CS216 - Final Project

Steven Yuan (szy2), Conner Byrd (ctb43), Zach Kinne (zpk), Sam Rivera (sfr11)

December 8, 2022

The full markdown, all code, and data can be accessed at the following Github repository (https://github.com/connerbyrd/CS-216-Final-Project).

I. Introduction

According to the Center for Disease Control, heart disease is one of the leading causes of death for people of most races in the United States. About half of all Americans have at least one of three key risk factors for heart disease defined by the CDC (high blood pressure, high cholesterol, and smoking), and detecting and preventing the factors that have the greatest impact on heart disease is very important to healthcare, as heart disease is treatable if there is quick access to equipped hospitals and early diagnosis, as well as key predictive screening to analyze the risk factors of patients. To provide preventative treatment to individuals from having heart disease, it is necessary to be able to determine risk factors which can be used to create actionable prevention plans. Currently, there are several different ways for physicians to diagnose patients that they believe to beat risk for heart disease, which often include measuring blood pressure, cholesterol level, and conducting further tests such as exercise stress tests and electrocardiograms. However, there are many issues with current diagnostic methods. A study of 500 patients found a false positive reading between 77 and 82 percent inpatients in patients at risk of heart disease screened by ECG, and a false negative reading between 6 to 7 percent in the same patient population. To more successfully prevent diagnose individuals that are at risk for heart disease, it is first necessary to be able to determine other strong risk factors which can be used to create actionable prevention plans.

In that aim, our research has two main questions. Our first question is what demographic and health factors tend to be the best at predicting the occurrence of heart disease? To answer this question, we plan on developing predictive models to assess the likelihood of a heart disease diagnostic for potential at-risk patients based on a number of factors. Our second question is what health and demographic factors tend to be associated with higher risk of having heart disease? We plan on creating models for the purpose of interpretation to answer this question and provide a greater understanding of signs that patients can analyze to check their risk for heart disease. Answering these questions requires an in-depth and substantial analysis of patient data. Doing so is relevant to the scientific community given the scope of the disease globally and within the United States. With the current metric of nearly half of Americans at risk of heart disease according to the CDC,diagnosing this issue and proposing a solution has the potential to save the lives and prosperity of millions of lives around the world.

II. Data Sources

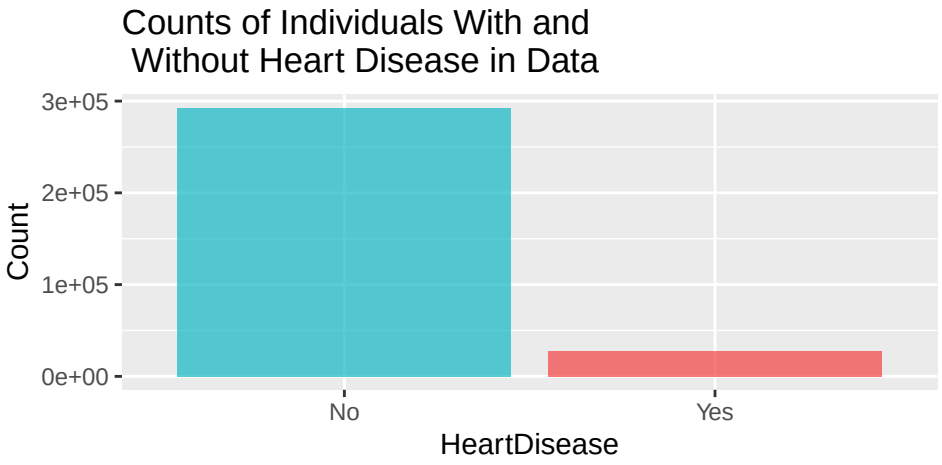
The dataset we utilize to answer our research questions is the 2020 CDC survey as part of the Behavioral Risk Factor Surveillance System, which conducts telephone surveys to gather data on the health status of US residents in all 50 states as well as the District of Columbia and three US territories, asking questions about the respondents' health status and demographic information (please see Works Cited for link to data).The original dataset of nearly 402,000 observations nearly 300 variables was reduced to 18 variables relating to various demographic health conditions, such as BMI, whether the respondent was a smoker, as well as the sex and race of the respondent. The categorical variable of whether the respondent has ever had heart disease

is also included in the dataset. We acquired the dataset from Kaggle, and the dataset has already been pre-processed and cleaned by both the CDC and Kaggle, with only observations where all 18 variables have been recorded kept in the cleaned data (See Appendix A for a description of the variables).

The dataset utilized for our project is relevant and appropriate for addressing our research question as the dataset is one of the largest datasets available on the health status of individuals that includes information on the prevalence of heart disease. The dataset, collected and curated by the CDC, also has a wide reach across the United States geographically. The datasets size and geographic breadth will allow us to generalize our findings across the United States with fewer issues and also has the potential to provide more predictive power and better data to train our models on than smaller datasets curated elsewhere.

Data Pre-Processing

The data was already pre-processed by both the CDC and Kaggle, and has no missing variables. Due to the size of the dataset and the high runtime with fitting models using such a large dataset, we trim down the dataset by sampling 10,000 observations randomly from the dataset, and split the sampled dataset into both a testing and a training set. The 18 variables included include 17 predictor variables as well as the dependent variable, `HeartDisease`. Some of the variables, such as `PhysicalHealth`, `MentalHealth`, and `SleepTime`, make more sense being a proportion of the number of days in the month or number of hours in the day, as `PhysicalHealth` and `MentalHealth` are limited to the number of days in a month and `SleepTime` is limited to 24 hours. Thus, we scale them to be in-between 0 and 1. Furthermore, `BMI` is transformed into a categorical variable defined to conventional medical norms (BMI of under 18.5 is underweight, 18.5 to 25 is considered normal, 25 to 30 is considered overweight, and over 30 is considered obese) as medically, there are very few differences in changes in BMI within a category (ex. a BMI of 21 and 20 are considered both medically to be normal BMIs), and most medical professionals utilize BMI as a categorical variable.

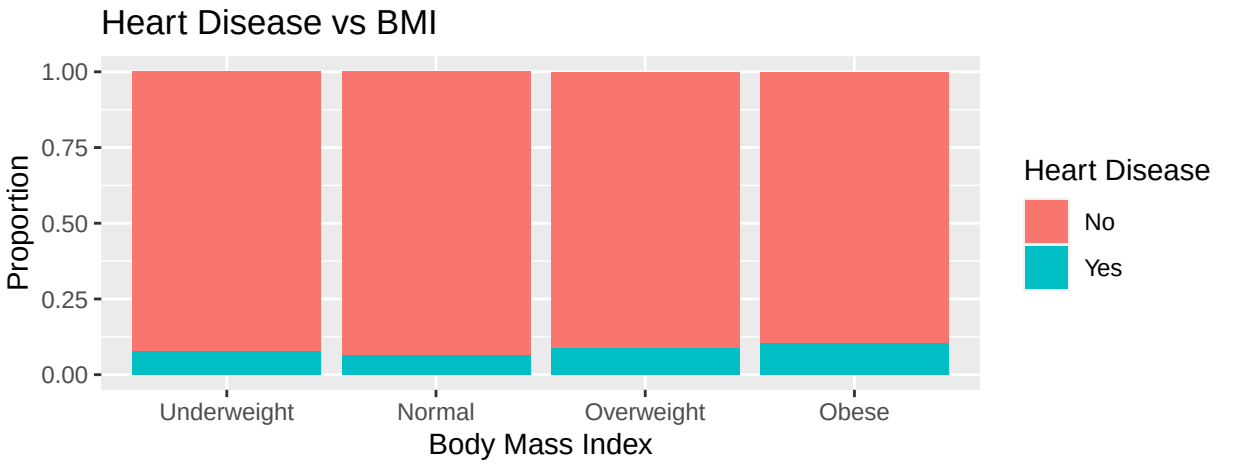


As shown above, the data is heavily imbalanced, with far more observations where heart disease was not observed than where it was observed. Imbalanced data poses a problem, as the model being trained would be dominated by the majority class, which in this case is where the patient does not have heart disease. The model would predict the majority class more effectively than the minority class, which is undesirable in this case as we want to ensure a high sensitivity rate as it is far more important to be able to correctly identify individuals with heart disease. A technique to reduce the negative impact of imbalanced datasets is by subsampling the data. Thus, we upsample the minority class by sampling with replacement so the two classes have the same size to create an upsampled training dataset.

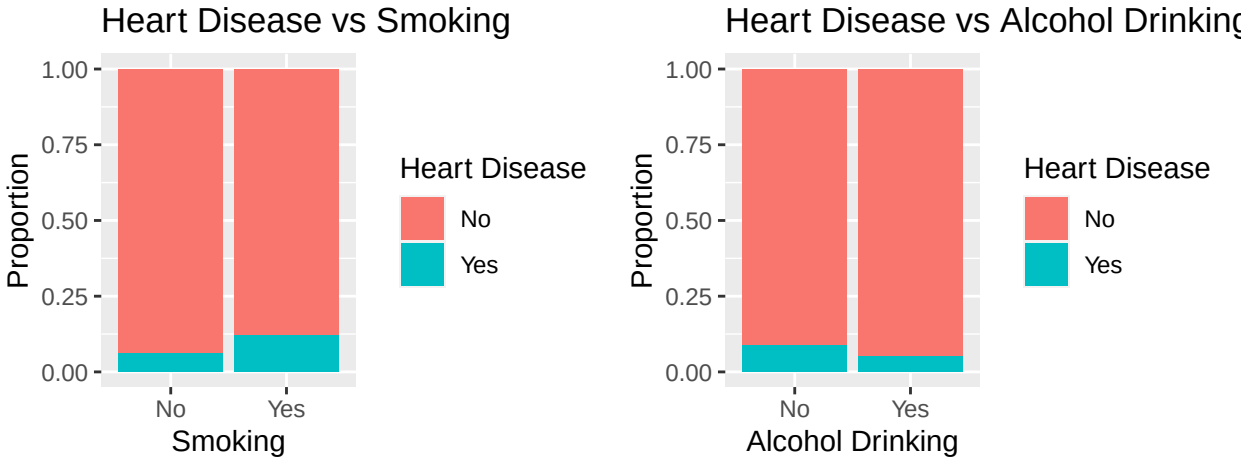
Exploratory Data Analysis

We first explore a few potential interesting relationships between the prevalence of heart disease and some general lifestyle and health factors that may be of interest, such as BMI and whether a person smokes or

drinks. We first decided to explore any potential relationship between BMI and the prevalence of Heart Disease, as although the usage of BMI is controversial, BMI is still popular in its medical usage in regards to association with heart disease.



The plot above displays the prevalence of heart disease by BMI classification. From the visual analysis, it appears that higher body mass may be correlated with a higher chance of heart disease than normal BMI. Underweight BMI may also be correlated with a higher chance of heart disease. This seems to reinforce the generally accepted medical concept that a normal BMI is most conducive to better general health.



The plots above display the prevalence of heart disease by whether or not the respondent is a smoker (defined as having smoked at least 100 cigarettes) and whether or not the respondent is a heavy drinker (defined as 14 drinks in the past week for males, 7 for females). From the plots, it appears that smokers may be positively associated with a higher chance of heart disease. Interestingly, it appears that heavy drinking may be associated with a lower chance of heart disease.

III. Modules Used

Being that our project will involve locating relevant predictors and using them to make inferences, Modules 3, 5, and 9 are all very applicable.

Module 3 - Probability

Probability is a foundational aspect of statistics and is used heavily when running experiments, making predictions or estimations, using distributions, and much more. Throughout the course of our project, we expect to use probability for numerous applications. The vast majority of these applications will come during the data investigation, data analysis, and final report portions of the project. When looking for significant predictors, probability is used in the form of p-values and is compared against a set alpha level in order to determine statistical significance. We will use probability in this manner to determine which patient variables lead to an increased risk of heart disease. Probability is also an important aspect of statistical distributions, which we expect to use substantially throughout our project. Noting which distributions to use based on the relationships between the response (heart disease indicator) and the various dependent variables will determine useful probabilities that can then be analyzed. Other probabilistic values such as the mean, variance, and standard deviation will also be useful for prediction, estimation, and even data imputation for certain patient lifestyle variables.

Module 5 - Statistical Inference

Much like probability, statistical inference is highly important for making conclusions about data. In our project, the majority of statistical inference will come in the form of hypothesis testing. Using hypothesis testing, we will be able to test certain variables for their significance when it comes to heart disease occurrence. From there, we can use the techniques described in Module 3 to determine exactly which health and body attributes most prominently lead to an increased risk in heart disease. Much like with probability, we expect to use statistical inference in the data investigation, analysis, and final report sections of the project.

Module 9 - Prediction & Supervised Machine Learning

Seeing as one of our research questions revolves entirely around prediction, we will be using concepts from Module 9 liberally throughout our project. Linear and Logistic Regression will be very useful to help us construct models of both numerical and categorical data and aid us in our predictive efforts. Other concepts learned in Module 9 will be helpful with checking for correlation between variables, performing stepwise regression, constructing a decision tree or support vector machine, and more. These techniques will allow us to further our understanding of the relationship between predictors and heart disease occurrence and to construct accurate models that we can then make sensible interpretations from. We will also likely end up splitting the data into training and testing cohorts in order to perform further analysis of variable selection and model fit. Prediction will be used in the data analysis and final report portions of the project.

IV. Preliminary Results and Methods

A higher-level discussion of important and notable findings are included in the "Discussion of Results" section below. See Appendices B and C for the full model outputs of the model for inference and for prediction, respectively.

Model for Inference

We decided to utilize a logistic regression model, with whether the person has heart disease (`HeartDisease`) as the response variable, to examine the relationship between the log-odds of heart disease and various predictors. The general logistic regression model form is formulated by the equation:

$$\log(\frac{P_i}{1-P_i}) = \beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + ... + \beta_k X_{k,i}$$

Using logistic regression lends itself well to both inference and prediction. For the purposes of inference, we decided to fit a model including all variables, as we are interesting in exploring the relationship between the predictors and the response variable. We do not utilize any automated variable selection method for our inference exploration, as variable selection methods such as stepwise selection may drop variables based on

its criteria that we are still interested in examining the relationship with the response variable. We utilize the training split from the 10,000-large sample as the data to train the model on, as mentioned in section II.

Model for Prediction

Whereas we utilize the full model for inference, we are interested in model performance for prediction. To select for the model, we utilize stepwise selection, which consists of iteratively adding and removing predictors, in the predictive model, to find the subset of variables resulting in the best performing model with the lowest prediction error. To select for variables, we began by performing both forwards and backwards selection on the full model. These forms of stepwise regression allow us to select for variables that result in the model lowest possible AIC value. AIC itself is an estimator of the predictive error of a model, allowing us to directly assess our model's performance.

Performing forwards and backwards selection resulted in the dropping of 6 variables from our model: BMI, $PhysicalHealth$, $MentalHealth$, $Race$, $PhysicalActivity$, and $SkinCancer$.

Discussion of Results

Full Model for Inference

Being that we used a logistic regression model to perform inference, we can interpret the coefficients in our results as the expected increase in log-odds of having heart disease for each predictor. Using $\alpha = 5$ as the significance level, we note that 10 predictor variables appear to be statistically significant. In the interest of Parsimony, we interpret these ten variables and their coefficient. The following interpretations are all mode assuming that all other variables are held constant:

The following age categories appeared to be significant in our model: 45-49, 50-54, 55-59, 60-64, 65 to 69, 70 to 74, 75 to 79, and 80+. Of these, the 80+ age category had the greatest coefficient of 2.3623. The odds that an individual who is 80 years old or older ($AgeCategory80orolder$) will develop heart disease is $e^{2.3623}$ = 10.6153 times greater than someone who is 24 years old or younger (the baseline category). The odds that an individual who is male ($SexMale$) will develop heart disease is expected to be $e^{0.7575}$ = 2.1329 times greater than an individual who is female.

The odds that an individual who has smoked at least 100 cigarettes in their lifetime ($SmokingYes$) will develop heart disease is expected to be $e^{0.1901}$ = 1.2094 times greater than an individual who hasn't. The odds that a male who drinks at least 14 alcoholic drinks per week or female who drinks at least 7 alcoholic drinks per week ($AlcoholDrinkingYes$) will develop heart disease is expected to be $e^{0.5441}$ = 1.7231 times greater than an individual who doesn't.

The odds than an individual who has had a stroke in their lifetime ($StrokeYes$) will develop heart disease is expected to be $e^{1.0265}$ = 2.7913 times greater than an individual who hasn't. The odds that an individual who is diabetic ($DiabetesYes$) will develop heart disease is expected to be $e^{.4200}$ = 1.5220 times greater than an individual who isn't. The odds that an individual who has asthma ($AsthmaYes$) will develop heart disease is expected to be $e^{0.2555}$ = 1.2911 times greater than an individual who doesn't. The odds that an individual who has kidney disease ($KidneyDiseaseYes$) will develop heart disease is expected to be $e^{0.7999}$ = 2.2253 times greater than an individual who doesn't. The odds that an individual who has difficulty walking or climbing stairs ($DiffWalkingYes$) will develop heart disease is expected to be $e^{0.3737}$ = 1.4531 times greater than an individual who doesn't.

All categories of self-reported general health appeared to be significant predictors of heart disease in our model. Of these, the self-categorized poor general health category had the greatest coefficient of 1.6418. The odds that an individual with poor general health ($GenHealthPoor$) will develop heart disease is expected to be $e^{1.6418}$ = 5.1645 times greater than those who aren't.

Overall, most of the findings are in line with our expectations and current medical literature, as generally variables that are associated with poorer health are also associated with higher odds of heart disease. However, interestingly, none of the BMI variables were significant in the model, perhaps reinforcing the growing idea that BMI itself is not a useful medical metric among medical professionals. Furthermore, `PhysicalHealth`

and `MentalHealth`, self-reported overall physical anad mental health, are not significant, which is interesting as self-reported general health is significant.

Predictive Model

To observe our model's initial predictive performance, we performed k-fold cross validation. K-fold cross validation starts by splitting the dataset into k equally sized "folds". Then, for each unique group, 1 fold is held out as the test dataset whereas the other remaining folds combine to form the training dataset. We found that for values $1 < k < 21$, $k = 5$ (5-fold cross validation) produced the highest model accuracy, at 91.28% (see Appendix D for full output).

For further interpretation and analysis of our model, we fit a random forest. A random forest combines multiple decision trees (in this case, 20) in order to yield the most common prediction class for certain values of the predictor variables (see Appendix E for the Random Forest error output table). The out-of-bag error from this model (the black line) is 0.1%, meaning we expect this model to classify patients incorrectly around 0.1% of the time. We can also observe the false positive rate (green) and false negative rate (red) that comprise this out-of-bag error. Overall, our predictive model has good predictive ability, especially when compared with the current methodologies utilized in the medical field, as mentioned in the introduction.

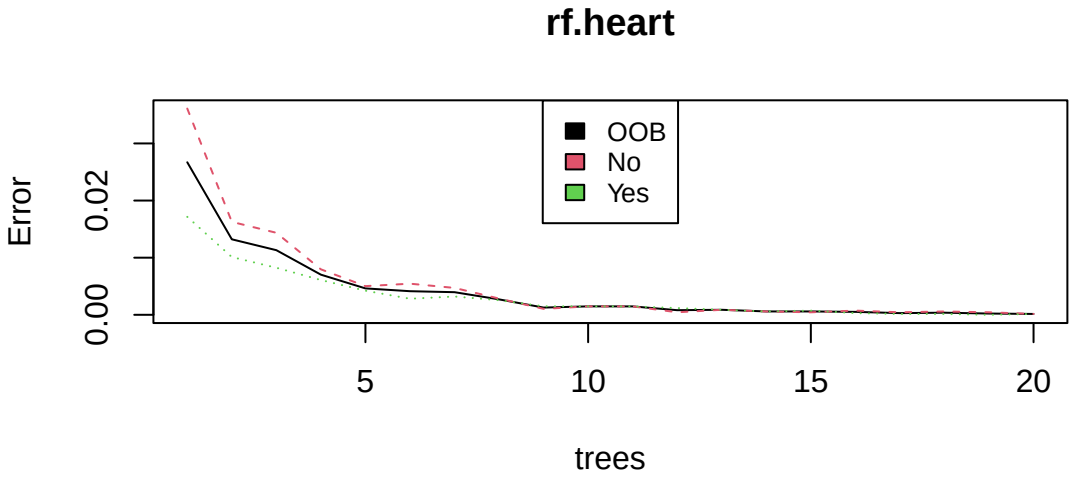


Figure 1: Error Lines for Random Forest

V. Limitations and Future Work

Below, we list out a number of potential limitations to our results, as well as future work that could be undertaken to improve on our analysis.

One limitation is model specification. The logistic regression model requires certain assumptions, including that there is a linear relationship between the logic of the outcome and each predictor, there are no influential values, and there are no high intercorrelations among the predictors. In choosing to utilize a logistic regression model, we assumed that there are no violations of these assumptions, which seemed reasonable. However, if any of these assumptions are violated, then a logistic regression model may not have been appropriate and thus another model class should have been utilized. In future work, we can more rigorously analyze these assumptions, such as creating smoothed scatter plots to analyze linearity, use Cook's distance to analyze influential values, and conduct Variance Inflation Factor analysis to analyze multicollinearity.

Another limitation is generalizability. While the dataset we used is relatively broad in its scope and application,

some of its aspects could have an impact on its generalizability. The data in the dataset comes from the Behavioral Risk Factor Surveillance System (BRFSS). Run by the CDC, this system collects health status data from individuals via telephone call. While this is likely the simplest way for this form of data collection, it does introduce a voluntary response bias that could affect the composition of the data. For example, someone with health issues could be more keen to sharing their health status data as opposed to a healthy individual who may believe the survey has no importance for them. Additionally, some individuals may not respond to the phone call altogether. These biases could skew the results of our methodology in a way we are not able to account for. In addition, the survey was conducted only on individuals within the United States, and thus the results are not generalizable to the global population. Potential future studies can include utilizing additional datasets that cover more geography and polling methods to improve the generalizability of our results.

Missingness may also be a limitation. The original dataset collected by the CDC consisted of over 400,000 observations of 300 different health and demographic variables. This was then reduced to the 18 variables in the dataset we used by its curator on Kaggle. Although this was done "to select the most relevant variables" and "be usable for machine learning projects", no in depth statistical detail was given for the removal of these variables. Therefore, we are unable to determine their possible influence on our results. It is possible that some of these variables could have been statistically significant predictors of heart disease in our inferential model or aided in the model accuracy for our predictive model. Potential future studies can include utilizing the original CDC data and all variables in our analysis, and mitigating the missingness with variable imputation methods.

Lastly, potential future studies may include adding interactions effects. Certain predictors may be related to other predictors. The incorporation of interaction effects may potentially allow for a more comprehensive model by indicating how the relationship between a predictor and the dependent variable changes depending on the value of another predictor (such as making gender an interaction effect so we can analyze how the other predictors' relationship with the dependent variable varies by gender).

VI. Conclusion

In this study, we sought to both understand the associations between a biological and demographic factors and the odds of having heart disease as well as create a predictive model for heart disease through an analysis of the CDC's 2020 annual health survey. To do so, we created two logistic regression models, where one was used to understand inferential relationships between the independent variables and the odds of having heart disease, and the other one was created using stepwise selection to act as as predictive model. We found that a number of biological and demographic factors are associated with a change in heart disease odds, and most of the factors we expected to have strong associations with a change in heart disease since they are also associated with changes in overall health. However, most interestingly, we found that BMI does not have a strong association with heart disease. We were also able to achieve high predictive accuracy with our predictive model that is able to match or surpass many commonly utilized methods in the real world currently.

Citations

Kaggle. (n.d.). Personal Key Indicators of Heart Disease. Kaggle. Retrieved October 15, 2022, from https://www.kaggle.com/datasets/kamilpytlak/personal-key-indicators-of-heart-disease

Appendix

Appendix A: Full Description of Variables

`HeartDisease`: An indicator variable that equals 1 if respondents indicate they have had coronary heart disease (CHD) or myocardial infarction (MI).

`BMI`: Body Mass Index (BMI)

`Smoking`: Indicator variable that indicates if the respondent has smoked at least 100 cigarettes, equal to 5 packs, in their life

`AlcoholDrinking`: Indicator variable that indicates if the respondent identifies as a heavy drinker (defined as having more than 14 drinks per week for adult men and more than 7 drinks per week for adult women).

`Stroke`: Indicator variable that indicates whether or not the respondent has had a stroke.

`PhysicalHealth`: The number of days over the past 30 days the respondent has had physical illness or industry.

`MentalHealth`: The number of days over the past 30 days the respondent has had poor mental health.

`DiffWalking`: Indicator variable that indicates if the respondent has had serious difficulty walking or climbing stairs.

`Sex`: Indicator variable that indicates the sex of the respondent as male or female.

`AgeCategory`: The age of the respondent, split into 14 levels.

`Race`: The imputed race/ethnicity of the respondent

`Diabetic`: Whether or not the respondent has had diabetes.

`PhysicalActivity`: Indicator variable indicating whether or not the respondent has done physical activity or exercise during the past 30 days other than their regular job.

`GenHealth`: The respondent's self-identification of their general health, split into five different levels, from "poor" to "excellent".

`SleepTime`: The average amount of sleep the respondent gets in a 24-hour period.

`Asthma`: Indicator variable indicating whether or not the respondent has had asthma.

`KidneyDisease`: Indicator variable indicating whether or not the respondent has had kidney disease, not including bladder infection or kidney stones.

`SkinCancer`: Indicator variable indicating whether or not the respondent has had skin cancer.

	estimate	std_error	z_value	pr_z
(Intercept)	-5.2663	0.6028	-8.7367	<0.001
BMIObese	-0.0027	0.1251	-0.0217	0.9827
BMIOverweight	0.1151	0.1173	0.9807	0.3268
BMIUnderweight	0.5892	0.3123	1.8868	0.0592
SmokingYes	0.1901	0.0924	2.0585	0.0396
AlcoholDrinkingYes	-0.5441	0.2357	-2.3083	0.021
StrokeYes	1.0265	0.1514	6.7805	<0.001
PhysicalHealth	-0.0694	0.1666	-0.4165	0.6771
MentalHealth	0.1042	0.1657	0.6287	0.5296
DiffWalkingYes	0.3737	0.1175	3.1800	<0.001
SexMale	0.7575	0.0942	8.0454	<0.001
AgeCategory25-29	-0.0191	0.6134	-0.0311	0.9752
AgeCategory30-34	0.8651	0.5040	1.7165	0.0861
AgeCategory35-39	0.5108	0.5186	0.9849	0.3247
AgeCategory40-44	0.4127	0.5132	0.8041	0.4214
AgeCategory45-49	0.9686	0.4730	2.0477	0.0406
AgeCategory50-54	1.1084	0.4625	2.3964	0.0166
AgeCategory55-59	1.6651	0.4418	3.7687	<0.001
AgeCategory60-64	1.7017	0.4386	3.8795	<0.001
AgeCategory65-69	1.8873	0.4361	4.3281	<0.001
AgeCategory70-74	2.1638	0.4358	4.9654	<0.001
AgeCategory75-79	2.2684	0.4406	5.1484	<0.001
AgeCategory80 or older	2.3623	0.4395	5.3743	<0.001
RaceAsian	-0.2517	0.5153	-0.4884	0.6253
RaceBlack	-0.1795	0.3951	-0.4543	0.6497
RaceHispanic	-0.2742	0.4018	-0.6824	0.495
RaceOther	0.0436	0.4355	0.1001	0.9203
RaceWhite	-0.0619	0.3612	-0.1715	0.8639
DiabeticNo, borderline diabetes	0.1267	0.2787	0.4547	0.6493
DiabeticYes	0.4200	0.1113	3.7751	<0.001
DiabeticYes (during pregnancy)	0.4535	0.5569	0.8143	0.4155
PhysicalActivityYes	-0.0226	0.1043	-0.2170	0.8282
GenHealthFair	1.6020	0.2061	7.7744	<0.001
GenHealthGood	0.8673	0.1871	4.6343	<0.001
GenHealthPoor	1.6418	0.2621	6.2650	<0.001
GenHealthVery good	0.5173	0.1880	2.7521	<0.001
SleepTime	-1.1975	0.7968	-1.5029	0.1329
AsthmaYes	0.2555	0.1223	2.0888	0.0368
KidneyDiseaseYes	0.7999	0.1566	5.1091	<0.001
SkinCancerYes	0.0803	0.1309	0.6131	0.5398

Appendix B: Full Model Output for Inferential Logistic Regression Model

	estimate	std_error	z_value	pr_z
(Intercept)	-5.3188	0.4822	-11.0294	<0.001
SmokingYes	0.1989	0.0916	2.1724	0.0299
AlcoholDrinkingYes	-0.5131	0.2345	-2.1878	0.0287
StrokeYes	1.0436	0.1507	6.9247	<0.001
DiffWalkingYes	0.3568	0.1115	3.1993	<0.001
SexMale	0.7537	0.0929	8.1125	<0.001
AgeCategory25-29	-0.0344	0.6128	-0.0562	0.9552
AgeCategory30-34	0.8367	0.5025	1.6651	0.0959
AgeCategory35-39	0.4905	0.5169	0.9489	0.3427
AgeCategory40-44	0.3904	0.5115	0.7631	0.4455
AgeCategory45-49	0.9759	0.4707	2.0733	0.0382
AgeCategory50-54	1.1053	0.4601	2.4026	0.0163
AgeCategory55-59	1.6716	0.4390	3.8080	<0.001
AgeCategory60-64	1.7229	0.4355	3.9566	<0.001
AgeCategory65-69	1.9007	0.4325	4.3943	<0.001
AgeCategory70-74	2.2025	0.4315	5.1041	<0.001
AgeCategory75-79	2.3048	0.4358	5.2884	<0.001
AgeCategory80 or older	2.4112	0.4338	5.5579	<0.001
DiabeticNo, borderline diabetes	0.1036	0.2769	0.3742	0.7083
DiabeticYes	0.3857	0.1082	3.5637	<0.001
DiabeticYes (during pregnancy)	0.4411	0.5563	0.7929	0.4279
GenHealthFair	1.5950	0.1989	8.0201	<0.001
GenHealthGood	0.8674	0.1855	4.6754	<0.001
GenHealthPoor	1.6536	0.2374	6.9667	<0.001
GenHealthVery good	0.5264	0.1872	2.8116	<0.001
SleepTime	-1.1752	0.7954	-1.4775	0.1396
AsthmaYes	0.2588	0.1218	2.1248	0.0337
KidneyDiseaseYes	0.8061	0.1554	5.1855	<0.001

Appendix C: Full Model Output for Prediction Logistic Regression Model

Appendix D: Predictive Model K-Fold Validation Output

```
## Generalized Linear Model
##
## 10000 samples
##    11 predictor
##     2 classes: 'No', 'Yes'
##
## No pre-processing
## Resampling: Cross-Validated (5 fold)
## Summary of sample sizes: 8000, 8001, 7999, 7999, 8001
## Resampling results:
##
##   Accuracy   Kappa
##   0.9127996  0.1339602
```

Appendix E: Output of the Optimal Random Forest Model in the First MICE Chain

Table 1: Output of the optimal Random Forest model in the first MICE Chain

Out-of-bag Error	**0.1%**		
Confusion Matrix			
	Predicted as 0	Predicted as 1	Class Error
Actually 0	6848	1	1e-04
Actually 1	1	6848	1e-04