**Cohort 12**

**Team One Project Proposal**

# **Team Members:**

# Arnab Acharya Arun Natarajan Robert Raney Iaroslava Guk Donald Dawson

**Project Coordinator:**

Robert Raney

**Project Domain:**

Healthcare

# **Project Topic:**

Numerous national studies of mortality, especially cardiovascular mortality, by occupation, industry, income, education or other socioeconomic factors prove that strong relationships exist between socioeconomic factors and mortality. These relationships may be influenced by and may account for differences in mortality by race, ethnic origin, or geographic factors. Thus, the project topic of this team is to investigate social, economic, demographic and occupational differentials in mortality (total and by cause) within a national sample of the U.S. population.

# **Project Hypothesis:**

Probability of mortality by all and some specific causes among U.S. population can be predicted based on a number of socio-economic factors.

# **Project Data Sources:**

National Longitudinal Mortality Study (NLMS) dataset is a series of cross-sectional, national samples of the U.S. population, as identified from the Annual Social and Economic Supplement (ASEC – formerly the Current Population Survey) of the Bureau of the Census. For the purposes of the project, the latest release of NLMS will be used, containing surveys with full follow-up for 6 years, centered around 2002 with full follow-up for 6 years (n=745,162). Patient records were matched to the National Death Index to identify the occurrence and cause of death through 2011. Extensive demographic, social, economic, and occupation information was collected in the NLMS. The linkage of the individual social and economic data with the mortality outcomes provides the resource for extensive predictive analytics.

**Project Objectives & Approach:**

Using NLMS national patient data collected during a 6-year follow-up period, this team will test the hypothesis that certain factors help explain variations in the probability of mortality.(~40, including Geographic Location, Age, Income, etc.). Statistical models such as Cox proportional hazards (Cox PH) regression analysis and Logistic regression will be used to predict probability of mortality. Although the model will be predictive, it will be derived through the hypothesis that certain factors are causal factors in influencing mortality at any given age. .

Project will be broken down and executed in the following stages in order to achieve the above stated objectives:

**Stage 1: Data Ingestion**. Available dataset will undergo data analysis and wrangling to remove outliers, normalize data and infer missing variables when possible. Relational database management system (RDMS) or NoSQL DB will be used at this stage to facilitate working with data.

**Stage 2: Modeling & Application**. Normalized data will be fitted to Cox PH and Logistic regression models using Python libraries and packages to identify statistical significance of available variables in relation to mortality outcome. Best fitting model will be trained using train part of the dataset, i.e. 80%, and then cross-validated by predicting mortality outcomes for the test part of the data, i.e. remaining 20% of the patient records.

**Stage 3: Reporting & Visualization**. Tableau visualization will be created based on selected model to plot dependency of mortality on a number of statistically significant variables. Visualization of geographical distribution of patient population, distribution by age, gender, income etc. will also be considered.

By modeling such a vast dataset with a variety of features, the outcome might provide care givers and lawmakers the ability to make more informed decisions as it relates to the healthcare spending and quality of care.