**BIOS735 Final Project Proposal**

**Group 2– Heart Disease**

**1 Introduction**

According to the [CDC](https://www.cdc.gov/heartdisease/risk_factors.htm), heart disease is one of the leading causes of death for people of most races in the US (African Americans, American Indians and Alaska Natives, and white people). We would like to use the 2020 annual CDC survey data to identify the risk factors that controbute to heart disease. 2020 annual CDC survey data of 400k adults related to their health status, It consists of 401,958 rows and 279 columns.

The survey dataset includes the response about the health status, which factors (questions) that directly or indirectly influence heart disease. One of the challenges of the dataset is the sparsity, which the responses contain a lot of missing data. We will propose statistical methods to handle the sparsity and missing issues.

**2 Aim**

Our study aims are as follows:

(1). We will assess the association and identfiy the risk factors that contribute to heart disease

(2) We will assess and compare the model fitting results (probability) for different methods (ie. EM, BFGS, Newton-Rapson (or IRSL), Stochastic Gradient Descent, etc). We will also do a simulation to compare the model fitting result.

**3 Method:**

We can apply the classic model, logistic regression, to access the relationship between heart disease and other variables.

**4 Analysis Plan**

We propose to use different methods to estimate parameters, ie. EM, BFGS, Newton-Rapson (or IRSL), Stochastic Gradient Descent, etc.