# Statistical Analysis

To perform all data processing and generate the statistical analysis components of this study R version 4.2.0 was used within Rstudio 2022.07.2 Build 576. The data files utilized come from the NHANES 2007-2008 Questionnaire Data public data access page. The data is provided as separate XPT files each containing the survey results from various questionnaire categories. Each survey participant is assigned a sequence number which is used to match their responses across multiple questionnaires. All survey participants are anonymous, and the data provided does not include any personally identifiable information.

Necessary pre-processing of the data includes renaming variables for clarity, creating custom factors from categorical variables, and standardizing continuous variables. Many of the categorical variables were inclusive and informative enough to be used in their provided format, but some had to be customized or engineered to effectively represent the element. Examples of variables that required custom classifications are tobacco use, alcohol consumption, education status, marital status, born outside US, number of healthcare visits, and religious attendance. Continuous variables that require standardizing include fasting glucose, age, and RFM. RFM or ratio of fat-to-mass is a derived feature calculated using the weight (kg) and waist circumference (cm). The independent variable cardiovascular disease is derived by checking if the participant had reported to have been told they had had either a heart attack or stroke by a physician. Lastly, the dataset was filtered to only include participants aged 40+ and several data subsets were made which correspond to six multiple logistic regression models to be used as starting points for the model evaluation and selection process. Logistic regression requires that the observations do not contain any missing values which is a significant challenge for this study due to the sparsity of NHANES datasets. By creating subsets of the available data, we can retain a larger proportion of observations after removing any records with missing values. The six base models’ features were determined by considering best practices regarding selecting groups of variables which share common characteristics. Table x describes the sample size and number of parameters in each of the base models.

In this analysis we are examining six logistic regression models with different feature sets including social determinants, health indicators, and religious attendance to determine the relationship between these features and the likelihood of the participant having had cardiovascular disease. Three base models are initially defined which incrementally include additional features, and then each of the base models are split by gender resulting in the six models of interest. Models 1f and 1m include only the social determinant variables, models 2f and 2m include social determinants and health indicators, and models 3f and 3m include all the previous elements with the addition of religious attendance.

An initial determination of model fit is evaluated by calculating the Bayesian Information Criterion (BIC). The BIC measures how well a statistical model is fit to its data. This metric is derived based on the number of parameters in the model, the sample size, and the log-likelihood of the model. Typically, a lower BIC value indicates a better fit when comparing similar models. An important component of BIC is that it applies a penalty that favors models with fewer parameters, so it is a useful tool to help determine which elements are the most important.

Another key indicator used in this analysis to compare model fit is McFadden’s R2 and adjusted pseudo R2. In the context of logistic regression, McFadden’s R2 measures how well a model predicts the binary outcome based on the predictor variables. McFadden’s R2 is derived by calculating the proportion of log-likelihood of the full model to the log-likelihood of a null model, where a higher R2 value indicates a better fit model. In this analysis, we look at both the R2 and the adjusted pseudo R2. The adjusted pseudo R2 considers the number of model parameters and applies a penalty which favors models with fewer parameters like the BIC does.

To measure how well the models can distinguish between the binary class we generate receiver operating characteristic plots (ROC) and calculate the area under the curve (AUC). The ROC plot shows the true positive rate on the y-axis and the false positive rate on the x-axis. A line is plotted using different values for the classification threshold which is the minimum estimated probability needed for the model to classify the ‘positive’ class. In the context of this analysis the ‘positive’ class is where a survey participant is told they had had a heart attack or stroke by a physician. AUC values range from 0 to 1 where 0.5 would indicate that the model performs as well as a random guess and a value of 1 would indicate perfect classification. An AUC value of 0.7 or higher is typically considered to be a well fit model.

The predictive capabilities of the models are evaluated by conducting out-of-sample testing and examining the sensitivity and specificity of the resulting estimations. Each of the models are trained using a random sample of 70% of the available records for that model, then the coefficient estimates are applied to generate probabilities using the remaining 30% of records using an initial threshold of 0.5 for classification.

The ROC plot generated in the statistical analysis can also be used to help determine what the optimal classification threshold should be for each model. For any given threshold value the resulting sensitivity and specificity can be determined. Sensitivity is the ‘true positive’ rate while specificity is the ‘true negative’ rate. For this analysis we use both the F1 and balanced accuracy scores to evaluate how well the models can make correct classifications. The F1 score is the harmonic mean of the precision and recall of the model where precision is the proportion of positive predictions that are correct, and recall is the proportion of positive cases that were correctly classified. Balanced accuracy is the average of sensitivity and specificity and provides a more realistic measure of how well the model fits, especially when the data is imbalanced. F1 scores can be artificially inflated if a model tends to over-classify the positive case. For classification testing, we choose a threshold value which provides the highest balanced accuracy in each model.

Each model is trained using 75% of its corresponding data setting the remaining 25% aside for classification testing. The estimated parameters are then applied to the test set using the optimal threshold that was derived by maximizing the balanced accuracy of the base model. The resulting classifications are evaluated and presented below.

# Results

Table x displays the resulting Chi Square significance and BIC of each of the models. Each of the models have a chi square significance much less than 0, indicating that they are all a better fit compared to the ‘empty model’. That being said, model 1m is the most significant by several degrees, and models 2f and 3f both had a relatively low significance compared to the others. Although models 2f and 3f had relatively lower significance, they did have the lowest BIC scores of the set, and models 1f and 1m had the relatively higher scores. Looking at these metrics alone would indicate that the addition of variables in models 2 and 3 were beneficial and did not significantly affect their goodness-of-fit.

Looking at the table of odds ratios and feature significance, we can make the following observations. All six models indicated as age increases the odds of cardiovascular disease increases, and was significant (p < 0.001) in all models. All six models also indicated that individuals who have more frequent healthcare visits are more likely to have experienced cardiovascular disease. Models 1m, 2m, and 3m indicate that an increased measurement of RFM in males is a significant indicator which increases one’s risk. This variable was not significant in any of the female model counterparts. Models 2f and 3f indicate that being a current tobacco user increases risk for CVD. Models 2m and 3m indicate that having difficulty walking increases one’s risk. Model 1f was the only model where ethnicity appeared to be a significant factor, indicating that being a Hispanic female reduced one’s risk of CVD. Finally, Religious attendance was found to be significant in Model 3m, indicating that individuals who attend more than weekly may have a reduced risk of CVD. Some features that were not significant within a 95% confidence interval, but close enough to still possibly informative include having high blood pressure and being a past drinker, where some of the models were trending towards indicating that these features increased one’s risk of CVD.

The resulting metrics produced by the classification testing indicate relatively similar predictive power for all six models. The models with the highest relative balanced accuracies are models 1f and 2m. These two models also resulted in the highest F1 scores, where Model 2m stands out as the highest score of the set. Each of the models had relatively similar resulting sensitivities and specificities with the exception of model 3m which had an extremely high specificity.