**STA 5167 Final Project Proposal**

Brianna Alston1 and Lucas Nolting2

balston@fsu.edu1

ln18s@fsu.edu2

Dr. Qing Mai

Florida State University

March 20, 2023

## **Data Overview**

We will use the Heart Disease Data compiled from the results of patients undergoing angiopathy at four different health clinics in 1988. The four participating clinics were the Cleveland Clinic, the Hungarian Institute of Cardiology, the Veterans Administration Medical Center, and the University Hospitals in Zurich and Basel. A coronary angiogram is a low risk X-ray procedure performed on patients to examine blood flow that can be used to diagnose heart conditions. The data consist of continuous variables such as blood pressure and cholesterol, and indicator variables such as sex and type of chest pain.

More information on the data and initial study can be found at:

Link to data: <https://www.kaggle.com/datasets/johnsmith88/heart-disease-dataset>

Link to initial publication: [pubmed.ncbi.nlm.nih.gov](http://pubmed.ncbi.nlm.nih.gov)

**Approach**

Using the Heart Disease Data, our goal is to examine the relationship between heart disease and indicators for the disease (such as cholesterol and resting heart rate) to see if the markers can be used to predict whether a patient has heart disease. Such a model would be beneficial for identifying patients that may be at a greater risk for disease and to implement preventative treatments to potentially reduce their risk. We will use multiple logistic regression to investigate the relationship.

**Potential Issues**

We expect to encounter some issues when working with this dataset. Firstly, the data were collected across four different locations meaning that there is potential for inconsistency in how the values were measured. The data include multiple factor variables with several levels, which can make it difficult to interpret the multiple logistic regression model due to possible interaction terms and dummy variables.

In addition to potential problems with the data, we may have issues with our methodology. We wish to use multiple logistic regression, which we have not yet covered in class. We believe that multiple logistic regression will be effective for predicting the likelihood of a patient having heart disease. However, there is potential that other methods, such as linear discriminant analysis, could lead to better results. There is also a chance that our model assumptions will not hold. If this happens, we will attempt data transformations to correct any issues.