Heart Disease Risk Analysis

Clay Davis[1], Catherine Koran[1], Kai Richter[1*] , Trevor Savage[1], Matthew Yakaboski[1]

[1] Clarkson University, Mathematics, Potsdam, NY, USA

*Corresponding Author
richteko@clakrson.edu

Keywords: Cardiovascular Disease$_1$, Indicators $_2$, Cholesterol $_3$, Gender$_4$, Medical Assessment $_5$, p-value $_6$.

Abstract

Heart disease has been the leading cause of death in the United States for 73 years, with many established factors. This study analyzed medical and lifestyle data from 70,000 patients with and without heart disease, including 11 objective, subjective, and examination variables contributing to heart disease. Using logistic regression, the p-values of each variable were determined to evaluate which had the most significant correlation with the presence of heart disease. Studies conducted by the CDC indicated that high cholesterol levels, smoking, and lack of physical activity are the highest indicators of heart disease. The results of the analysis of this dataset are inconsistent with this established knowledge. Cholesterol levels were one of five variables with the highest significance when predicting the presence of heart disease. Physical activity had the sixth highest significance, and smoking had the least significance when predicting the presence of heart disease. Further research using a different style of data collection for several of the factors analyzed may be necessary for a more accurate and conclusive regression model.

1. Introduction

Heart disease has existed for thousands of years. Studies done on mummies dating back to 1200 BC found that nine out of sixteen Egyptian pharaohs studied suffered from heart disease (Story). The famous painter Leonardo da Vinci is credited with the first scientific analysis of the heart. In 1507, he started researching the heart and is credited with discovering and documenting coronary arteries and valves for the first time and determining the four heart chambers (Roxby). In the early 1900s, modern research and understanding of heart disease began to take hold. In 1924, the American Heart Association was founded, and in 1929, the first successful cardiovascular catheterization procedure occurred, allowing the inside of the heart to be examined for the first time (Bourassa). In the 1950s, the first successful open heart surgery occurred, and the two forms of cholesterol (LDL and HDL) were discovered and linked to heart disease. The 1960s through 1980s held many scientific advances, including bypass surgery and stents, notably decreasing heart disease mortality rates (Story). Heart disease has not been cured; however, the CDC claims heart disease as the leading cause of death in the United States since 1950 (Centers for Disease Control and Prevention). Identifying those who are at risk of heart disease is critical to take preventive actions or begin early treatment to increase survivability chances. The dataset used for this analysis records a variety of medical and lifestyle-related variables, along with whether or not each patient listed has heart disease.

This investigation aims to determine how several variables indicate the presence of heart disease and which variables have the most significant weight in determining whether a person is at risk for heart disease. For an individual to reduce the risk of succumbing to heart disease, it is necessary to have a method for determining individual risk for heart disease and the contributing factors to implement the most effective preemptive lifestyle changes. As there is evidence of a probable relation between all the variables included in the dataset and the presence of heart disease, all variables recorded were used in the regression analysis. Studies done by the CDC include cholesterol levels, smoking, and inactivity as three of the leading risk factors for heart disease. Higher weight in relation to height (Held et al.), age ("Heart Health and Aging"), high diastolic and/or systolic blood pressure (Solan), and high glucose (when associated with diabetes) (National Institute of Diabetes and Digestive and Kidney Diseases), also have a medically established positive correlation with heart attacks and heart disease. Conversely, consistent exercise has been shown in previous studies to have a negative correlation with heart disease. Alcohol consumption also has an established correlation, though it has been shown to vary in its effects across age groups and consumption levels (Piano). Current medical opinion also alludes to a greater risk of cardiovascular disease among men than among women (Hajar). This analysis aims to answer the following question: "Out of these variables, which will contribute the most towards heart disease?". To better understand what contributes to cardiovascular disease, it is crucial to differentiate the risk factors for different populations. The other question that is answered in this investigation is: Is there a difference in the risk factors for cardiovascular disease between men and women?

2. Methods and Materials

2.1 Description of the Data

The dataset used for this investigation was found on Kaggle and used information collected from actual patients. It contains the information for 70,000 patients and their data regarding 11 variables possibly contributing to cardiovascular disease, as shown in **Table 1**. These variables can be grouped into three categories: objective (factual information about the patient), subjective (information given by the patient), and examination (results of a medical examination performed on the patient). The objective variables include age (days), height (cm), weight (kg), and gender (1-women or 2-men). The subjective variables are all binary–on a 0 or 1 scale (false or true, respectively)–and include smoking, alcohol intake, and physical activity. The examination variables include systolic blood pressure, diastolic blood pressure, cholesterol, and glucose. Blood pressure is a nominal variable, while cholesterol and glucose are on a 1 to 3 scale, where 1 is normal, 2 is above normal, and 3 is well above normal. The presence or absence of cardiovascular disease is binary and the target variable of the dataset.

TABLE 1 | Explanation of the variable names.

Variable Name	**Explanation of the Variable**
Age	Age of the patient in days when the data was recorded
Gender	Equals 1 if the patient is female and 2 if the patient is male
Height	Height of the patient in centimeters
Weight	Weight of the patient in kilograms
ap_hi	The systolic blood pressure of the patient in mm Hg

ap_lo	The diastolic blood pressure of the patient in mm Hg
cholesterol	Equals 1 if the cholesterol level of the patient is normal, 2 if the level is above normal, and 3 if the level is well above normal
gluc	Equals 1 if the glucose level of the patient is normal, 2 if the level is above normal, and 3 if the level is well above normal
smoke	Equals 0 if the patient does not smoke and 1 if the patient does smoke
alco	Equals 0 if the patient does not drink alcohol and 1 if the patient does drink alcohol
active	Equals 0 if the patient is not physically active and 1 if the patient is physically active
w_h_ratio	The ratio of a patient's weight to height
cardio	Equals 0 if the patient does not have cardiovascular disease and 1 if the patient does have cardiovascular disease

2.2 Preprocessing the Data

To begin the analysis, the data was reviewed, and variable values that may have been unrealistic and affected the regression were determined. In total, 1048 of the 70000 entries were removed, including unrealistically high and negative values for systolic and diastolic blood pressure and unrealistically large or small heights. Specifically, for systolic blood pressure, any values less than 1 and greater than 240 were removed; for diastolic blood pressure, any value less than 1 and greater than 190 were removed; and for height, any values greater than 207 cm and less than 100 cm were removed. The cut-offs for blood pressure were based on the interpretation of possible values from the CDC ("High blood pressure symptoms and causes"), and the cut-off for adult height values was based on the interpretation of reasonable adult heights. It is acknowledged that some of the patients removed are possible, but for this study, it made sense to narrow the dataset by eliminating these outliers. Due to medical research on the effects of weight-to-height ratios, patients' weights and heights were converted to a single variable by dividing their weight by height. After removing the outliers, the data was split into male and female. This is done to analyze males and females separately because of their genetic and hormonal differences ("Do men have a higher risk for heart disease?").

2.3 Coding Language and Libraries Used

R language (version 4.3.1) on a Windows Operating System was used for coding (64-bit). The following libraries were employed for the code: readxl, ggplot2, and pROC.

2.4 Analytical Approach

2.4.1 Descriptive Statistics and ROC Curves

To begin the data analysis, descriptive statistical values were calculated for each variable for male and female data. Receiver operating characteristic (ROC) curves were also created for the male and female data, and area under the curve (AUC) values were calculated to determine the accuracy of the models. To do this, the male and female data were each split into training and testing data, where 70% of the data was used for training and 30% for testing. The training data were used to create logistic regression models, and the test data were used to develop predictions of cardiovascular disease based on the models. The ROC curves were then plotted, and the AUC values were calculated.

2.4.2 Accuracy of the Logistic Regression Models Using the Entire Datasets

After the ROC curves were created and the AUC values were calculated, the logistic regression models were used to calculate their accuracy for the entire male and female datasets. A cutoff value of 0.5 was used to predict if a patient would have cardiovascular disease. If the patient had a value greater than 0.5, they received a value of 1, which means they are likely to have cardiovascular disease. If a patient had a value less than 0.5, they received a value of 0, which means they are unlikely to have cardiovascular disease. These predicted values were compared to the actual values from the data set, and the model's accuracy was calculated based on the percentage of correctly predicted values.

2.4.3 Logistic Regression Plots Using Cardiovascular Disease as a Response Variable

Two logistic regression models were created using the male and female data to show the relationship between the explanatory variables and the response variable. The explanatory variables used are age, height, weight, ap_hi, ap_lo, cholesterol, gluc, smoke, alco, and active. The response variable is whether a patient has cardiovascular disease or not. Logistic regression plots were created with cardiovascular disease on the y-axis and an explanatory variable on the x-axis. This was done for each explanatory variable for the male and female datasets.

3. Results

3.1 Descriptive Statistics and ROC Curves

Table 2 and **Table 3** show the descriptive statistics for the variables used in the male and female data, respectively. **Appendix Figure 1A and Figure 2A** show the male and female data's receiver operating characteristic (ROC) curves. The male curve has an area under the curve (AUC) of 0.7799, and the female curve has an AUC of 0.7907.

TABLE 2 | Descriptive Statistics for the Male Variables Used (n = 24,042)

Variable	Minimum	Maximum	Mean	Std. Deviation
Age	10,798	23,713	19,386	2,531
w_h_ratio	0.062	1.11	0.45	0.077
ap_hi	11	240	127.7	17.5
ap_lo	7	180	82.2	9.6
cholesterol	1	3	1.33	0.65
gluc	1	3	1.21	0.55
smoke	0	1	0.22	0.41
alco	0	1	0.11	0.31
active	0	1	0.81	0.40
cardio	0	1	0.50	0.50

TABLE 3 | Descriptive Statistics for the Female Variables Used (n = 44,910)

Variable	Minimum	Maximum	Mean	Std. Deviation
Age	10,859	23,701	19,506	2,433
w_h_ratio	0.13	1.57	0.45	0.087
ap_hi	7	240	125.6	17.7
ap_lo	1	190	80.9	9.9
cholesterol	1	3	1.38	0.69
gluc	1	3	1.23	0.58
smoke	0	1	0.018	0.13
alco	0	1	0.025	0.16
active	0	1	0.80	0.40
cardio	0	1	0.49	0.50

3.2 Accuracy of the Logistic Regression Models Using the Entire Datasets

The logistic regression models created with the training data for both the male and female datasets were used to calculate their accuracy with the entire male and female datasets. When comparing the predicted results for cardiovascular disease with the recorded values, the accuracy of the male logistic regression model is 0.725314, and the accuracy of the female model is 0.7277666. These results explain that the models accurately predicted the presence or absence of cardiovascular disease in 72.5% of the male and 72.8% of the female patients.

3.3 Logistic Regression Plots Using Cardiovascular Disease as a Response Variable

Logistic regression models were created for the full male and female datasets. Pr-values were obtained from the summaries of the regression models, which can be seen in **Table 4**. The Pr-values help determine how good of a predictor each variable is in the model, with a lower Pr-value indicating greater prediction. The Logistic regression curves were created for each explanatory variable with cardiovascular disease as the target variable with an output of 0 or 1. These curves can be seen in **Appendix Figure 1** and **Figure 2**.

TABLE 4 | P-Values for each variable based on the Logistic Regression Models

Variable	Male Pr(> |z|)	Female Pr(> |z|)
(Intercept)[1]	< 2e-16	< 2e-16
age	< 2e-16	< 2e-16
w_h_ratio	< 2e-16	< 2e-16
ap_hi	< 2e-16	< 2e-16
ap_lo	< 2e-16	< 2e-16
cholesterol	< 2e-16	< 2e-16
gluc	6.39e-07	8.28e-07
smoke	7.75e-05	0.21749
alco	2.48e-05	0.00822
active	6.85e-16	3.95e-12

[1] The Intercept is the log odds of outcome of having heart disease without the presence of any predictor variables

4. Discussion

The logistic regression indicates that age, weight to height ratio, diastolic blood pressure, systolic blood pressure, and cholesterol were the variables most indicative of the presence of heart disease, though all of the variables included in the regression had p-values that were statistically significant beyond the standard of .05. In females, the logistic regression graphs indicated a positive relationship between heart disease incidence versus diastolic blood pressure, systolic blood pressure, age, cholesterol, glucose, and weight-to-height ratio. A negative relationship between heart disease incidence was found with physical activity and smoking, though the negative relationship concerning smoking was very slight. The graph of the logistic regression showed an almost negligible negative relationship between alcohol consumption and heart disease. The positive relationship between heart disease incidence versus systolic blood pressure, diastolic blood pressure, glucose levels, age, and weight-to-height ratio was positive in males. Smoking, alcohol consumption, and physical activity had a negative correlation with heart disease incidence, though the correlations for smoking and alcohol consumption were almost negligible.

According to the CDC, the factors that have the highest contribution to heart disease are cholesterol, smoking, and inactivity. In this regression, smoking was indicative of a lower risk of heart disease, which is incongruous with current medical theory. This inconsistency could, in part, have been created in the process of collecting the data. As the data is set as a binary feature, and the data was self-reported by real patients, it is reasonable to believe that there is a degree of inaccuracy in reporting, either by the patients themselves or the medical professionals recording the data for them. Binary scales without precise definitions can be subjective and prone to error. If these factors were on a non-binary scale, they would likely be more accurate indicators of heart disease risk due to the vast differences between casual and regular smokers and drinkers. Additionally, due to the binary nature of the data collection method, casual users may not report themselves as drinkers or smokers. Based on the mean values of smoking and alcohol consumption seen in **Table 2** and **Table 3**, it can be concluded that not many patients recorded themselves as smokers or alcohol drinkers. This could be due to other factors outside the dataset and could be skewing the regressions. More patients may have been alcohol drinkers or smokers. However, due to the use of medication for heart disease, some patients may have recorded that they do not smoke or drink alcohol because the medication requires abstinence from smoking and drinking.

The male and female logistic regression results followed similar trends except for a few key variables. This study separated the data by gender to eliminate variables affected by biological differences between males and females concerning heart disease incidence. The logistic regression (**Table 4**) showed that smoking, alcohol consumption, and physical activity were better indicators of heart disease for males than females. This difference in the effectiveness of these indicators could be due to differences in willingness to report, accuracy of reporting, or biological differences between sexes.

5. Conclusions

After analyzing each variable’s significance, six variables contribute the most towards cardiovascular disease. Age, weight-to-height ratio, systolic blood pressure, diastolic blood pressure, and cholesterol are the greatest indicators of heart disease, with the smallest p-values in the model. There is a slight difference in risk factors for each gender. Smoking, alcohol use, and physical activity are greater indicators of cardiovascular disease for men than women, and the model for females was more accurate according to the ROC curve and accuracy test conducted.

For further analysis of heart disease risk, it is advisable to collect and utilize non-binary data for variables such as smoking, alcohol use, and physical activity. In reality, these variables are highly variable. Alcohol use, smoking, and exercise are lifestyle activities that do not exist on a zero or one binary scale. The measure of “yes” or “no” for these types of variables can be quite subjective, as patients are not likely to report substance use if it is not part of their daily life and may be more likely to report the presence of physical activity if they participate in any degree of physical activity, no matter the intensity. For example, a better measure of alcohol consumption could be the average amount of drinks per week a patient consumes or how many nights per week a subject drinks more than three drinks. Future studies using non-binary data for these variables may find different correlations and p-values and, therefore, may determine a different ranking for the various indicators of heart disease.

6. Data and Code Availability Statement

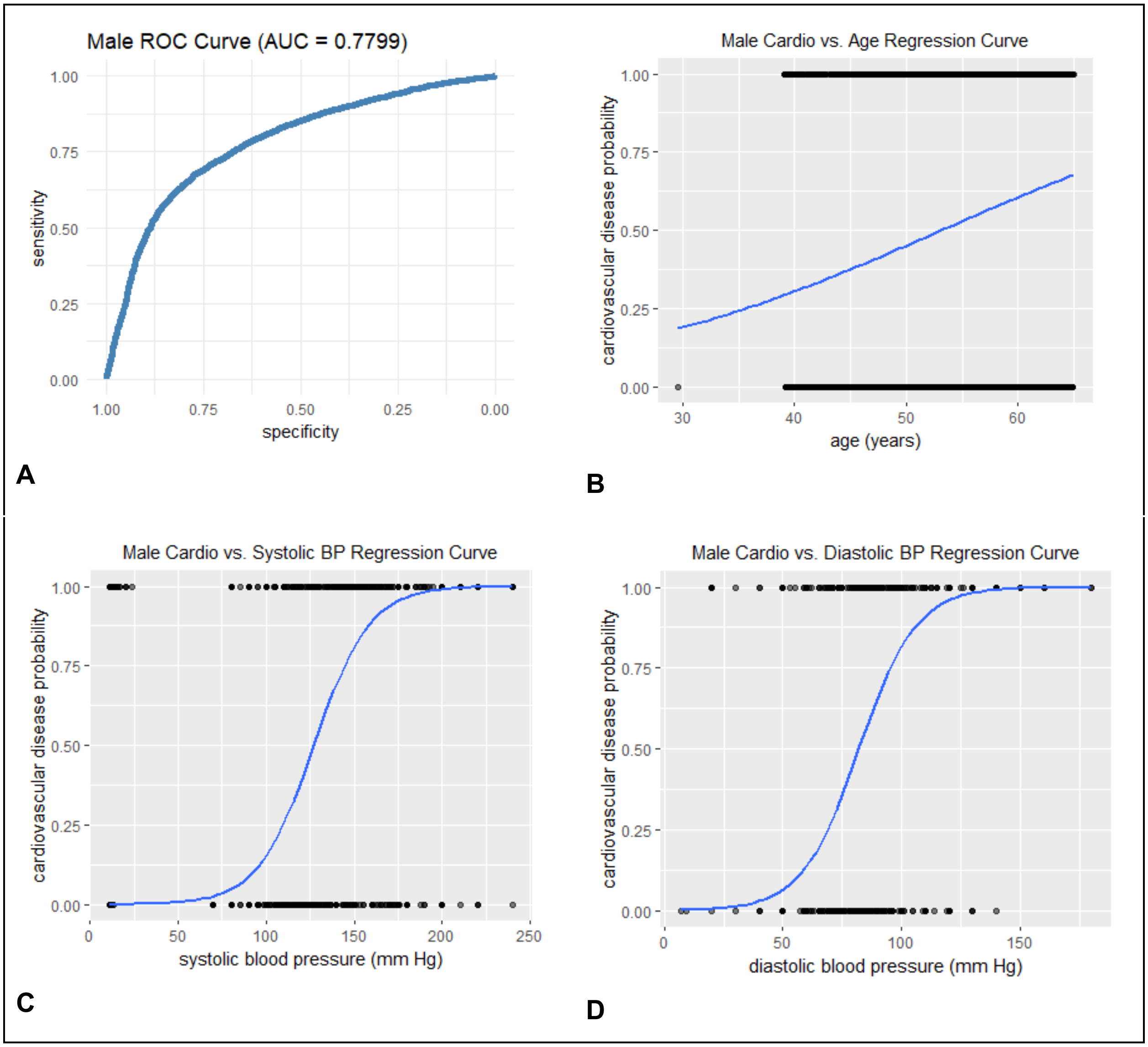
The datasets and code used for this study can be found in the GitHub repository https://github.com/TrevoratorSavage/Stat383. The original dataset can be found at https://www.kaggle.com/datasets/sulianova/cardiovascular-disease-dataset/data.

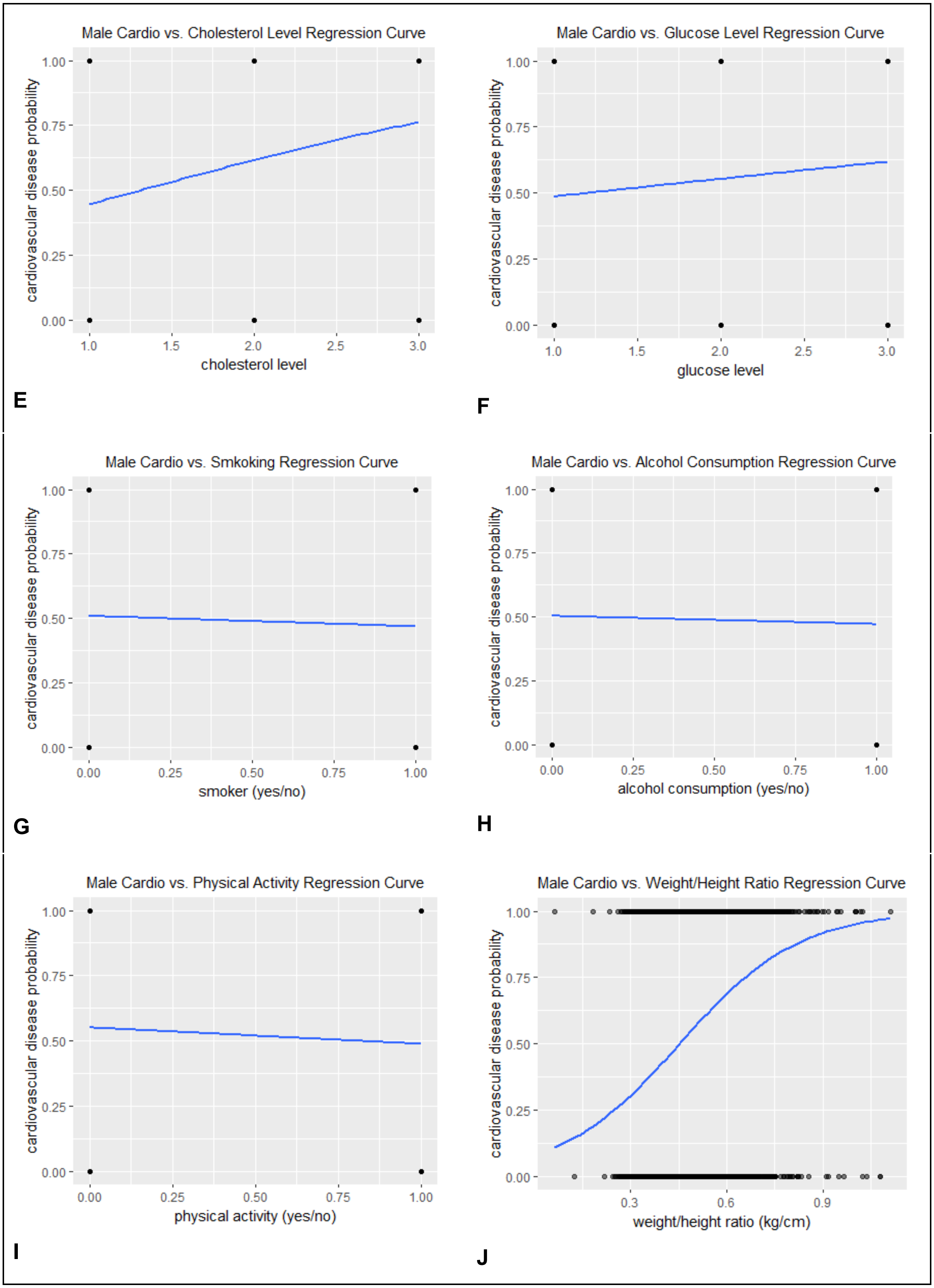
7. Author Contribution

C. D. writing proposal, writing report
C. K. writing proposal, writing report
K. R. code, writing proposal, writing report
T. S. writing proposal, writing report
M. Y. code, writing proposal, writing report

8. Appendix

Figure 1





Male Cardio vs. Cholesterol Level Regression Curve

Male Cardio vs. Glucose Level Regression Curve

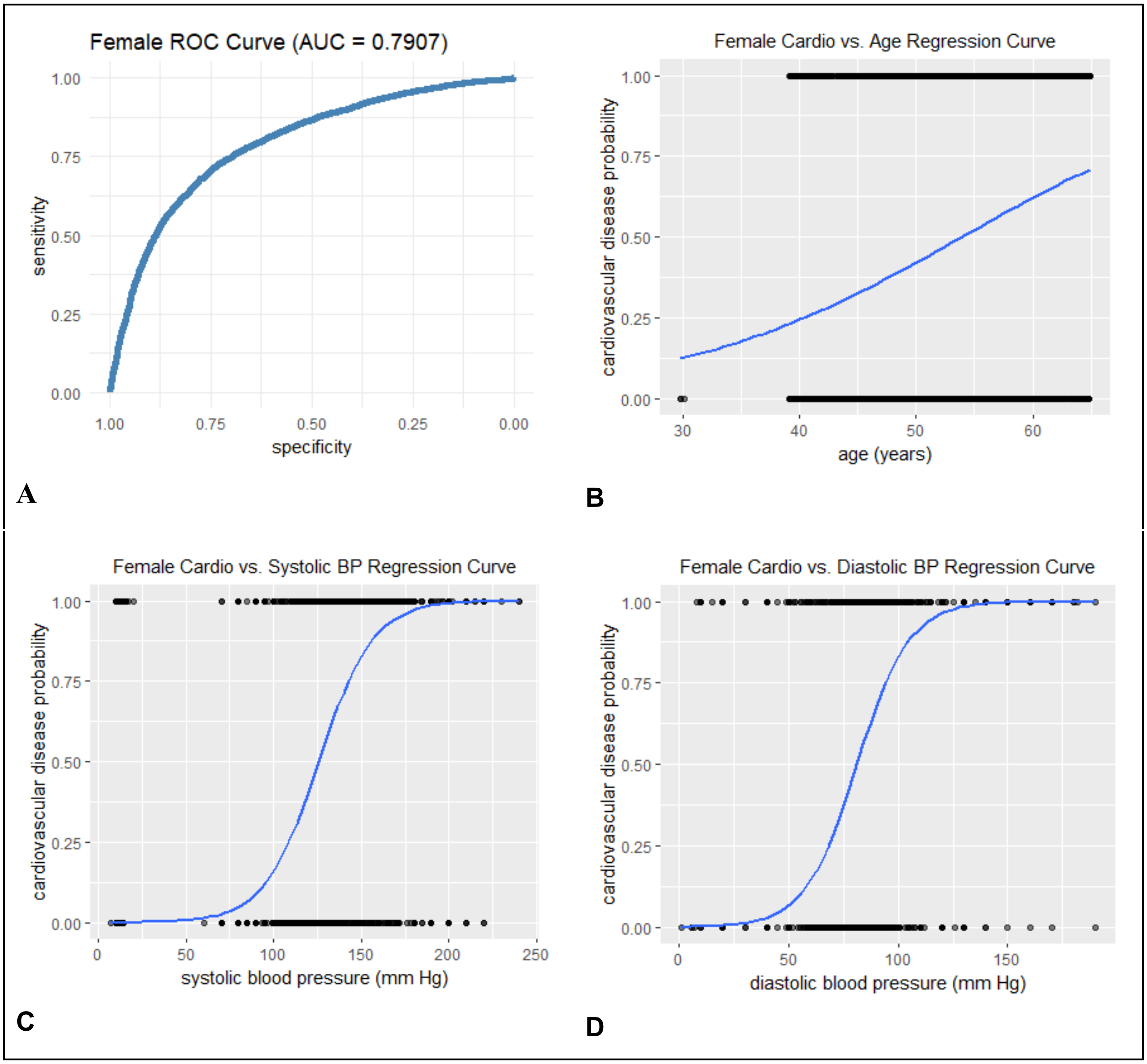
Male Cardio vs. Smkoking Regression Curve

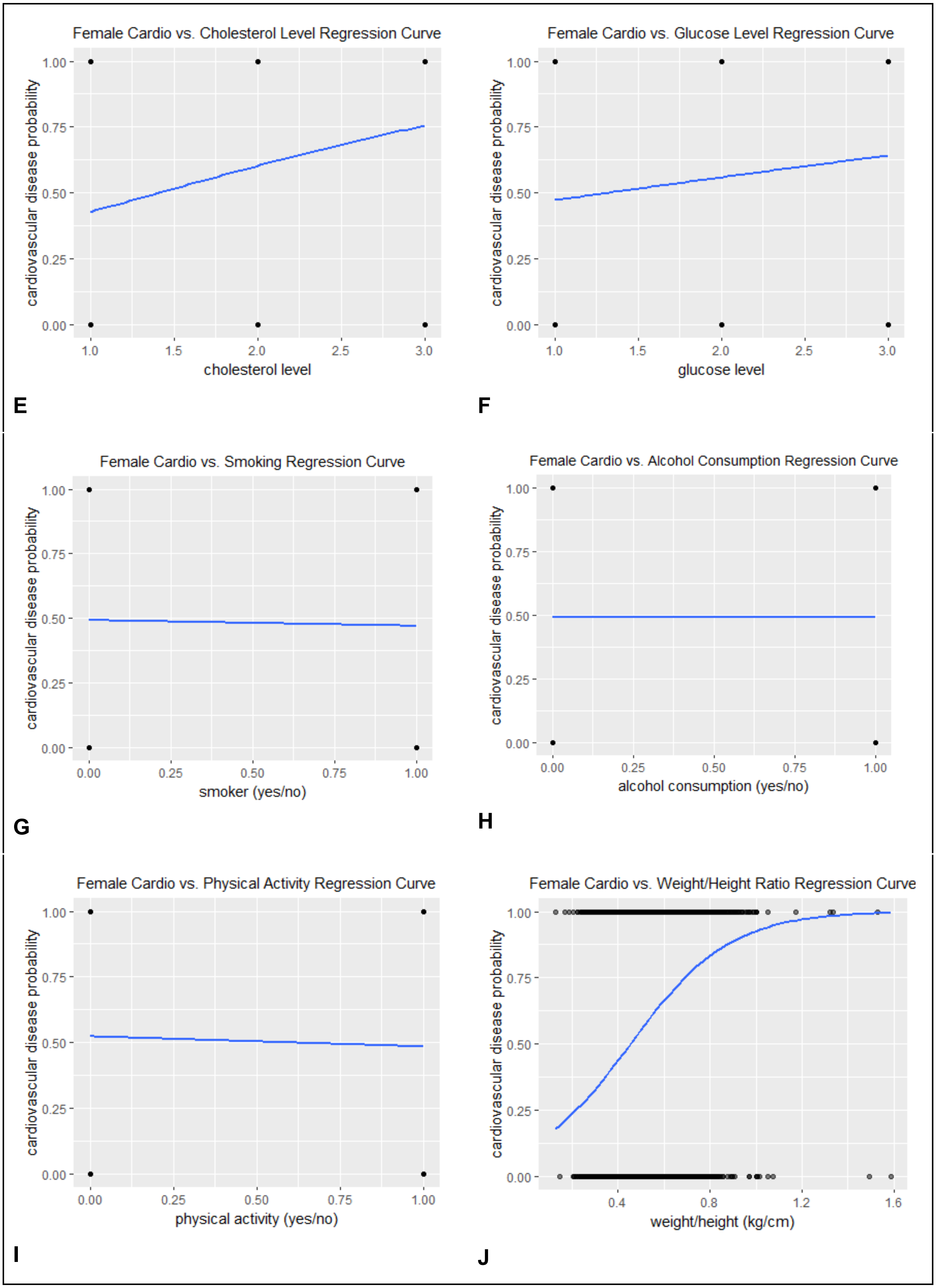
Male Cardio vs. Alcohol Consumption Regression Curve

Male Cardio vs. Physical Activity Regression Curve

Male Cardio vs. Weight/Height Ratio Regression Curve

Figure 2


Female ROC Curve (AUC = 0.7907)
sensitivity
1.00
0.75
0.50
0.25
0.00
1.00
0.75
0.50
0.25
0.00
specificity
A
Female Cardio vs. Age Regression Curve
cardiovascular disease probability
1.00
0.75
0.50
0.25
0.00
30
40
50
60
age (years)
B
Female Cardio vs. Systolic BP Regression Curve
cardiovascular disease probability
1.00
0.75
0.50
0.25
0.00
0
50
100
150
200
250
systolic blood pressure (mm Hg)
C
Female Cardio vs. Diastolic BP Regression Curve
cardiovascular disease probability
1.00
0.75
0.50
0.25
0.00
0
50
100
150
diastolic blood pressure (mm Hg)
D


Female Cardio vs. Cholesterol Level Regression Curve
cardiovascular disease probability
1.00
0.75
0.50
0.25
0.00
1.0
1.5
2.0
2.5
3.0
cholesterol level
E
Female Cardio vs. Glucose Level Regression Curve
cardiovascular disease probability
1.00
0.75
0.50
0.25
0.00
1.0
1.5
2.0
2.5
3.0
glucose level
F
Female Cardio vs. Smoking Regression Curve
cardiovascular disease probability
1.00
0.75
0.50
0.25
0.00
0.00
0.25
0.50
0.75
1.00
smoker (yes/no)
G
Female Cardio vs. Alcohol Consumption Regression Curve
cardiovascular disease probability
1.00
0.75
0.50
0.25
0.00
0.00
0.25
0.50
0.75
1.00
alcohol consumption (yes/no)
H
Female Cardio vs. Physical Activity Regression Curve
cardiovascular disease probability
1.00
0.75
0.50
0.25
0.00
0.00
0.25
0.50
0.75
1.00
physical activity (yes/no)
I
Female Cardio vs. Weight/Height Ratio Regression Curve
cardiovascular disease probability
1.00
0.75
0.50
0.25
0.00
0.4
0.8
1.2
1.6
weight/height (kg/cm)
J

9. References

Bourassa, Martial G. “The History of Cardiac Catheterization.” *The Canadian Journal of Cardiology*, vol. 21, no. 12, 1 Oct. 2005, pp. 1011–1014, pubmed.ncbi.nlm.nih.gov/16234881/.

Centers for Disease Control and Prevention. “Heart Disease Deaths - Health, United States.” *Www.cdc.gov*, 19 Sept. 2022, www.cdc.gov/nchs/hus/topics/heart-disease-deaths.htm#:~:text=Heart%20disease%20has%20been%20the.

“Do Men Have a Higher Risk for Heart Disease?” *Do Men Have a Higher Risk for Heart Disease?: Louisiana Heart and Vascular: Interventional Cardiologists*, www.louisianaheart.org/blog/do-men-have-a-higher-risk-for-heart-disease#:~:text=Men%20tend%20to%20develop%20heart,also%20boost%20blood%20vessel%20health. Accessed 6 Dec. 2023.

Hajar, Rachel. “Risk Factors for Coronary Artery Disease: Historical Perspectives.” *Heart Views*, vol. 18, no. 3, July 2017, p. 109, www.ncbi.nlm.nih.gov/pmc/articles/PMC5686931/, https://doi.org/10.4103/heartviews.heartviews_106_17.

“Heart Health and Aging.” *National Institute on Aging*, www.nia.nih.gov/health/heart-health/heart-health-and-aging#:~:text=People%20age%2065%20and%20older.

Held, Claes, et al. “Body Mass Index and Association with Cardiovascular Outcomes in Patients with Stable Coronary Heart Disease – a STABILITY Substudy.” *Journal of the American Heart Association*, vol. 11, no. 3, Feb. 2022, https://doi.org/10.1161/jaha.121.023667.

"High Blood Pressure Symptoms and Causes." Centers for Disease Control and Prevention, Centers for Disease Control and Prevention, 18 May 2021, www.cdc.gov/bloodpressure/about.htm.

National Institute of Diabetes and Digestive and Kidney Diseases. "Diabetes, Heart Disease, & Stroke | NIDDK." *National Institute of Diabetes and Digestive and Kidney Diseases*, Apr. 2021, www.niddk.nih.gov/health-information/diabetes/overview/preventing-problems/heart-disease-stroke#:~:text=High%20blood%20glucose%20from%20diabetes.

Piano, Mariann R. "Alcohol's Effects on the Cardiovascular System." *Alcohol Research : Current Reviews*, vol. 38, no. 2, 2017, pp. 219–241, www.ncbi.nlm.nih.gov/pmc/articles/PMC5513687/.

Roxby, Philippa. "What Leonardo Taught Us about the Heart." *BBC News*, 28 June 2014, www.bbc.com/news/health-28054468.

Solan, Matthew. "A Look at Diastolic Blood Pressure." *Harvard Health*, 1 Apr. 2022, www.health.harvard.edu/heart-health/a-look-at-diastolic-blood-pressure#:~:text=However%2C%20those%20with%20diastolic%20values. Accessed 5 Dec. 2023.

Story, Colleen. "The History of Heart Disease." *Healthline*, Healthline Media, 11 May 2018, www.healthline.com/health/heart-disease/history#the-future-of-heart-disease.