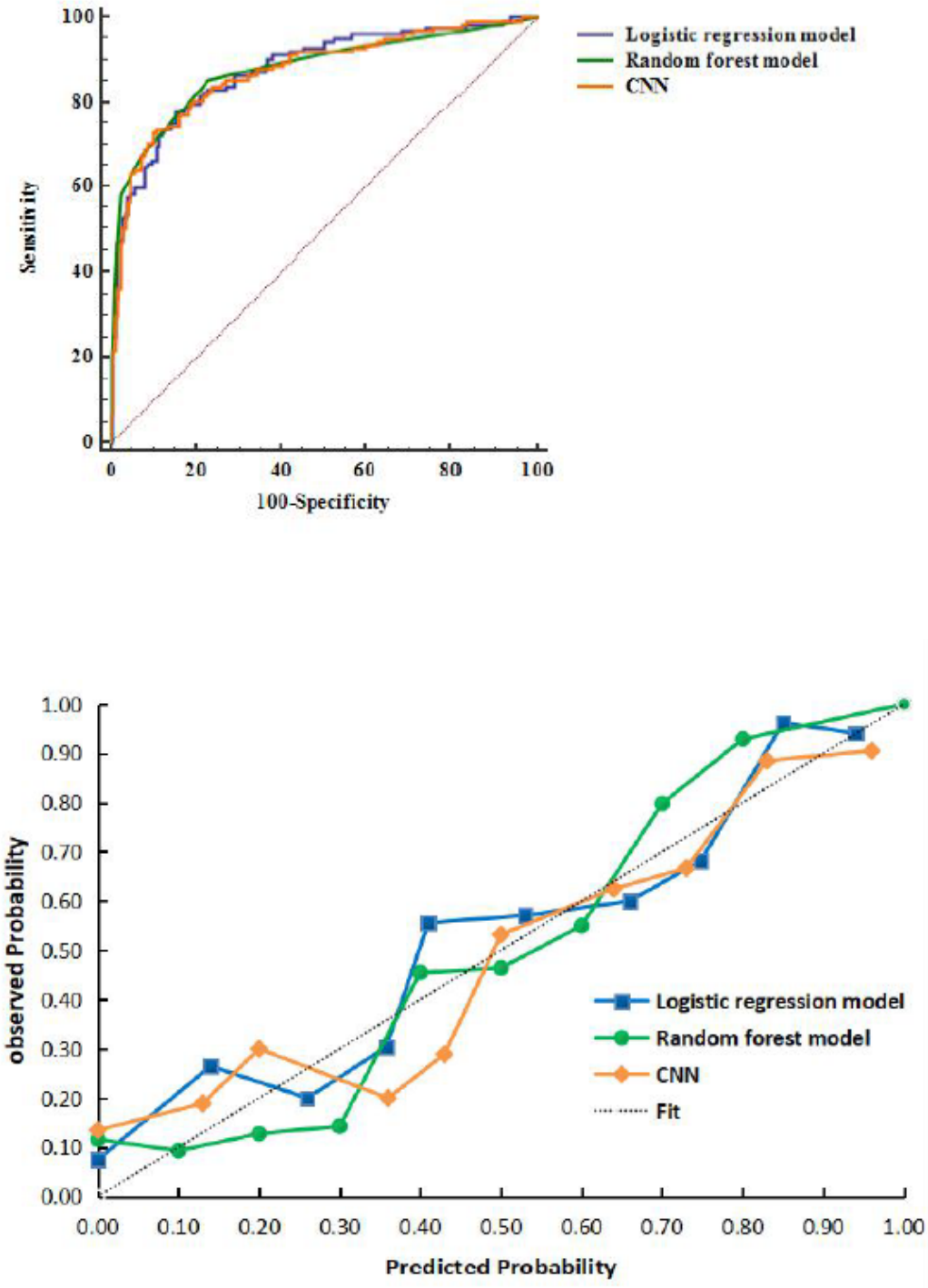

Figure 3

ROC curves and calibration curves of three predictive models in the validation set

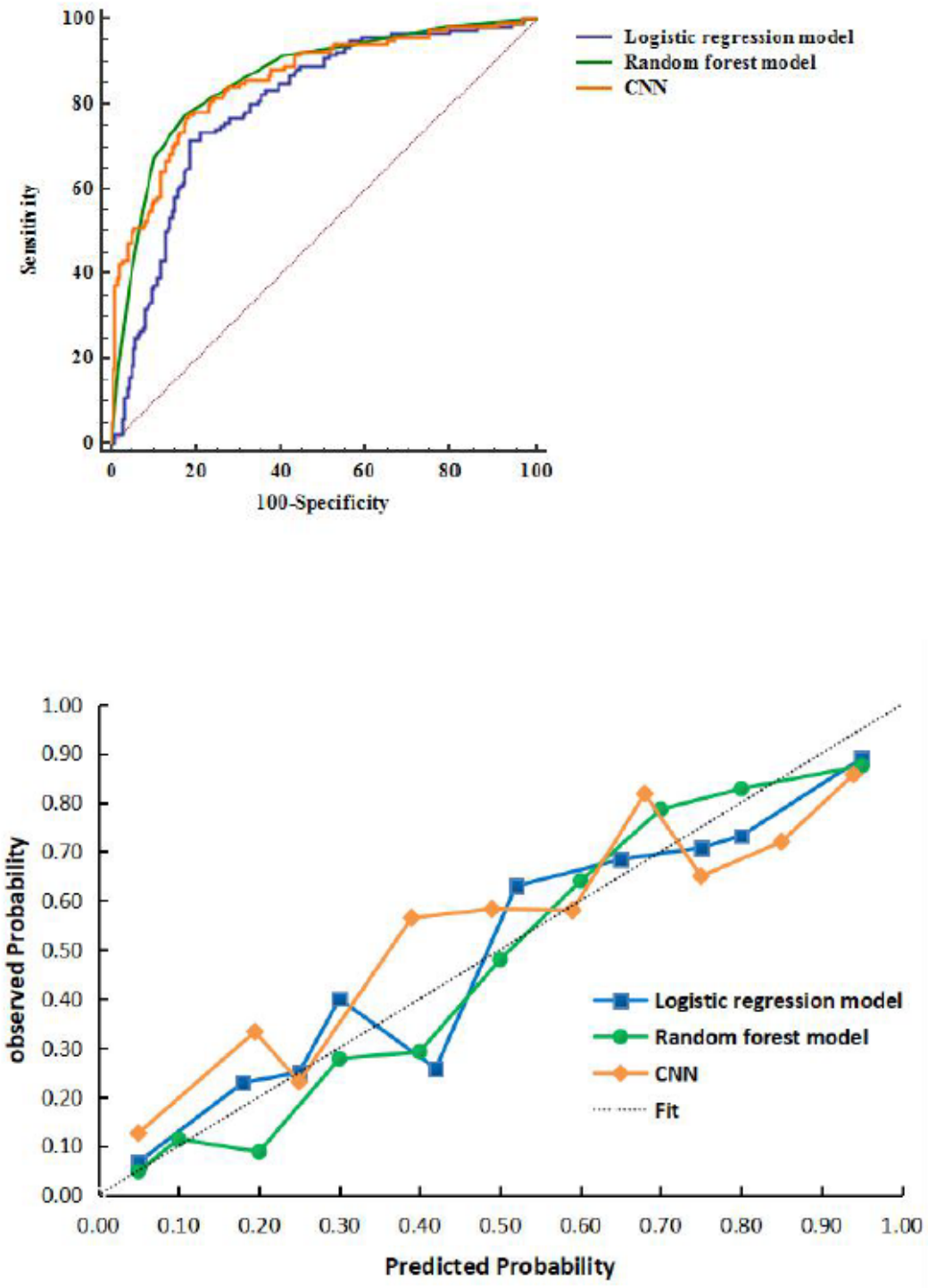

Figure 4

ROC curves and calibration curves of three predictive models in the test set