Heart Disease Trends by Demographics: A Statistical Analysis

Annelise Thorn
Department of Mathematics, University of Colorado at Boulder
STAT 5000: Statistical Methods and Applications I
Dr. Matt Reichenbach
May 1, 2025

Introduction

"The American Heart Association estimates up to 90% of cardiovascular diseases may be preventable with education and action," yet in the United States, one person dies every 33 seconds from heart disease, and it has remained the leading cause of death for Americans for over 10 years (Drinan, 2023; CDC, 2024). Not only does heart disease negatively impact patients' health, but also, it is a burden on the healthcare system. From 2019-2020, heart disease cost the United States $252.2 billion in healthcare services and medicines and lost productivity due to death (CDC, 2024). To improve patient outcomes and reduce strain on the healthcare system, it is imperative to understand heart disease and what factors increase individuals' risk of developing heart disease. Heart disease serves as an umbrella term for several diseases and conditions affecting the heart, such as coronary artery disease and heart valve disease (Mayo Clinic, 2024). Prior studies have found that age and sex influence heart disease risk (American Heart Association, 2022). Building on these findings, this study conducts a statistical analysis to examine how demographic and clinical characteristics differ between individuals with and without heart disease and to identify which factors are significantly associated with heart disease outcomes. While lifestyle and genetic factors are known contributors to heart disease, this analysis focuses specifically on which demographic and clinical conditions are most strongly associated with the presence of heart disease. This study intends to answer the following research questions: (1) How do patients' demographics and clinical conditions differ between patients with and without heart disease? (2) Are certain groups of people at higher risk of developing heart disease? The following sections describe the dataset, methods of analysis, results of statistical tests, and concluding insights.

Background

The data for this study comes from Kaggle and is a master dataset composed of five observational heart disease datasets from The Hungarian Institute of Cardiology, The University Hospital, Zurich, Switzerland, The University Hospital, Basel, Switzerland, and V.A. Medical Center, Long Beach, and Cleveland Clinic Foundation. The dataset has 918 observations and includes the following features: age, sex, chest pain type, resting blood pressure, cholesterol, fasting blood sugar, resting electrocardiogram results, maximum heart rate, exercise-induced angina, oldpeak, the slope of the peak exercise ST segment, and heart disease. These factors are key cardiovascular and overall health indicators and provide a basis for understanding potential risk factors for heart disease.

One of the most notable indicators of cardiovascular health is “ChestPainType.” “ChestPainType” represents different types of chest pain and is composed of “TA” (typical angina), “ATA” (atypical angina), “NAP” (non-anginal pain), and “ASY” (asymptomatic). Angina is chest pain caused by the heart receiving too little oxygen-rich blood (American Heart Association, 2022b). The most common type of angina is stable angina, which most often occurs during physical activity or when you experience strong emotions (American Heart Association, 2022b). Atypical angina is chest pain that occurs at rest and is usually caused by reduced blood flow to the heart from fatty deposits clogging the arteries (American Heart Association, 2022a). Non-anginal pain is recurring noncardiac chest pain and is most commonly related to esophagus issues but can also be related to lung issues and mental health issues (Cleveland Clinic, 2022a).

Other clinical variables included in the study are “RestingBP” (resting systolic blood pressure, measured in mmHg), “Cholesterol” (total blood cholesterol in mg/dL), “FastingBS” (a binary indicator of fasting blood sugar >120 mg/dL), and “MaxHR” (maximum heart rate achieved during exercise). Resting blood pressure reflects arterial pressure during cardiac contraction (American Heart Association, 2024), and fasting blood sugar indicates baseline glucose levels after an 8–12 hour fast (Cleveland Clinic, 2021). The dataset also records “ExerciseAngina” (presence of chest pain during exercise, noted as “Y” for yes and “N” for no) and “HeartDisease” (binary outcome: 0 indicating no heart disease and 1 indicating the presence of heart disease). These clinical markers serve as the primary variables for the statistical analyses conducted in this study.

Table 1. Normal and Abnormal Ranges of Variables

Variable	Description	Normal Ranges	Abnormal Ranges
Chest Pain Type	Typical Angina, Atypical Angina, Non-Anginal Pain, Asymptomatic	ASY	TA, ATA, & NAP
Resting Blood Pressure	Resting blood pressure [mm Hg]	< 120	Elevated: 120-129, Hypertension stage 1: 130-139, Hypertension stage 2: ≥ 140-180, & Hypertension crisis: ≥ 180
Cholesterol	Total cholesterol [mm/dl]	< 200	At risk: 200-239 & Dangerous: ≥ 240

Fasting Blood Sugar	[1: if FastingBS > 120 mg/dl, 0: otherwise]	0	1
Max Heart Rate	Maximum heart rate	60-202	< 60 or > 202
Exercise Angina	Exercise-induced angina [Y: Yes, N: No]	N	Y

Normal and abnormal ranges for key demographic and clinical variables used in the analysis. Variables outside the normal ranges may indicate increased cardiovascular risk.

Methods

Analytical Software & Data Cleaning

The dataset was read into a Google Colab R kernel, and the columns "RestingECG," "Oldpeak," and "ST_Slope" were removed to narrow the scope of the study. Duplicate entries and cells with NAs and 0's were checked. The dataset did not contain any duplicate entries or cells with NAs, however, it did contain 0's in the "RestingBP" and "Cholesterol" columns, which were used to mark cells that did not have an entry. The 0s in these columns were replaced with the column mean. The remaining data was clean and usable, so no additional changes were made, and the cleaned data frame was used to perform statistical analysis.

Figure 1. First 5 Rows of Cleaned Data Frame

	Age	Sex	ChestPainType	RestingBP	Cholesterol	FastingBS	MaxHR	ExerciseAngina	HeartDisease
	<int>	<chr>	<chr>	<dbl>	<dbl>	<int>	<int>	<chr>	<int>
1	40	M	ATA	140	289	0	172	N	0
2	49	F	NAP	160	180	0	156	N	1
3	37	M	ATA	130	283	0	98	N	0
4	48	F	ASY	138	214	0	108	Y	1
5	54	M	NAP	150	195	0	122	N	0

The dataset includes demographic and clinical variables such as age, sex, chest pain type, resting blood pressure, cholesterol level, fasting blood sugar status, maximum heart rate, exercise-induced angina status, and heart disease diagnosis.

Exploratory Data Analysis

To understand the data, a summary table (Table 2) was created to look at the average key variables for females with and without heart disease and males with and without heart disease. Upon the creation of the summary table, it was clear that the data disproportionately sampled men with heart disease and lacked data about women with heart disease by looking at the

"Count" column. From the summary table, the following trends appeared: the average of men and women with heart disease is higher than those without heart disease, the average resting blood pressure of men and women with heart disease is higher than those without heart disease, the average cholesterol of men and women with heart disease is higher than those without heart disease, and the average max heart rate of men and women with heart disease is lower than those without heart disease.

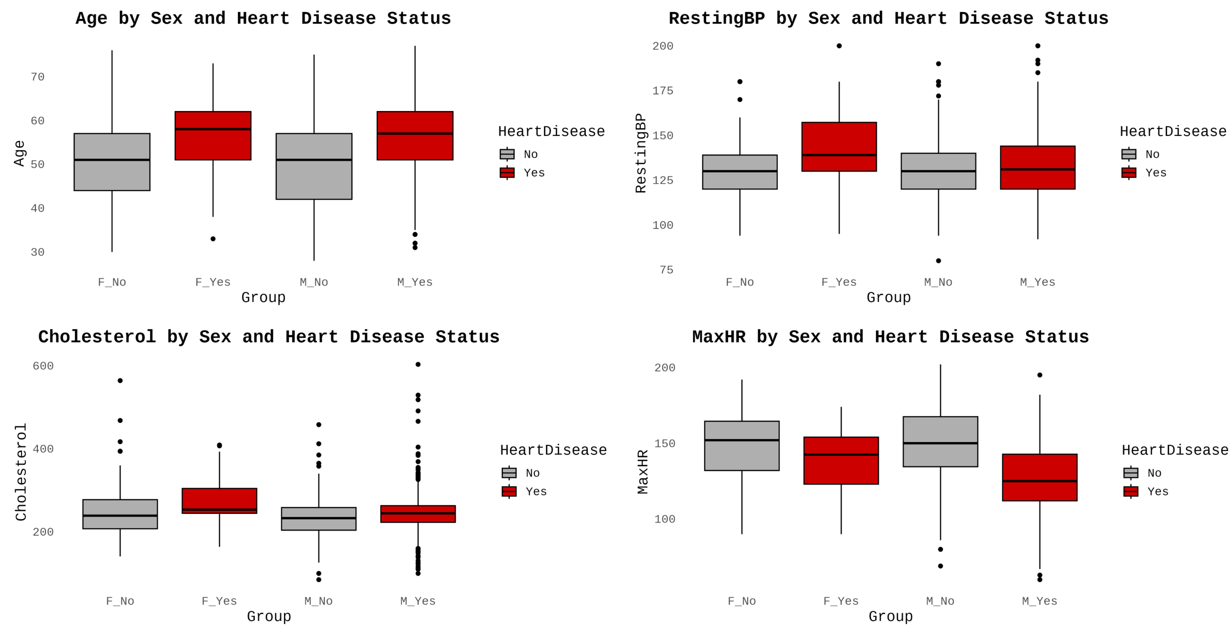
Table 2. Summary of Average Key Variables by Heart Disease Status and Sex

Sex	Heart Disease	Count	Age	Resting BP	Cholesterol	Max HR
Female	No Disease	143	51.2	128.8	249.2	149.0
	Disease	50	56.2	142.0	272.3	137.8
Male	No Disease	267	50.2	130.9	233.6	147.7
	Disease	458	55.9	133.6	246.6	126.5

The table displays sample counts, average age, resting blood pressure, cholesterol level, and maximum heart rate for male and female patients, grouped by presence or absence of heart disease.

The data was further explored using ggplot2 to create boxplots for continuous variables vs. heart disease status and bar plots for categorical variables vs. heart disease status. The boxplots examine how age, resting blood pressure, cholesterol, and maximum heart rate differ by sex and disease status. To create the boxplots, the "HeartDisease" variable was converted into factors; 0 was converted into "No" and 1 was converted into "Yes." A combined group variable was created so that the box plots could be split up by sex and heart disease status; "F_No" = Female, No Heart Disease, "F_Yes" = Female, Heart Disease, "M_No" = Male, No Heart Disease, and "M_Yes" = Male, Heart Disease. This expanded upon the initial summary table by visualizing key summary statistics of the dataset and information about the data's spread.

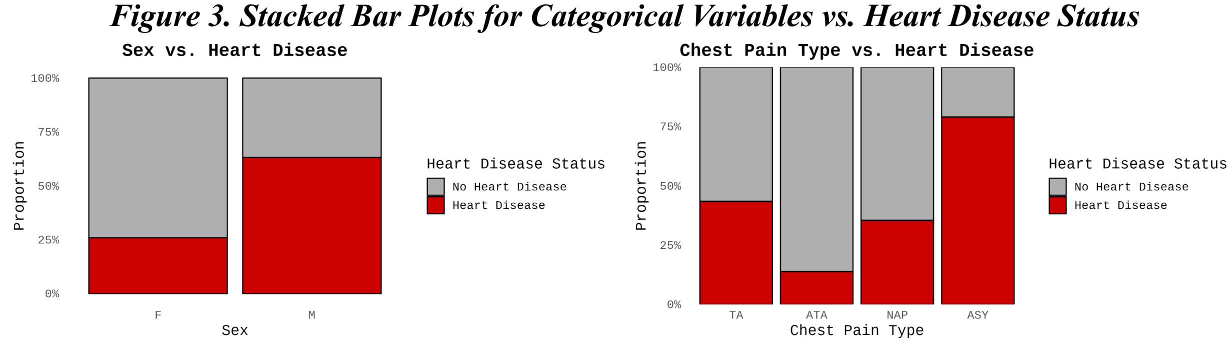
Figure 2. Boxplots for Continuous Variables vs. Heart Disease Status

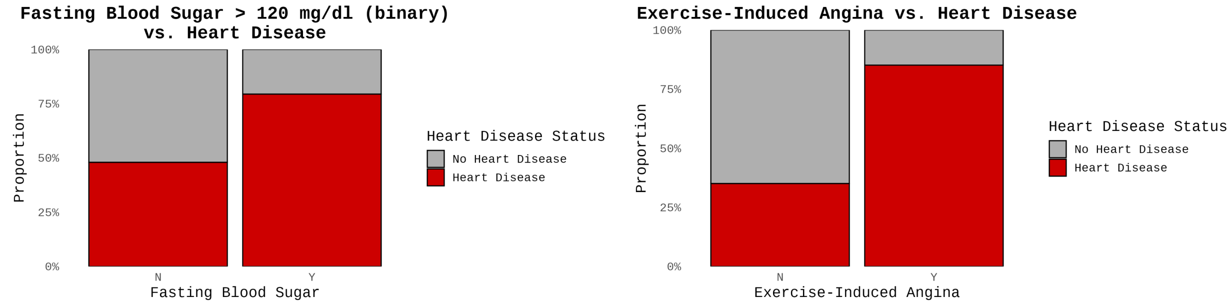


Each boxplot compares individuals with and without heart disease, highlighting differences in central tendency and variability across demographic groups.

Stacked bar plots were created to examine categorical variables and heart disease status using proportions of the dataset population to represent what percentage of men and women have heart disease vs. those that don't, what percentage of people with each chest pain type have heart disease vs. those that don't, what percentage of people with a fasting blood sugar greater than 120 mg/dL have heart disease vs. those that do not and what percentage of people with exercise-induced angina have heart disease vs. those that do not. From these stacked bar plots it was easy to see that being male, having asymptomatic chest pain, having high fasting blood sugar, and experiencing exercise-induced angina are all factors positively associated with heart disease.

Figure 3. Stacked Bar Plots for Categorical Variables vs. Heart Disease Status

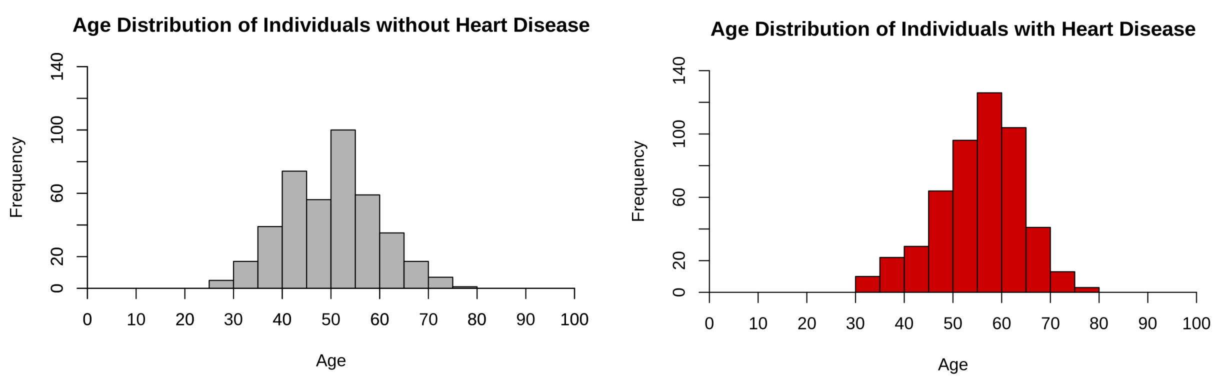




The plots display the proportion of individuals with and without heart disease within each category, highlighting patterns of association between these clinical and demographic factors and heart disease prevalence.

Histograms were also created to identify trends and compare distributions for heart disease based on age regardless of gender (Figure 5). The distribution of individuals without heart disease is normal and centered around 50-55 year olds. There is a wide spread from around 30-80 year olds, but there are few individuals below 40 and above 70 years old. The distribution of individuals with heart disease is centered more tightly around 55-65 years and is shifted slightly older. The distribution is standard, with a slight left skew indicating fewer young people have heart disease. The histograms furthered the idea that, on average, people with heart disease tend to be older than those without heart disease.

Figure 4. Histograms of the Ages of Individuals with and without Heart Disease



Individuals with heart disease tend to be older on average, with the distribution centered around ages 55–65, while individuals without heart disease show a broader distribution centered around ages 50–55.

Hypothesis Testing and Bootstrapping

After completing exploratory data analysis, three null hypotheses arose: H0_1: the mean age of patients with heart disease is less than or equal to the mean age of patients without heart disease, H0_2: the proportion of males with heart disease is less than or equal to the proportion of females with heart disease, and H0_3: the proportion of individuals with heart disease is less than or equal to for those with fasting blood sugar >120 mg/dL compared to those ≤120 mg/dL. For each of these null hypotheses is an alternative hypothesis: H1_1: the mean age of patients with heart disease is greater than the mean age of patients without heart disease, H1_2: the proportion of males with heart disease is greater than the proportion of females with heart disease, and H1_3: the proportion of individuals with heart disease is greater for those with fasting blood sugar >120 mg/dL compared to those ≤120 mg/dL. To test H0_1, a test was performed because average ages are being compared and the population standard deviation is unknown. Z-tests were performed to test H0_2 and H0_3 because proportions were being compared, and there was a large enough sample size. Bootstrapping was performed to check that the results of the hypothesis testing were statistically significant and not due to random chance. For the t-test, it is assumed that the samples are normally distributed and the sample size is large enough for the Central Limit Theorem to be applied. For the z-tests, it is assumed that the samples are independent and both np and n(1-p) are greater than 10 (there is a normal approximation).

Results

The hypothesis testing for H0_1, calculated t = 8.82 and the p-value is 0, which means we reject the null hypothesis, H0_1, and there is strong statistical evidence supporting the alternative hypothesis, H1_1, that on average, the age of patients with heart disease is greater than the age of patients. The mean of people with heart disease is ~56 years old, and the mean age of people without heart disease is ~51 years old. The bootstrapping for this hypothesis calculated the 95% confidence interval to be (~4.17, ~6.54), which means we are 95% confident that, on average, people with heart disease are about 4-7 years older than those without heart disease.

The hypothesis testing for H0_2 calculated a z-statistic of 10.27 and a p-value of 0, which means we reject the null hypothesis, H0_2, and there is strong statistical evidence supporting the alternative hypothesis, H1_2, that the proportion of males with heart disease is greater than the proportion of females with heart disease. In this study, ~74% of men had heart disease, and ~37% of women had heart disease. The bootstrapping for this hypothesis, calculated the 95% confidence interval to be (~0.30, ~0.44), which means we are 95% confident that men have about a 30-44% higher rate of heart disease than women.

The hypothesis testing for H0_3 calculated a z-statistic of 9.40 and a p-value of 0, which means we reject the null hypothesis, H0_3, and there is strong statistical evidence supporting the alternative hypothesis, H1_3, the proportion of individuals with heart disease is greater for those with fasting blood sugar >120 mg/dL compared to those ≤120 mg/dL. In this study, ~79% of people with heart disease had high fasting blood sugar, and ~48% of people without heart disease had high fasting blood sugar. The bootstrapping for this hypothesis calculated the 95% confidence interval to be (~0.25, ~0.38), which means we are 95% confident that people with high fasting blood sugar are 25% to 38% more likely to have heart disease than people with normal fasting blood sugar.

Conclusions

Old age, being a male, and having high resting blood sugar are important risk factors that significantly increase one's chances of developing heart disease during their lifetime. These results are consistent with previous studies and common medical thought. Understanding how demographic and clinical factors impact heart disease risk can help improve heart disease screening and detection. People who may not have previously thought they were at risk of heart disease may discover they are and get medical intervention, improving patient outcomes. This statistical analysis can also improve public health strategies aimed at lowering the rate of heart disease by helping public health organizations better allocate their time, attention, and money to populations most at risk of developing heart disease.

Further research could expand upon this study by looking at additional demographic and health conditions, like diet and exercise habits. Researchers could also look at the interplay of specific demographics and health conditions to see if they have a greater impact on heart disease risk. For example, the combination of high blood sugar and high cholesterol may put individuals at much higher risk of developing heart disease than those conditions on their own. Since this dataset is limited in population size and lacks race/ethnic information, it would be beneficial to examine how this data changes with a larger research group and how race/ethnic groups impact heart disease risk. Lastly, researchers could use machine learning models to predict whether or not patients are likely to develop heart disease and create a personalized risk score calculator to determine one's risk of developing heart disease.

References

American Heart Association. (2022a). *Unstable Angina*. www.heart.org. https://www.heart.org/en/health-topics/heart-attack/angina-chest-pain/unstable-angina

American Heart Association. (2022b, December 5). *Stable Angina*. Heart. https://www.heart.org/en/health-topics/heart-attack/angina-chest-pain/angina-pectoris-stable-angina

American Heart Association. (2024, May 17). *Understanding Blood Pressure Readings*. American Heart Association. https://www.heart.org/en/health-topics/high-blood-pressure/understanding-blood-pressure-readings

anneliset47. (2025). GitHub - anneliset47/Statistics5000FinalProject. GitHub. https://github.com/anneliset47/Statistics5000FinalProject/

CDC. (2024, May 15). *About Heart Disease*. Heart Disease; CDC. https://www.cdc.gov/heart-disease/about/index.html

Cleveland Clinic. (2021, October 17). *Fasting Blood Sugar: Screening Test for Diabetes*. Cleveland Clinic. https://my.clevelandclinic.org/health/diagnostics/21952-fasting-blood-sugar

Cleveland Clinic. (2022a, April 4). *Non-Cardiac Chest Pain*. Cleveland Clinic. https://my.clevelandclinic.org/health/diseases/15851-gerd-non-cardiac-chest-pain

Cleveland Clinic. (2022b, July 19). *Understanding Your Cholesterol Numbers*. Cleveland Clinic. https://my.clevelandclinic.org/health/articles/11920-cholesterol-numbers-what-do-they-mean

Drinan, K. (2023, February 26). *How to Prevent Heart Disease: 8 Simple Steps*. https://www.uchicagomedicine.org/forefront/heart-and-vascular-articles/heart-month-3-steps-to-a-healthier-heart-right-now

Janosi, A., Steinbrunn, W., Pfisterer, M., & Detrano, R. (2021). *Heart Failure Prediction Dataset*. Kaggle.com. http://www.kaggle.com/datasets/fedesoriano/heart-failure-prediction

Mayo Clinic. (2024). *Heart Disease*. Mayo Clinic; Mayo Clinic. https://www.mayoclinic.org/diseases-conditions/heart-disease/symptoms-causes/syc-20353118