Memorandum

To: Dr. Craig

From: Elijah Eberly

Date: December 10, 2021

Subject: What variables have the highest correlation to heart disease?

Introduction

Heart disease is the leading cause of death in the United States and is a problem that is not talked about enough. That is why I have chosen to do my statistical project on it to try and get a better understanding of heart disease and possible ways to control it. Heart disease can come from either unhealthy lifestyle or genetics so in this study I want to compare several variables and see which ones have the strongest correlation to heart disease. I am personally invested in this topic because my family has a long history of heart diseases, and it would be interesting to know if I can help prevent if possible. I try to live a healthy life, but if this test comes back more genetical it does not really matter. Personally, I think everyone should care about heart disease because it can save lives. Heart disease is very deadly but there are numerous ways to stop and control it and live a perfectly normal life, which I would think everyone wants to do.

Hypotheses Tests

Like stated previous my main goal in these tests is to find what variables are most correlated with heart disease. To do this the best way possible I will have to run several tests to compare which variables lead to a high chance of heart disease. Since I am limited on data, I want to focus on five different variables all related to heart disease. These variables include age, gender, cholesterol, exercise angina (a condition marked by severe pain in the chest caused by an inadequate blood supply to the heart), and resting blood pressure.

(1) Does being a certain gender have an effect on your heart disease chances?

Null Hypothesis: Gender is independent from heart disease.

Alternative Hypothesis: Gender is dependent of heart disease.

(2) Does age have an effect on your heart disease chances?

Null Hypothesis: There is equal chance of having heart disease at any age.

Alterative Hypothesis: Your chance of heart disease changes based on your age.

(3) Does having high cholesterol result in a higher chance of heart disease?

Null Hypothesis: The average cholesterol is equal to those who do and do not have heart disease.

Alterative Hypothesis: The average cholesterol id different from those who do and do not have heart disease.

(4) Does having exercise angina result in a higher chance of heart disease?

Null Hypothesis: Have heart disease and exercise angina are independent of each other.

Alternative Hypothesis: Having heart disease is dependent of having exercise angina.

(5) Does having a higher blood pressure result in a higher chance of heart disease.

Null Hypothesis: Blood pressure is the same among people with and without heart disease.

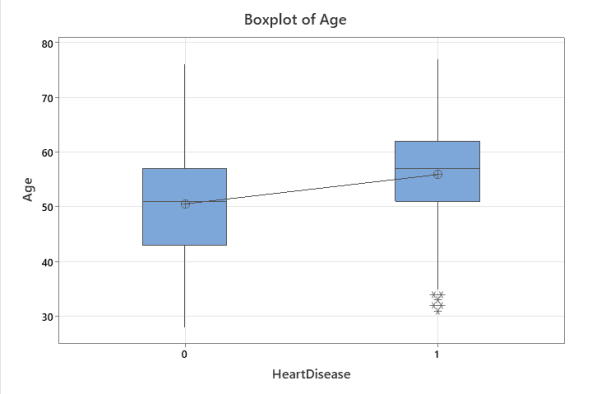
Alterative Hypothesis: Blood pressure is not the same among people with and without heart disease.

Data and Tests

The data gathered to test these hypotheses is from a database website called Kaggle which has a wide variety of different datasets. I specifically used the dataset labeled, "Heart Failure Prediction Dataset" (https://www.kaggle.com/fedesoriano/heart-failure-prediction/version/1). The dataset consists of 918 people all with a variety of conditions and whether they are heart disease or not. The main goal of this dataset is to compare and examine how certain variables effect your chances of heart disease. For the qualitative data I will be conducted a two sample T-test. I will be using this test because I want to know if the sample of people without heart disease is lower than the people with heart disease regarding the particular variable (age, cholesterol, and blood pressure) in the test. As for the quantitative data I will be using the chi-square test. This test is most effective because it will be able to compare the variables (gender and if they have exercise angina) of the sample to the sample of people with and without heart disease.

Results

The first test I ran was a two sample T-test comparing age and heart disease. The goal of this test was to see if the sample mean of the age of the people in the survey without heart disease was the same as the sample mean of the age of the people in the survey with heart disease. The average age without heart disease was 50.55 years old and the average age of people with heart disease was 55.90 years old. Obviously, there is a difference in the average, but by looking at the p-value I was able to determine if the difference was significant. The results were a p-value of 0.0% meaning there is significant difference between the age with heart disease and the age without heart disease. In other words, it says based off the two samples take (with and without disease) there is no chance the population average of age with and without heart disease could be equal. It also does not just show they are equal but also that heart disease is more common among older people, no surprise there. Below is the boxplot of the samples and also the test results including the hypotheses and p-value.



Test

Null hypothesis $H_0: \mu_1 - \mu_2 = 0$

Alternative hypothesis $H_1: \mu_1 - \mu_2 \neq 0$

T-Value	DF	P-Value
-8.82	843	0.000

The nest test performed was the chi-square test comparing the chances of heart disease being dependent of gender. Below the paragraph is the table from the test. The table means that out of the 918 total people in the sample 193 of them were females (21.0%) and 725 of them were males (79.0%). If the two are independent of each other than the percentage of females with heart disease should also be 21.0% of all the people with heart disease and 79.0% of people with heart disease are males. This also would be applied to the ratio of people without heart disease. Using these calculations, the chi-square test expects 86.2 females without heart disease and 106.8 with heart disease. As for the males it expects 323.8 without heart disease and 401.2 with heart disease. However, the expected values were not the actual amount. In the sample 143 females did not have heart disease and 50 did. For the males 267 did not have heart disease and 458 did. The actual amount was different than the expected, but the t-test is what determines if the difference is significant enough to deem the two are independent of one

another. The result of the p-value was 0.0% meaning heart disease is dependent of gender. Looking at the table the actual number of males with heart disease was 90.12% instead of the expected 79.0%. These means that males have a higher chance to have heart disease than females based off this sample.

Rows: Sex Columns: HeartDisease

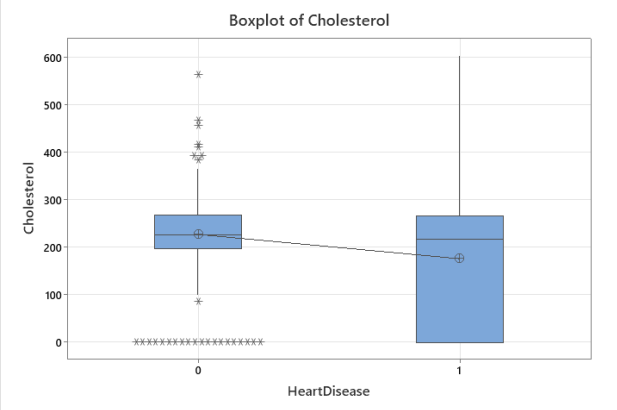
	0	1	All
F	143	50	193
	86.2	106.8	
M	267	458	725
	323.8	401.2	
All	410	508	918

Cell Contents
Count
Expected count

Chi-Square Test

	Chi-Square	DF	P-Value
Pearson	85.646	1	0.000
Likelihood Ratio	87.168	1	0.000

The next test I ran was another two sample T-test comparing cholesterol and heart disease. The null hypothesis since it is a two sample T-test is that there is no difference between the average sample cholesterol without heart disease and the average sample cholesterol with heart disease. Before even running the test, I expected to reject this null because I believed that people with heart disease have higher cholesterol. After running the test, I was correct that the null is rejected but for the opposite reason than I thought. The results came back that the people with heart disease had an average cholesterol of 176 and the people without heart disease had an average cholesterol of 227.1. Looking at the means alone does not prove a null rejected it only proves there is a difference of sample averages. The p-value is what determines if the difference is significant enough to reject or support the null. The p-value from this test was 0.0% rejecting the null. This means there is a zero percent that if the samples average is correct the difference between them is zero. Therefore, the final result of this test is that people with lower cholesterol have a higher chance of heart disease.



Test

Null hypothesis $H_0: \mu_1 - \mu_2 = 0$

Alternative hypothesis $H_1: \mu_1 - \mu_2 \neq 0$

T-Value	DF	P-Value
7.63	844	0.000

Like the first test I ran I ran another chi-square test but this time the test was to determine if having exercise angina is dependent of heart disease. Based on the sample of 918 people 547 do not have exercise angina and 371. That equivalates to 59.59% of the sample do not have it and 40.41% do. Based off these percentages the expected amount of people without heart disease and do not have exercise angina is 244.3 people and 165.7 without heart disease but do have exercise angina. As for the people with heart disease it was expected to have 302.7 people without exercise angina and 205.3 people do have exercise angina. However, after looking at the sample it turns out the expected values were not close. In reality out of the 410 people without heart disease 355 did not have exercise angina and 55 did. As for the 508 people with heart disease 192 did not have exercise angina and 316 did. Based off this information the p-value was 0.0% meaning there is no chance the variables are dependent. The two are independent of each other and having exercise angina increases your chance of having a heart disease.

Rows: Exercise Angina Columns: HeartDisease

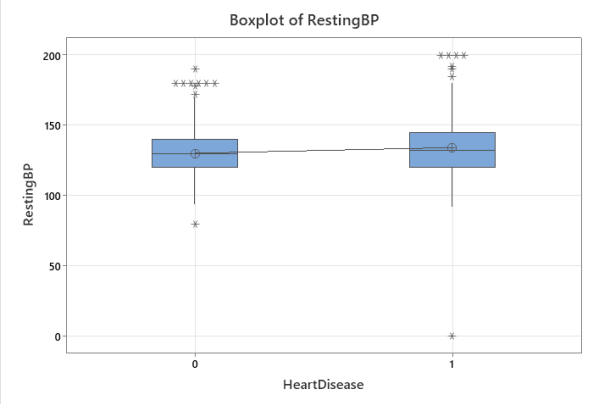
	0	1	All
N	355	192	547
	244.3	302.7	
Y	55	316	371
	165.7	205.3	
All	410	508	918

Cell Contents
Count
Expected count

Chi-Square Test

	Chi-Square	DF	P-Value
Pearson	224.281	1	0.000
Likelihood Ratio	241.765	1	0.000

The last test I conducted was another two sample T-test but this time looking at blood pressure and heart disease. The test is set up, so the null hypothesis is that the sample average of blood pressure for people with heart disease is equal to the sample average of blood pressure for people who do not have heart disease. Before running any test, I predict that the null will be rejected since I believe the sample average of blood pressure for people with heart disease will be higher than those without. After running the two sample T-test I was correct that we would reject the null since the p-value was only 0.1%. There was not enough evidence to support the null. The people with heart disease had a sample average of a blood pressure of 134.2 compared to the sample average of 130.4 for the people who do not have heart disease. This means the higher your blood pressure the higher chance of heart disease you have.



Test

Null hypothesis	$H_0: \mu_1 - \mu_2 = 0$
Alternative hypothesis	$H_1: \mu_1 - \mu_2 \neq 0$

T-Value	DF	P-Value
-3.34	915	0.001

Conclusion

The main problem I wanted to test was what variables can lead to a higher chance of heart disease. For the most part my question was answered. From these tests I was able to conclude that males that are older, have low cholesterol, have exercise angina, and have high blood pressure have the highest chance of heart disease. Obviously, there are a ton more variables that can lead to heart disease but with the dataset I had I am comfortable with my answer. There were more variables within the dataset that I could have tested but decided not to because I did not know much about them, and the answers would not really help because I would not be able to fully understand them. I also withed a could have run a fit regression model with all five variables I tested but since two of them were qualitative data I could not include all five in the test.