**Data File and Code Description - Group 4**

**Data Files:**

**framingham.csv** is the raw dataset.

**framingham\_filtered.csv** is the dataset where subjects who had only 2 measurements and had diabetes, MI or stroke at baseline are excluded. It is also used to compare the baseline characteristics of missing and nonmissing group.

**framingham\_nonafiltered.csv** is the dataset where the subjects who have missing values in BMI and totalChol variables are excluded. It is designed for further use in random sampling.

**Diabetes\_nona\_casecontrol** is the dataset for analysis of the effect of age on the diabetes (also used for baseline comparison between controls and cases).

**MISTRK\_nona\_casecontrol** is the dataset for analysis of the effect of age on the MI/stroke (also used for baseline comparison between controls and cases).

**Code Files:**

**Datapreprocess.r** cleans data and generates datasets proper for analysis based on exclusion criteria.

**BaselineCom.r** compares the baseline characteristics of missing data group and non-missing data group (for BMI or TotalChol).

**Baseline\_com\_case.r** compares the baseline characteristics of cases and control groups (MI/stroke or diabetes).

*Note:* The variable “**path**” has to be modified as per the location of the datasets on your computer.

**MainGEEModels.r**

1. 6 Binomial GEE models using logistic link function with AR1 correlation structure (R)
2. 6 Binomial GEE models using logistic link function with exchangeable correlation structure for sensitivity analysis (R)

*Note:* The variable “**path**” has to be modified as per the location of the datasets on your computer.

**MainGEEModels.sas**

1. 6 Binomial GEE models using logistic link function with AR1 correlation structure (SAS)
2. 6 Binomial GEE models using logistic link function with exchangeable correlation structure for sensitivity analysis (SAS)

**R Packages needed:**  
ggplot2  
geepack  
doBy

dplyr

tableone

\*Since the seed was not set when using random sampling to obtain control participants for comparison, thus this step for reproducibility is omitted. You can use the datasets as stated above for GEE analysis directly.