ISSN 2959-6157

Research on the Influencing factors of Heart Disease based on Binary Logistic Regression

Yuqi Guo

College of science, North China University of Technology, Beijing, 100043, China
Corresponding author: guoyq0509@mail.ncut.edu.cn

Abstract:

The main objective of this study was to analyze the multiple factors affecting heart disease using a binary logistic regression model. By examining the chart, this study draws many conclusions. Heart disease is at the forefront of mortality in China and even in the world. Thus, it is particularly important to explore its influencing factors. Through this study, chest pain type, resting electrocardiographic results, the slope of the peck exercise ST segment, and the maximum heart rate achieved were key factors affecting the occurrence of heart disease, and the prediction accuracy reached 86.05%, indicating that the conclusion is acceptable. The study provides more accurate strategies for the prediction of heart disease. This study fully explored the multiple influencing factors of heart disease; people should pay full attention to this indicator, maintain heart health, and be everyone's responsibility. Protecting cardiovascular health is everyone's responsibility. After all, health is the beginning of all work and good life.

Keywords: Heart disease; multiple linear regression; influencing factors.

1. Introduction

The heart is one of the most important organs in the human body. It is responsible for pumping blood to deliver oxygen and nutrients to various parts of the body while transporting metabolic waste to the lungs and kidneys for excretion. Normal heart function is crucial for maintaining the body's physiological activities. Heart disease is a disease in which the function of the heart is impaired or abnormal due to changes in the heart [1]. People who are chronically stressed are 1.27 times more likely to develop heart disease [2]. Any heart disease can lead to serious health problems or even life-threatening conditions. „China Cardiovascular Health and Disease Report 2020" pointed out that the incidence and mortality of heart disease in China are increasing with the development of the aging population, and the mortality rate of heart disease in rural and urban areas in 2018 was 1.6212×10-6 and 1.4634×10-6, respectively [3].

Globally, the number of deaths attributable to hypertensive heart disease reached 1.157 million in 2019, with a standardized mortality rate of 15.2 per 100,000, down 21.5% from 1990. In 2019, the number of deaths attributed to hypertensive heart disease in China reached 320,000, with a standardized mortality rate of 20.6 per 100,000, down 51.2% from 1990, but still higher than the global level [4]. The burden of Acute Rheumatic Fever (ARF) has not been adequately captured in South Africa, although illness notification is required by law. An examination of notifications has revealed that this legal duty has not been adequately emphasized, nor has notification been handled appropriately [5]. An important element of comprehensive Awareness and surveillance, the goal of the Advocacy and Prevention (ASAP) initiative is to raise awareness of the illness and the millions of people it affects globally, particularly in areas with few resources. It is common knowledge that interest in and understanding of this worldwide public health issue declined along with the incidence of RF in the industrialized world [5].

Moreover, Numerous studies have assessed the acute stage of the illness, but as COVID-19 is a relatively new diagnosis, there is little long-term follow-up data available. Fascinatingly, it has also been connected to typical cardiac problems [6]. On August 29, 2004, at the annual meeting of the European Society of Cardiology (ESC), Yusuf presented the results of a large study involving 262 medical centers in 52 countries (6,000 people in China), which found that through nine universal risk factors worldwide can predict more than 90.4% of acute myocardial infarction (AMI) These nine risk factors are: Abdominal obesity, high blood pressure, high blood sugar, dyslipidemia, smoking, excessive alcohol consumption, excessive stress, lack of exercise, and insufficient daily intake of fruits and vegetables, among the nine risk factors, "excessive stress"

has been recognized by the cardiovascular industry as a risk factor for serious heart disease for the first time [7].

Thus, it is of great importance to pay more attention to cardiovascular disease. The world has already built the Global Cardiovascular Academic Performance Evaluation (CAPE) system and ranked the academic impact of cardiovascular diseases in global medical institutions. Nine subdisciplines, including ischemic heart disease, hypertension, vascular disease, arrhythmia, pulmonary vascular disease, heart failure, congenital heart disease, cardiomyopathy, and valvular heart disease, were set up to realize the mapping of subdiscipline classification, cardiovascular terms, and entry words [8]. Another study showed that from 2004 to 2019, the death rate of heart disease in urban and rural China increased significantly, with more deaths in rural areas than in urban areas; it is necessary to strengthen health education and prevention and control of heart disease in rural areas [9]. Additionally, the core of health is harmony; health is the greatest harmony; the 21st century‘s health requirements to achieve the greatest harmony can be expressed in six words, that is, health, longevity, wisdom, happiness, beauty, and morality [10]. The research aims to use Binary Logistic Regression to analyze the factors that influence heart disease.

2. Methods

2.1 Data Source

The study used data from the website Kaggle. The dataset offers a including the predicted attribute, but all published experiments refer to using a subset of 14 of them. This dataset identifies a range of factors that may influence the pathogenesis of cardiovascular disease.

2.2 Variable Introduction

The analysis carefully chose specific indicators to deepen the understanding of the relationship between indicators and heart disease (Table 1). There are a total of 14 variables. Six of them are categorical variables, and eight are numerical variables. The variables identified in this study are strongly associated with cardiovascular health. There is a reliable basis and reference for the study of the subsequent influencing factors.

Table 1. Types of variables

Indicator	Type	Range
age	Numeric	29 to 77
sex	Categorical	0=male,1=female
chest pain type	Categorical	0-typical angina, 1- atypical angina, 2- non-anginal angina, 3-asymptomatic
resting blood pressure	Numeric	94 to 200 mmHg
serum cholesterol	Numeric	126 to 564 mg/dl
FBS over120	Numeric	0=false;1=true
resting electrocardiographic results	Numeric	0=normal,1=abnormality
maximum heart rate achieved	Numeric	71 to 202 beats
exercise-induced angina	Categorical	0=false, 1=true
ST depression	Numeric	0 to 6.2 mm
the slope of the peak exercise ST segment	Categorical	0=up;1=flat;2=down
number of major vessels	Numeric	0 to 3
Thallium	Categorical	0 = normal; 1 = fixed defect; 2 = reversable defect
target	Categorical	0=Confirmed heart disease;1=unconfirmed heart disease

2.3 Method Introduction

This study first conducted data screening, selecting variables that may be related and analyzing the data using binary logit regression by SPSS. The binary logit regression model is a form of logistic regression which is used to solve binary classification problems. The principle is based on a linear regression model, but classification is achieved by converting the output of a linear function into a logarithmic probability (logit).

The binary logit regression model sum's that the output variable (Y) of the classification problem follows a binomial distribution, i.e. the sum of the probabilities of the observed positive and negative classes is 1. The model establishes the relationship between the linear combination (z) and the logarithmic probability by:

$$\ln\frac{p}{1-p} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n \quad (1)$$

Where p represents the probability that the observed data point belongs to a positive class, represents the characteristic variable, and β_1 represents the characteristic coefficient.

Binary Logit regression analysis in this study was performed with sex, age, cp, resting blood pressure, serum cholesterol, fasting blood sugar, ST depression induced by exercise relative to rest, the slope of the peak exercise ST segment, number of major vessels, maximum heart rate achieved, normal or abnormal as independent variables and target as dependent variables.

3. Results and Discussion

3.1 Descriptive Analysis

Figure 1 is the histogram of treetops (resting blood pressure). The figure shows that for most patients with heart disease, their resting blood pressure is usually between 120-140mmHg, which differs from people without heart disease.

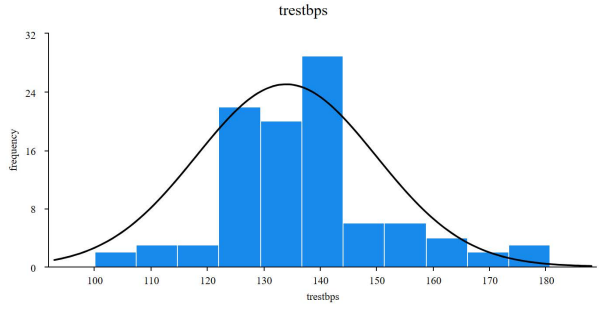

Fig. 1 The histogram of trestbps

Figure 2 is the histogram of thali (maximum heart rate achieved). In most patients with heart disease, their maximum heart rate is way too high and differs from normal people.

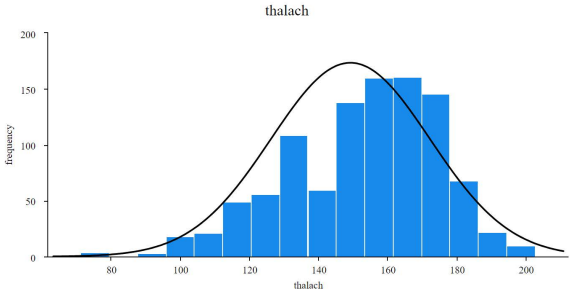

Fig. 2 The histogram of thalach

3.2 Logistic Regression Results

Binary Logit regression analysis was performed with sex, age, chest pain type, resting blood pressure, serum cholesterol, fasting blood sugar, ST depression induced by exercise relative to rest, the slope of the peak exercise ST segment, the number of major vessels, maximum heart rate achieved, normal or abnormal as the independent variable, and target as the dependent variable as can be seen from table 2, a total of 1025 samples were included in the analysis, and there were no missing data.

Table 2. Overview of Binary Logit Regression Analysis

Name	Options	Frequency	Percentage
Target	0	499	48.68%
	1	526	51.32%
	Total	1025	100.00%
Summary	Valid	1025	100.00%
	Hiatus	0	0.00%
	Total	1025	100.00%

From Table 3, it can be seen that Forecast Accuracy was 86.05%, which was high enough for this study. What's more, the predicting error rate was acceptable.

Table 3. Summary of Prediction Accuracy with Binary Logit Regression

0		Predicted Value		Forecast Accuracy	Predicting Error Rate
		1			
Ture value	0	412	87	82.57%	17.43%
	1	56	470	89.35%	10.65%
gather				86.05%	13.95%

From Table 4, it can be seen that the accuracy of the final model on the test set is 81.95%, the precision (compre-

hensive) is 82.81%, the recall rate (comprehensive) is 81.95%, and the F1-score (comprehensive) is 0.82. The model effect is acceptable. It can be concluded that the fault tolerance rate is very low, the method of investigation and analysis of cardiovascular health factors is very reliable, and the effectiveness of the model is high. This study can accurately predict the leading factors and indirect factors affecting cardiovascular health.

Table 4. Testing set model evaluation results

	Precision	Recall	F1-score	Samples
Absence	0.88	0.76	0.82	109
Presence	0.77	0.89	0.82	96
Accuracy	-	-	0.82	205
Average	0.82	0.82	0.82	205
Average(comprehensive)	0.83	0.82	0.82	205

4. Conclusion

The study shows that sex, chest pain type, resting electrocardiographic results, the slope of the peck exercise ST segment, and maximum heart rate achieved were key factors affecting the occurrence of heart disease, and the prediction accuracy reached 86.05%, indicating that the conclusion is acceptable. This study draws the following conclusions. In life and medical treatment, people should pay special attention to various factors affecting cardiovascular health, exercise more, eat more healthy food, and pay attention to cholesterol and other factors to enhance the possibility of cardiovascular disease. Only by paying more attention in life can the heart be in a healthy and vibrant state.

Apparently, cp, resting electrocardiographic results, slope, maximum heart rate achieved effect on heart disease is even more significant, and gender, resting blood pressure, serum cholesterol, ST depression induced by exercise relative to rest, number of major vessels, normal or abnormal has a significant negative effect on heart disease. By contrast, age did not affect heart disease.

References

[1]Zhang X H. Analysis of factors in diagnosis of heart disease based on Logistic regression and decision tree. Modern Information Technology, 2023, 7(7): 117-119.

[2]Cao Shufen. Pressure on the heart damage equal to five cigarettes a day. Journal of Shanxi old age, 2013 (3): 1.

[3]Zhou Xin, Li Jiahui, Xing Ying, et al. Research progress of exercise fear and its influencing factors in patients with heart disease. Journal of General Nursing, 2023, 21(26): 3633-3637.

[4]Zhang Ji sheng, Wang Meng long, Liu Jian fang, et al. Analysis of the change trend of hypertensive heart disease burden in China from 1990 to 2019 based on the data of the Global Burden of Disease Study in 2019. Chinese Journal of Hypertension, 2023, 31(2): 141-149.

[5]Carapetis J R J, Liesl J LJ Zühlke. Global research priorities in rheumatic fever and rheumatic heart disease. Annals of Pediatric Cardiology, 2011, 4(1): 4-12.

[6]Núñez-Gil I J, et al. POST-COVID-19 Symptoms and Heart Disease: Incidence, Prognostic Factors, Outcomes and Vaccination: Results from a Multi-Center International Prospective Registry (HOPE 2). J. Clin. Med., 2023, 706.

[7]Yang Juxian, Du Qin. Behavior of cardiology and the health promotion. Journal of preventive medicine, 2008.

[8]Yin Lu, et al. A methodological exploration of Global Cardiovascular Disease Academic Impact Assessment (CAPE) system. Chinese Journal of Circulation, 2019, 39(1): 3-16.

[9]Wei Qin, Shi Weiwei, Gao Jinchai, et al. Joinpoint regression analysis of heart disease death trends in urban and rural China from 2004 to 2019. Chinese Journal of Cardiology, 2019, 27(4): 371-376.

[10]Yang Juxian, Du Qin. The emergence of behavioral cardiology and its treatment model. Journal of Cardiovascular Rehabilitation Medicine, 2007, 421-425.