Group Members

Gina Peterson, Alex Jonas, Brandon Herrera, Claire Chen

Background

According to the Centers for Disease Control and Prevention (CDC), social determinants of health (SDOH) are nonmedical factors that influence health outcomes. SDOH factors include healthcare access, education, income, discrimination, and food insecurity. The World Health Organization (WHO) emphasizes the significance of SDOH research, citing that social factors may be of greater importance than healthcare or lifestyle choices in influencing health. Research cited by the WHO suggests that SDOH accounts for 30-55% of health outcomes. In recent years, government organizations and private sector companies have been focusing resources towards closing social gaps.

Objectives

- Evaluate the impact of social determinants on the following:
    - Unhealthy Behaviors (obesity, sleep, smoking, physical activity, etc.)
    - Use of Preventative Services (health insurance coverage, drug adherence, doctor's visits, etc.)
    - Health Outcomes (high blood pressure, diabetes, stroke, cancer, chronic kidney disease, mortality, etc.)
- Identify high risk populations to propose interventions and improve resource allocation.

Data

The 500 Cities-Places dataset is provided by the Centers for Disease Control and Prevention (CDC), Division of Population Health, Epidemiology and Surveillance Branch. The project for creation of the dataset was funded by the Robert Wood Johnson Foundation (RWJF) in conjunction with the CDC Foundation. The 500 cities data provides model-based small area estimates for 27 measures of chronic disease related to unhealthy behaviors (5), health outcomes (13), and use of preventive services (9). The dataset includes estimates for the 500 largest US cities and approximately 28,000 census tracts within the included cities. The estimates can be used to identify emerging health problems and to inform development and implementation of effective, targeted public health prevention activities.

The sample size of the CDC 500-Cities-Places dataset contains 810K values with 24 features. The dataset is open-source with no special permissions needed. The label for our data will be a combination of a few features provided, specifically: Measure or MeasureID (what exactly we are measuring i.e. "Obesity among adults aged >= 18 years"), Data_Value_Unit (i.e. "%"), and Data_Value (numerical value based upon Data_Value_Unit i.e. 24.2).

The goal will be to merge the 500-Cities-Places dataset with United States Census Bureau data in order to determine social determinants.

In conjunction with the CDCs 500 Cities-Places and Census Bureau databases we are also hoping to utilize the eicu-crd database (found here: https://eicu-crd.mit.edu/about/eicu/). The eicu-crd database provides numerous datatables with data collected from actual patients such as medication scheduling, past history, diagnosis, geographic information, nurse charting, etc. We are hoping the information provided by the eicu-crd database will help illuminate social determinant factors relating to disease onset for preventive care. The eicu-crd database will require permission in order to access.

Proposed Approach

We can split our approach into two components, prediction/analysis and visualization.

1. To preserve data interpretability, we avoid using PCA for feature reduction. Instead, we will conduct data preprocessing by computing the correlation matrix for all variables associated with social determinants of health, including the outcome variable.
2. We plan on using interpretable classification models in order to be able to analyze how much impact each social determinant has on the specified outcome. At the simplest level this would include using regression and assessing the coefficients to see the major contributors to the outcome. At a higher level we could use a neural network with self-attention to analyze cross-feature importance. After a model(s) has been trained, it can then be used to predict on unseen areas. This will predict the associated risks of certain locations so that help can be allocated to them accordingly.
3. After we train our model and predict on large amounts of areas, we will then put our data on an interactive map so that the government or private sector can utilize it to influence decision making.

Expected Outputs/Deliverables

1. A Correlation matrix for SDOH variables and outcomes.
2. An interpretation of model(s)' performance including classification metrics and the contribution of each SDOH factor.
3. A comparison between our model and a statistical model (e.g. mixed model in R)
4. An interactive map indicating the risk level of unhealthy behavior, use of prevention services, and health outcomes.