PM606 Health Data Science Practicum (Summer 2021)

Background

The new Master's of Science in Public Health Data Science at USC has a summer practicum where students are tasked with preparing a report and presentation based on their analysis of external, unseen data and a set of interesting research questions. The California Teachers Study (CTS) with a focus on OSHPD hospitalization records will serve as the inaugural dataset for the practicum. These data will provide students with a unique learning experience where they must access an external data source through a secure server, deal with a large number of study participant records, and apply analytic tools that they have learned through the core courses of the Data Science program including data wrangling, data visualization, regression analysis, and machine learning.

Aims

The overall objective of the practicum project is to predict the short-term risk of death based on prior in-patient hospitalization, accounting for subject-specific factors such as baseline characteristics from the CTS questionnaires 1-3 (e.g. age, race, ethnicity, height, weight, family history, physical activity, diet, alcohol, and tobacco use) and hospitalization information (e.g. co-morbidities, procedures, length of stay, discharge types, hospital location). Hospitalizations of CTS participants from 2000 through 2015 will be used.

Specific aims include:

1) To develop the best fitting model, either through regression or machine learning, that predicts the probability of death within a certain time window (e.g. 30 days, 180 days, etc.) based on patient-specific baseline and hospitalization characteristics.
2) To assess whether the time window after hospitalization plays a role in predicting the risk of death.
3) To assess whether certain phenotypes of co-morbidities (coded from the first through fourth hospitalization diagnosis code) play a role in predicting the risk of death.
4) To determine whether there are any temporal (seasonal) or spatial (zip code) trends in hospitalization-related deaths.

Methods

The students will explore different methods of analysis to answer the specific aims. Data exploration methods will include summary statistics of key variables, and graphical displays of interesting relationships. Modeling methods will draw from their courses on logistic regression and machine learning (including random forests, gradient boosting, clustering). The analyses will be conducted in R, and the results will be presented in a report that is published on their personal GitHub website.

Statistical analysis plan

Endpoint definitions: death after in-patient hospitalization within different time windows of discharge (e.g. 30 days, 180 days, etc.)

Primary exposures: length of stay in hospital, co-morbidities recorded during hospitalization

Covariates: baseline participant characteristics including age, race, ethnicity, height, weight, family history, physical activity, diet, alcohol, and tobacco use
Primary outcome measure: binary variable indicating death after hospitalization
Statistical methods: logistic regression, random forests, gradient boosting, lasso, ridge regression
Sensitivity analyses: validation using training and test sets to determine model performance