Goal

- Analyze NHANES data to create a model that can predict the cancer survival status for all NHANES participants from 1999-2010

Background

- The National Health and Nutrition Examination Survey (NHANES) is a cross-sectional, nationally representative survey that assesses demographic, dietary and health-related questions and can be used to better understand differences in health and nutrition across the life-span.
- Almost all survey data are made publicly available by the National Center for Health Statistics (NCHS). https://www.cdc.gov/nchs/nhanes/

For this study, I will use NHANES III data. The Third National Health and Nutrition Examination Survey (NHANES III), 1988-1994, contains data for 33,994 persons ages 2 months and older who participated in the survey.

Data

The data and corresponding documentation for the survey interview and examination components are found in four separate data files: - Demographic data https://wwwn.cdc.gov/nchs/nhanes/search/datapage.aspx?Component=Demographics - Dietary data https://wwwn.cdc.gov/nchs/nhanes/search/datapage.aspx?Component=Dietary - Examination data https://wwwn.cdc.gov/nchs/nhanes/search/datapage.aspx?Component=Examination - Laboratory Data https://wwwn.cdc.gov/nchs/nhanes/search/datapage.aspx?Component=Laboratory - Questionnaire data https://wwwn.cdc.gov/nchs/nhanes/search/datapage.aspx?Component=Questionnaire

Mortality data can be obtained from NCHS Data Linkage NDI Mortality Data Mortality Data homepage: https://www.cdc.gov/nchs/data-linkage/mortality-public.htm Mortality Data: ftp://ftp.cdc.gov/pub/Health_Statistics/NCHS/datalinkage/linked_mortality/ Mortality Data Dictionary: https://www.cdc.gov/nchs/data/datalinkage/Public_use_Data_Dictionary_11_17_2015.pdf The sequence number (SEQN) allows for linking the mortality data with the NHANES data.

Methods

Create Cox Proportional Hazards model (using sample weights) Use statistical shrinkage (e.g. lasso, ridge, boosting) and tree-based (e.g. random forest, extra

trees) methods to determine which variables are most important to the model.