universidade de aveiro theoria poiesis praxis

Statistical Analysis of the Cleveland Heart Disease Dataset

2024/2025

Afonso Duarte do Fundo Ruela Branco Carreira (107988)

Daniel Machado de Melo (107444)

Marta Francisca dos Santos Carvalho (107664)

Tomás Vasconcelos Branco Serras Geraldes (107508)

Master's in Clinical Bioinformatics

Master's in Medical Statistics

Fundamentals of Medical Statistics

Prof. Vera Afreixo

Abstract

Cardiovascular disease (CVD) is the main cause of death in the world, especially given the complex nature of this condition and the fact that is commonly present with other diseases. Diagnostics, prognostics and prevention techniques are not completely understood and/or are underdeveloped, and the most advanced procedures nowadays are invasive. Therefore, accounting for these problematics, cardioinformatics could be the key to potentiate health services, using the already available data from millions of patients which had little purpose until now, such as patterns in lifestyle and patients' characteristics that could be predictive of CVD. To achieve this, Machine Learning models coupled with the statistical studies of the populations for analysis and prediction of health trends, if used accordingly, would reduce invasive procedures and health costs, and possibly allow for a more precise and personalized therapy. In this paper, following Ciu and Oetama's work (2020), the Cleveland Heart Disease dataset was used, and a Logistic Regression predictive model was obtained, with accuracies of 90.0% (Confusion Matrix) 87.5% (10 fold-Cross Validation) and 85.8% (Leave One Out Cross Validation) for risk of CVD assessment, respectively, proving the capabilities of this tool on identification of patterns and links between factors that cause CVD.

1. Introduction

Cardiovascular disease (CVD) is the number one cause of death in the world, and even with the advancements in cardiovascular medicine resulting in the reduction of the rate of death due to heart disease and stroke, it is still projected that, in 2030, there will be approximately 25 million deaths worldwide (1).

Known obstacles in the treatment of the disease are the disparities in healthcare access that are influenced by racial, ethnic and socioeconomic differences. Patient genotype analysis is extremely important to establish possible underlying factors in heart disease, but this technique is not used often in healthcare. Besides this, traditional CVD risk factors such as smoking, high serum cholesterol levels, diabetes and hypertension are more prevalent in lower socioeconomic levels, also having a disproportionate burden of CVD morbidity and mortality.

Lately, there have been more associations of more genetic variants to human diseases thanks to genome-wide association studies (2), which not only allows for more personalisation of future treatments, but also the stratification of clinical trials and testing of treatments in more targeted subpopulations. In addition, another challenge in CVD diagnosis and research is the spectrum of the disease, with a transition from a generally healthy state, following by a progression of the disease and general worsening of the patient well-being. The monitorization of such progression is a particularly hard feat to achieve, underlining the need for the development of new data integration approaches, patient stratification, and near-constant surveillance of all CVD patients. Yet another source of complications is the very frequent co-morbidity with other disease phenotypes, such as diabetes and obesity. Being CVD a multifactorial disease, its diagnosis is time-consuming and complex, as well as treatment options. For this condition, there are no specific biomarkers and the symptoms in the early stage are transversal to several conditions, failing to recognise CVD as an underlying cause of death, but rather as a mediator, leading to under-detection of the disease and misattribution to conditions like hypertension or cardiomyopathy (3).

Recognizing the disadvantages that patient invasion has, the branch of bioinformatics and statistics aims to find patterns in the population that are predictive of CVD. Since bioinformatics is currently at the centre of precision medicine, efforts in bringing personalized medicine to CVD patients are being made. In more depth, Khomtchouk and

colleagues coined the term “cardioinformatics” to join the fields of cardiology and bioinformatics (2). With this, the authors aim to shed some light in often forgotten areas of medical bioinformatics and also to improve research and prediction of cardiovascular diseases, through the usage of techniques such as machine learning and integration of various areas of biology and medicine (such as multi-omics and genetics, the ladder with characteristics such as epigenetic changes, RNA and protein expression profiles, and genome sequences). Integration and coordination of all these areas and techniques is key to achieve the most personalized and precise treatment options possible.

The problem with cardioinformatics

All clinical trials and treatments have patient data included that can range from simple administrative or demographic data to highly confidential information, such as medical records or illnesses (4). Data sharing is often regarded as a tool to help scientific advancements and to help personalize patient treatments and machine learning techniques, but the line between patient data sharing and ethics concerns is very thin, with data sharing often being a source of compromise to the patient and even possibly subjecting them to discrimination. Nevertheless, there has been an explosion in the formation and research of biological datasets for machine learning purposes (usually very large in size), which are coupled with cutting-edge technologies such as high-throughput techniques (allowing for the simultaneous processing of very large datasets) and permitting the research, training and testing of predictive and analytical models. However, if we focus solely on CVD databases, it is found that many of the resources are discontinued, not well-maintained, or very archaic. In addition, most high-throughput data in CVD research has been linked to very broad omics (in essence, genomics).

Artificial intelligence and the advancement of cardiology

In the big data era, it is fundamental to adopt innovative technology to facilitate the retrieval and analysis of information. One of the main goals in the usage of artificial intelligence and machine learning in the healthcare area is the liberation of time-consuming activities (such as reading literature) or recognition of metabolite patterns in patients that can easily signal issues with the patient’s health.

An emerging trend in the cardioinformatics area is the development of heart failure prediction and classification models, which use various response variables (which include common factors associated with heart disease and categorical results of exams). After

proper training and validation, a predictive model like this is an excellent gateway for providing both personalized and precise medicine.

2. Data analysis in *Ciu and Oetama.*, 2020

In our reference paper , Ciu and Oetama (5) provide a comprehensive description of a logistic regression model that was trained from a set of patients' data from the Cleveland (Ohio, United States of America) region. The reference group used for derivation of the dataset comprehended 303 patients referred for angiography at a Cleveland clinic between May 1981 and September 1984. No patients had history of CVD, and a thorough examination, which included an electrocardiogram, fasting sugar and blood cholesterol level measurements, and non-invasive tests, was made to the patients. The original dataset is comprehended in 76 attributes: however, it is worth noticing that the authors and other machine learning researchers use only 14 variables out of the 76 attributes, since part of the excluded variables are of a confidential nature and others represent tests and exams that may be too specific for researchers outside of the cardiology area.

Four files are available from the repository, but the authors (and by proxy, us) opted for the Cleveland database, since it is the only fully-processed database by the creators, and, according to the authors of the original paper, is the only subset of the database that was derived from a population referred for angiography - thus, it is expected that the best performance in models would come from this subset (6).

3. Methodology

Data retrieval and database creation

As previously stated, we replicated the Cleveland Clinic dataset used by Ciu and Oetama, with the dataset being publicly available at the UCI data repository (7). All data analysis was done with recursion to the *RStudio* software, with *R* version 4.4.2.

The dataset contains 14 variables organized in 303 observations. Briefly, the database was organized in columns of various variables, with each line corresponding to a different patient. **Table 1** illustrates the structure of the original database, with the database also being provided in the supplementary file **"processed.cleveland.data"**.

Table 1. Head of the Cleveland dataset. Variable names were added after import to improve understanding. Variable names and their explanations are depicted in Table 2.

Age	Sex	CP	Trestbps	Chol	Fbs	Restecg	Thalach	Exang	Oldpeak	Slope	Ca	Thal	Num
63	1	1	145	233	1	2	150	0	2.3	3	0	6	0
67	1	4	160	286	0	2	108	1	1.5	2	3	3	2
67	1	4	120	229	0	2	129	1	2.6	2	2	7	1
37	1	3	130	250	0	0	187	0	3.5	3	0	3	0
41	0	2	130	204	0	2	172	0	1.4	1	0	3	0
56	1	2	120	236	0	0	178	0	0.8	1	0	3	0
62	0	4	140	268	0	2	160	0	3.6	3	2	3	3
57	0	4	120	354	0	0	163	1	0.6	1	0	3	0
63	1	4	130	254	0	2	147	0	1.4	2	1	7	2

Missing values on the database were marked by the authors with a question mark. We opted for the usage of the median to correct the missing values, since the usage of the average would not be possible due to the variable levels being coded as specific numbers. **Table 2** depicts the variables, type of variable, meaning, and levels of the variables used in our dataset, after data treatment.

Table 2. Names, types, meaning and levels of the variables used in the dataset after data treatment.

*Typical angina pectoris is a type of pain that occurs in the thorax, neck, shoulders, jaw or arms with exertion and is relieved within 20 minutes by rest: if not relieved in 20 minutes, the patient has atypical angina. Non anginal pain is pain not located on any of the locations mentioned, or not related to exertion.

**A fixed abnormality is any type of defects observed during exercise/ischemia that persisted after redistribution (that is, rest). Reverse abnormality regards any defects that were managed to be corrected.

Variable	Variable type	Meaning	Levels
Age	Continuous quantitative	Age of patient.	Not applicable.
Sex	Nominal qualitative	Gender of patient.	0 - Female 1 - Male
CP	Nominal qualitative	Chest pain type.	1 - Typical angina 2 - Atypical angina 3 - Non-anginal pain 4 - Asymptomatic
Trestbps	Continuous quantitative	Resting blood pressure, in mmHg, at admission to hospital.	Not applicable.
Chol	Continuous quantitative	Serum cholesterol in mg/dL.	Not applicable.
Fbs	Nominal quantitative	Fasting blood sugar.	0 if FBS≤120 mg/dL 1 if FBS>120 mg/dL
Restecg	Nominal quantitative	Resting electrocardiographic outcome.	0 - normal 1 - ST-T wave abnormality 2 - Probable or definitive left ventricular hypertrophy.
Thalach	Continuous quantitative	Maximum thalach heart rate.	Not applicable.
Exang	Nominal quantitative	Exercise-induced angina *.	0 - No 1 - Yes
Oldpeak	Continuous quantitative	ST variable depression caused by exercise relative to rest.	Not applicable.
Slope	Nominal quantitative	Slope of the peak exercise ST segment.	1 - Upsloping 2 - Flat 3 - Downsloping
Ca	Ordinal quantitative	Number of main vessels coloured by fluoroscopy.	0 - 0 vessels 1 - 1 vessel 2 - 2 vessels 3 - 3 vessels
Thal	Nominal quantitative	Thallium-201 stress scintigraphy **.	3 - Normal 6 - Fixed defect 7 - Reversible defect
Num	Nominal quantitative	Heart disease in patient. **Target variable.**	0 - No heart disease in patient 1 - Patient diagnosed with heart disease.

Analytical strategy

The data analysis was planned according to the type of variables for which the data was available. The analytical strategy used in our work is as follows:

a. Characterization of the sample using exploratory data analysis.

A brief exploratory data analysis is included in Ciu and Oetama, but it is lacking compared to more exploratory data analysis-focused approaches. In a perspective not related to the original article, we opted for the inclusion of several plots, adequate to the nature of the variables, that allowed a better understanding of the variables and possible correlations between these. The plots were made with recursion to the *ggstatsplot* package, which is considered as an extension of the *ggplot2* package, used very commonly for programming graphics in the *R* language, and with incorporation of statistical analysis methods (*statsExpressions* package) (8). The package, developed in 2021 by I. Patil, is a promising approach for the combination of visually appealing graphics combined with personalized statistics according to the type of plot and analysis preferred. All statistic tests are chosen based on the data inputted into the graphic: more information about the process of selection is available in (9).

b. Usage of descriptive statistics and statistical inference.

Various statistical tests are applied to our dataset in order to measure possible robustness of the data, alongside choosing the best methods to apply for further data treatment. In this case, the normality of all numeric groups was evaluated using the Shapiro-Wilk test. Since the normal distribution among the numerical variables was not satisfied, non-parametric methods were employed, such as the Mann-Whitney U test, to establish significant differences among the groups of people with and without heart disease. For the categorical variables, we opted for the Chi-squared test to explore their relationship with heart disease.

c. Assessment of possible correlations between characteristics of patients.

After the statistical analysis of the groups, the correlations between variables were assessed, in the form of correlation plots and Cramer's V statistic, to give clues about the variables which will be of importance in the logistic regression model.

d. Creation of a logistic regression model to predict the outcome of heart disease.

A logistic regression model was created and trained to infer and predict the outcome of heart disease. We modulated the relation between heart disease presence in patients (target variable), and variables of importance. To generate the model, all values from the 303 patients were used. A stepAIC approach was also used to select the best variables for inclusion in the logistic regression model, and multicollinearity was evaluated. An analysis of quality, meaning and further validation was also applied to the model, with recursion to a confusion matrix, a k-fold cross validation, and a Leave One Out Cross Validation (LOOCV) approach.

Statistical considerations

For statistical inference, the significance level used was a p-value of 0.05. All statistical analysis and graphics were done with *RStudio*, with minor formatting editions being performed after.

Since the sample size is larger than 10, we opted for the Shapiro-Wilk normality test, to determine whether the sample derived from a normally distributed population. When the test was pointed out in the direction of not normally distributed data, we opted for the non-parametric approach (Mann-Whitney U/Wilcoxon sum rank test).

4. Results and Discussion

a. Exploratory Data Analysis

Regarding qualitative variables (Sex, Fasting Blood Sugar, Exercise-Induced Angina, Chest Pain Type, Resting ECG levels, Slope, Major vessels coloured by fluoroscopy, Thallium-201 stress scintigraphy), we opted for the usage of a two-way table paired with a bar plot, since these variables are categorical. In terms of quantitative (and continuous) variables, a bar plot was chosen as the default plot to present data. Besides the statistical analysis present in the subtitles, these bar plots are also overlayed with violin plots and present the median values for each group.

All graphics regarding EDA are present in **Supplementary Information I** and were plotted with the variable in question across the target variable of heart disease presence (Num).

Regarding the Sex variable **(Figure 1A)**, a large discrepancy between genders and heart disease is noticeable: in the dataset, 82% of men (n=114) have heart disease,

compared to only 19% of women (n=25), on a total of 139 patients. Regarding absence of heart disease, the gender distribution is more balanced, with 56% of male patients and 44% of female patients, in a total of 164 patients.

Next, Fasting Blood Sugar **(Figure 1B)** does not seem to be related to the presence of heart disease, with 14% of people without heart disease and 16% of people with heart disease having a fasting blood sugar level of over 120 mg/dL (n=45).

In terms of Exercise-Induced Angina **(Figure 1C)**, a very disproportional distribution of patient groups is seen for instance, in the group of people who have exercise-induced angina (n=99), 76 of these 99 patients have heart disease, which amounts to approximately 75% of the group. People who do not have heart disease (excluding possible misdiagnosis) may have underlying conditions that induce angina in exercise, such as asthma. This is also noticeable in the Chest Pain Type **(Figure 1D)** bar plot, which sees 25 and 24 percent of patients with atypical and typical angina, respectively, but without heart disease. In the group of patients with heart disease, typical angina is present in 76% of the group (105 people out of 139). In terms of Resting ECG **(Figure 1E)**, there seems to have an overall very low ST-T wave abnormality in both groups of patients, and normal ECG and left ventricular hypertrophy seems to be balanced in both groups.

In terms of the slope of peak exercise ST segment **(Figure 1F)**, surprisingly, most patients without heart disease (65%) have an upsloping slope peak exercise ST, which is a common indicator in myocardial infarction (9).

Regarding the number of vessels coloured by fluoroscopy **(Figure 1G)**, the EDA seems to be on par with typical values: the majority of patients without heart disease have 0 vessels coloured (81%, n=133) since these patients usually do not need imaging techniques for the heart. The other patient groups may have been subjected to colouring of vessels since they may have an underlying condition. In the group of patients who have heart disease, the percentages of groups are more balanced, which is expected. In terms of Thallium-201 stress scintigraphy **(Figure 1H)**, patients with heart disease are significantly more likely to have any type of defect (fixed or reversible).

Regarding quantitative variables, the age groups **(Figure 2A)** in both heart disease groups seem to be balanced, which is corroborated by the median (52 years in non-HD patients vs. 58 years in HD patients, respectively). Surprisingly, resting blood pressure **(Figure 2B)** between groups had the same median (130 mmHg), but there are higher values of resting BP in patients with heart disease, with the BP values being higher than in the non-disease group which is expected. Regarding serum cholesterol **(Figure 2C)**, a

notable outlier is noticeable in the non-heart disease group, but there does not seem to have an association between serum cholesterol levels and heart disease.

In terms of heart rate achieved (*Thalach*) **(Figure 2D)**, a largely Gaussian distribution is presented. Also, ST variable depression **(Figure 2E)** is more noticeable in the group of patients with heart disease, which is on par with the slope analysis made previously. However, this only considers down sloping slopes, with is the minority of the ST slopes seen previously: thus, no further conclusions can be taken, since it is expected that most values are of around zero.

b. Hypothesis testing

Table 3 depicts the summary of all statistical tests made to the variables. Regarding the continuous variables, we used the Mann-Whitney U (also known as Wilcoxon Rank sum test) that allow us to verify significant differences with the following hypothesis:

H_0: "Individuals with CVD do not differ significantly from those without CVD in the studied variables."

H_1: "Individuals with CVD differ significantly from those without CVD in the studied variables."

For the categorical variables, we used a Chi-squared test, which was able to statistically assign the association between the variables and the CVD diagnostic with the following hypothesis:

H_0: "There is no significant association between the variable in study and the presence of CVD."

H_1: "There is a significant association between the variable in study and the presence of CVD"

With these tests, we were able to assess significant differences in risk factors evaluated between the individuals with and without CVD. Patients with CVD tend to have higher values in age, cholesterol levels and ST depression. Besides, in a demographic view, a higher proportion of males was observed with CVD propounding a significant relation between gender and CVD. As expected, symptoms in CVD, such as chest pain, abnormal ECG findings, exercise-induced angina and specific ST slopes were more commonly observed in affected individuals. Variables including Thallium-201 stress scintigraphy patterns and number of major vessels coloured by fluoroscopy also showed significant difference between CVD patient groups. In all the studied variables, only the Fasting Blood Sugar variable was not significant different (p-value of 0.8).

Table 3. P-values of statistical tests performed to the data variables, alongside frequency and proportions of the values in the groups.

Characteristic	No N = 164[1]	Yes N = 139[1]	p-value[2]
Age	52.00 (44.50, 59.00)	58.00 (52.00, 62.00)	**<0.001**
Sex			**<0.001**
Female	72 (44%)	25 (18%)	
Male	92 (56%)	114 (82%)	
Chest Pain Type			**<0.001**
Typical angina	39 (24%)	105 (76%)	
Atypical angina	41 (25%)	9 (6.5%)	
Non-anginal pain	68 (41%)	18 (13%)	
Asymptomatic	16 (9.8%)	7 (5.0%)	
Resting Blood Pressure	130.00 (120.00, 140.00)	130.00 (120.00, 145.00)	**0.026**
Serum Cholesterol	234.50 (208.50, 267.50)	249.00 (217.00, 284.00)	**0.035**
Fasting blood sugar > 120 mg/dL	23 (14%)	22 (16%)	0.8
Resting ECG			**0.007**
Normal	95 (58%)	56 (40%)	
ST-T wave abnormality	1 (0.6%)	3 (2.2%)	
Left ventricular hypertrophy	68 (41%)	80 (58%)	
Max Heart Rate	161.00 (148.50, 172.00)	142.00 (125.00, 157.00)	**<0.001**
Exercise Induced Angina	23 (14%)	76 (55%)	**<0.001**
ST Depression	0.20 (0.00, 1.05)	1.40 (0.50, 2.50)	**<0.001**
Slope			**<0.001**
Upsloping	106 (65%)	36 (26%)	
Flat	49 (30%)	91 (65%)	
Downsloping	9 (5.5%)	12 (8.6%)	
Major Vessels colored by Fluoroscopy			**<0.001**
0	133 (81%)	47 (34%)	
1	21 (13%)	44 (32%)	
2	7 (4.3%)	31 (22%)	
3	3 (1.8%)	17 (12%)	
Thallium-201			**<0.001**
Normal	130 (79%)	38 (27%)	
Reversible defect	28 (17%)	89 (64%)	
Fixed defect	6 (3.7%)	12 (8.6%)	

[1] Median (Q1, Q3); n (%)

[2] Wilcoxon rank sum test; Pearson's Chi-squared test

c. Analysis of correlations

Regarding possible associations between numeric variables, **Figure 1** depicts the correlation plot between all numeric variables on the dataset, after data treatment. Spearman correction was applied since the data is non-parametric. Comparisons with low correlation levels are noted with a cross in the plot.

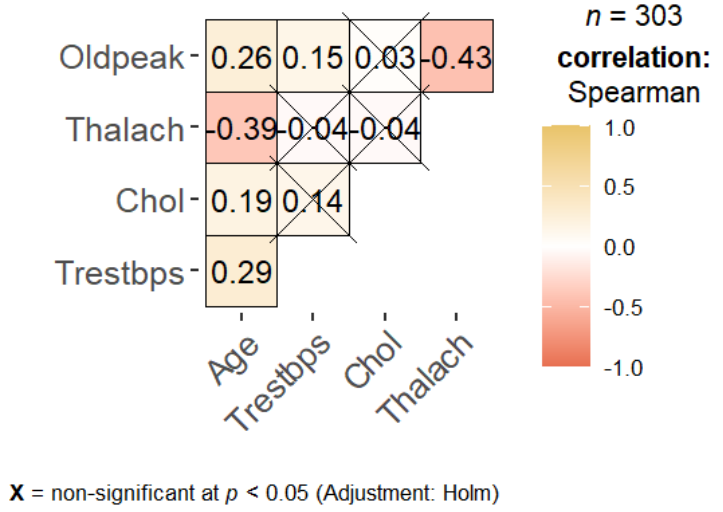


Figure 1. Correlation plot of numeric variables of the dataset. Spearman correlation was applied.

Looking further into **Figure 1**, a medium negative correlation between the Oldpeak and Thalach variables, alongside Thalach and Age variables, is noticeable. It is thus possible to infer that while one variable increments in value, the other one decreases in value. Also, a weak positive correlation is noticed between Oldpeak and Age, alongside Trestbps and Age, which infers that a small rise in one of these variables has the same effect on the other variable.

Regarding nominal variables, Cramer's V statistic was used to evaluate the strength of association between such variables. This value ranges from 0 to 1 (no association vs. perfect association). Cramer's V can be differently interpreted based on the number of degrees of freedom. **Supplementary Information II** depicts the table used for interpretation of Cramer's V (10). The Cramer's V statistic is present in every numeric variable plot in **Supplementary Information I**. An overall small association is seen in Sex, FBS, and Resting ECG, compared to heart disease diagnosis. Medium association is seen in Exercise-Induced Angina compared to heart disease diagnosis, and large

association is noticed in Chest Pain Type, Slope Peak ST segment, number of major vessels coloured by fluoroscopy, and Thallium-201 stress scintigraphy.

d. Logistic Regression Model

Similarly to Ciu and Oetama, we performed a logistic regression, with **Table 4A** presenting the summary of this regression, including the Odds Ratio (OR), Confidence Intervals (CI), and p-values for all variables, which were tested against the reference group in each variable (first line/lowest level of each variable category).

A stepAIC approach was used to evaluate which variables were of more importance for inclusion. It is inferred that the most suitable model for usage is HD ~ Sex + CP + Trestbps + Exang + Oldpeak + Slope + Ca + Thal, with an AIC of 222,24. Thus, we reformulated the logistic regression model, with its summary being presented in **Table 4B**.

Variables such as Sex, Trestbps, and Ca (number of vessels coloured by fluoroscopy) show a clear association with the presence of CVD (p-value < 0.05), while variables like Restecg and Age do not demonstrate statistical significance.

Regarding the OR of statistically significant variables, for males, the OR is 4.34, meaning the proportion of men at risk of heart disease is 4.3 times higher than that of women. In terms of the variable Trestbps, we infer that for each 1 mmHg increase in blood pressure (OR = 1.02), the probability of having CVD increases very slightly. Exercise-induced angina is associated with a 2 times higher risk of CVD (OR = 2.20), and certain slope categories are associated with a higher probability of CVD (4.30 times for flat waves and nearly 2 times for downsloping waves); Having 2 major vessels coloured by fluoroscopy (OR = 18.8) indicates a drastic increase in the likelihood of CVD diagnosis.

Table 4. Summary of the Logistic Regression models, for the model with all variables (A), and the most adequate model according to AIC (B), including Odds Ratio (and respective 95% confidence interval) and p-value for variable groups.

A)

Characteristic	OR[1]	95% CI[1]	p-value
Age	0.98	0.93, 1.02	0.327
Sex			
Female	—	—	
Male	5.50	1.94, 16.9	**0.002**
CP			
Asymptomatic	—	—	
Atypical angina	0.34	0.11, 1.01	0.058
Non-anginal pain	0.13	0.04, 0.34	**<0.001**
Typical angina	0.09	0.02, 0.33	**<0.001**
Trestbps	1.03	1.01, 1.05	**0.018**
Chol	1.00	1.00, 1.01	0.292
Fbs			
No	—	—	
Yes	0.69	0.22, 2.05	0.504
Restecg			
0	—	—	
1	2.82	0.05, 320	0.700
2	1.63	0.76, 3.55	0.216
Thalach	0.98	0.96, 1.00	0.108
Exang			
No	—	—	
Yes	2.09	0.87, 5.00	0.097
Oldpeak	1.49	0.94, 2.42	0.095
Slope			
Upsloping	—	—	
Flat	3.72	1.48, 9.77	**0.006**
Downsloping	1.80	0.26, 11.1	0.534
Ca			
0	—	—	
1	9.28	3.51, 26.4	**<0.001**
2	26.1	5.99, 133	**<0.001**
3	8.57	1.64, 63.0	**0.019**
Thal			
3	—	—	
6	0.76	0.16, 3.82	0.738
7	3.82	1.65, 9.09	**0.002**

[1] OR = Odds Ratio, CI = Confidence Interval

B)

Characteristic	OR[1]	95% CI[1]	p-value
Sex			
Female	—	—	
Male	4.34	1.67, 12.1	**0.004**
CP			
Asymptomatic	—	—	
Atypical angina	0.30	0.10, 0.86	**0.029**
Non-anginal pain	0.11	0.04, 0.27	**<0.001**
Typical angina	0.08	0.02, 0.28	**<0.001**
Trestbps	1.02	1.00, 1.05	**0.024**
Exang			
No	—	—	
Yes	2.20	0.95, 5.12	0.065
Oldpeak	1.60	1.04, 2.55	**0.038**
Slope			
Upsloping	—	—	
Flat	4.30	1.81, 10.7	**0.001**
Downsloping	1.87	0.31, 10.3	0.484
Ca			
0	—	—	
1	9.79	3.89, 26.5	**<0.001**
2	18.8	4.81, 84.8	**<0.001**
3	9.17	1.91, 65.2	**0.012**
Thal			
3	—	—	
6	0.78	0.18, 3.63	0.748
7	3.86	1.71, 8.95	**0.001**

[1] OR = Odds Ratio, CI = Confidence Interval

Simple Regression vs Multiple Regression

Comparing the results of simple and multiple regression (**Supplementary Information III**), some differences can be observed. First, regarding the Age variable, it is seen that in the multiple regression, age is not considered significant (p-value > 0.05), but in the simple regression, it is significant with an OR = 1.05, meaning that with age increase, individuals are more likely to develop CVD, which suggests that the significance of the variable in the simple analysis might be influenced by confounding factors or collinearity. The variable Exang is significant solely the simple regression, with its odds ratio (OR) being much higher in the simple regression (7.40) compared to the multiple regression (2.20). In other words, had we used the simple regression model, individuals with induced exercise angina have an almost 7.5 times increased likelihood of developing CVD, while in the multiple regression, the increase in likelihood would be of over 2 times. The lower OR value in the multiple regression reflects the effect of the Exang variable adjusted for all other variables in the model.

Training and testing of the logistic regression model

In order to train the logistic regression model, training and test datasets were prepared, with the training set being consisted of the first 250 patients and the other 53 having been part of the test dataset.

Figure 2 depicts the plots of the training and testing models, respectively.

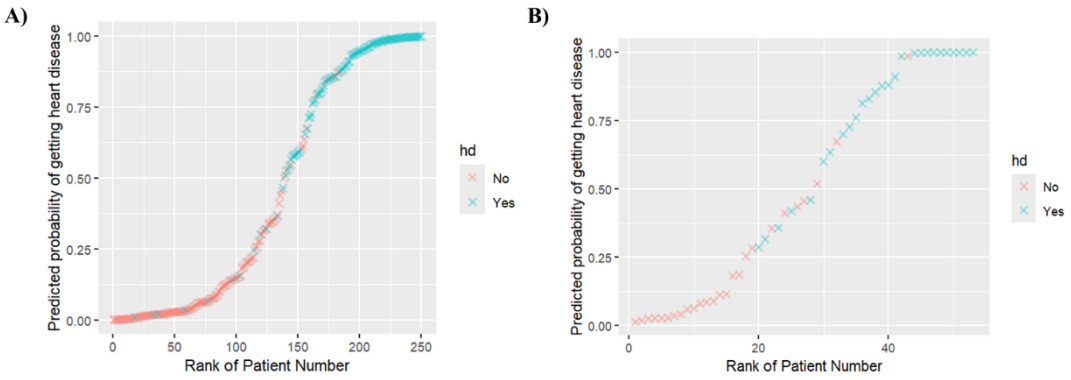


Figure 2. Logistic regression plots achieved with training (left) and testing datasets (right).

Multicollinearity was tested to infer possible correlation between variables. We reported no multicollinearity between variables, having Ciu and Oetama obtained the same results.

Three validation approaches were used: first, we opted for the confusion matrix, since it is one of the validation techniques used by the authors, with the other being AUC/ROC, which we also used. Lastly, a Leave One Out Cross Validation (LOOCV) was evaluated as an extra validation measure, since the authors recommended the usage of more than two validation methods. **Table 5** depicts the confusion matrix of our training dataset.

Table 5. Confusion Matrix of the training dataset.

		Reference	
		No	Yes
Prediction	No	126	12
	Yes	12	100

In comparison to Ciu and Oetama, we find that, overall, better results were obtained in terms of test statistics. However, direct comparison of the confusion matrices is not possible since the authors did not state the size of their training and testing set.

With this validation technique, we achieved an accuracy of 90.4%, which is an improvement compared to the 85,45% obtained by the authors. Specificity obtained is higher than the authors (0.8929 vs. 0.8041), as well as sensitivity (0.9130 vs. 0.8966). Both positive and negative predictive values are higher than the authors' values (PPV: 0.9130 vs. 0.8667; NPV: 0.8929 vs. 0.8455), with an overall balanced accuracy of 0.9030 for our confusion matrix and 0.8503 for the authors.

We also opted for the use of k-fold cross validation (k=10), and plotted a ROC curve, shown in **Figure 3**, which shows the relationship between sensitivity (probability of true positives) and 1-specificity (probability of false positives).

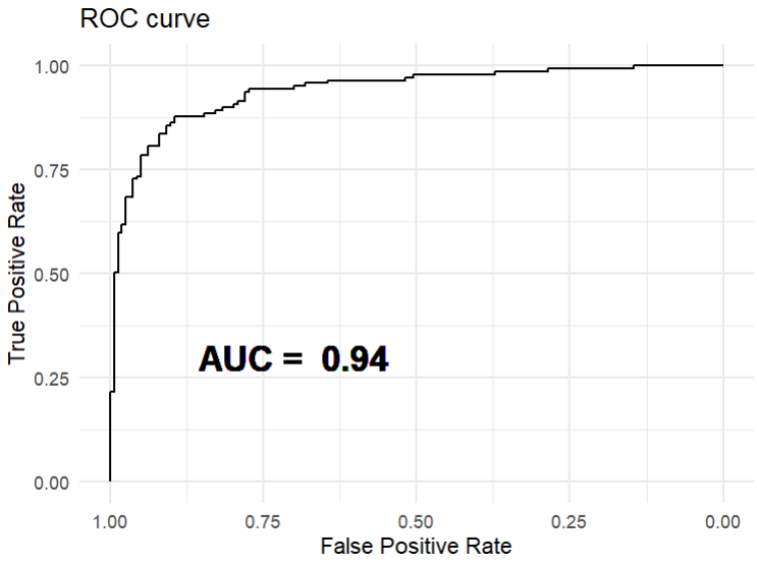


Figure 3. ROC curve and respective AUC.

Using the results from the 10-fold cross-validation (**Figure 4**), we observe that the model demonstrates strong overall performance, with a high AUC (0.94, DeLong 95% confidence interval: 0.9074-0.9644) close to 1 (indicating good discriminatory power), high sensitivity (85.3%) (accurate identification of positives), and a reduced number of false positives. The overall accuracy (0.891) is consistent with the other metrics, reinforcing that the model is reliable for correctly predicting both classes in approximately 89% of cases.

```
Generalized Linear Model

303 samples
 12 predictor
  2 classes: 'No', 'Yes'

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 272, 273, 273, 272, 273, 273, ...
Resampling results:

  ROC        Sens       Spec
  0.8936854  0.8595588  0.7978022

Average AUC:  0.894

Accuracy: 0.875
```

Figure 4. Results from 10-fold cross-validation.

To further improve performance, decision thresholds could be adjusted using the Youden index.

Lastly, the LOOCV (Leave One Out Cross Validation) method was also applied, since the dataset is of a small size, and may help in reducing bias and randomness. **Figure 5** depicts the results of the LOOCV model.

```
Generalized Linear Model

303 samples
  8 predictor
  2 classes: 'No', 'Yes'

No pre-processing
Resampling: Leave-One-Out Cross-Validation
Summary of sample sizes: 302, 302, 302, 302, 302, 302, ...
Resampling results:

  ROC        Sens       Spec
  0.9139323  0.8841463  0.8273381
```

Figure 5. Output of the LOOCV analysis.

From these results, we observe that the model demonstrates strong overall performance, with a sensitivity of 85.3% and a specificity of 79.1%, which is a better mark than the k-fold cross validation procedure. Besides this, specificity obtained was of 85.8%.

5. Conclusion

This study explored the Cleveland Heart Disease dataset to investigate risk factors and predictors for cardiovascular disease. Following the rejection of the normal distribution of our variables, tested by the Shapiro-wilk test, we employed the non-parametric test (Mann-Whitney U test), this analysis revealed a significant difference among the numerical continuous numerical variables (Age, Chol, Trestbps, Oldpeak and Thalach) in the patients with CVD and those without. The Chi-squared teste allowed us to conclude that in our categorical groups (Sex, CP, Fbs, Restecg, Ca, Exang, Thal and Slope), only Fasting Blood Sugar did not present a significant association within the individuals with CVD. The logistic regression model demonstrated strong predictive capabilities, as evidenced by an AUC of 0.94 in the 10-fold cross-validation, reinforcing its reliability for identifying individuals at risk. Variables such as Sex, blood pressure at rest (Trestbps), and number of major vessels coloured by fluoroscopy (Ca) exhibited strong associations with the CVD diagnosis, aligning with known clinical risk factors and

providing further validation of their importance. The logistic regression model achieved 90.4% accuracy in training, 87.5% accuracy during cross-validation, and 85.8% in LOOCV. While sensitivity and specificity were balanced, the model performed well in correctly predicting positive cases, suggesting its applicability in clinical screening scenarios.

6. References

1. Okwuosa IS, Lewsey SC, Adesiyun T, Blumenthal RS, Yancy CW. Worldwide disparities in cardiovascular disease: Challenges and solutions. Int J Cardiol. 2016 Jan 1;202:433–40.

2. Khomtchouk BB, Tran DT, Vand KA, Might M, Gozani O, Assimes TL. Cardioinformatics: the nexus of bioinformatics and precision cardiology. Brief Bioinform [Internet]. 2020 Dec 1 [cited 2024 Dec 8];21(6):2031–51. Available from: https://dx.doi.org/10.1093/bib/bbz119

3. Bozkurt B, Ahmad T, Alexander KM, Baker WL, Bosak K, Breathett K, et al. Heart Failure Epidemiology and Outcomes Statistics: A Report of the Heart Failure Society of America. J Card Fail [Internet]. 2023 Oct 1 [cited 2024 Nov 9];29(10):1412. Available from: https://pmc.ncbi.nlm.nih.gov/articles/PMC10864030/

4. Patient data and confidential patient information - NHS England Digital [Internet]. [cited 2024 Dec 8]. Available from: https://digital.nhs.uk/services/national-data-opt-out/understanding-the-national-data-opt-out/confidential-patient-information#patient-data-and-information

5. Ciu T, Oetama RS. Logistic Regression Prediction Model for Cardiovascular Disease. IJNMT (International Journal of New Media Technology). 2020 Jul 2;7(1):33–8.

6. Detrano R, Janosi A, Steinbrunn W, Pfisterer M, Schmid JJ, Sandhu S, et al. International application of a new probability algorithm for the diagnosis of coronary artery disease. Am J Cardiol. 1989 Aug 1;64(5):304–10.

7. Heart Disease - UCI Machine Learning Repository [Internet]. [cited 2024 Dec 8]. Available from: http://archive.ics.uci.edu/dataset/45/heart+disease

8. Patil I. Visualizations with statistical details: The “ggstatsplot” approach. J Open Source Softw. 2021 May 25;6(61):3167.

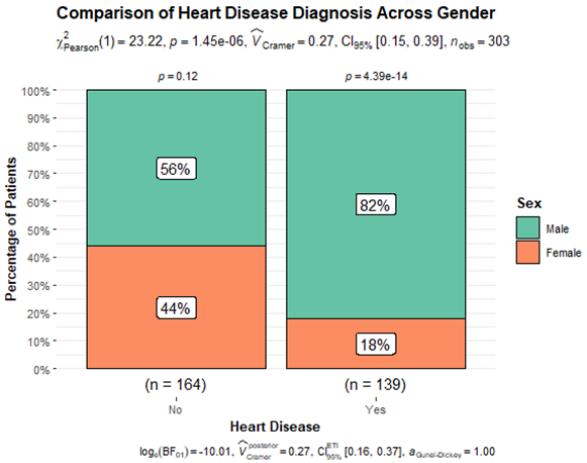
9. Patil I. statsExpressions: R Package for Tidy Dataframes and Expressions with Statistical Details. J Open Source Softw. 2021 May 20;6(61):3236.

10. How to Interpret Cramer’s V (With Examples) [Internet]. [cited 2024 Dec 12]. Available from: https://www.statology.org/interpret-cramers-v/

Supplementary Information I: Exploratory Data Analysis

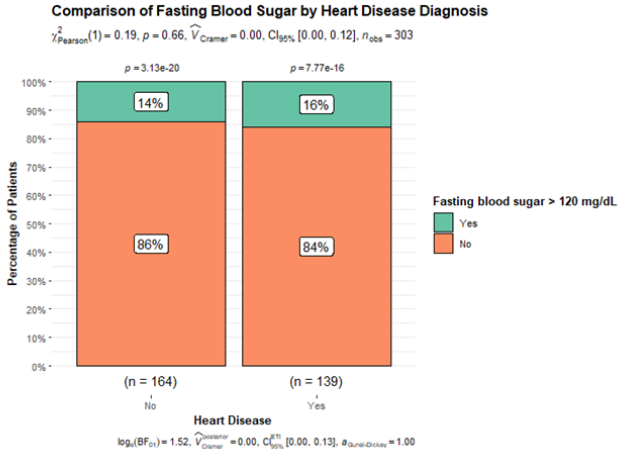
A)

	Sex		
	Female	Male	**Total**
Heart Disease			
No	72	92	164
Yes	25	114	139
Total	97	206	303



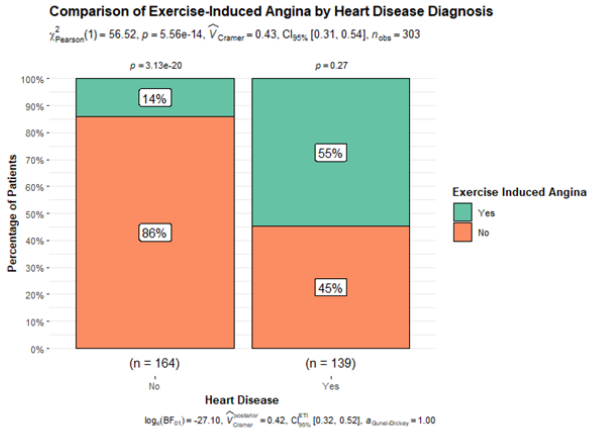
B)

	Fasting Blood Sugar > 120 mg/dL		
	No	Yes	**Total**
Heart Disease			
No	141	23	164
Yes	117	22	139
Total	258	45	303



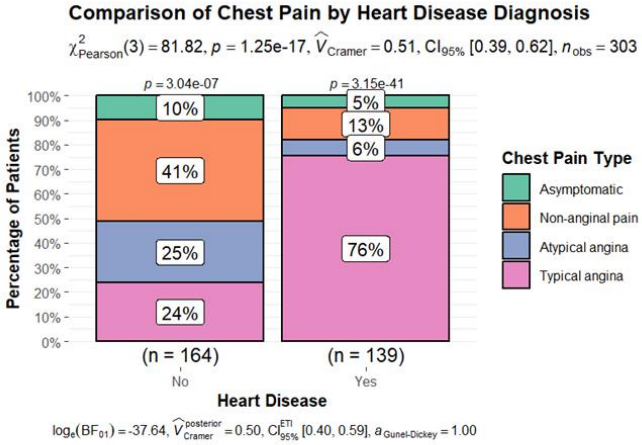
C)

	Exercise-Induced Angina		
	No	Yes	**Total**
Heart Disease			
No	141	23	164
Yes	63	76	139
Total	204	99	303



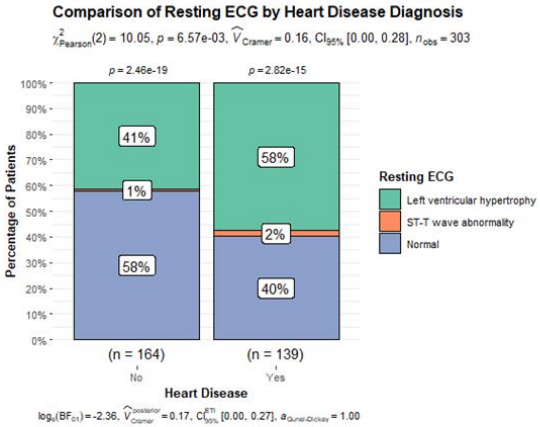
D)

	Chest Pain Type				
	Asymptomatic	Atypical angina	Non-anginal pain	Typical angina	**Total**
Heart Disease					
No	39	41	68	16	164
Yes	105	9	18	7	139
Total	144	50	86	23	303



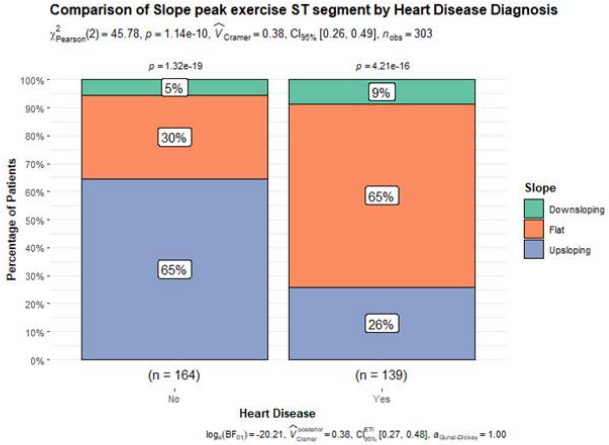
E)

	Resting ECG			
	Normal	ST-T wave abnormality	Left ventricular hypertrophy	**Total**
Heart Disease				
No	95	1	68	164
Yes	56	3	80	139
Total	151	4	148	303



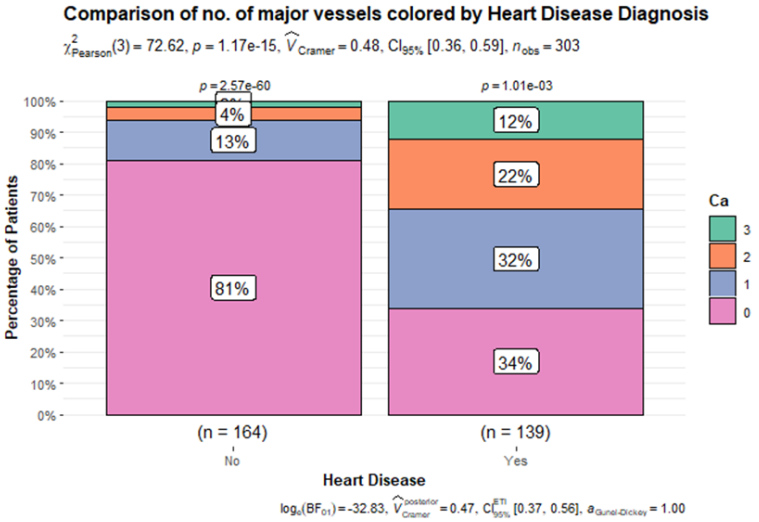
F)

	Slope of peak exercise ST segment			
	Upsloping	Flat	Downsloping	**Total**
Heart Disease				
No	106	49	9	164
Yes	36	91	12	139
Total	142	140	21	303



D)

	Number of major vessels colored by fluoroscopy				
	0	1	2	3	**Total**
Heart Disease					
No	133	21	7	3	164
Yes	47	44	31	17	139
Total	180	65	38	20	303



E)

	Thallium-201 Stress Scintigraphy			
	Normal	Reversible defect	Fixed defect	**Total**
Heart Disease				
No	130	28	6	164
Yes	38	89	12	139
Total	168	117	18	303

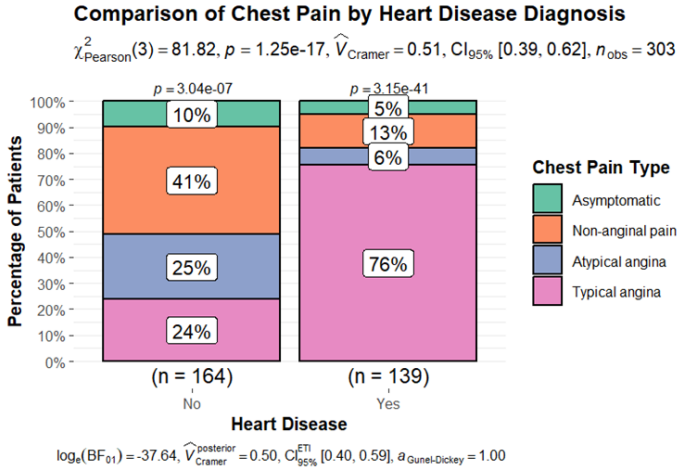


Figure I. Exploratory data analysis of qualitative variables. Association measures such as Cramer's V are present on top of the bar plots, and a frequency table is shown for every variable.

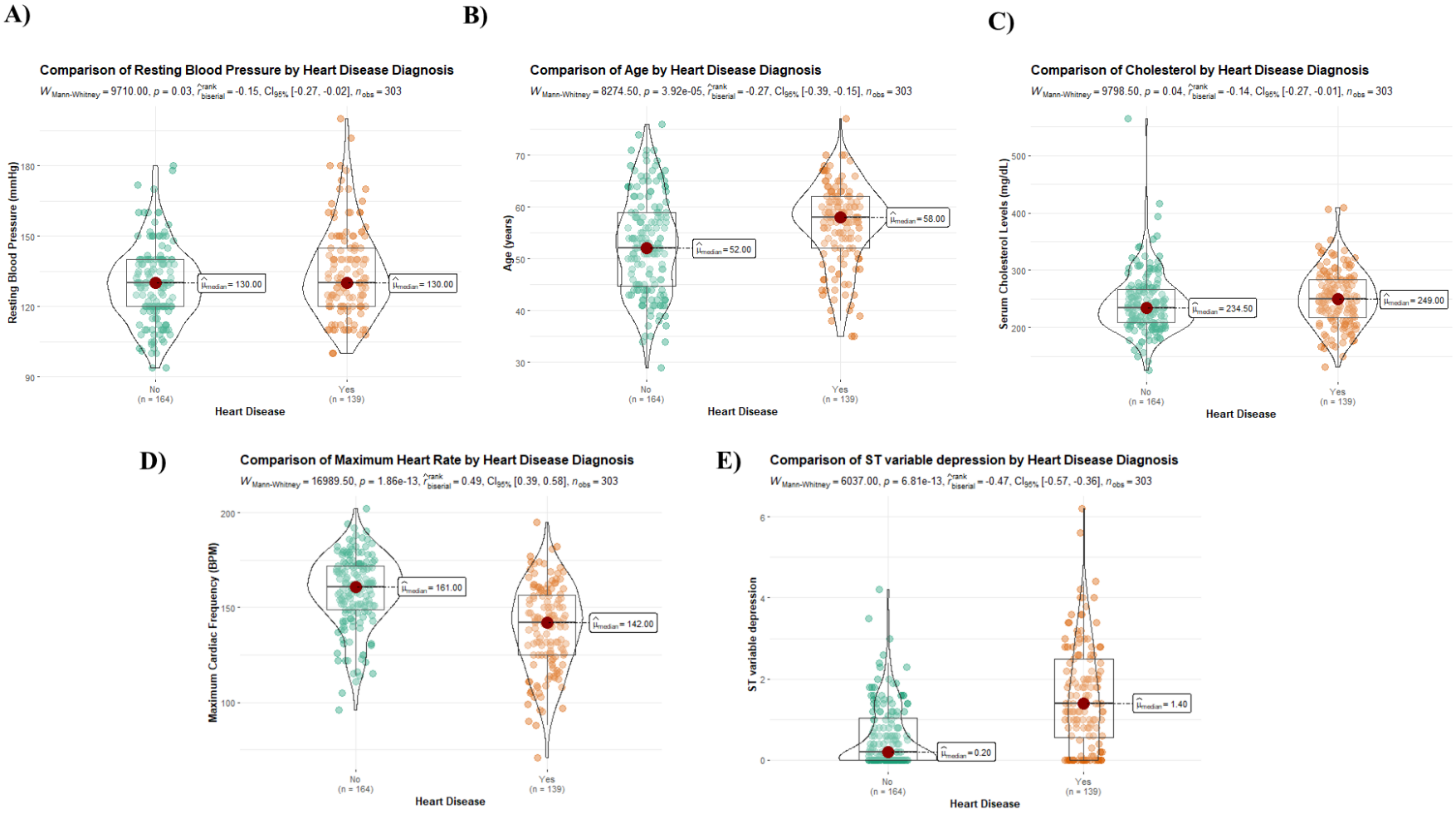


Figure II. Exploratory data analysis of quantitative variables.

Supplementary Information II: Cramer's V table

Table I. Cramer's V table used for measurement of association between nominal variables. Adapted from *(10)*.

Degrees of freedom	*Small*	*Medium*	*Large*
1	0.10	0.30	0.50
2	0.07	0.21	0.35
3	0.06	0.17	0.29
4	0.05	0.15	0.25
5	0.04	0.13	0.22

Supplementary Information III: Simple and Multiple Regression Tables

Table I. Simple Regression and Muliple Regression table obtained with the logistic regression model.

	Simples			Múltipla		
Characteristic	**OR**[1]	**95% CI**[1]	**p-value**	**OR**[1]	**95% CI**[1]	**p-value**
Age	1.05	03, 1.08	**<0.001**			
Sex						
Female	—	—		—	—	
Male	3.57	2.12, 6.16	**<0.001**	4.34	1.67, 12.1	**0.004**
CP						
Asymptomatic	—	—		—	—	
Atypical angina	0.08	0.03, 0.18	**<0.001**	0.30	0.10, 0.86	**0.029**
Non-anginal pain	0.10	0.05, 0.18	**<0.001**	0.11	0.04, 0.27	**<0.001**
Typical angina	0.16	0.06, 0.41	**<0.001**	0.08	0.02, 0.28	**<0.001**
Trestbps	1.02	1.00, 1.03	**0.010**	1.02	1.00, 1.05	**0.024**
Chol	1.00	1.00, 1.01	0.142			
Fbs						
No	—	—				
Yes	1.15	0.61, 2.18	0.660			
Restecg						
0	—	—				
1	5.09	0.63, 104	0.163			
2	2.00	1.26, 3.18	**0.003**			
Thalach	0.96	0.94, 0.97	**<0.001**			
Exang						
No	—	—		—	—	
Yes	7.40	4.31, 13.1	**<0.001**	2.20	0.95, 5.12	0.065
Oldpeak	2.51	1.94, 3.33	**<0.001**	1.60	1.04, 2.55	**0.038**
Slope						
Upsloping	—	—		—	—	
Flat	5.47	3.30, 9.23	**<0.001**	4.30	1.81, 10.7	**0.001**
Downsloping	3.93	1.54, 10.4	**0.004**	1.87	0.31, 10.3	0.484
Ca						
0	—	—		—	—	
1	5.93	3.24, 11.2	**<0.001**	9.79	3.89, 26.5	**<0.001**
2	12.5	5.45, 32.7	**<0.001**	18.8	4.81, 84.8	**<0.001**
3	16.0	5.11, 70.9	**<0.001**	9.17	1.91, 65.2	**0.012**
Thal						
3	—	—		—	—	
6	6.84	2.49, 20.8	**<0.001**	0.78	0.18, 3.63	0.748
7	10.9	6.31, 19.3	**<0.001**	3.86	1.71, 8.95	**0.001**

[1] OR = Odds Ratio, CI = Confidence Interval