

Research on the Influencing Factors of Heart Disease based on Binary Logit Regression

Yinuo Cheng¹, Haiyang Liang^{2,*}, Xirui Ma³ and Zizhou Yang⁴

¹Xinjiang Bingtuan No.2 Middle School, Urumqi, 300110, China

²School of bioengineering, Huzhou Normal College, Huzhou, 313000, China

³Wuhan Britain-China School, Wuhan, 430030, China

⁴Beijing No.80 High School, Beijing, 110108, China

* Corresponding Author Email: WAMAN30363@my.sunywc.edu

Abstract. The heart is an important organ of the human body and plays an important role in the functioning of the human body. Heart disease has the greatest impact on the circulatory system. Patients will suffer from insufficient circulation blood supply, which will affect their health, including obvious chest tightness, shortness of breath, palpitations, pain in the precardiac area, discomfort in the precardiac area, accompanied by fatigue and difficulty in breathing. The study looked at clinical factors affecting heart disease. The dataset included 819 clinical cases and 12 variables, such as blood pressure, type of chest pain, etc. The target variable was whether a person had been diagnosed with heart disease. In this study, the accuracy of the first training set was about 88%, so the logistic regression results were acceptable. This model examines the factors that influence heart disease and can be useful in predicting heart disease. The conclusion shows that the risk of heart disease may be related to age, oldpeak, ChestPainType, MaxHR, FastingBS, cholesterol, ST slope and exercise angina pectoris.

Keywords: Heart disease, influencing factors, binary logit regression.

1. Introduction

Nowadays, with the booming economy and technological advancement, more and more people are more concerned about their living standards and personal health. As a disease with a high mortality rate - heart disease - it has been given great attention by people from all walks of life. According to a statistical report, the probability of death due to heart disease among residents in urban China reached 148.51/100,000 in 2019, which means that about 148.15 people died due to heart disease in 100,000 RAMs, an increase of 2.2/100,000 compared with 2018. According to the analysis based on the above data, the death rate from heart disease is still not short of rising [1]. Therefore, the subject of exploring the factors affecting heart disease is given higher value as a key entry point for mankind in preventing heart disease and improving life expectancy.

Several scholars at home and abroad have conducted meticulous research work on the factors affecting heart disease. The study by Annika Bay investigated the prevalence of symptoms and associated factors during pregnancy in women with primiparous coronary heart disease. Using data from a national registry, the study found that 104 of 465 women assessed reported symptoms in late pregnancy, with palpitations and dyspnea being the most common. Low physical activity and higher NYHA classification were found to be associated with symptom onset. The results emphasize the prevalence of symptoms in women with coronary heart disease during pregnancy, which often precedes pregnancy. It is recommended that these findings be included in pre-pregnancy counseling and taken into account in monitoring during pregnancy to better attend to and manage the health of this special population [2]. Studies have shown that women have a lower risk of coronary heart disease than men. Estrogen protects the body's cardiovascular system, so women are less likely than men to develop coronary heart disease before menopause. As a result, women's risk of developing coronary heart disease increases after menstruation has stopped. women over the age of 65 and men over the age of 55 have a higher risk of developing coronary heart disease [3]. Studies have shown

that a major risk factor for heart disease in young people is elevated blood pressure levels. Even when blood pressure is not abnormally high, the likelihood of heart disease is elevated [4]. The data suggest that the risk of heart disease is proportional to the degree of depressive symptoms compared to no depressive symptoms and that the probability of heart disease is associated with mild anxiety symptoms compared to no anxiety symptoms [5]. Further studies have shown that people who sleep less than five hours a night have a higher risk for heart disease. In addition, scholars have found that too much sleep is unbeneficial for heart health. People who slept six hours or less had a lower risk of developing the disease than those who slept nine hours. Researchers found that people who had difficulty sleeping were 69% more likely to develop heart disease than those who slept easily [6]. Research analysis shows that people can help reduce their risk of cardiovascular disease by 20% by climbing five floors a day [7]. A serious risk factor for comorbid coronary heart disease in patients with T2DM has been found through research to be age [8].

The research problem of this paper is to analyze the influencing factors of heart disease and to discuss the eight aspects of the research in terms of Age, oldpeak, ChestPainType, MaxHR, FastingBS, Cholesterol, ST Slope, and ExerciseAngina. Through Logistic Regression analysis, the study found that Oldpeak, ChestPainType, MaxHR, FastingBS, Cholesterol, ST Slope, and ExerciseAngina are key influencing factors for heart disease. There are some limitations, such as the sample does not cover all age groups and races, which will affect the accuracy of the results, but we still have the goal of helping people detect heart disease earlier and take interventions by analyzing the study to find the influencing factors.

The use of this model to predict whether a patient is at risk of heart disease is a method that enables doctors to make better medical diagnoses on the one hand, and on the other hand, improves the efficiency of medical resources, maximizes the use of limited resources, and has an indispensable impact in reducing the risk of heart disease [9]. Sadiya Khan, assistant professor of medicine and preventive medicine at Northwestern University Feinberg School of Medicine, pointed that the leading cause of death in the U.S. is heart disease, and finding the best way to identify the risk of heart disease will help doctors identify specific treatments to reduce the risk of heart disease earlier, and as a result, people will be able to recognize and manage heart disease more quickly, reducing the probability of treatment failure due to failure to diagnose in time [10].

2. Methods

2.1. Data Source

The data for this literature is collected from the Kaggle website, which was compiled for the prediction of heart disease, exploratory data analysis (EDA) is used to explore the data set of heart disease, and classification algorithm is used for prediction (modeling). Prediction of heart disease was correlated with gender, Resting, FastingBS, RestingEC, MaxHR, ExerciseAngina, ExerciseAngina and ST_Slope.

2.2. Variable Selection

The data used in this paper count a total of 918 people, including those who have and do not have heart disease of whom 725 are male and 193 are women. The patients' ages range from 28 to 77 years, including people with and without heart disease, Data included 8 variables (Age, Oldpeak, ChestPainType, MaxHR, FastingBS, Cholesterol, ST_Slope, ExerciseAngina).

Table 1 shows the number of people and heart disease who have the disease. As shown in Table 1, for the convenience of writing, the logogram of the factors is as above. The sample of data is 508, of 410. There are 4 Categorical variables in the Experimental data.

Table 2 shows the number of females and males, and lung cancer patients in each age group. The majority of people in the data were aged 51-60 and 61-70 years old. Table 3 shows the ChestPainType distribution by group.

Table 1. Logogram and numbers of the 12 factors

Designation	Symbol	Minimum	Minimum
Age	X1	28.000	77.000
Sex	X2	0.000	1.000
RestingBP	X3	0.000	200.000
ChestPainType	X4	1.000	4.000
Oldpeak	X5	-2.600	6.200
FastingBS	X6	0.000	1.000
Re3ingECG	X7	1.000	3.000
Cholesterol	X8	0.000	603.000
HeartDisease	X9	0.000	1.000
ExerciseA1gi1a	X10	1.000	2.000
MaxHR	X11	60.000	202.000
ST_Slope	X12	1.000	3.000

Table 2. Gender distribution by age group

Age	[28-38]	[39-49]	[50-60]	[61-71]	[72-77]
Number 1	64	226	406	202	20
Cancer 1	22	84	297	92	13

*Number1: The number of people in each age group.

**Cancer1: The number of people in each age group who have heart disease

Table 3. ChestPainType distribution by group

ChestPainType	1	2	3	4
Cancer 2	386	24	72	20
Number 2	496	173	203	46

*Number 2: The number of people in each age group.

**Cancer 2: The number of people in each age group who have heart disease.

3. Results and Discussion

3.1. Descriptive Analysis

Most of the patients have ages between 45 to 60 years. However, there are still patients younger or older than that range, but no one was below 21 years old. This suggests that heart diseases happen in all ages and heart disease could have different classifications (Figure 1).

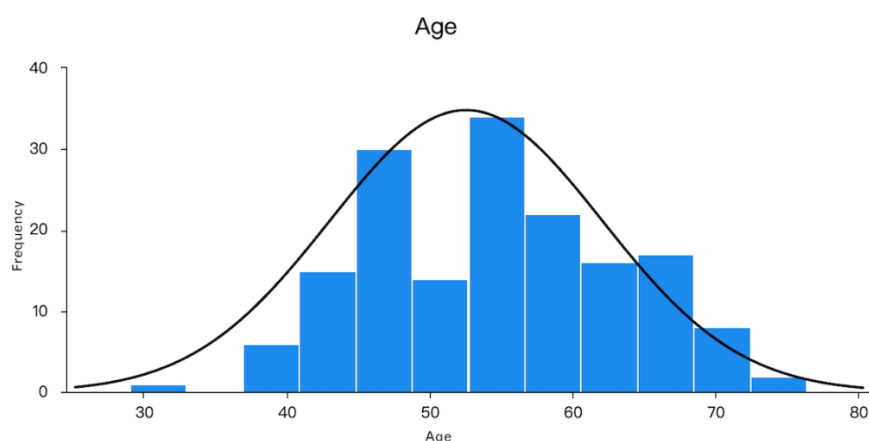


Fig. 1 The histogram of Age

Figure 2 shows the distribution of resting blood pressure of heart disease patients. Most patients have the level between 110 to 160, meaning many of them have heart rate that is too high compared to normal blood pressure which is less than 120 mmHg.

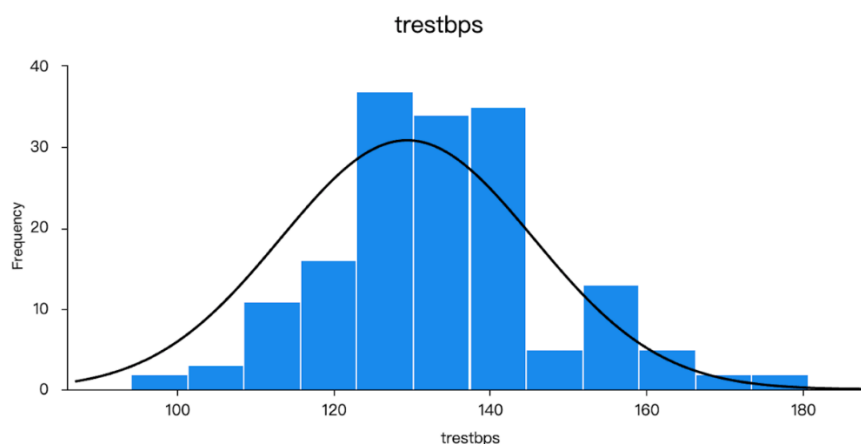


Fig. 2 The histogram of resting blood pressure

Cholesterol is a substance that travels through blood on proteins, high cholesterol level might lead to heart disease. As the figure 3 shows, the mean cholesterol level is around 250 mg/dl, and only few patients have normal cholesterol levels.

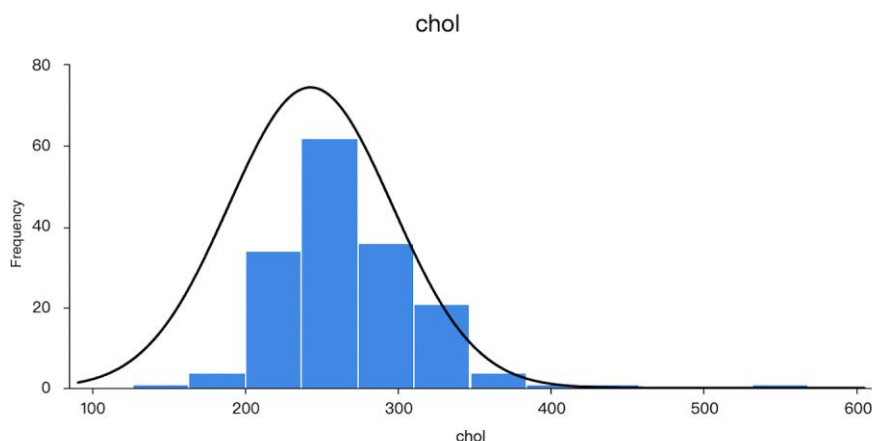


Fig. 3 The histogram of Cholesterol

The graph of maximum heart rate achieved per minute. The figure 4 shows slightly skewness with most of heart disease patients having the maximum heart rate over 140 which is not in healthy condition. The graph has a range over 100 with no significant outliers.

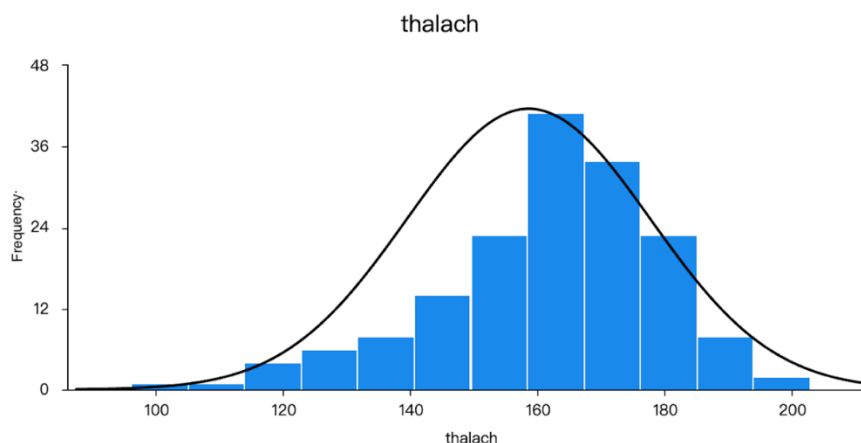


Fig. 4 The histogram of Maximum Heart Rate

3.2. Pearson Correlation

Table 4 shows the relevance between the 9 data categories (sex, age, old peak, pain type, fasting blood sugar, cholesterol, ST slope, exercise angina) collected from the heart disease patients. There exists a small negative relevance between ST slope and old peak, exercise angina. Other index shows no significant relationships.

Table 4. Correlation results

	Sex	Age	Oldpeak	ChestPain Type	MaxHR	FastingB S	Choles terol	ST_Slope	Exercise Angina
Sex	1								
Age	0.056	1							
Oldpeak	0.106	0.259	1						
ChestPainType	-0.127	-0.077	-0.177	1					
MaxHR	-0.189	-0.382	-0.161	0.289	1				
FastingBS	0.120	0.198	0.053	-0.073	-0.131	1			
Cholesterol	-0.200	-0.095	0.050	0.068	0.236	-0.261	1		
STSlope	-0.151	-0.268	-0.502	0.214	0.343	-0.176	0.111	1	
ExerciseAngina	0.191	0.216	0.409	-0.355	-0.370	0.060	-0.034	-0.429	1

3.3. Logistic Regression Results

Table 5 shows the various factors affecting heart diseases. ST-slope is a measure of segment shift in heart rate caused by frequent exercise, the z scores and p value show that it is strongly related to heart disease. Chest pain type related to exercise angina, which also refers to the discomfort of breath caused by such activities, also affects heart diseases, showing that exercise should be maintained in a suitable range for those patients. The maximum heart rate, however, did not show such relevance. Fasting BS, shortened for fasting blood sugar, is also a factors of heart diseases, meaning that diabetes also affects the function of the patient's heart. Old peaks as a p value of 3.226 could also affect the presence of heart diseases.

Table 5. Logistic model results

Term	Coefficient	Standard Error	z value	p value	95% CI
ST_Slope	-1.721	0.207	-8.332	0.000	0.119 ~ 0.268
ExerciseAngina	1.114	0.231	4.830	0.000	1.938 ~ 4.786
MaxHR	-0.009	0.005	-1.813	0.070	0.982 ~ 1.001
FastingBS	1.104	0.258	4.273	0.000	1.817 ~ 5.003
RestingECG	-0.196	0.167	-1.173	0.241	0.593 ~ 1.140
Oldpeak	0.372	0.115	3.226	0.001	1.157 ~ 1.818
Cholesterol	-0.004	0.001	-3.573	0.000	0.994 ~ 0.998
RestingBP	0.004	0.006	0.688	0.491	0.993 ~ 1.015
ChestPainType	-0.679	0.106	-6.404	0.000	0.412 ~ 0.624
Sex	1.348	0.258	5.228	0.000	2.322 ~ 6.377
Age	0.017	0.012	1.391	0.164	0.993 ~ 1.041
Constant Term	2.050	1.449	1.415	0.157	0.454 ~ 132.928

*Dependent Variable: HeartDisease

Compared to previous studies, the model shows similar results. As other paper suggests, sex, physical exercise, and even age have considerable impacts on patient's heart diseases. However, the data collected still have limitations that didn't include behavioral factors such as sleeping time or smoking condition, overall physical and mental health are also not evaluated. According to previous

data analysis, sample size does also impact the logistic regression model results, but the current sample size is still restricted due to the lack of data sources. More practical survey or experiment results could be investigated to improve the overall heart disease analyze.

Based on Table 6, the model has been used effectively in computing the results of factors affecting heart diseases. The overall accuracy had reached 85.95% and it can be accepted, and it can be concluded that heart disease is dependent to part of its index.

Table 6. Prediction accuracy

		Predicted Value		Accuracy	Error Rate
		0	1		
True value	0	341	69	83.17%	16.83%
	1	60	448	88.19%	11.81%
Total				85.95%	14.05%

4. Conclusion

The current study selected a variety of data and focused on influencing factors that may be related to the development of heart disease. The conclusion shows that the risk of heart disease may be related to age, oldpeak, ChestPainType, MaxHR, FastingBS, cholesterol, ST slope and exercise angina pectoris. Most of these factors have not been noticed before.

Admittedly, due to the limited amount of data, there may be errors in the model in addition to these factors, and the sample does not cover all ages and ethnicities, so there may be differences, which may also affect the accuracy of the results. However, the research still has a lot of value and value. First of all, the method adopted in this study is innovative. On the one hand, the graphical method is used to intuitively analyze the difference in the proportion of various factors between the heart disease population and the non-heart disease population. This allows the experiment to be visualized, making the results clearer and more intuitive. On the other hand, this experiment did not use single factor analysis like many previous experiments, but chose a more comprehensive multiple linear regression. Secondly, it has a positive effect on the treatment of heart disease. In addition to known factors related to heart disease, such as drinking status and smoking status, there are more factors that may be related to heart disease and deserve attention and consideration, such as congenital abnormalities, overweight, emotional agitation and so on. Whether these factors are linked to the development of heart disease will require further medical research, which means that these findings will point the way to further related research in the future. Once new risk factors are identified in addition to those already identified, this could help people to detect heart disease earlier and treat it as soon as possible, improving survival rates and patients' quality of life.

Authors Contribution

All the authors contributed equally and their names were listed in alphabetical order.

References

- [1] Lin Yi, Zhang Yuejuan, Feng Jin, et al. Application of homemade mint solution in postoperative oral care of children with congenital heart disease Observe the effect. *Introduction to Traditional Chinese Medicine*, 2019, 25(6): 125-127.
- [2] Zhao Bo. Coronary heart disease, heart disease Comprehensive understanding of coronary heart disease. *Food and Health*, 2024, 36(01): 4-5.
- [3] Li Wenjuan, Wan Yongchao, Guo Lu, et al. The effect of different baseline blood pressure levels on early onset heart disease in young people under 40 years old. *Chinese Journal of Circulation*, 2023, 38(12): 1267-1273.

- [4] Qi Ran, Sheng Lily, Zhang Xinyu, et al. A study on the correlation between depression and anxiety symptoms in elderly people in China and heart disease and chronic non communicable respiratory system diseases. *Prevention and Control of Chronic Diseases in China*, 2023, 31(12): 881-886.
- [5] Yang Guangping. Insomnia increases the risk of heart disease. *Family Medicine*, 2023, 12: 33.
- [6] Huang Jinfa, Huang Shicong. Analysis of influencing factors of angina pectoris in elderly type 2 diabetes patients with coronary atherosclerotic heart disease. *Chinese and Foreign Medical*, 2023, 42(33): 39-42.
- [7] Qiao Hui. Research on heart disease prediction methods based on data mining classification techniques. Lanzhou University of Technology, 2023.
- [8] Li Muzi. A new method for predicting heart disease in middle-aged people has emerged. *China Science Daily*, 2023.
- [9] Yafasova A, Fosboel E L, Gustafsson F, et al. Sarcoidosis is associated with a higher long-term risk of ischaemic heart disease, venous thromboembolism, and ischaemic stroke: a nationwide study. *European Heart Journal*, 2023.
- [10] Ritter A, Leonard J, Gray C, et al. MYH7 variants cause complex congenital heart disease. *American journal of medical genetics*, 2022, 188(9): 2772-2776.