Proceedings of ICBioMed 2025 Symposium: Computational Modelling and Simulation for Biology and Medicine
DOI: 10.54254/2753-8818/2025.LD26436

Influencing Factors and Prediction of Incidence of Coronary Heart Disease

Jing Wang

School of Mathematical Sciences, University of Liverpool, Liverpool, United Kingdom
Jing.Wang2403@student.xjtlu.edu.cn

Abstract. Coronary heart disease (CHD) is a disease of myocardial ischemia and hypoxia caused by coronary atherosclerosis, which is one of the main causes of death. This article selects relevant data from the Kaggle dataset and conducts rigorous preprocessing, such as data cleaning and outlier handling. Then, the classic Logistic regression model and random forest model were selected to conduct a comprehensive and in-depth analysis of many potential factors affecting the incidence of coronary heart disease. The results showed that age, BMI, education level, total cholesterol, blood pressure, diabetes, and lifestyle (smoking, drinking) have significant effects on the incidence of CHD. Then, the logistic regression model was used to predict the incidence rate of CHD. In addition, the numerical results of the OC-ROC Curve showed that the logistic regression model had good predictive performance. Finally, based on the research results, a series of targeted measures are proposed to enhance people's awareness of preventing CHD, effectively reduce medical expenditure, and improve the quality of life of the public.

Keywords: Coronary heart disease, incidence, logistic regression model, prediction.

1. Introduction

Coronary heart disease (CHD) is one of the common cardiovascular diseases [1]. It is a kind of heart disease caused by coronary atherosclerosis, which not only leads to narrowing or blocking of blood vessels and changes in coronary artery function but also leads to myocardial ischemia or necrosis [2]. Globally, approximately 5% of adults over the age of 20 have CHD. In 2021 alone, 19.39 million people died from cardiovascular and cerebrovascular diseases, among which CHD is the most common and important type [3]. Therefore, CHD has been identified as the world's highest mortality disease, which undoubtedly brings a heavy burden to patients' families and society. Its pathogenesis is a highly complex process [4]. In addition to uncontrollable factors such as gender, age, and genes, controllable factors such as blood pressure, smoking, cholesterol, and diabetes also increase the risk of CHD [5]. Of course, with the advancement of medical methods, effective prevention and treatment strategies can be developed by studying the pathogenesis and risk factors of CHD. This not only promotes the development of cardiology but also advances in other fields such as molecular biology and genetics.

There have been many researches conducted relevant studies on the influencing factor of CHD around the world. For instance, the study by Johansen, Vedel-Krogh, Nielsen, Afzal, Davey Smith,

DOI: 10.54254/2753-8818/2025.LD26436

& Johansen et al. included 104,867 individuals from the Copenhagen General Population study and conducted a mediation analysis using VanderWeele's method [6]. It was found that elevated residual cholesterol largely explained the increased risk of myocardial infarction and coronary heart disease in individuals with unhealthy lifestyles. Cao & Wu found that healthy lifestyle changes based on exercise could improve multiple risk factors and prevent CHD in the elderly [7]; The use of KNN, decision tree and SVM methods not only confirmed the known influencing factors of CHD, but also found that self-evaluation level of health, income level and education level have potential effects on CHD. Moreover, it is recommended to combine XGBoost and stepwise logistic regression (LR) analysis after data balancing in the CHD risk prediction models [8].

In this article, CHD data, the LR model, and the random forest (RF) model were used to analyze the influencing factors and predict the incidence of CHD. By analyzing the research results, the aim is to put forward relevant suggestions for the prevention of CHD.

2. Method

2.1. Data set

This article uses Kaggle's Heart Disease Dataset, which has been reviewed and verified by the community, and the data annotation is relatively clear, therefore, it has a certain level of reliability [9]. The dataset contains 4238 sets of data. After deleting 595 sets of data with missing values, this study conducts research based on the remaining 3643 sets of data. For the occurrence of coronary heart disease, convert “yes” and “no” to 0 and 1, respectively, to meet the requirements of LR and RF models for binary classification variables.

2.2. Model

LR is a multiple regression method used to analyze the relationship between binary or classification results and multiple influencing factors [10]. A logical function maps the result of linear regression to a probability value to determine the likelihood that the sample belongs to a certain class. LR model based on linear regression model

$$z = w_0 + w_1x_1 + w_2x_2 + \ldots + w_nx_n \tag{1}$$

Where x_i is the feature, w_i is the corresponding weight, and w_0 is the bias term. Using the logical function $g(x) = \frac{1}{1+e^{-z}}$, convert z to a probability p , that is, $p = g(z)$, where p represents the probability that the sample belongs to the positive class, and $1 - p$ is the probability that the sample belongs to the negative class. For this, it can be seen that the principle of the LR model is simple, the calculation efficiency is high, the probability output is meaningful and stable, and the overfitting phenomenon is not easy.

The RF model is a classifier or regressor that contains multiple decision trees, utilizing the difference and diversity among decision trees to improve the accuracy and stability of decisions. It can effectively reduce the risk of overfitting, have a good tolerance for data anomalies, and evaluate the importance of each feature to the prediction results during the training process.

DOI: 10.54254/2753-8818/2025.LD26436

2.3. Experimental design

2.3.1. LR model

Table 1. Estimation results of the LR model

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-8.4511	0.7074	-11.947	0.0000***
Gender/Male	0.5489	0.1101	4.984	0.0000***
Age	0.0614	0.0068	9.088	0.0000 ***
Education post graduate	0.1350	0.1979	0.682	0.4950
Education primary school	0.0109	0.1643	0.066	0.9473
Education uneducated	0.2073	0.1508	1.375	0.1691
Current Smoker	0.0725	0.1567	0.463	0.6436
Cigs Per Day	0.0174	0.0062	2.794	0.0052 **
Blood pressure medicines (BPMeds)	0.2051	0.2324	0.882	0.3776
Prevalent Hypertension (PrevalentHyp)	0.2425	0.1382	1.755	0.0793
diabetes	0.0360	0.3156	0.114	0.9093
total cholesterol (totChol)	0.0025	0.0011	2.183	0.0291 *
systolic blood pressure (sysBP)	0.0152	0.0038	3.999	0.0001***
Diastolic blood pressure (diaBP)	-0.0036	0.0064	-0.551	0.5815
BMI	0.0039	0.0128	0.303	0.7619
Heart Rate	-0.0026	0.0042	-0.623	0.5335
glucose	0.0071	0.0022	3.189	0.0014 **

Note：***‘p<0.001’；**‘p<0.01’; *‘p<0.05’; . ‘p<0.1’

The summary information output from LR models constructed without standardized data, including model coefficient estimates, standard errors, Z-values, p-values, etc. These statistics can be used to determine whether each independent variable has a significant impact on the dependent variable (CHD attack), in which the estimated value reflects the degree of impact and positive negative relationship, the standard error measure is stable, and the Z-value test coefficient is significantly non-zero. When $p<0.05$, the independent variable has a significant impact on the probability of CHD attack.

Table 1 shows the results of the LR model, where variables such as gender, age, Education post graduate, Cigs Per Day, prevalentHyp, totChol, sysBP, and glucose have significant effects on the onset of CHD. However, Education primary school, Education uneducated, Current Smoker, BPMeds, diabetes, diaBP, BMI, and Heart Rate have no significant effect on the onset of CHD.

Then, on this basis, the abstract of the LR model constructed with standardized data was output and compared with the unstandardized model to more intuitively compare the relative size of the influence of different independent variables on the incidence of CHD and exclude the impact of variable dimension differences.

DOI: 10.54254/2753-8818/2025.LD26436

2.3.2. RF model

Table 2. Variables %IncMSE and IncNodePurity

	%IncMSE	IncNodePurity
age	19.2325	51.1663
education	1.2850	13.0113
CurrentSmoker	5.6541	3.8298
Cigs Per Day	8.4717	21.6815
BPMeds	4.3185	2.6729
prevalentHyp	15.7722	7.3107
diabetes	3.4415	2.2347
totChol	7.4441	52.5294
sysBP	28.6078	57.8281
diaBP	28.2540	50.6748
BMI	9.0417	56.4464
Heart Rate	2.4016	39.2694
glucose	6.4964	51.3779

Table 2 shows the importance indicators of the variables in the RF model, including %IncMSE (percentage increase in mean square error) and IncNodePurity (increase in node purity). The greater the value of both, the more important the feature is to the model. Table 2 shows that sys BP, diaBP, and age have a significant impact on the incidence of CHD.

2.3.3. Visualized analysis

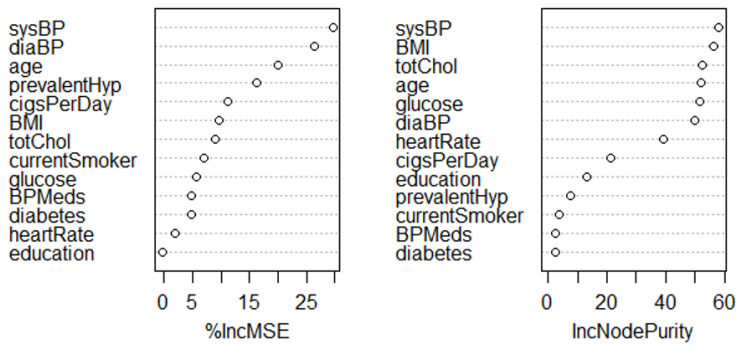

Figure 1. Scatter plots of variables vs %IncMSE and IncNodePurity in RF model (photo/picture credit: original)

Figure 1 shows the relationship between %IncMSE and IncNodePurity and various variables respectively in the form of scatter plots by using a RF model. Figure 1 measures the importance of variables from the perspectives of model error and node purity and clearly displays the corresponding indicator values for each variable.

DOI: 10.54254/2753-8818/2025.LD26436

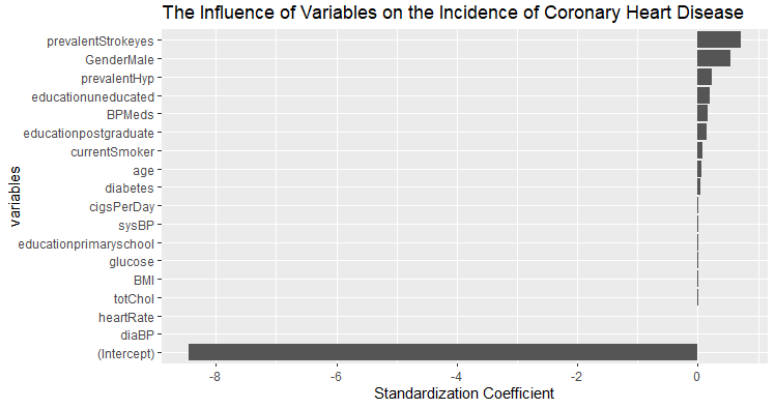

Figure 2. The influence of variables on the incidence of coronary heart disease (photo/picture credit: original)

After standardizing the data in Figure 2, the coefficients obtained by fitting the regression model are plotted as the horizontal axis, and each variable is a vertical coordinate, creating a bar chart that eliminates the influence of independent variable dimensions and value ranges. This allows for a direct comparison of the degree of influence of different independent variables on the dependent variable, intuitively reflecting the degree and direction of their influence.

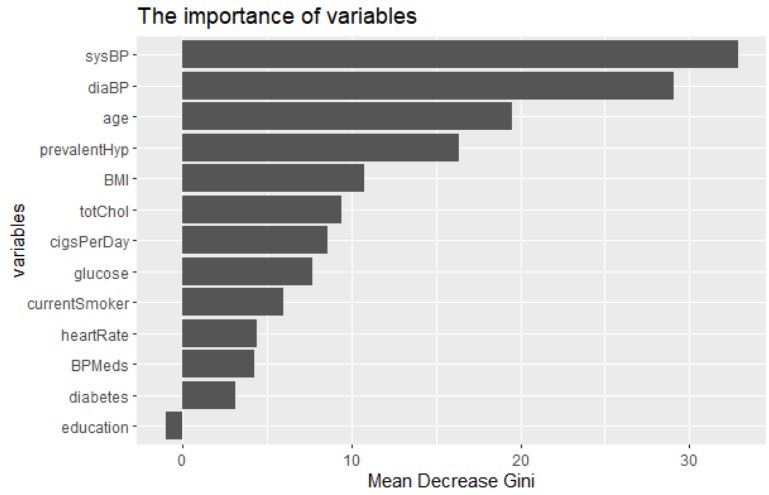

Figure 3. The importance of variables (photo/picture credit: original)

Figure 3 reflects the importance of variables to the incidence of CHD. The ordinate represents the type of variables, and the abscissa represents the average reduction of the Gini index. It is the average value of the Gini index reduction caused by a certain feature as a splitting feature in all decision trees. The larger the value, the stronger the ability of this feature to divide data and reduce impurity, which means the higher the importance of this feature to the model. It can be seen from the above table that the importance of sysBP is relatively important.

2.4. Prediction

First, the experimental data was fitted to the LR model, and then the data set was split according to the proportion of 70% and 30%, among which 70% was the training set and 30% was the prediction

DOI: 10.54254/2753-8818/2025.LD26436

set. Then, the constructed model was used for prediction, and prediction probability was obtained. Finally, draw an ROC curve based on the prediction results, as shown in Figure 4.

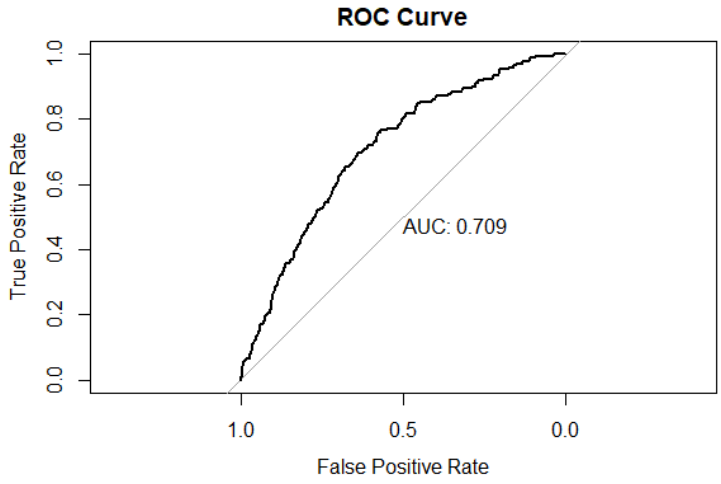

Figure 4. Analysis of the ROC curve for predicting CHD influenced by LR model (photo/picture credit: original)

As can be seen from Figure 4, this model has good prediction performance, with the area under the curve (AUC) of 0.709, which is close to 1. Therefore, it can be considered that the model has a strong fitting effect, but it can be further optimized.

3. Conclusion

In this study, it can be seen that CHD is the result of a combination of many factors. In addition to age, gender, and somatic mutations in specific cells (such as DNMT3A, TET2), metabolic abnormalities and unhealthy lifestyles also increase the risk of CHD significantly. Therefore, to reduce the impact of CHD on people's quality of life, controlling blood pressure, glucose, BMI, and total cholesterol, and smoking cessation are the core strategies to prevent its incidence rate. This article demonstrates good performance in predicting CHD based on the LR model. The results show that the area under the ROC curve (AUC) is 0.709, which indicates that the model has a certain ability to distinguish. However, there is still room for optimization in this model, such as introducing more characteristic variables, adjusting data preprocessing methods, or experimenting with nonlinear models to further improve prediction accuracy and generalization.

The Kaggle dataset used in this article has some limitations, which do not indicate the race and source of the experimental samples, so it may affect the ranking of the importance of variables to the incidence of CHD. Some research reports have clearly pointed out that a series of irresistible factors such as race will also have an impact on the incidence of CHD, which may be closely related to their genes and living environment. For example, black people are more likely to suffer from CHD than white people. In the future, further exploration of gene-environment interactions and social factor interventions is also needed to reduce the burden of diseases on individuals, families, and society.

References

[1] Li, T., Shi, W.T., Wang, G.R., & Jiang, Y.L. 2025. Prevalence and risk factors of frailty in older patients with coronary heart disease: A systematic review and meta-analysis. Archives of Gerontology and Geriatrics, 130, 105721. ISSN 0167-4943.

DOI: 10.54254/2753-8818/2025.LD26436

[2] Vrints, C., Andreotti, F., Koskinas, K.C., Rossello, X., Adamo, M., Ainslie, J., … ESC Scientific Document Group. 2024. 2024 ESC Guidelines for the management of chronic coronary syndromes. European Heart Journal, 45(36), 3415–3537.
[3] Ortiz-Ospina, E., & Roser, M. 2016. Global Health. Data adapted from IHME, Global Burden of Disease. https: //ourworldindata.org/grapher/deaths-from-cardiovascular-disease
[4] Du, J., Wu, W., Zhu, B., Tao, W., Liu, L., Cheng, X., … Pei, K. 2023. Recent advances in regulating lipid metabolism to prevent coronary heart disease. Chemistry and Physical Lipids, 255, 105325.
[5] National Cholesterol Education Program (US). Expert Panel on Detection and Treatment of High Blood Cholesterol in Adults, 1989. Report of the expert panel on detection, evaluation, and treatment of high blood cholesterol in adults (No. 89). US Department of Health and Human Services, Public Health Service, National Institutes of Health.
[6] Johansen, M.Ø., Vedel-Krogh, S., Nielsen, S.F., Afzal, S., Davey Smith, G., & Nordestgaard, B.G. 2025. Association of remnant cholesterol with unhealthy lifestyle and risk of coronary heart disease: a population-based cohort study. The Lancet Regional Health – Europe, 10, 101223. ISSN 2666-7762.
[7] Cao, F., & Wu, X.P. 2024. Lifestyle intervention for the prevention of coronary heart disease in the elderly. Journal of Clinical Cardiology, 40(10), 785–789.
[8] Yue, H.T., He, C.C., Cheng, Y.Y., Zhang, S.C., Wu, Y., & Ma, J. 2025. Coronary Heart Disease Risk Prediction Model Based on Machine Learning. Chinese General Practice, 28(04), 499–509.
[9] Kaggle. n.d. Heart disease dataset. https: //www.kaggle.com/datasets/mirzahasnine/heart-disease-dataset
[10] Wang, Q.Q., Yu, S.C., Qi, X., Hu, Y.H., Zheng, W.J., Shi, J.X., & Yao, H.Y. 2019. Overview of LR model analysis and application. Zhonghua Yu Fang Yi Xue Za Zhi, 53(9), 955–960.