# Heart Disease Data Analysis

Winnie Louh, Jackson Savage

11/20/2020

# Introduction

## Motivation

For this project, we wanted to use our statistical skills to study issues impacting human life. Considering Jackson is a biochemistry major, naturally, our ideation process led us to explore healthcare-related data. We began our search by investigating common illnesses and diseases. To our surprise, many of the top causes of death are preventable through lifestyle changes, with heart disease being the leading cause of death in the United States. Around 655,000 Americans die from heart disease each year[2]–for context, around 234,000 have died from COVID-19.[3] Yet, much less attention is placed on heart disease compared to other illnesses since it takes longer to onset and is mostly preventable. Of course there is a significant fiscal impact on the healthcare system, with about $219 billion allocated to treatment from year 2014 to 2015.[2]

## Data Gathering

Once we decided to pursue heart disease, we looked at the causes and symptoms to ensure we find data that captured these variables. Starting with the symptoms, the most common traits are chest pain, shortness of breath, and angina (chest pain) after exercise. The risk factors are much more diverse, and they include smoking, high blood sugar, high cholesterol, high blood pressure, diabetes, poor diet, alcohol use, and lower maximum heart rate output.[1] While both of these lists are not exhaustive, together they cover the majority of factors for patients. In addition to these factors, we wanted data that included basic information like age and sex but also individuals who did not have heart disease to compare. These criteria led us to our current data set, obtained on Kaggle.com.[4]

## Objective

The goal of our project is to provide patients and healthcare professionals with better information on what traits put individuals at risk of developing heart disease. We plan to do this through understanding more about possible differences between males and females as well as those diagnosed and not diagnosed with heart disease through confidence intervals and hypothesis tests. Most importantly, we plan on finding a regression model through backwards selection to predict the probability of someone being diagnosed with heart disease given some of their characteristics. With access to better data,

patients and professionals can take proactive steps to prevent diagnosis of the disease and ultimately lessen the financial burden on our healthcare system.

# Data Description

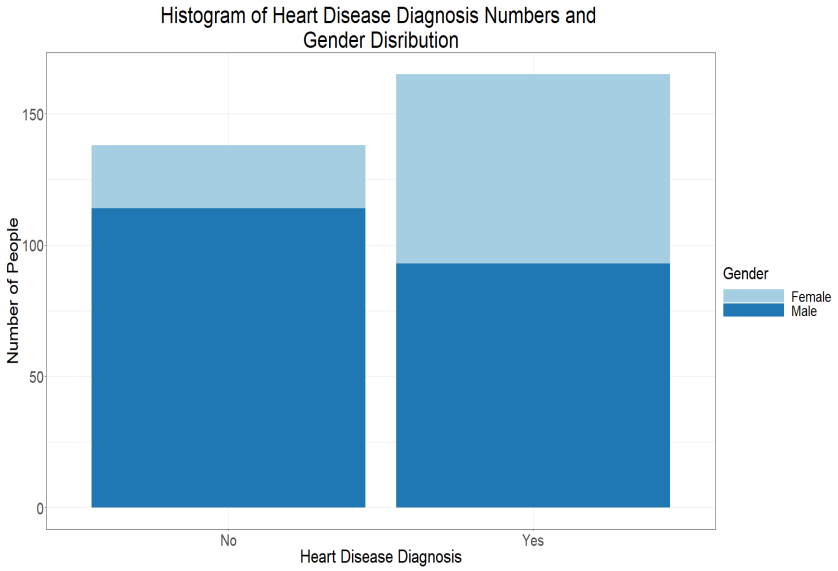
Our data comprises of fourteen variables which we will detail further. To begin, our data contains the basic variables of age, sex, blood pressure, cholesterol, blood sugar (categorical: 1 if $>$ 120 mg/dl, 0 if $<$ 120 mg/dl), max heart rate, and a diagnosis of heart disease (catagorical: 1 if yes, 0 if no). Several other categorical variables are included as well. First, there is a Resting Electrocardiographic Result, which is a test used to check the heart's rhythm and electrical activity. In layman terms, it's the machine you hear "beeping" in movies and shows and indicates if a person has flatlined. A 0 value represents normal results, while a 1 value indicates wave inversions (a common but sometimes life-threatening heart condition), and a 2 value indicates left ventricular hypertrophy (thickening of the heart's left pumping chamber). Additionally, the data contains information on the presence of exercise-induced angina (chest pain caused by reduced blood flow to the heart itself), with 0 meaning no pain and 1 indicating pain. There is also the presence of chest pain at rest, with 0 being typical angina, 1 being atypical angina, 2 being non-anginal pain, and 3 being no pain or asymptomatic. One unique aspect of our data compared to others is that it includes fluoroscopy information. Fluoroscopies are a form of X-ray imaging used to obtain real-time moving images. From this procedure, cardiologists were able to determine the number of major vessels with limited blood flow, classified from 0 to a max of 3 vessels clogged. Additionally, during fluoroscopic exams the activities of the ventricles can be observed. Our data possesses this information, called the ST Segment, and lists both the ratio of ST depressions induced by exercise relative to rest and the slope of the peak exercise ST segment.[4] Our data also included information on thalassemia, but because the data description was inconsistent with the values in the data, this column was excluded in the analysis.

To better understand our data, we began with a summary of each of the variables in the data, shown in the table below.

Table 1: Summary of Variables in the Data

| **Variable** | **Min** | **Quartile 1** | **Median** | **Mean** | **Quartile 3** | **Max** |
|---|---|---|---|---|---|---|
| Age | 29 | 47.5 | 55 | 54.37 | 61 | 77 |
| Sex | 0 | 0 | 1 | 0.6832 | 1 | 1 |
| Chest Pain | 0 | 0 | 1 | 0.967 | 2 | 3 |
| Resting Blood Pressure | 94 | 120 | 130 | 131.6 | 140 | 200 |
| Cholesterol | 126 | 211 | 240 | 246.3 | 274.5 | 564 |
| Fasting Blood Sugar | 0 | 0 | 0 | 0.1485 | 0 | 1 |
| Resting ECG | 0 | 0 | 1 | 0.5281 | 1 | 2 |
| Max Heart Rate | 71 | 133.5 | 153 | 149.6 | 166 | 202 |
| Angina | 0 | 0 | 0 | 0.3267 | 1 | 1 |
| ST Depressions | 0 | 0 | 0.8 | 1.04 | 1.6 | 6.2 |
| ST Exercise Slope | 0 | 1 | 1 | 1.399 | 2 | 2 |
| Fluorouscopy | 0 | 0 | 0 | 0.7294 | 1 | 4 |
| Diagnosis | 0 | 0 | 1 | 0.5446 | 1 | 1 |

Following this, to further understand our data, we employed some preliminary data visualizations.

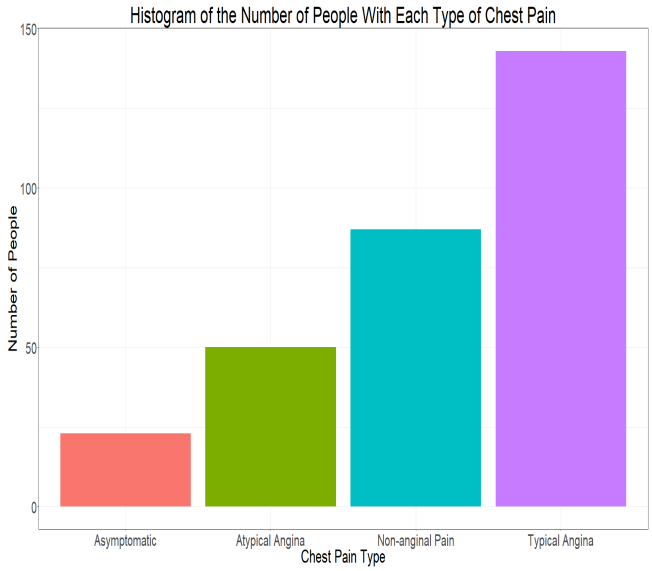



This first graph shows us the distribution of those diagnosed and not diagnosed with heart disease within our data sample. With the total number of people in our data being 303, we can see that a little more than 50 percent of the sample have been diagnosed and a little less than 50 percent have not been diagnosed. In addition, out of those who have been diagnosed with heart disease, over 60% of them are male. Out of those without heart disease, over 70% of them are male.

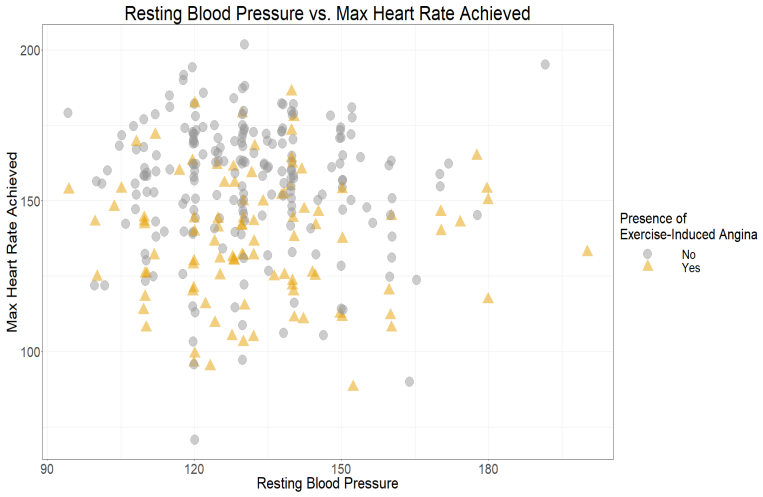
Continuing on with our preliminary data visualization, the next figure shows the distribution of age throughout our data set as well as the proportion of the total sample who do or do not have heart disease at that particular age.


From the figure above, we can observe that within our sample data set, the average age of people without heart disease, denoted by the dotted, vertical, and pink line, is slightly higher than the average age of people with heart disease, denoted by the dotted, vertical, and blue line.

Since it is known that chest pain and angina after exercise are symptoms of heart disease, we decided to visualize these variables as well.




From the figure above, we can see that out of the total 303 people in this data set, almost half have chest pain from typical angina. About 80 have non-anginal pain, about 50 have atypical angina, and almost 25 are asymptomatic.




From the figure above, we can see that there is not a clear trend between resting blood pressure and maximum heart rate achieved. However, we can see that those who do not experience exercise-induced angina appear to, on average, achieve a higher maximum heart rate.

# Data Analysis and Results

## Confidence Intervals

We employed confidence intervals for two studies. First, we found a 95% confidence interval for the mean age of people with heart disease (51.0-54.0). Next, we found a 95% confidence interval for the mean resting blood pressure of people with heart disease (126.8-131.8 mm HG).

## Hypothesis Tests

We also ran two hypothesis tests. First we examined the mean difference in resting blood pressure between men and women and found that we do not reject the null hypothesis with a p-value of 0.325. This result implies that men and women with heart disease do not have a significant difference in resting blood pressure, so our confidence interval found above can, for the most part, be applied to both men and women. Next, we ran a test on the mean difference in number of vessels colored by fluoroscopy. With a p-value of 0.040, we rejected the null hypothesis and concluded that men and women do not have similar fluoroscopic results; thus, they need to be studied on a gendered basis.

## Regression Model

For our regression, we ran two different tests: backwards AIC and backwards BIC. The first method yielded a model with the variables shown below. The model started with an AIC of 247.36 and was reduced to an AIC of 243.17.

Model selected by backwards AIC:

$y$: Diagnosis of Heart Disease

$x_1$: Sex

$x_2$: Chest pain

$x_3$: Resting blood pressure

$x_4$: Cholesterol level

$x_5$: Maximum heart rate achieved

$x_6$: Exercise-induced angina

$x_7$: ST depressions induced by exercise relative to rest

$x_8$: Slope of the peak exercise ST segment

$x_9$: Number of vessels colored by fluoroscopy

The latter yielded a model with only 6 regressors which are shown below. The model started with a BIC of 295.64 and ended with a BIC of 274.21

Model selected by backwards BIC:

$y$: Diagnosis of Heart Disease

$x_1$: Sex

$x_2$: Chest pain

$x_3$: Maximum heart rate achieved

$x_4$: Exercise-induced angina

$x_5$: ST depressions induced by exercise relative to rest

$x_6$: Number of vessels colored by fluoroscopy

This discrepancy is expected, as the BIC formula places a higher penalty for model complexity. We ran a Variance Inflation Factor (VIF) test to see if any of the variables in the two models are correlated. The results are displayed below:

Table 2: Variance Inflation Factor Results for the Backwards AIC Model

| **Sex** | **Chest Pain** | **Resting BP** | **Cholesterol** | **Max HR** |
|---|---|---|---|---|
| 1.305 | 1.2246 | 1.0664 | 1.1792 | 1.2185 |

| **Angina** | **ST Depressions** | **ST Slope** | **Fluoroscopy** |
|---|---|---|---|
| 1.1433 | 1.411 | 1.4866 | 1.0848 |

Table 3: Variance Inflation Factor Results for the Backwards BIC Model

| **Sex** | **Chest Pain** | **Max HR** | **Angina** | **ST Depressions** | **Fluoroscopy** |
|---|---|---|---|---|---|
| 1.086 | 1.1625 | 1.1221 | 1.0931 | 1.1192 | 1.0177 |

In general, the AIC model had higher VIF's compared to the BIC model. This result persuaded us to choose the BIC model since it has less correlation between regressors.

Looking more specifically at the final logistic regression model we chose, the following table summarises the estimates of the coefficients corresponding to each regressor as well as the intercept, the standard errors of the coefficients, the confidence intervals for each coefficient, each coefficient's p-value, and whether or not each coefficient is significant. If the coefficient is significant, the level at which it is significant is specified.

Table 4: Regressor Coefficient Values and Statistics

| **Regressor** | **Estimate** | **Std. Error** | **95% CI** | **P-value** | **Significance** |
|---|---|---|---|---|---|
| Intercept | -1.2988 | 1.368 | (-3.9801, 1.3825) | 0.3424 | No |
| Sex | -1.5317 | 0.3904 | (-2.2969, -0.7665) | 8.74e-05 | 0.001 |
| Chest Pain | 0.8059 | 0.1718 | (0.4691, 1.1427) | 2.73e-06 | 0.001 |
| Max HR | 0.0225 | 0.0086 | (0.0056, 0.0394) | 0.0092 | 0.01 |
| Angina | -1.1026 | 0.3779 | (-1.8432, -0.3619) | 0.0035 | 0.01 |
| ST Depr. | -0.7587 | 0.1762 | (-1.104, -0.4134) | 1.66e-05 | 0.001 |
| Fluoroscopy | -0.7174 | 0.1680 | (-1.0468, -0.3881) | 1.96e-05 | 0.001 |

Consequently, we compute the odds ratio for our regression. Our fitted model looks as follows

$$\hat{\theta}(x) = \frac{1}{1 + e^{-(\hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3 + \hat{\beta}_4 x_4 + \hat{\beta}_5 x_5 + \hat{\beta}_6 x_6)}}$$

$$\hat{\theta}(x) = \frac{1}{1 + e^{-(-1.2988 - 1.5317 x_1 + 0.8059 x_2 + 0.0225 x_3 - 1.1026 x_4 - 0.7587 x_5 - 0.7174 x_6)}}$$

As a result, the odds ratio is

$$\left( \frac{\hat{\theta}(x)}{1 - \hat{\theta}(x)} \right) = e^{-(-1.2988 - 1.5317 x_1 + 0.8059 x_2 + 0.0225 x_3 - 1.1026 x_4 - 0.7587 x_5 - 0.7174 x_6)}$$

When the odds ratio is greater than one, that means the person has a higher chance of being diagnosed with heart disease than not being diagnosed.

To help with the interpretation of our $\hat{\theta}(x)$ equation, let us look at a few individuals and their probability of being diagnosed with heart disease based on their characteristics.

Table 5: Characteristics of Indivudals

| **Regressor** | **Person 1** | **Person 2** | **Person 3** | **Person 4** |
|---|---|---|---|---|
| Sex | 1 | 0 | 1 | 0 |
| Chest Pain | 0 | 0 | 0 | 0 |
| Max HR | 150 | 150 | 160 | 150 |
| Angina | 0 | 0 | 0 | 0 |
| ST Depr. | 1 | 1 | 1 | 1 |
| Fluoroscopy | 0 | 0 | 0 | 1 |

We can see that Person 1 is a male with no chest pain, a maximum heart rate of 150, no exercise-induced angina, has an ST depression value of 1, and has no vessels colored by fluoroscopy. Person 2 has exactly the same characteristics with respect to these regressors except Person 2 is a female. Person 3 is the same as Person 1 except their maximum heart rate achieved is ten more. Person 4 is the same as Person 2 except they have 1 more vessel colored by fluroscopy.

Plugging in these values into the equation for $\hat{\theta}(x)$, we get the following probabilities for each person to be diagnosed with heart disease.

Table 6: Probabilities of Heart Disease Diagnosis

| | **Person 1** | **Person 2** | **Person 3** | **Person 4** |
|---|---|---|---|---|
| **Probability** | 0.44665 | 0.78876 | 0.50270 | 0.64568 |

As a result, we can see that between Person 1 and Person 2, compared to males, females' chances of getting heart disease if they have the same characteristics regarding chest pain, maximum heart rate achieved, angina, ST depressions, and fluoroscopy, increase by 0.34211. Between Person 1 and Person 3, the maximum heart rate achieved increased by 10. Consequently, the probability of being diagnosed with heart disease increased by 0.05605. Between person 2 and person 3, the number of vessels colored by fluoroscopy increased by one which caused the probability of being diagnosed with heart disease to decrease by 0.14308 which appears to be the opposite of what we would expect.

# Summary

Our study of heart disease data points to sex, chest pain, maximum heart rate achieved, exercise-induced angina, ST depressions induced by exercise, and number of vessels colored by fluorosocpy as the leading indicators of heart disease. With this information, doctors and patients can better understand if they or a patient are at risk and can take proactive measures. Additionally, we learned that on average heart disease strikes between ages 51-54 and strikes for individuals with resting blood pressures

of 126.8-131.8 mm HG. Based on a hypothesis test, we know that on average, men and women with heart disease have similar blood pressures, so our confidence interval applies, for the most part, to men and women. Additionally, a second hypothesis test found that gender needs to be taken into account when examining fluoroscopic results. Going forward, future research can be done to investigate the effects of certain lifestyle choices (smoking, alcohol consumption, lack of exercise, poor diet) on the likelihood of developing heart disease. Similarly, our data set did not encompass all risk factors (such as diabetes, family history, and obesity), so future studies containing more variables would be beneficial. In regards to our final regression model, we could work on verifying that the logistic regression met all of its assumptions and was, indeed, a good fit. As we can see from our application to the four individuals, our logistic regression model responded the opposite of what we anticipated would happen when the number of vessels colored by fluoroscopy increased, so this, specifically, could also be explored more as the model is improved in future work.

# Appendix

Link to Code and Data Set: https://drive.google.com/drive/folders/1MPBZazY1VI$_H Rd - Eb_{TX} b0bgs42Wz7W?usp = sharing$

# References


From this source, we used facts about heart disease.
1. Felman, Adam. "Everything You Need to Know about Heart Disease." Medical News Today, Healthline Media UK Ltd, 29 Sept. 2020, www.medicalnewstoday.com/articles/237191.

From this source, we used facts about heart disease.
2. "Heart Disease Facts." Centers for Disease Control and Prevention, Centers for Disease Control and Prevention, 8 Sept. 2020, www.cdc.gov/heartdisease/facts.htm.

From this source, we used a fact about coronavirus.
3. "Provisional Death Counts for Coronavirus Disease 2019 (COVID-19)." Centers for Disease Control and Prevention, Centers for Disease Control and Prevention, 19 Nov. 2020, www.cdc.gov/nchs/nvss/vsrr/covid19/index.htm.

Our data file comes from this source.
4. VolodymyrGavrysh. "HeartDisease." Kaggle, 17 Jan. 2020, www.kaggle.com/volodymyrgavrysh/heart-disease.
Authors that contributed to the data set include Andras Janosi, William Steinbrunn, Matthias Pfisterer, and Robert Detrano.


# Self-Reflection

For this project, I put in an extremely large amount of time. In the process of completing the project, one problem Jackson and I ran into was that we first tried to do a hypothesis test comparing males and females where we included both a male and female column in the design matrix. As a result, we could not invert the matrix X'X. After

speaking to the TA, we learned that X'X was non-invertible because of the high correlation between the male and female indicator columns. Another problem we ran into was the logistic regression model. At first, we thought that there was a severe problem with the model since the residual plot showed a pattern and the QQplot showed non-Normal behavior. However, after speaking with the TA, I was reminded that logistic regression does not have the same assumptions that are usually used for linear regression such as normality. Something else I ran into that was a problem was the data visualization graph for age and heart disease diagnosis. I encountered a lot of trouble trying to display the desired legend. However, after consulting some past code I had written, I eventually figured out how to format the legend in the way I wanted.

Something I would do differently next time is have better time management. Although I worked on the project early overall to meet the previous deadlines, they were often bursts of work time where right after a deadline, I did not work on the project for a few days. Some advice I would give to future students is to start early and work in even time periods throughout the week. The project analysis and visualization took longer than I expected, so a good idea is to not wait until the last minute, especially since the project is towards the end of classes which, for some, may mean they have more time, but for others, means that they will be under a lot of pressure to finish on time.