# PREDICTING HEART ATTACK RISK

## Using Medical and IoT Data

Charan Sankaran

# Table of Contents

## Acknowledgements

First and foremost, I would like to thank my parents for their endless love, support, and encouragement. Secondly, I would like to thank William Ellis, my project advisor, for guiding me throughout the entirety of this project. I would also like to thank Siobhan Curran, my STEM teacher, for constantly motivating me and providing me with helpful suggestions for my project. Finally, I would like to thank Ramviknesh Ramanathan and Aditya Hoque for engaging in insightful conversations with me about statistics and programming.

## Abstract

Although heart attacks have a high prevalence in the United States, many people are not consistently monitored for heart attack risk. Several studies have demonstrated tremendous potential in heart attack risk prediction. However, these studies have not incorporated both medical and IoT device data in their predictive models. The goal of this project was to engineer an algorithm that used medical and IoT device data to predict the risk a person has of experiencing a heart attack. Prior to the construction of this algorithm, the major risk factors for heart attacks were identified. To form the algorithm, many versions of multivariate regression analysis were used such as linear regression and polynomial regression. R, a statistical software, was used to calculate the coefficients in the algorithm, which yielded a final equation. Each version of the algorithm was tested for accuracy and efficiency. The final algorithms chosen had RMSE values of 2.45 for male accuracy and 0.97 for female accuracy. Overall, the implementation of this algorithm as a monitoring tool for patients could decrease the number of deaths due to heart attacks.

## Introduction

Heart attacks are the reason for many deaths worldwide, we must do something to minimize the risk of death from heart attacks. One of the major reasons heart attacks become fatal are because the symptoms and risk factors of heart attacks are not immediately identified and tested for. To decrease the possibility of a heart attack, people can deal with using predictive analytics on medical and IoT data to predict if a person will have a heart attack based on their health trends. This technology will reduce both the chance of a person having a heart attack as well as the response time to a person having a heart attack if they use IoT devices. The data from the IoT device will be constantly sent to a computer that will run the algorithm on the data. The doctor will be able to have full access of the data and the results of the algorithm.

# Literature Review

## What is a Heart Attack?

A heart attack occurs when the blood flow to and from the heart is obstructed. This may be caused by various different factors. Heart attacks are common among the elderly and are not prevalent among younger people. Although it is very unlikely, it is still possible for young children to experience a heart attack. Heart attacks can be deadly and are quite prevalent in the world. In fact, heart attacks and heart related arrhythmias are ranked first in the leading cause of death in the United States (American Heart Association).

## Causes of a Heart Attack

As stated earlier, a heart attack occurs when the blood flow to and from the heart is obstructed. This is typically caused by three different factors: (good place to list them, then you expand upon them below). The first factor is the buildup of material inside the coronary arteries (Mayo Clinic). The buildup is usually due to an excess of fat and cholesterol, which builds up and forms plaque (Mayo Clinic). The plaque clots the arteries obstructing the flow of blood in the cardiovascular system (Mayo Clinic).

There is also a second cause of blood flow stoppage. The stoppage is due to spasms that cause the blood to stop flowing to and from the heart (Mayo Clinic). Illicit drugs such as cocaine typically cause these spasms (Mayo Clinic).

Lastly heart attacks can be caused by tears in the artery, which prevent blood from traveling to and from the heart (Mayo Clinic).

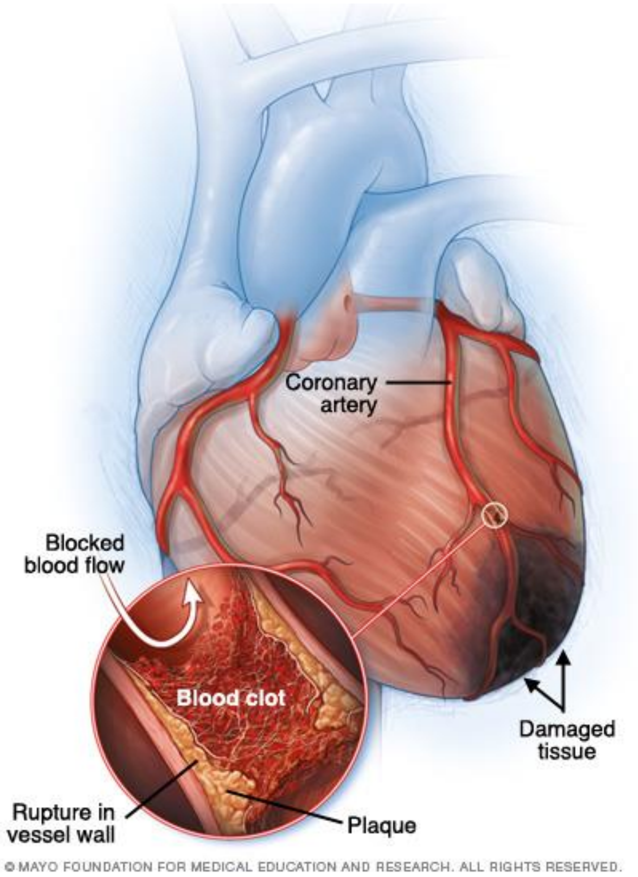


**Figure 1.** This image shows the heart along with the coronary arteries, these are the arteries, which typically cause heart attacks because of blockage and other factors described above. The close-up shows a blood clot in one of the arteries and shows how the plaque blocks the blood flow. (Mayo Clinic)

## Risk Factors and Symptoms of Heart attacks

There are countless risk factors and symptoms for heart attacks. Below some of the major risk factors have been listed along with small excerpts about each risk factor.

### Age

The older people get the more likely it is for them to experience a heart attack. Men over the age of 45 and woman over the age of 55 have a much greater risk of a heart attack than others. (Mayo Clinic)

## Tobacco

Smoking tobacco, or even enduring second hand smoking can lead to an increased chance of a heart attack. (Mayo Clinic)

## High blood pressure

Over time, high blood pressure can damage arteries that feed your heart by accelerating atherosclerosis. High blood pressure that occurs with obesity, smoking, high cholesterol or diabetes increases your risk for heart attacks even more.

## High blood cholesterol or triglyceride levels

The more bad-cholesterol in the body the more a person has the risk of experiencing a heart attack (Mayo Clinic). On the other hand, “a high level of high-density lipoprotein (HDL) cholesterol (the "good" cholesterol) lowers your risk of heart attack (Mayo Clinic).” Triglyceride levels also increase the chance of a heart attack.

## Diabetes

Insulin, a hormone secreted by your pancreas, allows your body to use glucose, a form of sugar. Having diabetes causes your body's blood sugar levels to rise, which can increase the chance of heart attack. (Mayo Clinic)

## Family history of heart attack

If any of your direct family or any relatives of yours has had a heart attack, you have an increased risk for having one as well. (Mayo Clinic)

## Lack of physical activity

If a person is very inactive they have an increased risk for heart attack. On the other hand, the increase of aerobic exercises can greatly decrease the chance of having a heart attack. (Mayo Clinic)

### Obesity

Being overweight contributes to increased blood cholesterol, triglyceride levels, and blood pressure, which all contribute to an increased likelihood of experiencing a heart attack. (Mayo Clinic)

### Stress

Many people have responses to stress that can increase their risk of a heart attack. (Mayo Clinic)

### Illegal drug use

As discussed before, the use of illicit drugs such as cocaine can lead to spasms that can stop the flow of blood to and from the heart. (Mayo Clinic)

### A History of preeclampsia

Preeclampsia, which can last for your lifetime, is a disease that produced high blood pressure during pregnancy. (Mayo Clinic)

### A History of an autoimmune condition

Autoimmune conditions can also increase your risk of having a heart attack. (Mayo Clinic)

## Data Analysis

### Multivariate Regression Analysis

The analysis of data is essential for further understanding the implications of the data associated with heart attacks. Multivariate regression analysis is used for analyzing statistical data with multiple variables. The process of multivariate regression analysis begins with developing regression patterns between each independent variable and the single dependent variable (Leitmeyer). The process of creating a single regression line is referred to as single variable regression analysis. Single variable regression analysis is the way of extracting the most optimal line of fit from the data given (Leitmeyer). To calculate which line of fit is most optimal we can use different methods. One simple method for finding the line of best fit is to find the amount of error a certain line of fit has using the RMSE value

(Origin Labs). Taking every value of error in the graph, squaring them, taking the mean of those values and finally square-rooting the mean will produce this value (Origin Labs). The line with the lowest RMSE value would therefore be the most optimal line of fit. The RMSE value is a very basic method of checking average error. There are other more complicated ways such as finding residual value and estimating the standard deviation of the residual (Leitmeyer). These methods must be used a second time for multivariate regression as explained below.

When analyzing a multivariate regression, finding the correlation or best-fit line for each dependent variable to the independent variable is done as described, but combining these values is where the use of the average error value comes in a second time. This value can tell us how much of a weight we should put toward each variable to show the probability for that independent variable (Leitmeyer). This is the overall process of multivariate regression, but there are many different ways to calculate average error and weightage of each variable beyond what is said above.

## Decision Trees

Decision trees are a unique way to analyze data. When analyzing a set of data, there are almost always nominal variables, which are variables with a binary answer (Decision Tree). Other data in which only a limited number of outcomes are possible are also great types of data to use for decision tree analysis (Decision Tree). When creating a decision tree a person will be creating many different nodes and branches from one major aspect that they are analyzing (Decision Tree). For example, when looking into the equation for BMI of a person, there are typically two equations based off of sex. The grouping between sexes was created through a two-node decision tree. Although this is a very simple example, when there are multiple nominal variables there will be many different outcomes. This way, the equation will not have to account for the nominal variables.

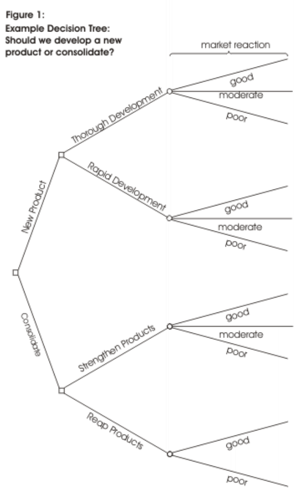
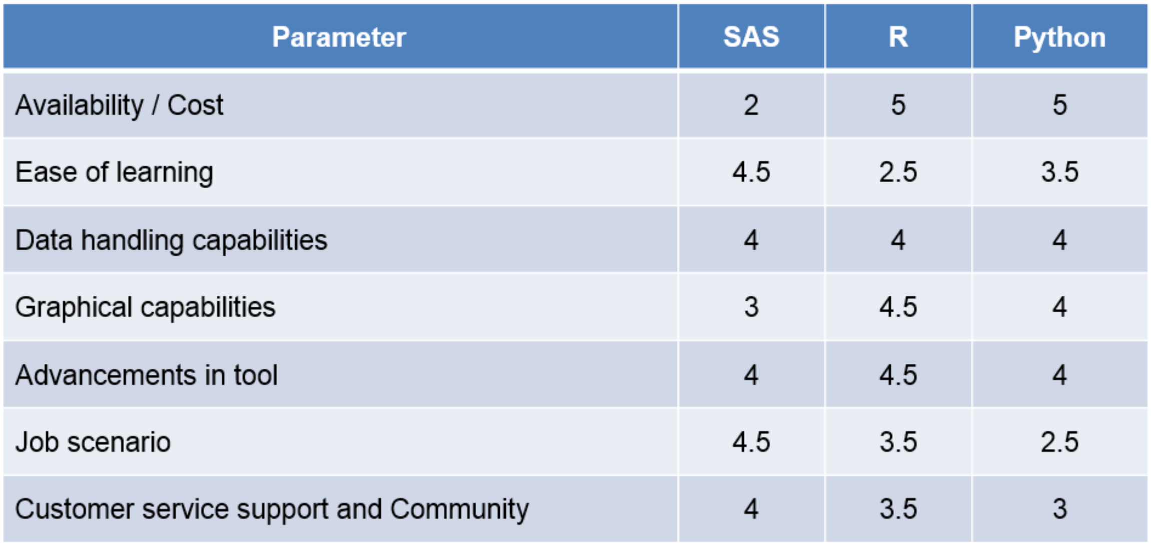


**Figure 2.** The above figure is a simple decision tree unrelated to the project topic, but provided for a visual understanding of decision trees. (Wrote)

## Programming Language

The programming languages R, SAS, and Python were all available for statistical programming. Each program has its unique advantages. Overall, R seems to encompass the most beneficial aspects of statistical programming needs as shown in the table below (Jain).

**Table 1.** This figure shows the comparison between three statistical programming languages available for statistical analysis. (Jain)



| Parameter | SAS | R | Python |
|---|---|---|---|
| Availability / Cost | 2 | 5 | 5 |
| Ease of learning | 4.5 | 2.5 | 3.5 |
| Data handling capabilities | 4 | 4 | 4 |
| Graphical capabilities | 3 | 4.5 | 4 |
| Advancements in tool | 4 | 4.5 | 4 |
| Job scenario | 4.5 | 3.5 | 2.5 |
| Customer service support and Community | 4 | 3.5 | 3 |

# Engineering Plan

## Engineering Problem Statement

There is a pressing clinical need to identify heart attacks as quickly and efficiently as possible. Heart attacks have a high prevalence in the United States and many people are not consistently monitored for heart attack risk.

## Engineering Goal

The goal of this project was to engineer an algorithm that utilizes medical data and IoT device data to accurately and efficiently predict the risk a person has of experiencing a heart attack.

## Procedure

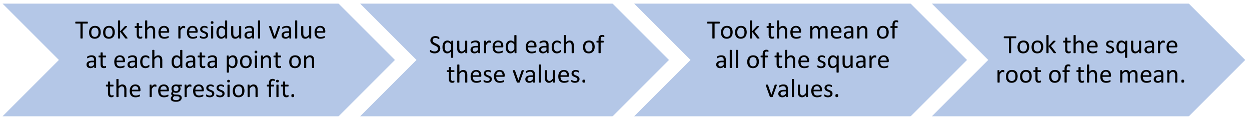
Identify the leading factors that contribute towards heart arrhythmias and heart attacks. Gather patient data with history of heart disease with the details of the contributing factors listed. Build regression models and generate correlation coefficients of these factors using these patient data. Transfer data into R and build predictive algorithms using R program. Explore existing algorithms and data analysis methods. Use multivariate regression and data analytic techniques to design and create multiple algorithms for predicting heart attacks. Test the algorithms for both accuracy and efficiency to determine the best algorithm.

## Criteria

The criteria that the algorithm was tested for were accuracy and efficiency.

## Methodology

The leading risk factors and symptoms that contribute toward heart attacks and heart arrhythmias were identified through research. Existing algorithms and data analysis methods were researched and explored. Data was then obtained from the heart disease data set in the UCI machine learning repository. R was the programming language used. Data was imported as a csv file and was then transferred into R to begin building and creating a predictive algorithm. The csv file was also transferred into excel, for testing and exploring how regression models work. Regression models for each independent variable were then built using different types of regression fitting in R. These models include linear multivariate regression fitting and polynomial multivariate regression fitting. The correlation coefficients of each risk factor were generated through the use of existing R libraries. The correlation coefficients were used to create a final equation for the algorithm (Different equations for each type of regression). The error value or RMSE value was taken for each regression equation (Figure 1). The average run time of each algorithm was also obtained. Minor improvements and debugging of each algorithm was then done. Each algorithm was tested multiple times. Compared both accuracy and efficiency to determine which algorithm worked best. User interface and interactive software was finally added for ease of use.



Took the residual value at each data point on the regression fit.

Squared each of these values.

Took the mean of all of the square values.

Took the square root of the mean.

**Figure 3.** This figure describes the method for finding the RMSE value, which is one method for obtaining the overall error of a regression fit.

## Results

Each variation of the algorithm was tested for accuracy and efficiency. The most accurate algorithm for males had an RMSE value of 2.45, but an average run time of 0.01258 seconds. The most accurate algorithm for females had an RMSE value of 0.97, but an average run time of 0.01030 seconds. The most efficient algorithm for males had an average run time of 0.00631 seconds, but an RMSE value of 3.08. The most efficient algorithm for females had an average run time of 0.00625 seconds, but an RMSE value of 2.29. Although these algorithms had very quick run times, their accuracy shows that increasing the exponents of certain risk factors will significantly increase their accuracy while not decreasing efficiency a large amount. The final algorithm chosen was actually the algorithm that had the highest accuracy. The reason this algorithm was chosen over others was because the equations coefficients showed that all values are significant. As shown in the decision matrices below the best algorithm was chosen by ranking them from best to worse in accuracy and efficiency and weighting the accuracy eight times as much as the efficiency.

Note: For results on every regression tested refer to logbook

| Criteria | Scale | Version | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 |
| Accuracy | 32 | 2 | 4 | 6 | 8 | 18 | 22 | 20 | 24 | 10 | 14 | 12 | 16 | 26 | 30 | 28 | 32 |
| Efficiency | 4 | 3.75 | 3.00 | 3.25 | 1.50 | 1.75 | 3.50 | 2.00 | 0.25 | 4.00 | 1.25 | 2.75 | 1.00 | 0.50 | 2.50 | 2.25 | 0.75 |
| Total | 36 | 5.75 | 7.00 | 9.25 | 9.50 | 19.75 | 25.50 | 22.00 | 24.25 | 14.00 | 15.25 | 14.75 | 17.00 | 26.50 | 32.50 | 30.25 | 32.75 |

**Figure 4.** Final Engineering Matrix for all male versions of the algorithm. The matrix was done based off of rank of RMSE and average run time.

| Criteria | Scale | Version | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 17 | 18 | 19 | 20 | 21 | 22 | 23 | 24 | 25 | 26 | 27 | 28 | 29 | 30 | 31 | 32 |
| Accuracy | 32 | 2 | 4 | 10 | 14 | 4 | 8 | 12 | 16 | 18 | 26 | 22 | 28 | 20 | 30 | 24 | 32 |
| Efficiency | 4 | 3.00 | 3.25 | 2.00 | 0.75 | 4.00 | 1.75 | 1.50 | 0.50 | 2.50 | 3.50 | 3.75 | 0.25 | 2.75 | 2.25 | 1.25 | 1.00 |
| Total | 36 | 5.00 | 7.25 | 12.00 | 14.75 | 8.00 | 9.75 | 13.50 | 16.50 | 20.50 | 29.50 | 25.75 | 28.25 | 22.75 | 32.25 | 25.25 | 33.00 |

**Figure 5.** Final Engineering Matrix for all female versions of the algorithm. The matrix was done based off of rank of RMSE and average run time.

## Analysis

The accuracy and efficiency of each regression was tested. Accuracy was tested through RMSE value and efficiency was tested through run time. The coefficients of the equation were also tested for significance. Significance was determined by checking if the coefficient was less than $10^{-6}$. If a coefficient was determined not to be significant then the regression with the insignificant term removed would be used instead. The accuracy and efficiency were both taken to find the most optimal regression. The accuracy was most important so 1/RMSE was used and was weighted on a scale of zero to thirty-two to represent accuracy. The efficiency was not as important so the run time was only weighted on a scale of zero to four to represent efficiency. Lastly, each regression was checked for visual correlation. If the graph did not correlate with the trend of each singular risk factors versus heart attack risk, then the regression was not considered.

There are a few sources of error which could skew the data. The run time varies depending on each run and the tasks being run in the background. To lower the risk all background programs and tasks were closed during run time and the program ran five times to find a better average run time. This could still be an issue because the CPU could have general tasks constantly running in the background. The accuracy could be skewed due to the size of the data. With only 300 data point, certain regressions could yield higher or lower accuracies than they would if large amounts of data were used. Overall the sources of error are very minimal, and error would be largely decreased in the use of this concept in a real world scenario.

## Conclusion

The final algorithm chosen had a high accuracy as well as a high efficiency showing that the algorithm would be useable in the real world. The application of the techniques used in this project on larger databases and with more advanced processing computers would exponentially increase the real world usability of this algorithm. The current algorithm could still be used in the real world as a consistent heart attack risk monitoring tool, but the effects of specifically increasing the amount of data used to create the algorithm would largely increase the accuracy of the algorithm. Advancements could also be made by adding data like PQRST complex that is available in some wearable chest devices. The implementation of the algorithm as a monitoring tool for patients could not only decrease the number of deaths due to heart attacks, but could decrease the amount of heart attacks that occur.

## References

American heart association - building healthier lives, free of cardiovascular diseases and stroke. Retrieved December 8, 2016, from http://www.heart.org/HEARTORG/

CDC. (2015, August 5). Heart attack. Retrieved December 8, 2016, from http://www.cdc.gov/heartdisease/heart_attack.htm

Copyright Office. Creative commons. . 2016. http://guides.lib.umich.edu/creativecommons

Corporation, O. Algorithm (multiple linear regression). Retrieved December 8, 2016, from http://www.originlab.com/doc/Origin-Help/Multi-Regression-Algorithm

Decision tree: Introduction. Vol 1. ; 2009:323-328.

Jain, K., Shaikh, F., Kaushik, S., & Kashyap, S. (2016, December 7). SAS vs. R (vs. Python) – which tool should I learn? Retrieved from https://www.analyticsvidhya.com/blog/2014/03/sas-vs-vs-python-tool-learn/

Lathauwer LD. A short introduction to tensor-based methods for factor analysis and blind source separation. ISPA. 2011:558-563. http://ieeexplore.ieee.org/document/6046668.

Leitmeyer, K., & Adlhoch, C. (2016). Review article. Epidemiology, 27(5), 743-751. doi:10.1097/EDE.0000000000000438

Staff, M. C. (2014). Heart attack definition. Mayoclinic. Retrieved from http://www.mayoclinic.org/diseases-conditions/heart-attack/basics/definition/con-20019520

Wrote, Y. Decision tree analysis: Choosing by projecting. Retrieved December 8, 2016, from https://www.mindtools.com/dectree.html

## Appendix A: Limitations and Assumptions

There is a pressing clinical need to identify heart attacks as quickly and efficiently as possible. The goal of this project is to engineer an algorithm that utilizes medical data and IoT data to accurately and efficiently predict the risk a person has of experiencing a heart attack. The project is limited to data from public databases with de-identified data. The project is also limited by the amount of data available in these database. The project assumes that the algorithm will be transferable from data in a database to the real world data and problems. The project also assumes that the data is representative of real, accurate, and precise data on heart attacks. Assumptions were made that the R software works as stated. The library and classes present in the R software are also assumed to work as defined.

## Appendix B: Literature Review Search Terms

The terms used to search for information on this project were: Heart Attacks, Heart Attack Risk Factors, Multivariate Regression, Predictive Analytics, IoT Data, Heart Attack Symptoms, and Modeling.

## Appendix C: Program Code

### Regression Formation Code:

```
mydatam = read.csv("rawdatamale.csv")
mydatam
mydataf = read.csv("rawdatafemale.csv")
mydataf

# Version 1
fit1 <- lm(Risk ~ Age + Race + BP + HDLChol + TChol + Diab + Smk + HTT, data=mydatam)
summary(fit1)

# Version 2
fit2 <- lm(Risk ~ Age + Race + BP + HDLChol + poly(TChol, 2, raw=TRUE) + Diab + Smk + HTT,
data=mydatam)
summary(fit2)

# Version 3
fit3 <- lm(Risk ~ Age + Race + BP + poly(HDLChol, 2, raw=TRUE) + TChol + Diab + Smk + HTT,
data=mydatam)
summary(fit3)

# Version 4
fit4 <- lm(Risk ~ Age + Race + BP + poly(HDLChol, 2, raw=TRUE) + poly(TChol, 2, raw=TRUE) +
Diab + Smk + HTT, data=mydatam)
summary(fit4)

# Version 5
fit5 <- lm(Risk ~ Age + Race + poly(BP, 2, raw=TRUE) + HDLChol + TChol + Diab + Smk + HTT,
data=mydatam)
summary(fit5)

# Version 6
fit6 <- lm(Risk ~ Age + Race + poly(BP, 2, raw=TRUE) + HDLChol + poly(TChol, 2, raw=TRUE) +
Diab + Smk + HTT, data=mydatam)
summary(fit6)

# Version 7
fit7 <- lm(Risk ~ Age + Race + poly(BP, 2, raw=TRUE) + poly(HDLChol, 2, raw=TRUE) + TChol +
Diab + Smk + HTT, data=mydatam)
summary(fit7)
```

```
# Version 8
fit8 <- lm(Risk ~ Age + Race + poly(BP, 2, raw=TRUE) + poly(HDLChol, 2, raw=TRUE) + poly(TChol,
2, raw=TRUE) + Diab + Smk + HTT, data=mydatam)
summary(fit8)

# Version 9
fit9 <- lm(Risk ~ poly(Age, 2, raw=TRUE) + Race + BP + HDLChol + TChol + Diab + Smk + HTT,
data=mydatam)
summary(fit9)

# Version 10
fit10 <- lm(Risk ~ poly(Age, 2, raw=TRUE) + Race + BP + HDLChol + poly(TChol, 2, raw=TRUE) +
Diab + Smk + HTT, data=mydatam)
summary(fit10)

# Version 11
fit11 <- lm(Risk ~ poly(Age, 2, raw=TRUE) + Race + BP + poly(HDLChol, 2, raw=TRUE) + TChol +
Diab + Smk + HTT, data=mydatam)
summary(fit11)

# Version 12
fit12 <- lm(Risk ~ poly(Age, 2, raw=TRUE) + Race + BP + poly(HDLChol, 2, raw=TRUE) +
poly(TChol, 2, raw=TRUE) + Diab + Smk + HTT, data=mydatam)
summary(fit12)

# Version 13
fit13 <- lm(Risk ~ poly(Age, 2, raw=TRUE) + Race + poly(BP, 2, raw=TRUE) + HDLChol + TChol +
Diab + Smk + HTT, data=mydatam)
summary(fit13)

# Version 14
fit14 <- lm(Risk ~ poly(Age, 2, raw=TRUE) + Race + poly(BP, 2, raw=TRUE) + HDLChol +
poly(TChol, 2, raw=TRUE) + Diab + Smk + HTT, data=mydatam)
summary(fit14)

# Version 15
fit15 <- lm(Risk ~ poly(Age, 2, raw=TRUE) + Race + poly(BP, 2, raw=TRUE) + poly(HDLChol, 2,
raw=TRUE) + TChol + Diab + Smk + HTT, data=mydatam)
summary(fit15)

# Version 16
fit16 <- lm(Risk ~ poly(Age, 2, raw=TRUE) + Race + poly(BP, 2, raw=TRUE) + poly(HDLChol, 2,
raw=TRUE) + poly(TChol, 2, raw=TRUE) + Diab + Smk + HTT, data=mydatam)
summary(fit16)
```

```
# Version 17
fit17 <- lm(Risk ~ Age + Race + BP + HDLChol + TChol + Diab + Smk + HTT, data=mydataf)
summary(fit17)

# Version 18
fit18 <- lm(Risk ~ Age + Race + BP + HDLChol + poly(TChol, 2, raw=TRUE) + Diab + Smk + HTT,
data=mydataf)
summary(fit18)

# Version 19
fit19 <- lm(Risk ~ Age + Race + BP + poly(HDLChol, 2, raw=TRUE) + TChol + Diab + Smk + HTT,
data=mydataf)
summary(fit19)

# Version 20
fit20 <- lm(Risk ~ Age + Race + BP + poly(HDLChol, 2, raw=TRUE) + poly(TChol, 2, raw=TRUE) +
Diab + Smk + HTT, data=mydataf)
summary(fit20)

# Version 21
fit21 <- lm(Risk ~ Age + Race + poly(BP, 2, raw=TRUE) + HDLChol + TChol + Diab + Smk + HTT,
data=mydataf)
summary(fit21)

# Version 22
fit22 <- lm(Risk ~ Age + Race + poly(BP, 2, raw=TRUE) + HDLChol + poly(TChol, 2, raw=TRUE) +
Diab + Smk + HTT, data=mydataf)
summary(fit22)

# Version 23
fit23 <- lm(Risk ~ Age + Race + poly(BP, 2, raw=TRUE) + poly(HDLChol, 2, raw=TRUE) + TChol +
Diab + Smk + HTT, data=mydataf)
summary(fit23)

# Version 24
fit24 <- lm(Risk ~ Age + Race + poly(BP, 2, raw=TRUE) + poly(HDLChol, 2, raw=TRUE) +
poly(TChol, 2, raw=TRUE) + Diab + Smk + HTT, data=mydataf)
summary(fit24)

# Version 25
fit25 <- lm(Risk ~ poly(Age, 2, raw=TRUE) + Race + BP + HDLChol + TChol + Diab + Smk + HTT,
data=mydataf)
summary(fit25)
```

```
# Version 26
fit26 <- lm(Risk ~ poly(Age, 2, raw=TRUE) + Race + BP + HDLChol + poly(TChol, 2, raw=TRUE) +
Diab + Smk + HTT, data=mydataf)
summary(fit26)

# Version 27
fit27 <- lm(Risk ~ poly(Age, 2, raw=TRUE) + Race + BP + poly(HDLChol, 2, raw=TRUE) + TChol +
Diab + Smk + HTT, data=mydataf)
summary(fit27)

# Version 28
fit28 <- lm(Risk ~ poly(Age, 2, raw=TRUE) + Race + BP + poly(HDLChol, 2, raw=TRUE) +
poly(TChol, 2, raw=TRUE) + Diab + Smk + HTT, data=mydataf)
summary(fit28)

# Version 29
fit29 <- lm(Risk ~ poly(Age, 2, raw=TRUE) + Race + poly(BP, 2, raw=TRUE) + HDLChol + TChol +
Diab + Smk + HTT, data=mydataf)
summary(fit29)

# Version 30
fit30 <- lm(Risk ~ poly(Age, 2, raw=TRUE) + Race + poly(BP, 2, raw=TRUE) + HDLChol +
poly(TChol, 2, raw=TRUE) + Diab + Smk + HTT, data=mydataf)
summary(fit30)

# Version 31
fit31 <- lm(Risk ~ poly(Age, 2, raw=TRUE) + Race + poly(BP, 2, raw=TRUE) + poly(HDLChol, 2,
raw=TRUE) + TChol + Diab + Smk + HTT, data=mydataf)
summary(fit31)

# Version 32
fit32 <- lm(Risk ~ poly(Age, 2, raw=TRUE) + Race + poly(BP, 2, raw=TRUE) + poly(HDLChol, 2,
raw=TRUE) + poly(TChol, 2, raw=TRUE) + Diab + Smk + HTT, data=mydataf)
summary(fit32)
```

## Efficiency Testing Code:

```
mydatam = read.csv("rawdatamale.csv")
mydatam
mydataf = read.csv("rawdatafemale.csv")
mydataf

# Version 1
start <- Sys.time()
for(i in 1:5){
fit1 <- lm(Risk ~ Age + Race + BP + HDLChol + TChol + Diab + Smk + HTT, data=mydatam)
}
(Sys.time()-start)/5

# Version 2
start <- Sys.time()
for(i in 1:5){
fit2 <- lm(Risk ~ Age + Race + BP + HDLChol + poly(TChol, 2, raw=TRUE) + Diab + Smk + HTT,
data=mydatam)
}
(Sys.time()-start)/5

# Version 3
start <- Sys.time()
for(i in 1:5){
fit3 <- lm(Risk ~ Age + Race + BP + poly(HDLChol, 2, raw=TRUE) + TChol + Diab + Smk + HTT,
data=mydatam)
}
(Sys.time()-start)/5

# Version 4
start <- Sys.time()
for(i in 1:5){
fit4 <- lm(Risk ~ Age + Race + BP + poly(HDLChol, 2, raw=TRUE) + poly(TChol, 2, raw=TRUE) +
Diab + Smk + HTT, data=mydatam)
}
(Sys.time()-start)/5

# Version 5
start <- Sys.time()
for(i in 1:5){
fit5 <- lm(Risk ~ Age + Race + poly(BP, 2, raw=TRUE) + HDLChol + TChol + Diab + Smk + HTT,
data=mydatam)
}
```

```
(Sys.time()-start)/5

# Version 6
start <- Sys.time()
for(i in 1:5){
fit6 <- lm(Risk ~ Age + Race + poly(BP, 2, raw=TRUE) + HDLChol + poly(TChol, 2, raw=TRUE) +
Diab + Smk + HTT, data=mydatam)
}
(Sys.time()-start)/5

# Version 7
start <- Sys.time()
for(i in 1:5){
fit7 <- lm(Risk ~ Age + Race + poly(BP, 2, raw=TRUE) + poly(HDLChol, 2, raw=TRUE) + TChol +
Diab + Smk + HTT, data=mydatam)
}
(Sys.time()-start)/5

# Version 8
start <- Sys.time()
for(i in 1:5){
fit8 <- lm(Risk ~ Age + Race + poly(BP, 2, raw=TRUE) + poly(HDLChol, 2, raw=TRUE) + poly(TChol,
2, raw=TRUE) + Diab + Smk + HTT, data=mydatam)
}
(Sys.time()-start)/5

# Version 9
start <- Sys.time()
for(i in 1:5){
fit9 <- lm(Risk ~ poly(Age, 2, raw=TRUE) + Race + BP + HDLChol + TChol + Diab + Smk + HTT,
data=mydatam)
}
(Sys.time()-start)/5

# Version 10
start <- Sys.time()
for(i in 1:5){
fit10 <- lm(Risk ~ poly(Age, 2, raw=TRUE) + Race + BP + HDLChol + poly(TChol, 2, raw=TRUE) +
Diab + Smk + HTT, data=mydatam)
}
(Sys.time()-start)/5

# Version 11
start <- Sys.time()
```

```
for(i in 1:5){
fit11 <- lm(Risk ~ poly(Age, 2, raw=TRUE) + Race + BP + poly(HDLChol, 2, raw=TRUE) + TChol +
Diab + Smk + HTT, data=mydatam)
}
(Sys.time()-start)/5

# Version 12
start <- Sys.time()
for(i in 1:5){
fit12 <- lm(Risk ~ poly(Age, 2, raw=TRUE) + Race + BP + poly(HDLChol, 2, raw=TRUE) +
poly(TChol, 2, raw=TRUE) + Diab + Smk + HTT, data=mydatam)
}
(Sys.time()-start)/5

# Version 13
start <- Sys.time()
for(i in 1:5){
fit13 <- lm(Risk ~ poly(Age, 2, raw=TRUE) + Race + poly(BP, 2, raw=TRUE) + HDLChol + TChol +
Diab + Smk + HTT, data=mydatam)
}
(Sys.time()-start)/5

# Version 14
start <- Sys.time()
for(i in 1:5){
fit14 <- lm(Risk ~ poly(Age, 2, raw=TRUE) + Race + poly(BP, 2, raw=TRUE) + HDLChol +
poly(TChol, 2, raw=TRUE) + Diab + Smk + HTT, data=mydatam)
}
(Sys.time()-start)/5

# Version 15
start <- Sys.time()
for(i in 1:5){
fit15 <- lm(Risk ~ poly(Age, 2, raw=TRUE) + Race + poly(BP, 2, raw=TRUE) + poly(HDLChol, 2,
raw=TRUE) + TChol + Diab + Smk + HTT, data=mydatam)
}
(Sys.time()-start)/5

# Version 16
start <- Sys.time()
for(i in 1:5){
fit16 <- lm(Risk ~ poly(Age, 2, raw=TRUE) + Race + poly(BP, 2, raw=TRUE) + poly(HDLChol, 2,
raw=TRUE) + poly(TChol, 2, raw=TRUE) + Diab + Smk + HTT, data=mydatam)
}
```

```
(Sys.time()-start)/5

# Version 17
start <- Sys.time()
for(i in 1:5){
fit17 <- lm(Risk ~ Age + Race + BP + HDLChol + TChol + Diab + Smk + HTT, data=mydataf)
}
(Sys.time()-start)/5

# Version 18
start <- Sys.time()
for(i in 1:5){
fit18 <- lm(Risk ~ Age + Race + BP + HDLChol + poly(TChol, 2, raw=TRUE) + Diab + Smk + HTT,
data=mydataf)
}
(Sys.time()-start)/5

# Version 19
start <- Sys.time()
for(i in 1:5){
fit19 <- lm(Risk ~ Age + Race + BP + poly(HDLChol, 2, raw=TRUE) + TChol + Diab + Smk + HTT,
data=mydataf)
}
(Sys.time()-start)/5

# Version 20
start <- Sys.time()
for(i in 1:5){
fit20 <- lm(Risk ~ Age + Race + BP + poly(HDLChol, 2, raw=TRUE) + poly(TChol, 2, raw=TRUE) +
Diab + Smk + HTT, data=mydataf)
}
(Sys.time()-start)/5

# Version 21
start <- Sys.time()
for(i in 1:5){
fit21 <- lm(Risk ~ Age + Race + poly(BP, 2, raw=TRUE) + HDLChol + TChol + Diab + Smk + HTT,
data=mydataf)
}
(Sys.time()-start)/5

# Version 22
start <- Sys.time()
for(i in 1:5){
```

```
fit22 <- lm(Risk ~ Age + Race + poly(BP, 2, raw=TRUE) + HDLChol + poly(TChol, 2, raw=TRUE) +
Diab + Smk + HTT, data=mydataf)
}
(Sys.time()-start)/5

# Version 23
start <- Sys.time()
for(i in 1:5){
fit23 <- lm(Risk ~ Age + Race + poly(BP, 2, raw=TRUE) + poly(HDLChol, 2, raw=TRUE) + TChol +
Diab + Smk + HTT, data=mydataf)
}
(Sys.time()-start)/5

# Version 24
start <- Sys.time()
for(i in 1:5){
fit24 <- lm(Risk ~ Age + Race + poly(BP, 2, raw=TRUE) + poly(HDLChol, 2, raw=TRUE) +
poly(TChol, 2, raw=TRUE) + Diab + Smk + HTT, data=mydataf)
}
(Sys.time()-start)/5

# Version 25
start <- Sys.time()
for(i in 1:5){
fit25 <- lm(Risk ~ poly(Age, 2, raw=TRUE) + Race + BP + HDLChol + TChol + Diab + Smk + HTT,
data=mydataf)
}
(Sys.time()-start)/5

# Version 26
start <- Sys.time()
for(i in 1:5){
fit26 <- lm(Risk ~ poly(Age, 2, raw=TRUE) + Race + BP + HDLChol + poly(TChol, 2, raw=TRUE) +
Diab + Smk + HTT, data=mydataf)
}
(Sys.time()-start)/5

# Version 27
start <- Sys.time()
for(i in 1:5){
fit27 <- lm(Risk ~ poly(Age, 2, raw=TRUE) + Race + BP + poly(HDLChol, 2, raw=TRUE) + TChol +
Diab + Smk + HTT, data=mydataf)
}
(Sys.time()-start)/5
```

```
# Version 28
start <- Sys.time()
for(i in 1:5){
fit28 <- lm(Risk ~ poly(Age, 2, raw=TRUE) + Race + BP + poly(HDLChol, 2, raw=TRUE) +
poly(TChol, 2, raw=TRUE) + Diab + Smk + HTT, data=mydataf)
}
(Sys.time()-start)/5

# Version 29
start <- Sys.time()
for(i in 1:5){
fit29 <- lm(Risk ~ poly(Age, 2, raw=TRUE) + Race + poly(BP, 2, raw=TRUE) + HDLChol + TChol +
Diab + Smk + HTT, data=mydataf)
}
(Sys.time()-start)/5

# Version 30
start <- Sys.time()
for(i in 1:5){
fit30 <- lm(Risk ~ poly(Age, 2, raw=TRUE) + Race + poly(BP, 2, raw=TRUE) + HDLChol +
poly(TChol, 2, raw=TRUE) + Diab + Smk + HTT, data=mydataf)
}
(Sys.time()-start)/5

# Version 31
start <- Sys.time()
for(i in 1:5){
fit31 <- lm(Risk ~ poly(Age, 2, raw=TRUE) + Race + poly(BP, 2, raw=TRUE) + poly(HDLChol, 2,
raw=TRUE) + TChol + Diab + Smk + HTT, data=mydataf)
}
(Sys.time()-start)/5

# Version 32
start <- Sys.time()
for(i in 1:5){
fit32 <- lm(Risk ~ poly(Age, 2, raw=TRUE) + Race + poly(BP, 2, raw=TRUE) + poly(HDLChol, 2,
raw=TRUE) + poly(TChol, 2, raw=TRUE) + Diab + Smk + HTT, data=mydataf)
}
(Sys.time()-start)/5
```

## Appendix D: Graphs and Figures

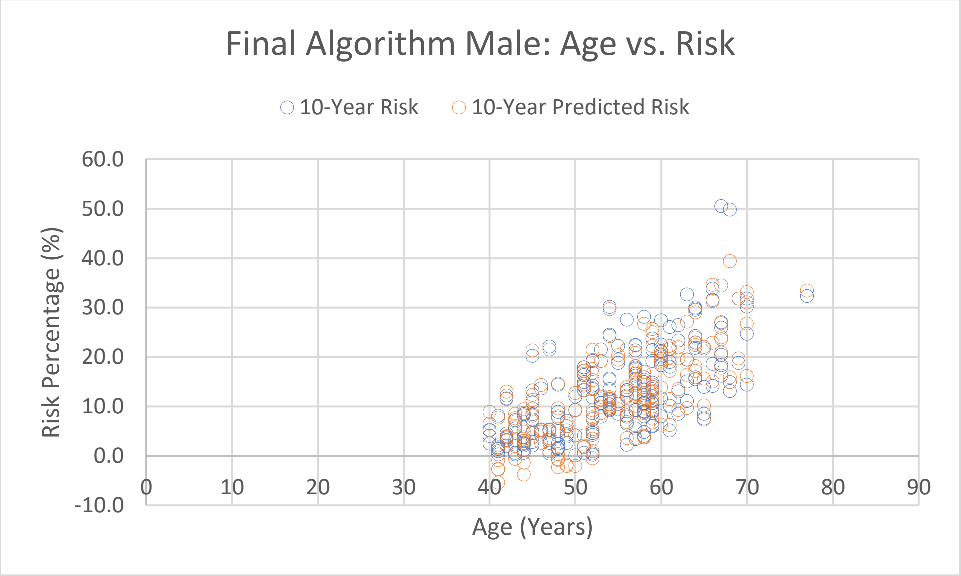


**Figure 6.** Graph comparing the relationship of Age vs. Risk in both 10-Year Risk and the 10-Year Predicted Risk.

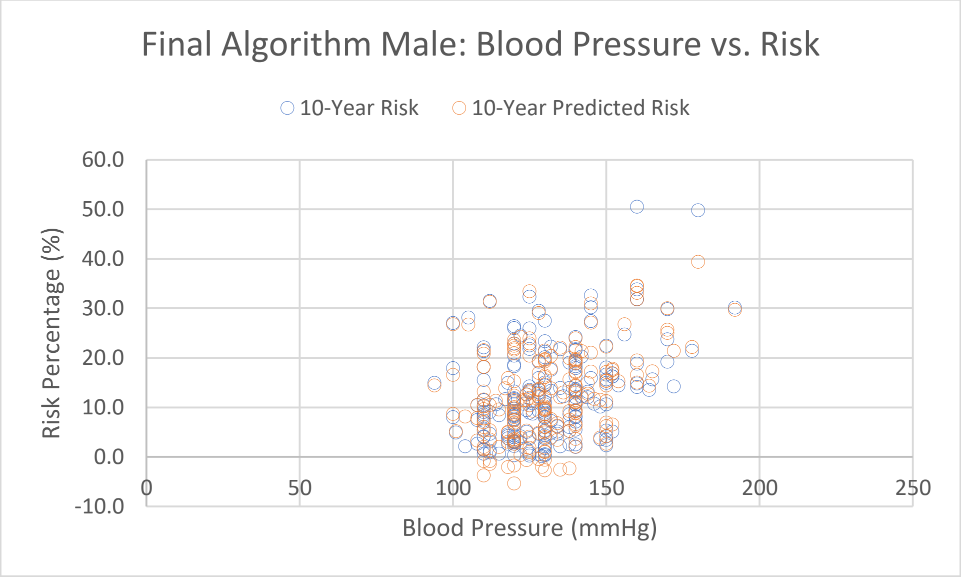


**Figure 7.** Graph comparing the relationship of Blood Pressure vs. Risk in both 10-Year Risk and the 10-Year Predicted Risk.

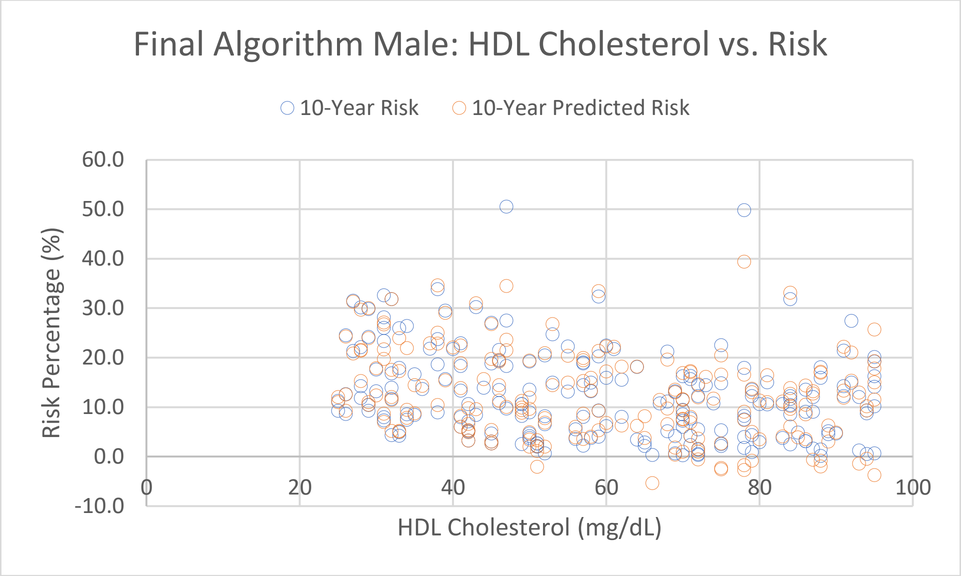


**Figure 8.** Graph comparing the relationship of HDL Cholesterol vs. Risk in both 10-Year Risk and the 10-Year Predicted Risk.

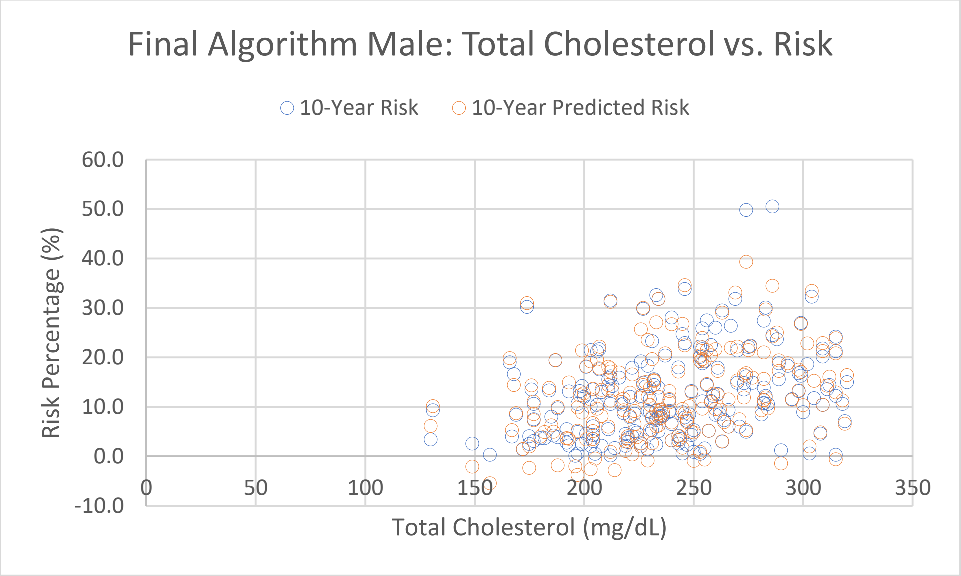


**Figure 9.** Graph comparing the relationship of Total Cholesterol vs. Risk in both 10-Year Risk and the 10-Year Predicted Risk.

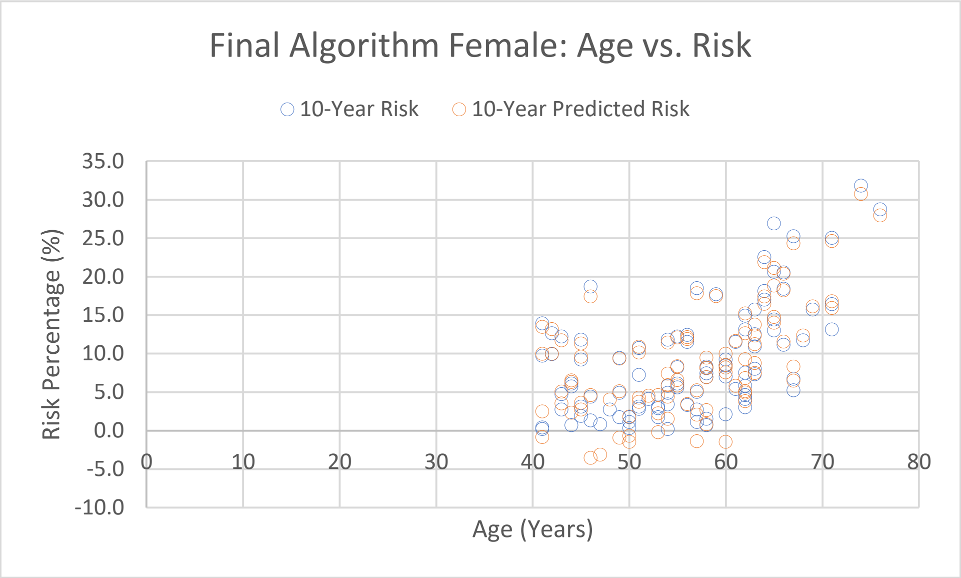


**Figure 10.** Graph comparing the relationship of Age vs. Risk in both 10-Year Risk and the 10-Year Predicted Risk.

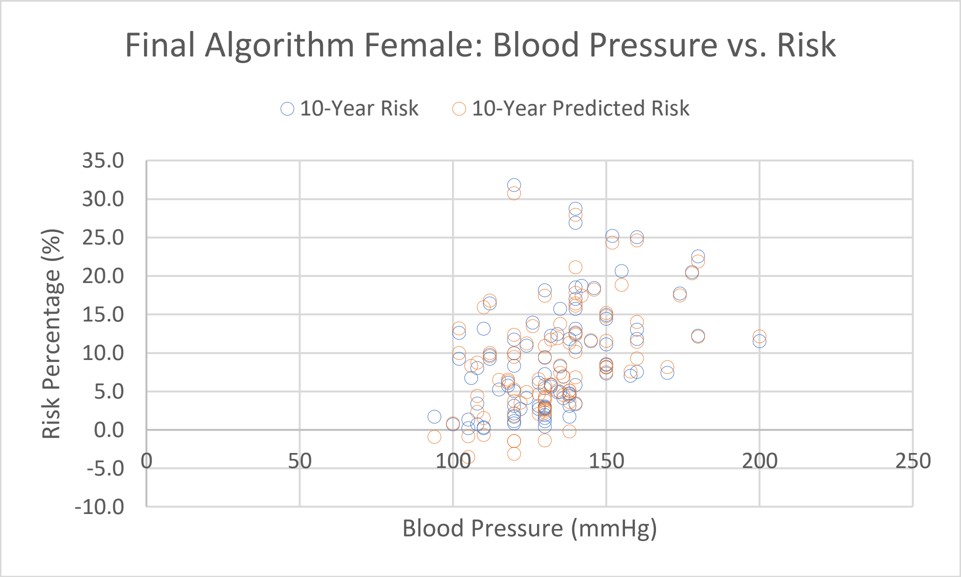


**Figure 11.** Graph comparing the relationship of Blood Pressure vs. Risk in both 10-Year Risk and the 10-Year Predicted Risk.

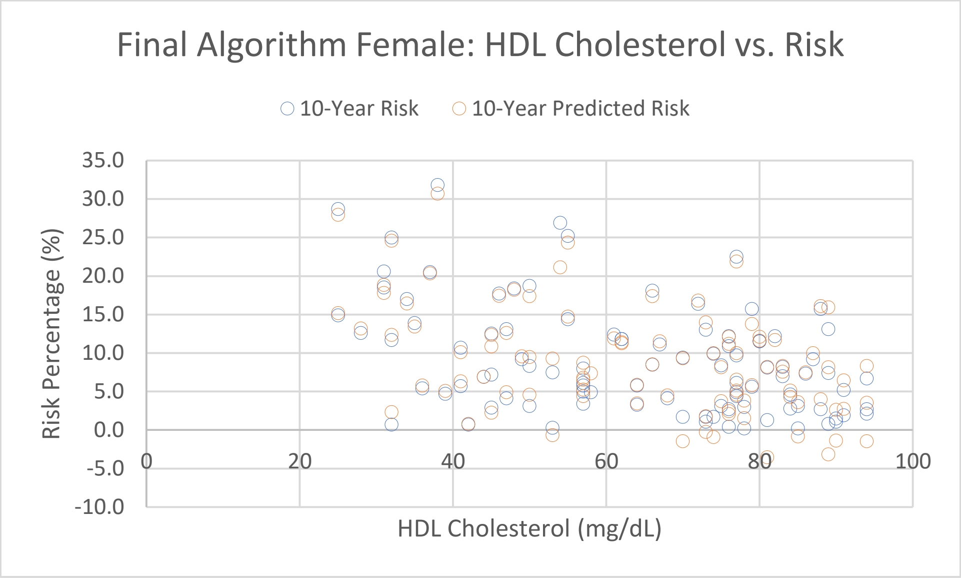


**Figure 12.** Graph comparing the relationship of HDL Cholesterol vs. Risk in both 10-Year Risk and the 10-Year Predicted Risk.

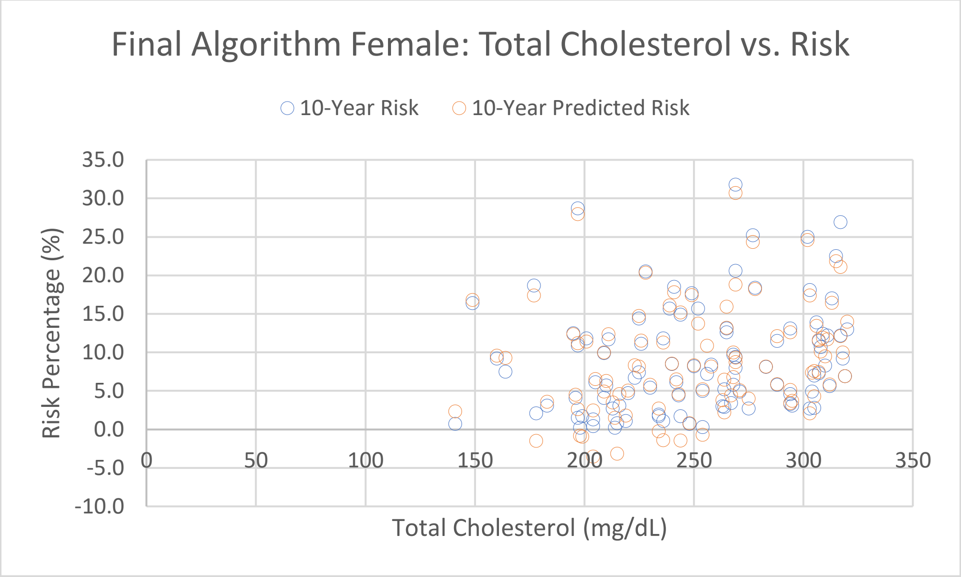


**Figure 13.** Graph comparing the relationship of Total Cholesterol vs. Risk in both 10-Year Risk and the 10-Year Predicted Risk.

# Appendix E: Background Notes

## Janke, Exploring the Potential of Predictive Analytics and Big Data in Emergency Care

Goal of the Paper:

Analyzing the potential of using predictive analytics and big data in emergency care.

Major Findings:

There is a large potential for predictive analytics in the field of healthcare. The potential in emergency health care is lower, but still significant.

Notes on the paper:

The paper focuses on the potential of predictive analytics in the medical field.

Lot of technical terms and discusses major concept that needs to be researched in depth.

Biases of Authors:

There is no apparent bias

My Opinions on the paper:

This paper seems like a great starting point for my project as it does not go too in detail but still gives me a lot of ideas and branches off into more specific and detailed topics.

Follow up Questions and Ideas:

Research Patient Protection and Affordable Care Act.

What are resource intensive casual inferences?

What are heuristics for risk stratification?

Predictive analytics requires an abundance of data, think about where I can get data.

Keywords: Potential, Predictive Analytics, Big Data, Emergency Care

## Jowaheer, Statistical Analysis of Medical Data – Doing it Right

Goal of the Paper:

Discuss data analysis and techniques for data mining.

Major Findings:

No Major Findings as it is an informative paper.

Notes on the paper:

The paper focuses on only the statistical analysis of medical data, which can be very useful for my project.

Focus on Diagram *

Biases of Authors:

There is no apparent bias

My Opinions on the paper:

This paper gives me a rough idea about statistical analysis of medical data, but does not give me anything that concrete to work with as I am researching Predictive analytics.

Follow up Questions and Ideas:

How can some of these techniques and methods be adapted for predictive analytics?

What can I take away from this that will still be valid for my project?

Keywords: Statistical Analysis, Medical Data

## Amarasingham, Implementing Electronic Healthcare Predictive Analytics: Considerations and Challenges

Goal of the Paper:

Discuss data analysis and techniques in health care. Discusses the considerations and challenges.

Major Findings:

No Major Findings as it is an informative paper.

Notes on the paper:

The paper focuses on only the statistical analysis of medical data, which can be very useful for my project.

The paper discusses the triple aim system where you improve outcomes, and improve patient experience, for less cost.

Biases of Authors:

There is no apparent bias

My Opinions on the paper:

This paper gives me a rough idea about statistical analysis of medical data. It gives me a great view of predictive analytics as well.

Follow up Questions and Ideas:

How can some of these techniques and methods be adapted for predictive analytics?

What can I take away from this that will still be valid for my project?

Keywords: Statistical analysis, Medical Data, Predictive Analytics, Healthcare

## Reddy, Predictive big Data analytics in Healthcare

Goal of the Paper:

Discuss big data in healthcare and the statistical analysis of this data.

Major Findings:

No Major Findings as it is an informative paper.

Notes on the paper:

The paper focuses on only the statistical analysis of medical data, which can be very useful for my project.

The paper discusses the intricacy of healthcare analytics and how that can be transferred into cures.

Biases of Authors:

There is no apparent bias

My Opinions on the paper:

This paper gives me a rough idea about statistical analysis of medical data. It gives me a great view of predictive analytics as well a very in-depth view into what I need to consider for my project.

Follow up Questions and Ideas:

How can I incorporate all of these in my project?

Will I be able to add a transferable cure finding addition to the algorithm I am creating?

Keywords: Predictive Analytics, Healthcare, Big Data

## Alexopoulos, Introduction to Multivariate Regression Analysis

Goal of the Paper:

Explain and discuss the basics of Multivariate Regression Analysis. Explain ways to quantify regressions between single variables and relate to overall equation of multi-regression.

Major Findings:

No Major Findings as it is an informative paper.

Notes on the paper:

The paper focuses on using Multivariate regression analysis, which is exactly what I have to do in order to compile all the data into one equation or algorithm.

This is the basis of my project and I will be researching this in further detail later.

Biases of Authors:

There is no apparent bias

My Opinions on the paper:

This paper gives me a rough idea about statistical analysis of multivariate regression.

This is the entirety of my project, but I must focus on different ways to express the variables.

Follow up Questions and Ideas:

How to derived the equation for the standard error of b?

Why the standard deviation of the residual could be estimated by the equation: $\sqrt{\frac{\sum(Y_i-Y_{fit})^2}{n-2}}$? What ANOVA tables are and how they can be used for other statistical analysis methods?

Keywords: Multivariate Regression Analysis

## SAS vs. R (vs. Python) – Which tool should I learn?

Goal of the Paper:

Informs people about which programing language they should use depending on the position they are in and what type of project they are attempting to complete.

Major Findings:

No Major Findings as it is an informative paper.

Notes on the paper:

SAS: SAS has been the undisputed market leader in commercial analytics space. The software offers huge array of statistical functions, has good GUI (Enterprise Guide & Miner) for people to learn quickly and provides awesome technical support. However, it ends up being the most expensive option and is not always enriched with latest statistical functions.

R: R is the Open source counterpart of SAS, which has traditionally been used in academics and research. Because of its open source nature, latest techniques get released quickly. There is a lot of documentation available over the internet and it is a very cost-effective option.

Python: With origination as an open source scripting language, Python usage has grown over time. Today, it sports libraries (numpy, scipy and matplotlib) and functions for almost any statistical operation / model building you may want to do. Since introduction of pandas, it has become very strong in operations on structured data.

Look at Table*

Biases of Authors:

Does not seem like there is much if not any bias.

My Opinions on the paper:

This paper gives me a good idea about what software I should use.

Follow up Questions and Ideas:

How easy will learning R be?

Keywords: R, SAS, Python

## Wilson, Prediction of Coronary Heart Disease Using Risk Factor Categories

Goal of the Paper:

A study on predictive analytics for predicting coronary heart failure

Major Findings:

Around 28 to 29% of coronary heart disease is attributable to high blood pressure levels.

Notes on the paper:

Although the paper does not use the exact same statistical analysis methods it does give a good idea and procedure for my project.

The project uses multivariate regression for calculating risk, which is the same thing I have to do.

Biases of Authors:

Does not seem like there is much if not any bias.

My Opinions on the paper:

This paper will be very useful for showing me multivariate regression in action and will be a big help to my project.

Follow up Questions and Ideas:

Should I contact a doctor of professional to help me weight which symptoms cause heart attacks easier based on other data?

Keywords: Coronary Heart Disease, prediction, hypertension, cholesterol

## Lathauwer, A Short Introduction to Tensor-Based Methods for Factor Analysis and Blind Source Separation

Goal of the Paper:

Explain and discuss the basics of Factor Analysis and Blind Source Separation.

Major Findings:

No Major Findings as it is an informative paper.

Notes on the paper:

The paper focuses on using Factor Analysis, which is another way in which I could analyze my data. This method can be exponentially more difficult than other methods such as Multivariate regression analysis.

Biases of Authors:

There is no apparent bias

My Opinions on the paper:

This paper gives me a rough idea about statistical analysis of factor analysis and blind source separation. These ideas could be essential to my project if I decide to use this method over multivariate regression.

Follow up Questions and Ideas:

In the example of the boxes, why is the number set from 1 to R, shouldn't R be the same value as derived from above?

Keywords: Tensor, Factor Analysis, Blind Source Separation.

## Stata Data Analysis Examples – Multivariate Regression Analysis

Goal of the Paper:

Explain and discuss the basics of Multivariate Regression Analysis. Also provides examples that can be mimicked in R.

Major Findings:

No Major Findings as it is an informative paper.

Notes on the paper:

The paper focuses on using Multivariate regression analysis, which is exactly what I have to do in order to compile all the data into one equation or algorithm.

This is the basis of my project and I will be researching this in further detail later.

Biases of Authors:

There is no apparent bias

My Opinions on the paper:

This paper gives me a rough idea about statistical analysis of multivariate regression.

Follow up Questions and Ideas:

Will I be able to use the example to code in R so that I get a rough understanding of what I need to do?

Do I have time for doing multiple example programs in R?

Keywords: Multivariate Regression Analysis, Stata data analysis

## Xue, introduction to Path Analysis

Goal of the Paper:

This is a presentation by professor Xue on the basics of path analysis.

Major Findings:

No Major Findings as it is an informative paper.

Notes on the paper:

This presentation teaches me about path analysis, which could be vital in my algorithm. I can do a path analysis on my dependent and independent variables to see which variables I can use in my algorithm.

Biases of Authors:

There is no apparent bias

My Opinions on the paper:

This paper gives me a rough idea about statistical analysis of path analysis. This paper is very important because I need to learn as much possible on analytic methods.

This is more physiologic, however, so I may not use it.

Follow up Questions and Ideas:

Will I be able to combine different analysis methods. For example would I be able to do a regression between each independent variable and the dependent variable and then do path analysis on that data?

Keywords: Path Analysis

## Hengel, Hierarchical Model Fitting to 2D and 3D Data

Goal of the Paper:

Teach the user how to fit hierarchical models to data.

Major Findings:

No Major Findings as it is an informative paper.

Notes on the paper:

This presentation teaches me about modeling. This can be a very key step in the future of my project when I want to model the data and the regressions between the data that I have found.

The Information in this document is essential to teaching me how to model this.

Biases of Authors:

There is no apparent bias

My Opinions on the paper:

This is a perfect way for be to graphically and cleanly show my data to the judges and present my project.

Follow up Questions and Ideas:

Will I be able to do all of this modeling in R itself?

If not, will I be able to transfer information from the R program back to a different program and back? (Connector)

Keywords: Model Fitting, 2D and 3D data, Hierarchical data