Daniel Belkin, Catherine Collier, Ethan Emile, Lauren Joyce

STAT 355 Final Report

1. Problem Description:

Cardiovascular (heart) disease is currently the global leading cause of death. Heart disease is an all-encompassing term given to a variety of conditions which affect the structure and function of the heart and surrounding blood vessels. This project looks at the risk-factors commonly associated with the development of heart disease. Analyzing the risk factors can determine what risk factors more significantly influence the likelihood of developing heart disease. This is important because if the risk factors that most commonly lead to heart disease are known, then we are better able to limit those specific factors. Primary prevention of heart disease is limiting the high risk factors for people that are likely to develop the disease. Overall, limiting risk factors will decrease the number of people with heart disease.

2A. A summary of variables and their distribution:

Target - Whether the disease is present or not

chol - Cholesterol level : Slight negative skew to distribution (Mean 246.3 ,Median 240.0, range 125-564)

age - Age of subject : distributed normally (Mean age 54.37 , range 29-77)

fbs - Whether blood sugar is above 120 mg/dl(258 subjects with normal sugar level, 45 subjects with high level)

sex - Patient sex : (207 male subjects, 96 females)

restecg - Electrocardiogram result at rest (4 subjects with abnormal results, 152 normal, 147 with left ventricular hypertrophy)

trestbps - Resting blood pressure : Normally distributed (Mean 131.6, range 94-200)

exang - Whether the subject had angina during exercise : mostly no angina during exercise 204, compared with 99 yes angina during exercise

slope - The slope of the ST segment during the most demanding part of the exercise : descending: 21, flat :140 ,ascending :142

oldpeak - A decrease of ST-segment during exercise (negative skew, Mean 1.04 , Median 0.80, range 0- 6.2)

thal - Results of the blood flow observed via the radioactive dye (2 null, 18 fixed defect, 166 normal, and 117 reversible defect)

ca - Number of main blood vessels coloured by the radioactive dye due to narrow structure (175 with 0 narrow blood vessels, 65 with 1 narrow blood vessel, 20 subjects with at least 2 narrow blood vessels, 5 null)

cp - Chest pain (143 no pain, 45 no pain- atypical angina, 4 pain but no angina,23 typical anginal)

thalach - Maximum heart rate during a stress test. Positive skew to distribution (Mean 149.6 , Median 153.0, range 71-202)

2B. **Exploration Data Analysis of Variables:**

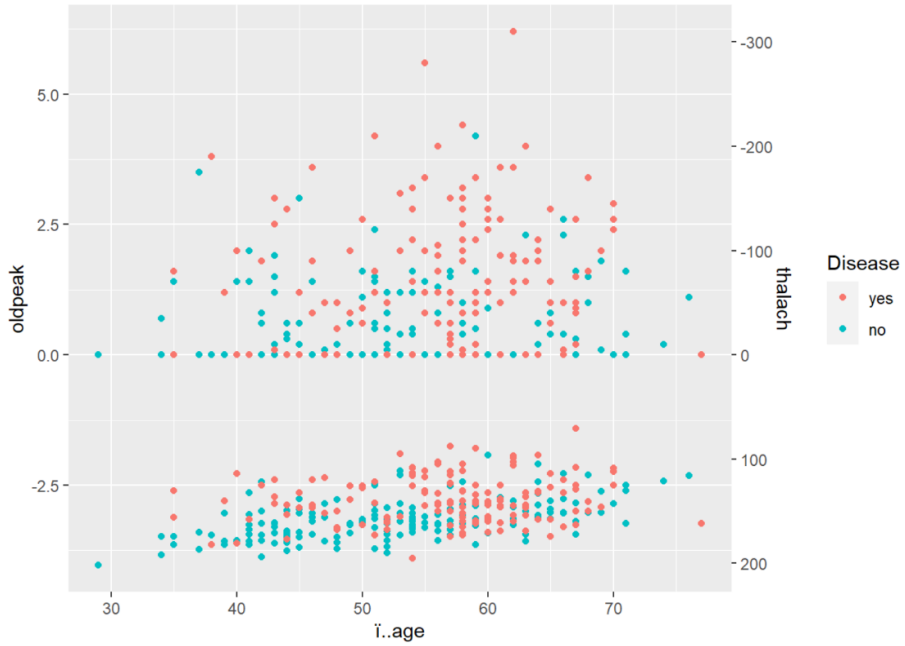


Figure 1

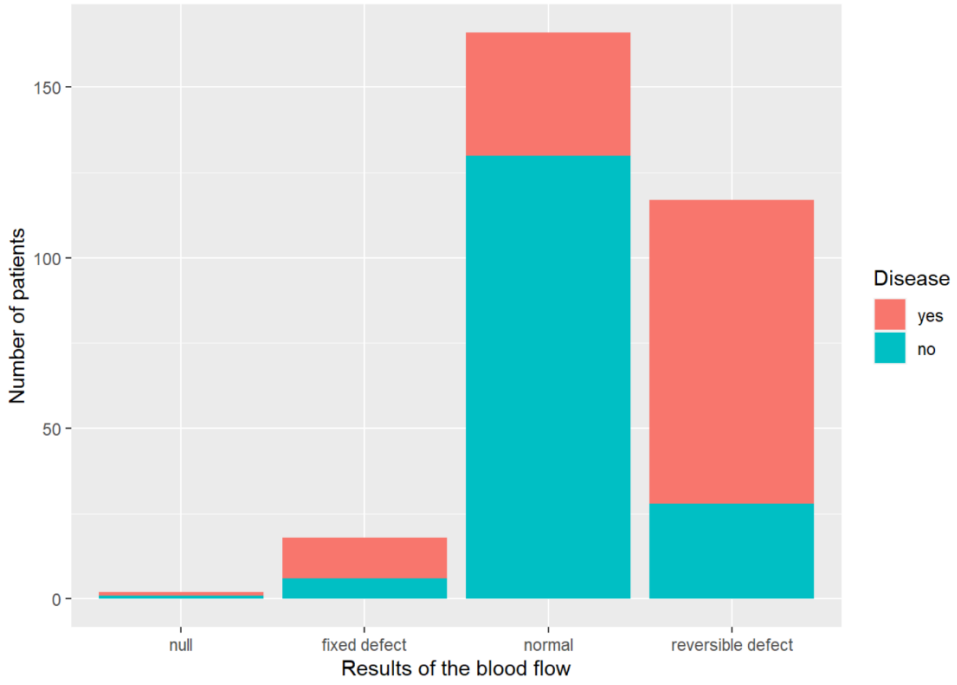
Scatterplots of Relationship of "thalch", "oldpeak", and "age" With Disease Outcome

As seen Figure 1, two plots are illustrated. The top plot illustrates the relationship between age and oldpeak (left y-axis) while also illustrating if heart disease is present or not. In the top plot, there is a trend for more subjects to have heart disease with an increase in oldpeak, and fewer subjects having heart disease with a decrease in old peak.

In the bottom plot of the figure above, the plot illustrates the relationship between age, thalach (right y-axis) and while also illustrating if heart disease is present or not. The plot shows that as thalach level drops from 200 to 100 more of the data points appear as having the disease. Both plots show that as age increase so does presence of disease in the subjects,

Figure 2

Disease Presence With Relation to Condition of Blood Flow (Thal)



As seen in the figure above there are clearly more subjects with heart disease both in quantity and proportion in the reversible defect group than the normal group (and in proportion between fixed defect and normal although sample is small for fixed defect).

3. **Logistic Regression Model Analysis:**

At first our objective was to try and determine a single categorical variable which had the highest statistical significance for determining presence of disease. For this we used the Chi-squared test . We found that the best single statistically significant association between target and categorical variables is for thal. with a p-value of 2.2e-16 which is statistically significant at the 99.9 confidence interval, and the correlation has a strength of 0.54 (using Cramer's V method).

We used this single variable to build our first logistic regression model. p(y) = 3.2016-1.2916x , Where p(y) is the probability of NOT developing heart disease, 3.2016 is the y-intercept, and -12916 is the slope. It's important to note that the higher the probability the more likely the person is NOT to have heart disease as the target is coded as 1= no disease, and 0 = disease. This is reflected in figure 3 as the probability of NOT developing disease drops as we

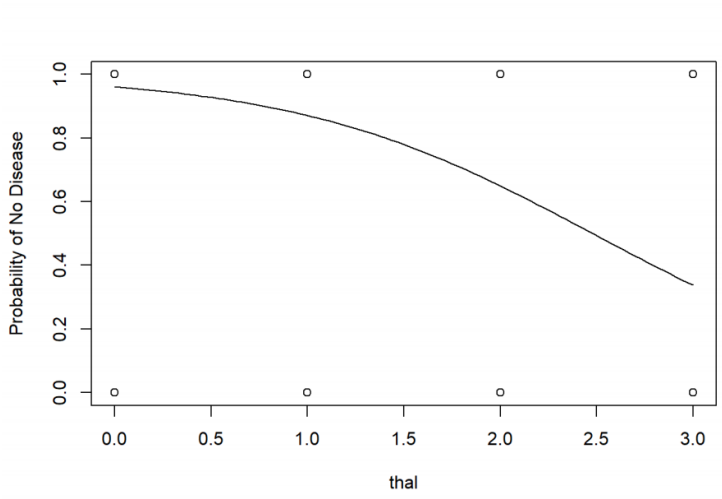
move towards having a non fixed reversible defect. Both intercept and slope were significant at the 99.9% confidence interval.

We then wanted to build a logistic model around the single significant variable of Thal that will include other continuous and categorical variables and which would account for any interactions with Thal. using multiple logistic regression we settled on the following model:

p(y)= 3.95429-1.354833(thal)-2.371512(ca)+0.665827(slope)-0.910823(exang)+0.939324(cp)-1.621221(sex)-0.588705(oldpeak)+0.023139(thalach)-0.021921(trestbps)+0.698071(that)(ca)

All slope coefficients, intercept, and interaction were at least statistically significant at the 95% confidence interval

Figure 3



Probability of NO *Disease in Relation to Condition of Blood Flow (Thal)*

Table 1

Multiple Logistic Regression Mode

```
##
## Call:
## glm(formula = target ~ thal + ca + slope + exang + cp + sex +
##     oldpeak + thalach + trestbps + thal * ca, family = binomial,
##     data = Heart)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.6293  -0.4109   0.1444   0.5500   2.5856
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  3.395429   1.998530   1.699 0.089326 .
## thal        -1.354833   0.359013  -3.774 0.000161 ***
## ca          -2.371512   0.802078  -2.957 0.003109 **
## slope        0.665827   0.345518   1.927 0.053975 .
## exang       -0.910823   0.410931  -2.216 0.026658 *
## cp           0.939324   0.190802   4.923 8.52e-07 ***
## sex         -1.621221   0.446693  -3.629 0.000284 ***
## oldpeak     -0.588705   0.211950  -2.778 0.005477 **
## thalach      0.023139   0.009659   2.396 0.016588 *
## trestbps    -0.021921   0.010132  -2.164 0.030499 *
## thal:ca      0.698071   0.326046   2.141 0.032272 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

4. **Logistic Regression Model Interpretation**

Once the final model has been settled , the only variables left out of the model were : Cholesterol, Fbs, restecg, and age. This was surprising as we expected age to play a large role in our model based of our initial data exploration, this goes to show that perhaps age is just a number (although many of the factor that play a part in our model could worsen as age advance due to deterioration and a more sophisticated model might be able to better show this interaction).

The significant variables, interaction, and intercept all had at least p-values of less than 0.05 (with some variables even more significant at the 99% confidence interval), indicating a strong relationship with disease presence. The only variable to end up significantly interacting with Thal, the variable which we centered this model around, was CA with a coefficient of 0.698071with the interaction being statistically significant at the 99% confidence interval. This interaction could be due to the reason that the presence of narrow or abnormal blood vessels could contribute to blood flow defects or vice versa.