The Study of the Factors Influencing the Risk of Being Diagnosed with Coronary Heart Disease

Qing Wen

2020/12/21

Abstract

Cardiovascular disease has been a leading cause of death globally and it is not until the late 20th century that people develop a more and more sophisticated understanding of its underlying cause and possible early prevention. In this study, we focus on investigating several potential risk factors that contribute to the determination of future Coronary Heart Disease, which is a type of cardiovascular disease. A logistic model is built using a binary response variable that determines whether the patient has Coronary Heart Disease. As a result, factors like smoking, blood pressure medication, and Body Mass Index all contribute to the development of the disease to certain extent.

Key Words

Coronary Heart Disease, Risk Factor, Framingham Heart Study, Longitudinal Study

Introduction

There has been a rising awareness globally towards improving human health and promoting the importance of tracking one's own health status. As an increasing amount of innovative technology comes out around the world, enhancement in medical equipment and new breakthroughs within the medical field has provided a strengthened understanding of diseases that were once thought untreatable. Doctors and scientists are also utilizing their expertise to convince people what should be done in their normal life as early prevention for any preventable diseases.

Among all the diseases that could cause the early death of an individual, cardiovascular disease(CVD) accounts for the number one reason for such deaths. According to the World Health Organization, an estimated 17.9 million people died from CVDs in 2016, which accounted for approximately 31% of the global deaths(*Cardiovascular Diseases (Cvds)* 2017). Looking more closely, CVD is a group of disorders of the heart and blood vessels, including disorders like cerebrovascular disease, coronary heart disease(CHD), and congenital heart disease. However, even with the high death rate, CVD is considered to be preventable if individuals could address their behavioral risk factors.

With no special preference, CHD is picked as the objective of our study and we want to find potential factors that help determine the risk of individuals being diagnosed with CHD. According to the Mayo Clinic, CHD is a disorder caused by damage to the major blood vessel. Cholesterol-containing deposits accumulate within the arteries and gradually narrow the blood vessels, making it hard for them to deliver blood and essential nutrients to the heart(*Coronary Artery Disease* 2020). As the risk of blocking the blood vessels likely to develop over decades, it becomes difficult for one to diagnose the problem early, probably not until one got a complete block of the arteries which causes a heart attack.

Therefore, in this study, potential risk factors contributing to the development of CHD is analyzed, and a logistic model is fitted to predict the probability of one being diagnosed with CHD in the future. In the end, we want to generalize the results from the model output to develop the importance of paying attention to certain health indications to help people adapt to a healthier lifestyle in the effort of decreasing the risk of having a CHD.

The study will be conducted using the data obtained from a well-known longitudinal study regarding heart disease, namely the Framingham Heart Study which was first launched in 1948. In the Methodology section, more detailed information on the data would be provided and the underlying theory for the logistic model will be presented. Results regarding the analysis of the variables and the model output will be included in the Result section. Finally, further inferences regarding the results and any weaknesses associated with the study will be included in the Discussion section. The relevant codes and the pdf version of the report can be found in the following repo: https://github.com/QingWen-0310/10-year-risk-of-future-coronary-disease.

Methodology

Data

Framingham Heart Study(FHS)

The Framingham Heart Study(FHS) is an ongoing longitudinal cohort study on heart disease. A longitudinal cohort study is a type of study that follows a group of people over time to determine the natural development and the behaviors of the disease, and potentially finds factors that explain the behavior. With a history of 72 years, FHS has successfully altered public approaches to treat CVDs. Back in 1948, little was known about heart disease despite its cause for the increasing death rate. What FHS has accomplished is revolutionary, as it contributes to uncovering general patterns and trends which is meaningful and applicable to the general population(Papanicolaou, n.d.).

As the study began in 1948, the first generation of the participants consisted of 5,209 men and women aged between 30-62 years from the town of Framingham, Massachusetts. The participants had not yet develop any noticeable symptoms of heart disease and have not had a stroke or heart attack beforehands(*About the Framingham Heart Study*, n.d.). The participants had their physical statuses extensively examined and recorded, and were expected to return for another round of detailed examination and lab tests every 5 to 6 years since then. In 1971, the next generation of these participants was enrolled in the study, and in 2002, the grandchildren of the original participants were examined for the new phase of the study.

Hence, the target population of the study is the general public who has the chance to develop the heart disease. Since the study recruited voluntary participants, there is no sampling frame in theory. However, there is restriction on entry as participants need to be within the 30-62 years age bound and they do not have a heart attack or stroke before. The sample ended up being the 5,209 participants for the original round of cohort study.

Strengths and Weaknesses

The FHS has been greatly appreciated for its scope and duration of the study. Statistics are taken and examined carefully throughout the conduction of the study. The data set obtained for this study has a large sample size(around 4000 observations), and contains predictors that are highly relevant to the purpose of the study. Though missing values are taking out which reduces the sample size, the reduction is not large enough to decrease the statistical power of the analysis.

We then turn to some of the weaknesses the study and the data have. First of all, the study recruited voluntary participants, who were mainly doctors and workers in the health organization back then. There is a potential voluntary response bias as the participants who were interested in the target of the study are more likely to take part in the study, which seemed to be the case for FHS. Secondly, the study focused on the participants in Framingham, and then generalized the study results to the general public. There is debate recently that the study over-estimated the risk for some lower-risk groups, for instance the British population(Peter Brindle, n.d.).

Data for Study

The data set used in this study is obtained from Kaggle for which the author requested from the original study. However, little details have been provided regarding which round of the study these observations are taken from, so we could not generalize the results to the specific study population. But the aim of this study is to determine the potential risk factors that could serve for the prediction of the risk of having a CHD. With the 4238 observations and 16 variables within the data set after cleaning, we can use the prescribed model to yield predictions and provide analysis regarding some important variables in determining risks.

The original data set is almost clean and does not require too much manipulation. We removed the *education* and the *glucose* variable in the original data set as they contain many missing values. Also, the rest of the observations that contain missing values are removed. In the end, 4088 observations are present in the data for our latter analysis, which is a considerable sample size for our goal of study.

Table 1: Baseline Characteristics of the Data Set

Variable Name	Meaning	Minimum	Mean	Maximum
male	whether the patient is male	0.00	0.4347	1.0
age	the age of the patient	32.00	49.5	70.0
currentSmoker	whether the patient is currently smoking	0.00	0.4902	1.0
cigsPerDay	cigarettes per day if the patient is smoking	0.00	8.992	70.0
BPMeds	whether the patient is on blood pressure medication	0.00	0.02935	1.0
prevalentStroke	whether the patient had a stroke before	0.00	0.005382	1.0
prevalentHyp	whether the patient is hypertensive	0.00	0.3092	1.0
diabetes	whether the patient has diabetes	0.00	0.02544	1.0
totChol	total cholesterol level	113.00	236.7	696.0
sysBP	systolic blood pressure	83.50	132.2	295.0
diaBP	diastolic blood pressure	48.00	82.89	142.5
BMI	Body Mass Index	15.54	25.80	56.8
heartRate	the patient's heart rate	44.00	75.84	143.0
TenYearCHD	whether the patient has a 10-year-risk of CHD	0.00	0.1495	1.0

Table 1 presents the baseline characteristics of the data for each of the variables present in the data set. The variable *TenYearCHD* indicates whether the patient has a 10-year-risk of developing a CHD, so we will use this as the binary response in the logistic model. Among all the potential predictors, variables *male*, *age*, *whether the patient is a current smoker*, *blood pressure medication*, *total Cholesterol level*, and *BMI* are selected as predictors in this study. As we have previously noted, the CHD develops as the main blood vessels are blocked by cholesterol-containing deposits, which is a huge indication of examining a patient's lifestyle and physical statuses like total cholesterol value and the Body Mass Index(BMI). Studies have shown that smoking changes blood chemistry, which contaminates the oxygen-rich blood that flows into the heart. Also, sex and age are often common factors that are involved in disease development. Hence, these variables are selected for further study.

Model

To accomplish the objective of predicting the risk of patients having CHD, we adopt a logistic model. We are trying to predict whether a patient has the risk of being diagnosed with CHD which is a binary response variable. A Multiple Linear Regression(MLR) model may be suggested as an alternative model. However, in the context of the study and the data collected, we are interested in a binary outcome of whether the patient has CHD or not. It would make less sense to investigate how many heart diseases the patient would have in the future than looking at purely the determination of the presence of the disease. Therefore, a logistic model is the most appropriate model under the circumstances.

Specifically, the theoretical model with our pre-selected predictors and response will be defined as follows:

$$log(\frac{p}{1-p}) = \beta_0+\beta_1 x_{male}+\beta_2 x_{age}+\beta_3 x_{currentSmoker}+\beta_4 x_{BloodPressureMedication}+\beta_5 x_{totalCholesterol}+\beta_6 x_{BMI}$$

The variable on the left side of the equation represents the log odds, which is a log transformation on the odds that contains the probability we target in finding. Any changes resulted from the significant predictor variables will result in the average change in the log odds. However, in order to form more understandable and more meaningful conclusions from the model output, we should apply the exponential transformation to the log odds in order to obtain our final probability. To give a simple but concrete example to demonstrate this mechanism, suppose we have a simplified version of our defined logistic model that only consists of one predictor: x_{age}. The model can be spelled out like:

$$log(\frac{p}{1-p}) = \beta_0 + \beta_1 x_{age}$$

, and the estimate for $\beta_0 = -10.8$, and the estimate for $\beta_1 = 0.23$. Suppose we have a patient that is 52 years old, and we want to predict the probability of this patient having CHD. Using the estimated model, we would have $log(\frac{\hat{p}}{1-\hat{p}}) = -10.8 + 0.23 * 52 = 1.16$. In order to find the probability, we apply an exponential transformation as follows: $e^{log(\frac{\hat{p}}{1-\hat{p}})} = \frac{\hat{p}}{1-\hat{p}} = e^{1.16}$. By isolating for $\hat{p}$, we can find the probability of the patient having CHD is 0.76.

With this calculation in mind, we will define other parameters in the model in its naive meanings(i.e. their meaning as regard to the log odds):

p: the predicted probability of the patient being diagnosed with CHD

β_0: the intercept of the model, which indicates the average value for log odds when all the predictors have the value 0

Holding other variables constant, then:

β_1: the average difference in log odds resulted from the patient being a male or not a male

β_2: the expected change in log odds resulted from an one unit increase in age

β_3: the average difference in log odds resulted from whether the patient is a smoker

β_4: the average difference in log odds resulted from whether the patient is on blood pressure medication

β_5: the expected change in log odds resulted from an one unit increase in the total cholesterol level

β_6: the expected change in the log odds resulted from an one unit increase in the BMI

As seen from above, variables like *age*, *total cholesterol level* and *BMI* are all numerical variables, while the rest of the variables are categorical. Usually, categorical variables are treated as dummy variables in the model, which have different levels. For example, the *current smoker* variable would have two levels: whether the patient smokes or does not smoke. If the patient is a smoker, then the variable $x_{currentSmoker}$ would have a value of 1, and 0 otherwise. The categorical variables in our analysis all consist of two levels and are coded as 1 if in the presence of the condition defined by the variable name, and 0 otherwise.

The model is built using R Studio(R Core Team 2020), and the relevant packages used in this study are: tidyverse(Wickham et al. 2019), and knitr(Xie 2020).

Result

According to the model output, the fitted model is written as follows:

$$log(\frac{\hat{p}}{1-\hat{p}}) = -7.7645 + 0.5656x_{male} + 0.0777x_{age} + 0.4154x_{currentSmoker} + 0.6946x_{BloodPressureMedication}$$

$$+0.0025x_{totalCholesterol} + 0.0351x_{BMI}$$

Table 2: Statistics for the Coefficient Estimates from the Logistic Model

Coefficient	Estimate	Standard Error	p value
intercept	-7.7645	0.4850	< 0.001
male	0.5656	0.0956	< 0.001
age	0.0777	0.0058	< 0.001
current smoker	0.4154	0.0984	< 0.001
blood pressure medication	0.6946	0.2116	0.001
total cholesterol level	0.0025	0.0010	0.0157
BMI	0.0351	0.0112	0.0018

From Table 2, we see that all the coefficient estimates for the slope parameters are positive.

Also, we see that all the coefficient estimates have a p-value much less than 0.05, suggesting that these predictors are all statistically significant. Among all the predictors, the table shows that the coefficient estimates for *male*, *current smoker*, and *blood pressure medication* are relatively bigger than the rest of the estimates, not accounting for the intercept estimate.

Table 3: Proportion of Patients with Different Sex Having CHD

Sex	proportion
female	0.1194288
male	0.1885200

Table 4: Proportion of Patients Taking Blood Pressure Medication Having CHD

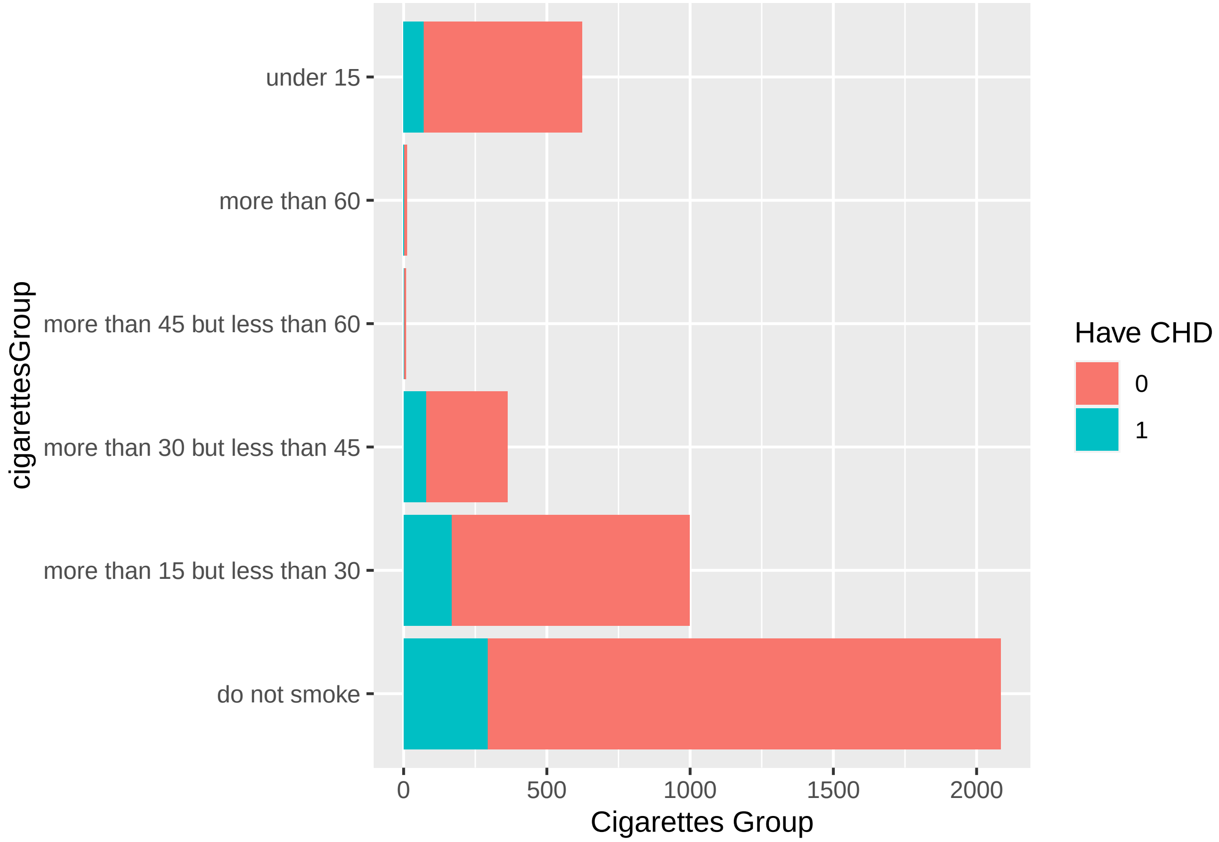
Blood Pressure Medication	proportion
not on medication	0.1441532
on medication	0.3250000

Table 3 and 4 each gives a simple proportion summary for patients with certain characteristics having CHD. From Table 3, around 18.9% of the males in the data have CHD while around 11.9% of females in the data have CHD. From Table 4, around 32.5% of the patients who take blood pressure medications have CHD, and around 14.4% of the patients who do not take blood pressure medications have CHD.

Table 5: Proportion of CHD Patients Smoking Different Number of Cigarettes Per Day

Cigarettes Group	proportion
do not smoke	0.4795417
more than 15 but less than 30	0.2733224
more than 30 but less than 45	0.1276596
more than 45 but less than 60	0.0016367
more than 60	0.0032733
under 15	0.1145663

Graph 1: Proportion of Patients in Each Cigarettes Group Having CHD



Graph 1 shows the proportion of patients in each cigarettes group that have CHD. The cigarettes group are divided into different categories regarding to the different number of cigarettes the patients smoke per day. Patients diagnosed with CHD in each group represents a relatively small proportion of each group.

In addition, Table 5 shows that among all the patients who are diagnosed with CHD, what proportion of them smokes the number of cigarettes as categorized by each cigarette group presented in the table. The table shows that patients who do not smoke constitute around 48% of patients who are diagnosed with CHD. Among the rest of the patients who do smoke, people who smoke 15 to 30 cigarettes per day constitutes the largest part of the patients who have CHD. On the other hand, only a small proportion of people diagnosed with CHD smokes around 45-60+ cigarettes per day.

Discussion

Summary

In this study, we investigate the potential factors that influence the probability of one being diagnosed with CHD, a heart disease. We started by looking at the data set and variable definitions, then justifies the subset of variables we choose to include in this study. We then moved on to describe the logistic model and its interpretations. After that, results regarding the model output is included in the result section, and all the p-value for the coefficient estimates are statistically significant in a 0.05 significance level. Further interpretations of the results and remarks will be discussed next.

Conlcusion

To interpret the results given by the model output, taking the predictor variable *current smoker* as an example. Table 2 shows the estimate for it is 0.4154. It means that holding other variables constant, if the patient is a current smoker, then the odds would increase multiplicatively by $e^{-7.7645+0.4154} = e^{-7.349}$. To get the probability of a patient who smokes being diagnosed with CHD, we get that $\hat{p} = \frac{e^{-7.349}}{1+e^{-7.349}} = 6.4282179 \times 10^{-4}$. On the other hand, holding other variables constant, if the patient is not a current smoker, then the odds would increase multiplicatively by $e^{-7.7654}$. Following a similar argument, the probability of the patient who does not smoke being diagnosed with CHD would be 4.2398009×10^{-4}.

Also, from the previous Result section, we found that all the predictor variables seem to be significant, as the p-values for their coefficient estimates are all smaller than 0.05. Hence, we would not further consider variable selection techniques to try to reduce our existing model. However, looking closely at the estimates, variable *total cholesterol level* has a small estimate for its coefficient, namely 0.0025. When transformed back to the probability, we found that with an one unit increase in total cholesterol level, the expected change in log odds is 0.0025, holding other variables constant. Exponentiating it, we get that $e^{0.0025} = 1.0025031$. This means that we expect to see a 0.25% increase in the odds of being diagnosed with CHD, which is not too big a change.

On the other hand, looking at the variable *blood pressure medication*, it seems to be the predictor variable that has the greatest impact on the response variable, as it has the largest coefficient estimate among all the other selected predictors. Specifically, the odds ratio for patients who take blood pressure medication is 2.0029, which means that patients taking blood pressure medication would be 2 times more likely to have a CHD than patients who do not take the medication.

Following the same constructs, The other coefficient estimates can be interpreted accordingly. Despite the magnitude of the impact each predictor has on the response variable, we also want to comment on the fact that all the coefficient estimates are positive. To elaborate on this, patients who are male, is currently smoking and is on blood pressure medication would have a higher probability of being diagnosed with CHD. Furthermore, with older age, larger BMI, and higher total cholesterol levels, the probability of a patient being diagnosed with CHD also increases. Table 3 and 4 in the Result section also confirms the fact that with the data collected, males and patients who are on blood pressure medication seems to more likely to be diagnosed with CHD.

From Table 5 of the Result, we see that patients who smoke constitute a slightly larger proportion of the patients who are diagnosed with CHD. We anticipate that people who smoke would have a higher chance of developing CHD. Looking back to the data set, there are a lot fewer observations collected from people who smoke more than 30 cigarettes per day than from those who smoke less than 30 cigarettes per day. But the data observed seem to align with the idea that smoking increases the risk of developing CHD. Relevant studies have shown that smoking leads to long-term increases in blood pressure and heart rate(*Smoking and Cardiovascular Disease*, n.d.). More importantly, it reduces the blood flow to the heart and decreases the amount of oxygen carried in the blood. Despite the people who smoke, people suffering from second-hand smoke also have an increased risk of developing diseases. Quitting smoking would not only benefit the patients themselves by reducing their risk of being diagnosed with serious diseases like CHD but also protect the public from inhaling second-hand smoke and improve their health status as well.

Next, we would briefly justify some of the observed positive relationships between the outcome and the predictors other than smoking. Firstly, a study conducted on the four southern states and Colorado in the United States has confirmed a strong association between BMI and CHD(Akil, n.d.). Study shows that a high BMI indicates obesity and which in turn increases blood pressure. Furthermore, CHD mortality is shown to be elevated in patients who are overweight. Total cholesterol level is said to have a direct impact on an increased risk of having CHD. Cholesterol is necessary for building healthy cells inside the human body, but high cholesterol levels could lead to the accumulation of fatty deposits in the blood vessels and make it difficult for the blood to flow through the arteries(*Coronary Artery Disease* 2020). Although high cholesterol could be inherited, in often cases it is a result of unhealthy living habits. Lastly, according to a study conducted by the research team led by George Howard, a professor in the Department of Biostatistics in the UAB School of Public Health, with each medication the patient takes to treat hypertension, the risk of having heart disease and a stroke increases by about 33%(*Blood Pressure Medications Can Lead to Increased Risk of Stroke*, n.d.).

Therefore, it ties back to the fact that the development of CHD hugely relies on the living habits the patients have. CHD can be preventable as patients take control of these controllable risk factors. With the potential risk factors identified, it justifies why the doctors check routinely about the presence of high blood pressure, high cholesterol level, smoking, and unhealthy weight. It is important to maintain a healthy lifestyle that could serve as early prevention of various diseases.

Weaknesses

Though we obtained a model with all significant covariates and yield meaningful results, there are also weaknesses that we fail to cover in the study.

To start with, the data we used in the study does not indicate which round of the FHS study it was collected from, so we lack the ability to generate a specific set of conclusions and descriptions of the pattern pertaining to the population of which the data is collected from. What we did accomplish is to try to generalize the result to the general public and develop some tips for early-prevention of the disease. Nonetheless, if more information could be extracted from the data source, and perhaps including data from other rounds of the study could yield more solid and reliable conclusions and potentially investigate the family pattern of the disease development.

Furthermore, as mentioned previously in the *strengths and weaknesses* subsection of the Data section, the FHS may have voluntary response bias as the first round of study recruited participants that were mainly doctors and health organization workers. However, the diversity of the population did come in at the second round of study, which involved the children of the first generation.

Next Steps

FHS is a longitudinal study and is thereby costly and time-consuming to conduct. So far, only 3 to 4 rounds of study has been carefully carried out. Though many studies and researches have been successful in extracting information from the data and provide the public with a much enhanced understanding on the CHD, the data collection process is still taking place and conclusions drawn previously in the study could be elevated or altered depending on the richer data available. For further development of the study, the FHS keeps a full record of all the observations from each round of the study. Their current goal is not only to keep tracking of the families' medical histories from the first generation of the study, they now aim at the more detailed objective of the studies since basic patterns of the CHD has been revealed in the past century. FHS now design and modify their studies to encounter the effect that race and ethnicity have on the study outcome, which could improve the richness of the data and allow more analysis targeting narrower topics to be conducted.

Also, it may be valuable to investigate more of the variables present in the current data set to find any other significant determinants for the outcome. Our goal would still be to obtain a parsimonious model. Assessing different subsets of predictors by building multiple models and comment on the conclusions could enrich the arguments we made in this study.

Concluding Remarks

To conclude, we want to first assert that the conclusions drawn from this study are based on an observational study, and no causal relationship is inferred from the context. However, with observational data, we still yield meaningful results to determine the potential risk factors for a patient being diagnosed with CHD. The important message to take away is to develop healthy living habits which necessarily decreases the risk of developing serious diseases. FHS has certainly altered the way people approach heart disease. Despite the ongoing longitudinal heart study, it also currently carries out studies that benefit the general public. The valuable and rich databases provide the researchers with great opportunities to perform analysis, and we look out for more inspiring findings in the future.

References

About the Framingham Heart Study. n.d. National Heart, Lung,; Blood Institute. https://framinghamheartstudy.org/fhs-about/.

Akil, Luma. n.d. *Relationships Between Obesity and Cardiovascular Diseases in Four Southern States and Colorado*. Environmental Science, College of Science, Engineering; Technology at Jackson State University. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3250069/.

Blood Pressure Medications Can Lead to Increased Risk of Stroke. n.d. University of Alabama at Birmingham. https://www.sciencedaily.com/releases/2015/05/150529193554.htm.

Cardiovascular Diseases (Cvds). 2017. World Health Organization. https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds).

Coronary Artery Disease. 2020. Mayo Clinic. https://www.mayoclinic.org/diseases-conditions/coronary-artery-disease/symptoms-causes/syc-20350613.

Papanicolaou, George. n.d. *Framingham Heart Study (Fhs)*. National Heart, Lung,; Blood Institute. https://www.nhlbi.nih.gov/science/framingham-heart-study-fhs.

Peter Brindle, Fiona Lampe, Jonathan Emberson. n.d. *Predictive Accuracy of the Framingham Coronary Risk Score in British Men: Prospective Cohort Study*. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC286248/.

R Core Team. 2020. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Smoking and Cardiovascular Disease. n.d. Johns Hopkins Medicine. https://www.hopkinsmedicine.org/health/conditions-and-diseases/smoking-and-cardiovascular-disease.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

Xie, Yihui. 2020. *Knitr: A General-Purpose Package for Dynamic Report Generation in R*. https://yihui.org/knitr/.