Investigation of risk factors for the diagnosis of heart disease

Tingyi Li[1, *, †], Ruiying Zhang[2, a, †]

[1]College of Letters and Sciences University of Wisconsin Madison Madison, Wisconsin, United States

[2]Huazhong University of Science and Technology Wuhan, Hubei, China

* Corresponding Author Email: tli343@wisc.edu, [a]u201912677@hust.edu.cn

†These authors contributed equally.

Abstract. Introduction: Limited studies have investigated the association between diagnosis of heart disease and its relating risk factors and haven't met with robust results. We hypothesized that all or part of the risk factors are correlated with the heart disease. Method: This dataset published in 1998 was obtained from UCI Machine Learning Repository and was collected at University of Irvine. The dataset has 303 instances with 76 attributes, but all published experiments refer to using a subset of 14 of them. Our study focuses on 11 parameters specifically, including body health conditions, historic medical records, and habits. Logistic regression analyses were conducted to assess the relative risks of heart disease. Results: Both chest pain type ($p < 0.01$) and ST depression ($p < 0.05$) are positively correlated with the incidence of heart disease. Maximum heart rate, on the other hand, are negatively correlated with the diagnosis of heart disease. Conclusion: Our study suggested that chest pain type, ST depression and maximum heart rate are saliant contributors to indicate the occurrence of heart disease. The findings from our study have implications for the heart disease and call for future studies to explore the underlying prevention strategies of this findings.

Keywords: heart disease, risk factors, logistic regression, ST depression, chest pain type.

1. Introduction

Heart disease can be categorized into several conditions, which are coronary artery disease, congenital heart defects, arrhythmia, dilated cardiomyopathy, myocardial infarction, heart failure, hypertrophic cardiomyopathy, mitral valve regurgitation, mitral valve prolapse, and aortic stenosis. In research from Ross and Stier, CHD is the leading cause of death for adults worldwide and is expected to be the leading cause of death in western countries as well in the 21st century [4]. Coronary artery disease (CHD) is the most common type of heart disease in the United States, which influences the blood flow to the heart. Decreasing blood flow is likely to cause heart attack, consequently, developing coronary artery disease. In the latest report, CDC demonstrated that high blood pressure, high cholesterol, and smoking are the key risk factors. Specifically, approximately half of people (47%) in the U. S have at least one of these three risk factors [1]. Generally, there is a sex difference in heart disease. Physical inactivity is associated with an increased risk of developing heart disease in men [7]. In contrast, a study conducted by Blair and colleagues reported that an increasing level of exercise was related to a decreasing risk of developing in men but not in women [8]. Women in their midlife years, aged 45-54, reported stroke prevalence was double that of similarly aged men [5]. In other words, women are so superior regarding cardiovascular disease. Gender differences could also be shown in the diagnosis and treatment of heart disease. For example, males and females respond differently to many drugs that treatment response varies according to these two genders. Also, difficulties in diagnosing heart disease in women might another difference. Specifically, women with chest pain are more likely to be considered benign, thus, it's hard to diagnose and analyze the early sign of heart disease [6]. Cholesterol level is considered to be an indicator of predicting heart disease. The study conducted by Menotti et al. showed that late coronary heart disease death rates are largely 'explained' by changes in blood cholesterol levels [3]. Chest pain is consistently the initial symptom of acute myocardial infarction.ST-segment depression is one of the indicators in the diagnosis of heart disease, its depression indicates a poor prognosis. Reference standards for maximal heart rate (HR)

are important to help interpreting the adequacy of physiologic stress during graded exercise testing. It can also identify the presence of chronotropic incompetence, and prescribe an exercise training regimen when measured maximal HR is not available.

Although heart disease has been well studied and discussed, there are still some problems that exist. To be specific, it is still obscure for people to predict the risks of heart disease in the early stage. The occurrence and development of heart disease have a gradual evolutionary process, which can take from several years to decades. A huge body of evidence indicates that chronic heart attack or heart failure has the potential to develop into heart disease. Therefore, it is necessary to explore risk factors related to heart disease. Consequently, it is useful for people to find this disease earlier and have a better way to prevent disease than it used to be.

Considering the limited investigation of early diagnosis of heart disease, we investigated the risk factors of heart diseases by analyzing the data from University of California Irvine.

2. Method

2.1. Data source

After searching a plenty of datasets, we choose one film from UCI Machine Learning Repository because it includes risk factors of heart disease that helps predict the occurrence of heart disease. This dataset was collected at the University of California Irvine, the United States. Due to privacy concerns, the names and social security numbers of the patients were recently removed from the database.

A study of the data collection was performed in Irvin, the United States, and published in 1998. The dataset contains 303 instances with 76 attributes, but all published experiments refer to using a subset of 14 of them. Our study focuses on 11 parameters specifically, including body health conditions, historic medical records, and habits.

2.2. Risk factors

We considered eleven parameters for our research, which are age, chest pain type (cp), maximum heart rate (Max. HR), fast blood sugar > 120 mg/dl (FBS over 120), ST depression, cholesterol (chol), sex, resting blood pressure (trestbps), EKG results, exercise-induced angina (exang), and slope of ST. specifically, the range of the age is from 29 to 77. A persons' cholesterol measurement in mg/dl in dataset is from 126 to 564. The maximum heart rate people reached ranges from 71 to 202. The ST segment is an interval between ventricular depolarization and ventricular repolarization. It is identified as the end of the QRS complex to the beginning of the T wave. FBS over 120 is a kind of categorical data that we use 1 represents "true" and 0 represents "false." Similarly, chest pain type is also a kind of categorical variable that value 1 means no symptom, value 2 means atypical angina, value 3 means non-anginal pain and value 4 means typical angina.

2.3. Statistical analysis

Chi-square test is used to evaluate the correlation between each categorical variable and the dependent variable. Wilcox test is used to evaluate the correlation between each continuous variable and the dependent variable to the preliminary judgment of the relevance between independent variables and dependent variable.

Logistic regression (LR) is a multivariable method that was devised for dichotomous outcomes. It is particularly appropriate for models involving disease state (diseased/healthy) and decision making (yes/no), and therefore is widely used in studies in the health sciences. In LR one obtains the logarithm of the odds of a positive outcome (where "positive" is defined by the encoding of

the outcome variable, that is, Y=1); a straightforward algebraic manipulation transforms this into the outcome's probability.

We use Logistic regression to analyze the probability among age, chest pain type, chest pain type, maximum heart rate achieved, ST depression and serum cholesterol (mg/dl) and diagnosis of heart disease. Then we stepwise regression to remodel to reduce multicollinearity.

We used train set to perform logistic regression and train set to train the model. We took 0.5 as the threshold. True Positive, True Negative, False Positive, and False Negative were used to establish the confusion matrix to evaluate their sensitivities and accuracy. We also drew the ROC curve and calculated the AUC value. AUC means "Area under the ROC Curve". An ROC curve plots TPR vs. FPR at different classification thresholds. Lowering the classification threshold classifies more items as positive, thus increasing both False Positives and True Positives.

3. Results

Table 1: Distribution of Selected Characteristics (N=270)

Continuous Variables	Mean	Range
Age	54.43	29-77
Cholesterol (mg/dl)	249.7	126-564
Maximum heart rate	149.7	71-202
ST depression	1.05	0.0-6.2

Characteristics	Type	total
Diagnosis of heart disease		
	Yes	120
	No	150
Chest pain type		
	1	20
	2	42
	3	79
	4	129
fasting blood sugar > 120 mg/dl		
	TRUE	40
	FALSE	230

Chi Square Test		
Variable	chest pain type	fasting blood sugar > 120 mg/dl
p-value	1.561e-15	0.9237

Table 2: Correlation tests between heart disease and each selected risk factor

Wlicox test	
Variable	p-value
Age	0.0002
Cholesterol(mg/dl)	0.0077
Maximum heart rate	5.841e-12
ST depression	2.979e-11

It was found that, holding all other predictor variables constant, the odds of diagnosis of heart disease occurring increased by 0.19 for a one-unit increase in Age, by 1.33 for a one-unit increase in chest pain type, by 0.48 for a one-unit increase in ST depression.

It was also suggested that, holding all other predictor variables constant, the odds of diagnosis of heart disease occurring decreased by 0.027 for a one-unit increase in maximum heart rate, by 0.006 for a one-unit increase serum cholesterol (mg/dl).

$$\begin{aligned}\text{logit}(y) = {} & 0.017x_1 + 2.752x_2 + 1.344x_3 + 0.029x_4 + \\ & 0.006x_5 - 0.057x_6 + 0.377x_7 - 0.027x_8 + \\ & 0.377x_9 + 0.482x_{10} + 0.379x_{11} - 10.512\end{aligned}$$

x_1 : Age; x_2 : Sex; x_3 : Chest pain type; x_4 : BP;
x_5 : Cholesterol; x_6 : FBS; x_7 : EKG results; x_8 : Max HR;
x_9 : Exercise angina; x_{10} : ST depression; x_{11} : Slope of ST

BP stands for resting blood pressure (in mm Hg on admission to the hospital), EKG results means the results of the blood flow observed via the radioactive dye. And Exercise angina stands for whether exercise induced angina, 1 for true, 0 for false. The slope of ST also represents variable in the dataset.

Table 3: The result of logistic regression

	Estimate	P value	Signif.
(Intercept)	-10.512	0.0027	**
Age	0.017	0.5277	
Sex	2.752	0.0000	***
Chest pain type	1.334	0.0000	***
BP	0.029	0.0341	*
Cholesterol	0.006	0.1496	
FBS over 120	-0.575	0.3556	
EKG results	0.377	0.0886	
Max HR	-0.027	0.0244	*
Exercise angina	0.761	0.1075	
ST depression	0.482	0.0545	
Slope of ST	0.379	0.4104	

Signif code:0<=’***’<0.01’**’<0.05’*’<0.1’’<1

Then we used stepwise regression to remodel. After remodeling, the AIC value increased significantly.

It was found that, holding all other predictor variables constant, the odds of diagnosis of heart disease occurring increased by 1.30 for a one-unit increase in chest pain type, by 0.60 for a one-unit increase in ST depression.

It was also showed that, holding all other predictor variables constant, the odds of diagnosis of heart disease occurring decreased by 0.0068 for a one-unit increase serum cholesterol (mg/dl), by 0.031 for a one-unit increase in maximum heart rate.

After stepwise regression, the age was removed. Formula follows:

$$\text{logit}(y) = 2.651x_1 + 1.301x_3 + 0.029x_4 + 0.007x_5 + 0.029x_6 + 0.382x_7 - 0.032x_9 + 0.753x_{10} + 0.603x_{11} - 8.519$$

Table 4: The result of stepwise regression

	Estimate	P value	Signif.
(Intercept)	-8.519	0.0021	**
Sex	2.651	0.0000	***
Chest pain type	1.301	0.0000	***
BP	0.029	0.0293	*
Cholesterol	0.007	0.1004	
EKG results	0.382	0.0824	
Max HR	-0.032	0.0040	**
Exercise angina	0.753	0.1089	
ST depression	0.603	0.0032	**

Signif code:0<=’***’<0.01’**’<0.05’*’<0.1’’<1

The association between heart disease and risk factors were presents in Figure 1. From the bar plot, ST depression induced by exercise relative to rest and serum cholesterol in mg/dl are all greater in confirmed heart disease patients, which partly proved the previous study about the positive correlation between cholesterol level and heart disease. However, healthy people have higher maximum heart rates compared to heart disease patients. From the bar chart, the first graph indicated that the number of healthy people is greater than people with heart disease. The second one illustrated that the most severe chest pain occurred frequently in confirmed patients than that of healthy people. The last one, on the other hand, indicated a smaller number of fast blood sugar instances in heart disease patients.

Figure 1: The distribution of different factors

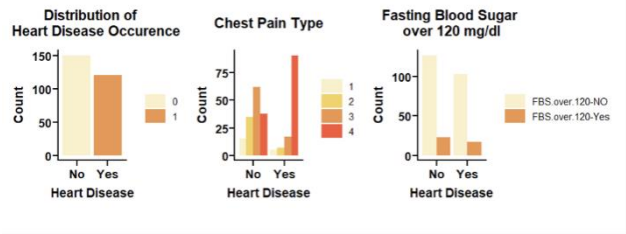

Figure 2: Distribution in different factors

4. Discussion

Our results support the previous studies that there is a consistent positive correlation between ST depression ($p < 0.05$), chest pain type ($p < 0.01$) and diagnosis of heart disease, and there is a negative correlation between maximum heart rate ($p < 0.05$) and diagnosis of heart disease. It is apparent that ST depression, Chest pain type and maximum heart rate are strong contributors to predict the occurrence of heart disease. Other risk factors, on the other hand, are insignificant to be considered in the diagnosis of heart disease due to their p values.

We compare the model with different manipulations. We first used logistic regression to analyze the data. We found that AIC value is quite high and almost no risk factors are significant to predict the incidence of heart disease. Then, we used stepwise regression to examine the data. After using stepwise logistic regression, we found that AIC value has a significant decreasing trend, as a result, we got better p values for each risk factor.

Our research still has some limitations to be considered. From the process of logistic regression, it is apparent that there exists a negative correlation between cholesterol level and the occurrence of

heart disease. However, CDC demonstrated that high cholesterol is one of the key factors that increase the risk of developing heart disease. By considering the difference that cholesterol level has a negative correlation in our dataset and a positive correlation generally, we think it's probably due to different measures of cholesterol. To be specific, there are total cholesterol, low-density lipoprotein (LDL) cholesterol, ratio of total cholesterol to high-density lipoprotein (HDL) cholesterol, and the ratio of LDL to HDL. In research from Kinosian, the total cholesterol/HDL ratio seems to be a superior factor that predict the heart disease [2]. However, our dataset doesn't specify the types of cholesterol. Therefore, it is possible to explain the negative relationship between cholesterol and diagnosis of heart disease according to our data and positive association in general trend. We still need more further evidence to eliminate this difference. Moreover, the risk factors in the dataset are not comprehensive and there are some important risk factors that are not included. For example, smoking (years) and different types of heart diseases could also be relating risk factors that influence the diagnosis of heart disease. Since our dataset is derived from the U.S, a high-income country, the data may only represent the conditions in high-income countries rather than other developing countries. Therefore, the data might not be representative, and we still need more general data from different locations.

5. Conclusion

Our study supports the hypothesis that part of the risk factors are strong contributors to the diagnosis of heart disease. We discover that chest pain type, ST depression and maximum heart rate are the salient risk factors for diagnosing the heart disease. Specifically, chest pain type and ST depression have strong positive relationship with the incidence of heart disease, whereas maximum heart rate has negative correlation with the diagnosis of heart disease. In other words, more severe chest pain type and ST depression are correlated with a high probability of heart disease diagnosis. Lower maximum heart rate also correlated with high possibility of developing heart disease. Future studies are needed to explore more general association between heart disease and their relating risk factors. We can also use other statistical methods like random forest which trained on the different parts of our dataset to reduce the variance.

References

[1] Heart Disease Resources | cdc.gov. (2021, September 27). Centers for Disease Control and Prevention. https://www.cdc.gov/heartdisease/about.htm

[2] Kinosian, B. (1994). Cholesterol and Coronary Heart Disease: Predicting Risks by Levels and Ratios. Annals of Internal Medicine, 121(9), 641. https://doi.org/10.7326/0003-4819-121-9-199411010-00002

[3] Menotti, A., Blackburn, H., Kromhout, D., Nissienen, A., Fidanza, F., Giampaoli, S., Buzina, I., Mohacek, I., Nedeljkovic, S., Aravanis, C., & Toshima, H. (1997). Changes in population cholesterol levels and coronary heart disease deaths in seven countries. European Heart Journal, 18(4), 566–571. https://doi.org/10.1093/oxfordjournals.eurheartj.a015298

[4] Ross, G., & Stier, J. (1999). Lifetime risk of developing coronary heart disease. The Lancet, 353(9156), 924. https://doi.org/10.1016/s0140-6736(99)00046-x

[5] Towfighi, A. (2009). Sex-Specific Trends in Midlife Coronary Heart Disease Risk and Prevalence. Archives of Internal Medicine, 169(19), 1762. https://doi.org/10.1001/archinternmed.2009.318

[6] Xhyheri, B., & Bugiardini, R. (2010). Diagnosis and Treatment of Heart Disease: Are Women Different From Men? Progress in Cardiovascular Diseases, 53(3), 227–236.https://doi.org/10.1016/j.pcad.2010.07.004

[7] Powell, K. E., Thompson, P. D., Caspersen, C. J., & Kendrick, J. S. (1987). Physical activity and the incidence of coronary heart disease. Annual review of public health, 8(1), 253-287.

[8] Blair, S. N. (1993). Evidence for success of exercise in weight loss and control. Annals of internal medicine, 119(7_Part_2), 702-706.