Investigating the Association between Thalassemia and Heart Disease using Logistic Regression Modeling

Data 621, Winter 2023 Course Project

Emma Bogner, Meghana Kompally, Jason Wong, and Serena Sun

Faculty of Science and Cumming School of Medicine, University of Calgary

Abstract word count: 332
Manuscript word count: 4493
Figures: 2
Tables: 3
References: 7 (178 words)

Table of Contents

Abstract

Background

Heart disease affects millions of people globally and is influenced by various factors. Thalassemia, a hereditary condition that results in inadequate amounts of hemoglobin in the body, can cause heart damage due to excess iron in the bloodstream.

Methods

This paper presents a cross-sectional study assessing the association between thalassemia and heart disease using data collected by the Cleveland Clinic in 1988. The primary objective of the study was to determine whether the presence of thalassemia is associated with heart disease and—if so—whether it is associated with increased heart disease severity. In addition, the association between thalassemia and serum LDL cholesterol levels was assessed. These analyses were carried out using logistic and linear regression.

Results

Of a total of 301 patients, 135 (44.85%) had thalassemia and 166 (55.15%) did not. Thalassemia had the greatest effect on the presence of heart disease with an odds ratio of 5.2845 (95% CI [2.7118,10.5854], p-value < 0.0001). This was followed by age group and exercise-induced angina, with odds ratios of 3.2539 (95% CI [1.7012, 6.3953], p-value=0.0005) and 2.8801 (95% CI [1.4119,5.9481], p-value=0.0038), respectively. Similar results were found when testing the association between thalassemia and heart disease severity, with a cumulative odds ratio of 4.9941 (95% CI [2.7846,9.0925], p-value < 0.0001), although this model was not valid. No significant association was found between thalassemia and serum cholesterol (p-value= 0.7510).

Conclusion

The results of this study indicate that patients with thalassemia have higher odds of developing heart disease, while there is no association between thalassemia and serum cholesterol. In addition, individuals above the age of 56 and those who experience exercise-induced angina have significantly higher odds of having a positive heart disease diagnosis. Other factors such as sex, chest pain, and abnormal resting electrocardiogram results are also significantly associated with heart disease.

Keywords: Thalassemia, Heart Disease, Cholesterol Levels, Logistic Regression Analysis, Linear Regression Analysis

1. Introduction

1.1 Background

Heart disease is a complex condition that affects millions of people globally and is one of the leading causes of death across all ethnic groups. Several factors have been associated with heart disease, including age, sex, chest pain, and serum cholesterol levels. Women, for instance, develop coronary heart disease later in life than men; however, data from the past decade show that they have a higher 30-day mortality rate, and this difference is mainly attributed to clinical differences at presentation (Mosca et al. 2011). Hypercholesterolemia, a condition characterized by high levels of cholesterol in the bloodstream, is a known risk factor for heart disease. Interestingly, hypercholesterolemia has been linked to anginal chest pain (chest pain caused by a lack of blood flow to the heart), which is also a risk factor for the development of heart disease (Cleveland Clinic, 2022).

Thalassemia is a hereditary condition that has previously been shown to be associated with heart disease (Assopos et al., 2007). Thalassemia is characterized by gene mutations that result in inadequate amounts of hemoglobin in the body. Many thalassemia patients present with excess iron levels in the bloodstream, either from treatment or the disease itself, which can lead to heart damage (Cohen et al., 2004). Interestingly, some studies show that patients with thalassemia have reduced levels of low-density lipoprotein (LDL) cholesterol (Ricchi, Ammirabile, & Maggio, 2011). Although excess LDL cholesterol is known to be significantly associated with the development of heart disease, the balance between the potential protective effect of reduced LDL levels and the negative effect of cardiac iron overload in thalassemia patients remains unclear. Understanding the risk factors associated with heart disease, including thalassemia, can help in the early diagnosis and appropriate management of the condition, leading to better health outcomes for affected individuals.

1.2 Research Question

What is the association between thalassemia diagnosis and the risk of heart disease among adults over 30 years of age who were hospitalized in 1988, and does the severity of heart disease differ in individuals with thalassemia compared to those without thalassemia? Additionally, is a thalassemia diagnosis associated with lower serum LDL cholesterol levels in this population?

1.3 Objectives

The primary objective of this study was to determine whether the presence of thalassemia is associated with a diagnosis of heart disease and, if so, whether it is associated with increased severity of heart disease. The secondary objective of this study was to determine if a thalassemia diagnosis is associated with decreased serum LDL cholesterol levels.

1.4 Data Source

The dataset used in this study containing information collected at the Cleveland Clinic in 1988 is publicly available from the UCI Machine Learning Repository, Center for Machine Learning and Intelligent Systems (Dua & Graff, 2019). These data were originally collected by Robert Detrano,

M.D., Ph.D., and his team. The dataset contains the demographic information and clinical features of 303 unique patients, including age (years; continuous), sex (binary), chest pain type (categorical), resting blood pressure on admission (mmHg; continuous), cholesterol (mg/dl; continuous), resting electrocardiogram (ECG) results (categorical), exercise-induced angina (binary), and thalassemia diagnosis (categorical). Additionally, information on the severity of each patient's heart disease is also provided as a level from 0 (absence) to 4 (highest severity). All patient identifiers were previously removed from the dataset, ensuring that any ethical considerations concerning patient privacy have already been addressed.

2. Methods

2.1 Study Design

2.1.1 Study Population

This cross-sectional study includes adult patients aged 30 and over who were treated at the Cleveland Clinic in 1988. Of the 303 patients included in the dataset, two were excluded from the study because they did not have a valid observation for thalassemia diagnosis, leaving a total of 301 participants as illustrated in Figure 1. It should be noted that in this dataset, information was only available for adults over 30 years of age; for this reason, we could not generalize our targeted study population as all adults (i.e., above 18).

2.1.2 Study Power

In addition, the dataset used here was not generated for a specific study or experiment; rather, it contains data from a convenience sample of patients who were referred to the Cleveland Clinic for chest pain and confirmed or suspected heart disease. Therefore, the sample size for the present study could not be calculated the same way as for a typical research study, and it remains unclear how this specific sample size was determined. Most likely, the sample size was purely based on how many patients were referred to the Cleveland Clinic in 1988, rather than a formal sample size calculation. Consequently, since this data was not systematically or randomly sampled, it may not be representative of the larger population. Since our outcome variable is categorical, we decided to use Cramer's V in the rcompanion package in R to calculate the effect size, which resulted in an effect size of 0.2991. We used this value to calculate the power of our study using the pwr.2p.test function from the pwr package in R, giving us a power of 0.9563, at a sample size of 301 and a significance level of 0.05. (Champely, 2020)

2.3 Study Outcomes

2.3.1 Objective 1

2.3.1.1. Primary Outcome

For the first objective of this study, investigating the association between thalassemia and heart disease, the primary outcome was the presence or absence of heart disease. Patients were classified as either having heart disease, indicated by greater than 50% narrowing in any major vessel as determined by angiographic testing, or not having heart disease, indicated by less than 50% narrowing in any major vessel as determined by angiographic testing. Since the original dataset had heart disease encoded as an ordinal variable but the documentation

provided outcome definitions for a binary variable (given above), the ordinal variable was transformed into a binary variable for use as the primary outcome.

2.3.1.2. Secondary Outcome

The secondary outcome was the association between thalassemia and the severity of heart disease, as determined by the degree of vessel narrowing. Since this outcome was not well-documented in the dataset, no specific measurement cutoffs can be provided for each level; however, 0 indicates the absence of heart disease while levels 1, 2, 3, and 4 indicate increasing severities. No calculations or transformations were performed on this outcome.

2.3.2 Objective 2

2.3.2.1. Primary Outcome

For the secondary objective exploring the association between thalassemia and cholesterol, serum cholesterol levels in mg/dL were analyzed as the primary outcome. No calculations or transformations were performed on this outcome.

2.5 Statistical Methods

In order to address our objectives, we used logistic regression and linear regression models. All hypothesis testing was conducted using a predetermined significance level (α) of 0.05; all results with p-values lower than this limit were considered statistically significant. In addition, 95% confidence intervals are reported for all estimated regression coefficients and odds ratios to provide a range of values in which the true population parameter is likely to fall. These confidence intervals were assessed to determine if they included 0 (for regression coefficients) or 1 (for odds ratios), which would indicate no significant effect.

2.5.1 Objective 1

2.5.1.1. Primary Outcome

For the primary outcome of objective 1, we used descriptive analysis and a binary logistic regression model, because the heart disease outcome was transformed from an ordinal to binary variable. The main exposure assessed was thalassemia, and the model was adjusted for age, sex, chest pain type, resting ECG results, resting blood pressure, and exercise-induced angina. These covariates were chosen based on a theoretical understanding of the factors that may influence the presence of heart disease reported in previous studies'. Out of these covariates, two potential confounding variables were identified: serum cholesterol and resting blood pressure. Serum cholesterol levels and high blood pressure are known risk factors for heart disease and are therefore likely to be associated with both thalassemia and the presence of heart disease. The estimated strength of the association between thalassemia and heart disease is presented as an odds ratio and corresponding confidence interval.

2.5.1.2. Secondary Outcome

For the secondary outcome of objective 1, we used descriptive analysis and an ordinal logistic regression model to assess the association between thalassemia and the severity of heart disease, determined using the degree of vessel narrowing as a proxy (the original ordinal format of the outcome variable). The model was adjusted for the same covariates as the primary

outcome model and thus shared the same potential confounders. The effect of thalassemia on the degree of heart disease severity is presented as an adjusted cumulative odds ratio and corresponding confidence interval for greater vessel narrowing (i.e., more severe heart disease).

2.5.2 Objective 2

The association between thalassemia and cholesterol levels was also explored using a linear regression model. Here, the outcome was serum cholesterol levels (mg/dL) and the main exposure was thalassemia. This model was also adjusted for the same covariates as those used in the above analyses. The effect of thalassemia on serum cholesterol levels is presented as an estimated beta coefficient and corresponding confidence interval for the cholesterol variable.

2.5.3 Effect Modification

No potential effect modifiers were identified in any of the analyses carried out in this study, and no subgroup analyses were planned.

2.5.4 Model Assumptions

All models were assessed to determine if they met their respective assumptions required for regression analysis. Briefly, the two logistic regression models were assessed to determine if they met the assumptions of independence (through inspection of the original data) and linearity (using a Box-Tidwell test). For the ordinal logistic regression model, the proportional odds assumption was also validated using the *brant* package in R. Finally, the linear regression model was originally going to be assessed for the assumptions of linearity (), independence (through inspection of the original data), normality (using a normal qq plot and Shapiro–Wilk test), and equal variance (using a residual plot). However, since the model was found to be insignificant, these tests were not carried out and thus no results are presented in this paper.

2.5.5 Missing Values

There were no missing values for the 301 observations used in this study.

2.5.6 Software

Python (v. 3.10.2) and RStudio (v. 2022.12.0+353) were used to perform all necessary data cleaning and statistical analyses.

3. Results and Discussion

3.1 Descriptive Statistics

Table 1 shows the descriptive statistics for study participants stratified by the main exposure (i.e. with thalassemia N=135 and without thalassemia N=166). There were considerably more male patients in the study overall (68.1%) compared to female patients (31.9%). There were also more male patients (88.1%) in the exposed group than in the unexposed group (51.8%) and fewer female patients (11.9%) in the exposed group than in the unexposed group (48.2%). The

mean age of people who did not have thalassemia was 53.4 years old (SD=9.80), which was slightly lower than that of the thalassemia-diagnosed patients (55.8, SD=7.82). The mean resting blood pressure was also slightly higher in the exposed group (134, SD=19.2) compared to the unexposed group (130, SD=16.0). Patients diagnosed with thalassemia were more likely to experience exercise-induced angina (49.6%) compared to those who did not have thalassemia (18.7%). It is interesting to note that patients in the exposed group were more likely to be asymptomatic with respect to chest pain (66.7%) rather than experiencing non-anginal pain, atypical angina, or typical angina, whereas in the unexposed group, more patients reported having non-anginal pain. Both groups, however, showed very few patients in the typical anginal pain category (7.4% and 7.8% in exposed vs. unexposed). There was about an equal number of patients with normal resting ECG results versus those showing probable or definite left ventricular hypertrophy in both the exposed (normal=48.1% and hypertrophy=50.4%) and unexposed groups (normal=51.2% and hypertrophy=47.6%). Finally, most of the patients with thalassemia were also diagnosed with heart disease (74.8%) whereas most patients without thalassemia were not diagnosed with heart disease (77.7%). When the heart disease outcome was ordered from 0 to 4, we saw that there were an equal number of patients in the 0–3 categories (i.e. without heart disease (0) = 25.2%, with level 1 heart disease = 24.4%, with level 2 heart disease = 20.7% and with level 3 heart disease = 31.5%) for the exposed group and only 8.1% in the level 4 heart disease category. However, in the unexposed group, 77.1% of the patients did not have heart disease, while 13.3% of them had low-severity (level 1) heart disease.

3.2 Objective 1 Results

3.2.1. Primary Outcome

A binary logistic regression model was first created to test the association between thalassemia and heart disease. The final model included five significant predictors, including thalassemia (p-value<0.0001), age (p-value=0.006), sex (p-value=0.011), chest pain (p-values= 0.002,0.001,0.001), exercise-induced angina (p-value=0.007), and resting ECG (p-values=0.274,0.038). The insignificant variables cholesterol and resting blood pressure were found to be non-confounding (magnitude of confounding < 10%) and were therefore excluded from the model. The OR of this model will not be interpreted and reported as the linearity assumption of the model was not satisfied: a Box–Tidwell test indicated that the relationship between the log odds of the dependent variable, thalassemia, and age was not linear (p-value = 0.021). The linearity assumption was satisfied for the rest of the predictors and main exposure due to their categorical nature. To address this issue of non-linearity, we carried out a subgroup analysis where age was categorized into two groups, “Under or Equal to 56” and “Over 56”. The split was implemented at 56 years of age in order to have an approximately equal percentage of patients in each group. Splitting at 56 put 160 (53.2%) patients in the “Under or Equal to 56” category and 141 (46.8%) patients in the “Over 56” category.

After categorizing age, our final binal logistic regression model included the same five significant predictors, thalassemia (p-value < 0.0001), age group (p-value=0.0005), sex (p-value=0.006), chest pain type (p-values= 0.001, 0.001, 0.0001), exercise-induced angina (p-value=0.004) and resting ECG (p-value=0.145, 0.025). Cholesterol and resting blood pressure were found to be

non-confounding at a magnitude of 1.45% and 2.78%, respectively, and were thus omitted from the final model.

At a significance level of 0.05, the OR for thalassemia with respect to no thalassemia is 5.2845 (95% CI [2.7118,10.5854], p-value < 0.0001). This means that patients with thalassemia are 5.2845 times more likely to develop heart disease compared to patients without thalassemia. The 95% confidence interval does not contain 1, indicating that the association between thalassemia and heart disease is statistically significant. Female patients have approximately 0.3318 (95% CI [0.1476,0.7264], p-value=0.006) times lower odds of developing heart disease compared to males after adjusting for the included covariates. Individuals aged 56 years and over have 3.2539 (95% CI [1.7012, 6.3953], p-value=0.0005) times higher odds of developing heart disease compared to those under or equal to 56 years while controlling for the other variables in the model. Individuals with chest pain of any type (typical angina, atypical angina, and non-anginal pain) have lower odds of developing heart disease compared to those with no chest pain while controlling for the other variables in the model (typical angina: OR=0.1537, 95% CI [0.0470,0.4613], p-value=0.001; atypical angina: OR=0.2031, 95% CI [0.0734,0.5234], p-value=0.001; non-anginal pain: OR=0.1748, 95% CI [0.0762,0.3847], p-value < 0.0001). The OR for exercise-induced angina is 2.8801 [95% CI [1.4119,5.9481], p-value=0.0038). This means that individuals who experience exercise-induced angina have 2.88 times higher odds of developing heart disease compared to those without angina while controlling for the other variables in the model. Finally, individuals with ST-T wave abnormalities in their resting ECG results have 8.3578 times (95% CI [0.5527, 239.0209, p-value=0.145) higher odds of developing heart disease compared to those with no resting ECG abnormalities, while controlling for the other variables in the model. However, this OR is not reliable due to its insignificance (p-value > 0.05) and very large confidence interval. On the other hand, individuals with left ventricular hypertrophy identified by ECG have 2.10 (95% CI[4.0555, 1.1069],p-value= 0.0245) times higher odds of developing heart disease compared to those with no resting ECG abnormalities while controlling for the other variables in the model. The confidence intervals for all the predictors are relatively large, especially that of resting ECG type 1 (ST-T wave abnormalities), which may mean that the results are not very precise and there is a greater degree of uncertainty about the true association between these variables and outcome. Overall, these odds ratios suggest that thalassemia has the greatest effect on heart disease in this cohort, followed by age group and then exercise-induced angina. All the results of this analysis are shown in Table 2.

Primary Outcome Assumptions

Independence Assumption: This assumption was satisfied for all models constructed in this study given that each observation in the dataset used represented a single unique patient, with no dependence between observations.

Linearity Assumption: The linearity assumption does not apply in this binary logistic regression model as there is no continuous predictor variable in the model. The linearity assumption assumes a straight line between the natural logs of the odds of the outcome variable and the continuous predictor variables. However, since this model does not have any continuous

variables, it assumes that the relationship between the categorical predictor variables and the log odds of the outcome variable is additive. This would mean that the effect of each category of the categorical variable is linearly related to the log odds of the outcome variable.

Multicollinearity Assumption: Multicollinearity refers to correlations between the predictor variables of the model. The presence of correlation would mean that there is instability in the model. The multicollinearity in a logistic regression model is typically assessed by examining the variance inflation factor (VIF) for each predictor variable. After running the VIF method on our final binary logistic regression model, we found no collinearity between any of the predictor variables (i.e. all VIF values were below 10 and around 1).

Influential Points and Outliers: To check for any influential outliers in our data, we assessed each data point using a Cook's distance plot. Based on this plot (Figure 2), we see that no points fall outside of a Cook's distance of 0.1. Therefore, there are no influential points in the data.

3.2.2. Secondary Outcome

An ordinal logistic regression model was also created to test the association between thalassemia and the severity of heart disease, determined by the degree of vessel narrowing. The final model after the categorization of the predictor 'age' had the following significant predictors: thalassemia (p-value < 0.0001), age group (p-value=0.0007), sex (p-value=0.015), chest pain type (p-values= 0.001, 0.001, 0.0001), exercise-induced angina (p-value=0.009) and resting ECG (p-value=0.037, 0.035). Cholesterol and resting blood pressure were again found to be non-confounding at a magnitude of 4.11% and 0.07%, respectively, and were thus omitted from the final model.

At a significant level of 0.05, the OR for thalassemia with respect to no thalassemia is 4.9941 (95% CI [2.7846,9.0925], p-value < 0.0001). This means that patients with thalassemia are 4.9941 times more likely to have more severe heart disease compared to patients without thalassemia. The 95% confidence interval does not contain 1, indicating that the association between thalassemia and heart disease is statistically significant. Individuals over 56 years have 2.5899 [95% CI 1.5126,4.4867, p-value=0.0007] times higher odds of having more severe heart disease compared to those under or equal to 56 years while controlling for the other variables in the model. The ORs of the other predictor variables yielded a similar result to that of the primary outcome analysis, where female patients and patients without chest pain are less likely to have a higher severity of heart disease, while patients who experience exercise-induced angina or have a resting ECG abnormalities have a higher likelihood of having more severe heart disease. The confidence intervals for all the predictors are again relatively large for this model, especially that for ST-T wave abnormalities, which may mean that the results may not be very precise and there is a greater degree of uncertainty about the true association between the variable and outcome. Overall, these odds ratios suggest that thalassemia has the greatest effect on heart disease, followed by age group and then exercise-induced angina. All the results of this analysis are shown in Table 3.

Assumptions
The independence, linearity, and multicollinearity assumptions were considered to be satisfied for the same reasons as those listed above for the primary outcome analysis. Additionally, the proportional odds assumption, which assumes that the effect of the predictors on the odds of being in a higher category of the response variable is the same across all levels of the response variables, was assessed for this model. The results of a Brant test showed that this assumption was potentially violated by the model, as indicated by an error message indicating empty cells, which meant that some combinations of the dependent variable and independent variables did not have any observations. Given this result, we attempted to reclassify the outcome variable into four or three categories instead of five; however, the same error was encountered each time. Since this assumption was not met, the results of this model may not be considered valid.

3.3. Objective 2 Results
A linear regression model was created to test the association between thalassemia and serum LDL cholesterol levels. After fitting the model without adjusting for any covariates, we found that a thalassemia diagnosis was not significantly associated with serum cholesterol levels, with a p-value of 0.969 at a significance level of α= 0.05. The result was the same even after adjusting for the covariates included in the previous models (age group, sex, chest pain type, exercise-induced angina, resting blood pressure, and resting ECG results), with a p-value of 0.7510. Due to the insignificance of the main exposure variable in this objective, any further analysis was discontinued.

4. Conclusion
Heart disease is a common and multifactorial condition that represents a major health concern. This cross-sectional study had two objectives. The first was to investigate the relationship between thalassemia and heart disease, examining not only the presence of thalassemia as a risk factor for heart disease but also its association with the severity of heart disease. The second was to explore the link between thalassemia and cholesterol levels to better understand the potential impact on heart disease. The descriptive statistics showed that patients with thalassemia were more often male and had a higher mean age and resting blood pressure than patients without thalassemia. Patients with thalassemia also experienced more exercise-induced angina and less chest pain. The binary logistic regression analysis found that thalassemia, age group, sex, chest pain type, exercise-induced angina, and resting ECG were significant predictors of heart disease. Patients with thalassemia were 5.2845 times more likely to develop heart disease compared to patients without thalassemia. Although the ordinal logistic regression model provided a similar conclusion, its results are considered to be invalid due to the violation of the proportional odds assumption. We also found that there was no significant difference in the serum LDL cholesterol levels between the exposed (thalassemia) and unexposed (no thalassemia) groups.

4.1 Limitations and Future Directions
In this study, two regression models were used to analyze the data; however, the proportional odds assumption was not met for the ordinal logistic regression model. This raises concerns

about the reliability of the ordinal regression model results. Furthermore, it can be seen from Table 1 that three out of five of the ordinal outcome groups had small sample sizes. This suggests that the ordinal regression model may not be appropriate for this particular dataset, despite the initial coding of the data. Therefore, alternative statistical methods may need to be explored to analyze the data effectively, and caution must be exercised in interpreting the results of the ordinal regression model. Overall, careful consideration of the assumptions and limitations of statistical models is critical in ensuring the validity and reliability of data analysis in research.

The dataset used for the study did not differentiate between the two types of thalassemia, major and minor, which may have impacted the results obtained. Therefore, we opted to consider the presence vs. absence of the disease instead. However, if the dataset had differentiated between these two types of thalassemia, the findings for cholesterol levels may have been different.

In regards to our statistical power calculation, the calculated power of 0.9563 was high despite the fact that the estimated confidence intervals were quite wide, indicating a large level of uncertainty in our results. We point to the fact that this sample was a convenience sample, which is non-random and may introduce sampling bias. This can impact the precision and generalizability of the effect size of our study. As the effect size is based on the sample data used for analysis, if the convenience sample used in the analysis is not representative of the target population and contains higher variability within the data, the estimated effect size may not accurately reflect the true effect size in the population. As a result, the confidence interval based on this estimated effect size may be wider, accounting for the uncertainty associated with estimating the effect size from a non-representative sample.

5. Contribution Statement

Emma Bogner fitted models for both objectives; carried out data cleaning; created result tables and figures; and carried out an overall review of this paper. Meghana Kompally adjusted models with respect to the age variable and refitted the models for objectives 1 and 2; created the flow chart; and contributed to the Results, Abstract, and Methods sections of this paper. Jason Wong carried out a literature review; completed the power calculation; checked model assumptions; and contributed to the Introduction, Abstract, and Methods sections of this paper. Serena Sun fitted the ordinal regression models with collapsed categories; checked assumptions; and contributed to the Introduction, Methods, and Conclusion sections of this paper.

6. References

Aessopos, A., et al. *Heart Disease in thalassemia intermedia: A Review of the underlying pathophysiology*. *Haematologica*, vol. 92, no. 5, May 2007, pp. 658–65. DOI.org (Crossref), https://doi.org/10.3324/haematol.10915.

Champely, S. (2020, March 16). PWR: Basic functions for power analysis. Retrieved April 16, 2023, from https://cran.r-project.org/web/packages/pwr/pwr.pdf

Cleveland Clinic. (2022). *Understanding Cholesterol Levels and Numbers*. Retrieved March 6, 2023, from https://my.clevelandclinic.org/health/articles/11920-cholesterol-numbers-what-do-they-mean

Cohen, A. R., Galanello, R., Pennell, D. J., Cunningham, M. J., & Vichinsky, E. (2004). *Thalassemia*.

Dua, D. and Graff, C. (2019). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA.

Mosca, Lori, et al. “Sex/Gender Differences in Cardiovascular Disease Prevention: What a Difference a Decade Makes.” Circulation, vol. 124, no. 19, Nov. 2011, pp. 2145–54. DOI.org (Crossref), https://doi.org/10.1161/CIRCULATIONAHA.110.968792.

Ricchi, P., Ammirabile, M., & Maggio, A. (2011). *Hypocholesterolaemia in Thalassaemia – Pathogenesis, Implications and Clinical Effects*. European Haematology. https://doi.org/10.17925/EOH.2010.04.0.20

7. Tables and Figures

Figure 1. Flow Chart of screening and inclusion process

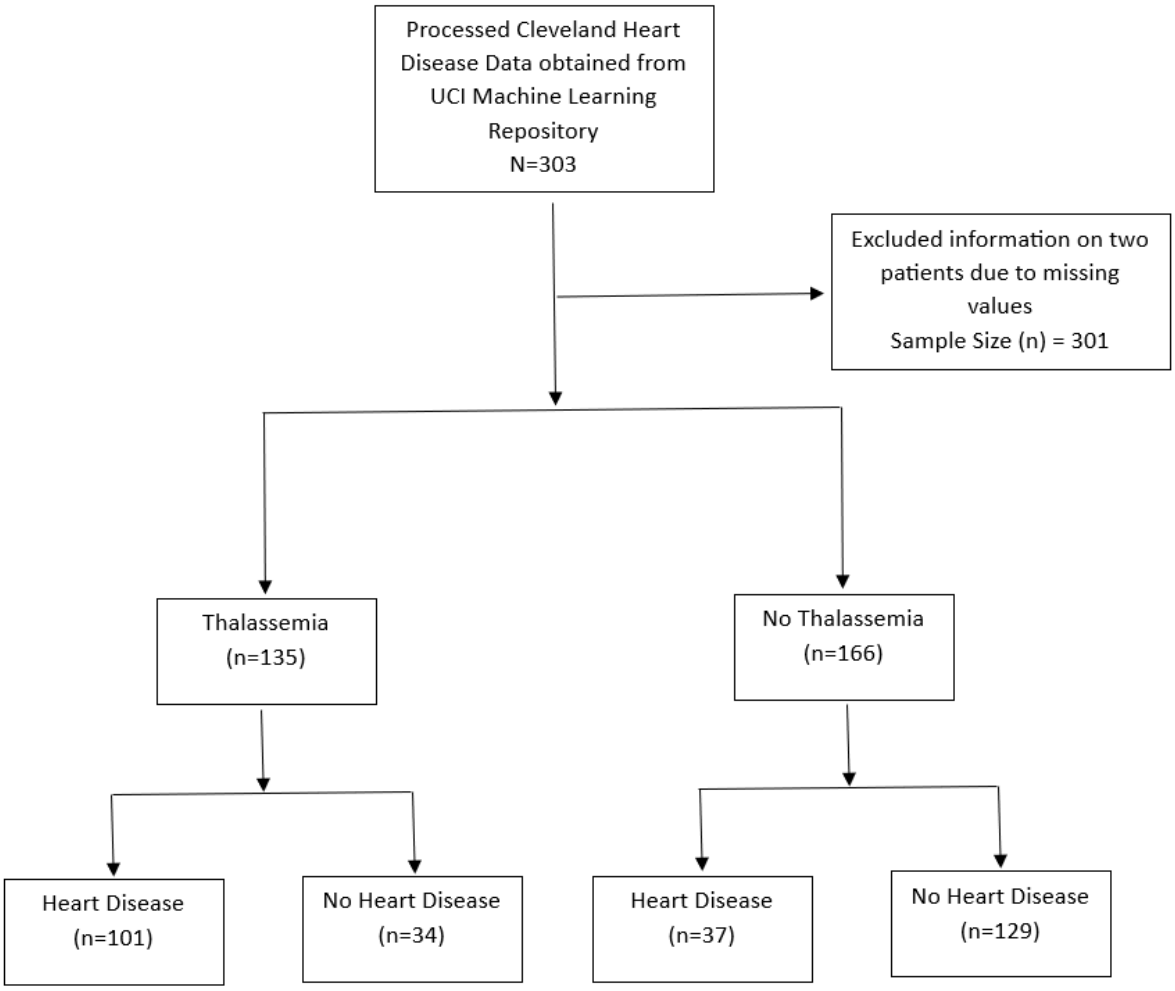

Table 1. Descriptive statistics for selected study participants grouped by exposure (thalassemia) status.

	No (N=166)	Yes (N=135)	Overall (N=301)
Heart Disease			
No (<50% vessel narrowing)	37 (22.3%)	101 (74.8%)	138 (45.8%)
Yes (>50% vessel narrowing)	129 (77.7%)	34 (25.2%)	163 (54.2%)
Heart Disease Severity*			
0	129 (77.7%)	34 (25.2%)	163 (54.2%)
1	22 (13.3%)	33 (24.4%)	55 (18.3%)
2	7 (4.2%)	28 (20.7%)	35 (11.6%)
3	6 (3.6%)	29 (21.5%)	35 (11.6%)
4	2 (1.2%)	11 (8.1%)	13 (4.3%)
Age (Years)			
Mean (SD)	53.4 (9.80)	55.8 (7.92)	54.5 (9.07)
Serum Cholesterol (mg/dL)			
Mean (SD)	247 (48.7)	247 (55.7)	247 (51.9)
Resting Blood Pressure (mmHg)			
Mean (SD)	130 (16.0)	134 (19.2)	132 (17.7)
Sex			
Male	86 (51.8%)	119 (88.1%)	205 (68.1%)
Female	80 (48.2%)	16 (11.9%)	96 (31.9%)
Resting ECG Results			
Normal	85 (51.2%)	65 (48.1%)	150 (49.8%)
Left ventricular hypertrophy**	79 (47.6%)	68 (50.4%)	147 (48.8%)
ST-T wave abnormality***	2 (1.2%)	2 (1.5%)	4 (1.3%)
Chest Pain (Type)			
Asymptomatic	53 (31.9%)	90 (66.7%)	143 (47.5%)
Atypical angina	39 (23.5%)	11 (8.1%)	50 (16.6%)
Non-anginal pain	61 (36.7%)	24 (17.8%)	85 (28.2%)
Typical angina	13 (7.8%)	10 (7.4%)	23 (7.6%)
Exercise-Induced Angina			
No	135 (81.3%)	68 (50.4%)	203 (67.4%)
Yes	31 (18.7%)	67 (49.6%)	98 (32.6%)

* 0 indicates absence and 5 indicates highest severity (greatest vessel narrowing).
** T wave inversions and/or ST elevation or depression of > 0.05 mV.
*** Showing probable or definite left ventricular hypertrophy by Estes' criteria.

Table 2. Estimated odds ratios for heart disease (presence vs. absence) determined using binary logistic regression.

Significant Predictors	**OR**	**95% CI Lower Bound**	**95% CI Upper Bound**	**p-value**
Thalassemia (Presence vs. absence*)	5.2845	2.7118	10.5854	< 0.0001
Age Group (Over 56 vs. under 56*)	3.2539	1.7012	6.3953	0.0005
Sex (Female vs male*)	0.3318	0.1476	0.7264	0.0064
Chest Pain (Typical angina vs asymptomatic*)	0.1537	0.0470	0.4613	0.0012
Chest Pain (Atypical angina vs. asymptomatic*)	0.2031	0.0734	0.5234	0.0014
Chest Pain (Non-anginal pain vs. asymptomatic*)	0.1748	0.0762	0.3847	< 0.0001
Exercise-Induced Angina (Presence vs. absence*)	2.8807	1.4119	5.9481	0.0038
Resting ECG (ST-T wave abnormality vs. normal*)	8.3578	0.5527	239.0209	0.1454
Resting ECG (Left ventricular hypertrophy vs normal*)	2.1000	1.1069	4.0555	0.0245

* Reference group.

Figure 2. Cook's distance plot for binary logistic regression model.

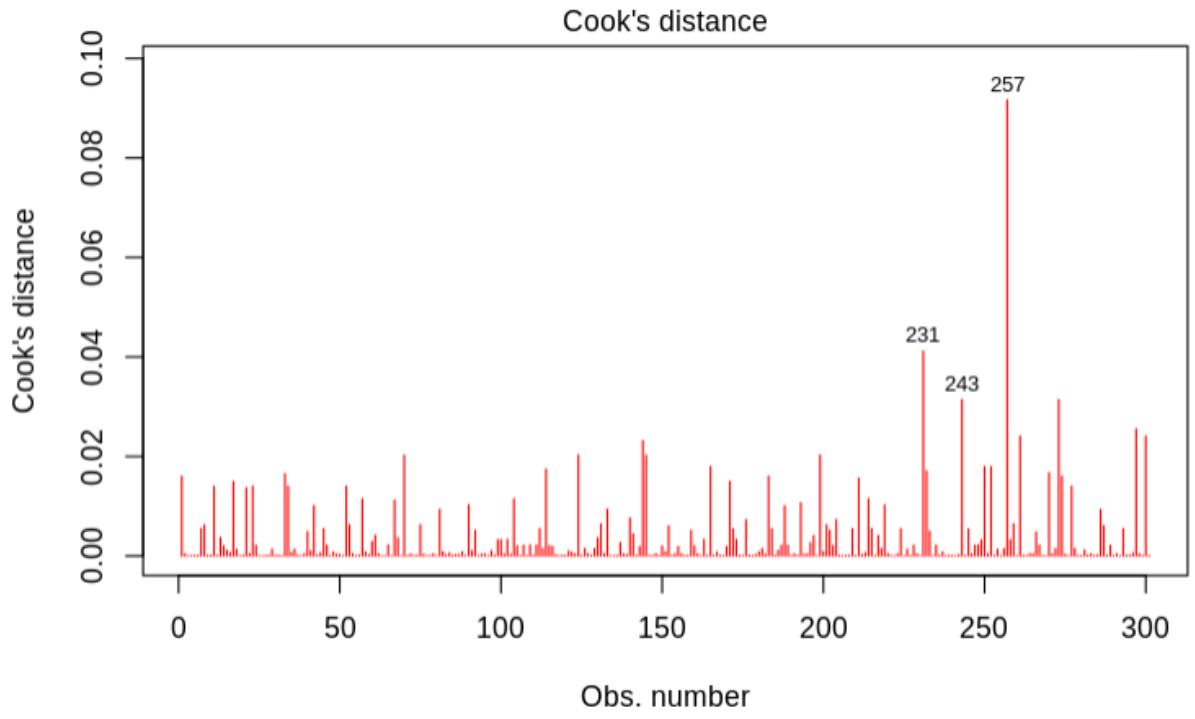

Table 3. Estimated odds ratios for the severity of heart disease determined using ordinal logistic regression.

Significant Predictors	**OR**	**95% CI Lower Bound**	**95% CI Upper Bound**	**p-value**
Thalassemia (Presence vs. absence*)	4.9941	2.7846	9.0925	< 0.0001
Age Group (Over 56 vs. under 56*)	2.5899	1.5126	4.4868	0.0007
Sex (Female vs male*)	0.4296	0.2154	0.8411	0.0153
Chest Pain (Typical angina vs asymptomatic*)	0.1715	0.0563	0.4747	0.0012

Chest Pain (Atypical angina vs. asymptomatic*)	0.2102	0.0822	0.4989	0.0007
Chest Pain (Non-anginal pain vs. asymptomatic*)	0.2065	0.0992	0.4163	< 0.0001
Exercise-Induced Angina (Presence vs. absence*)	2.2189	1.2267	4.0059	0.0086
Resting ECG (ST-T wave abnormality vs. normal*)	14.8768	1.3797	347.7926	0.0374
Resting ECG (Left ventricular hypertrophy vs normal*)	1.7829	1.0435	3.0667	0.0359

* Reference group