# Impact of Behaviors on Chronic Diseases

## Team members: Yu Fu, Hao Li, Nina Nguyen

## Problem statement

Chronic disease in the United States is a considerable burden for patients and health insurance payers. Chronic diseases can lead to high medical bills, impaired physical mobility, and increased risk of sudden death, ultimately causing a decline in the quality of life of patients. The factors that cause chronic diseases are the complex effects of genes and acquired environments, which leaves room for interventions to prevent or reduce chronical disease. Human daily behaviors, including diet, sleep, and exercise, are important environmental factors for chronic diseases. The goal of this project is to study how could interventions of daily behaviors improve health conditions regarding chronic diseases in the population level and quantify the improvement by treatment cost saved.

For this project, we choose diabetes and cardiovascular diseases (CVD) as target chronic health conditions. We will analyze possible interventions on behaviors and quantify the improvement of health conditions of the population in Seattle, on both city level and zipcode level.

## Data sources

- The Behavioral Risk Factor Surveillance System (BRFSS)
    - The primary source for demographics, behavior and health condition variables
    - 2018 BRFSS data https://www.cdc.gov/brfss/annual_data/annual_2018.html
    - Potential to merge data from previous years
- National Health and Nutrition Examination Survey (NHANES)
    - Mash with BRFSS to incorporate extra diet data, examination data, and laboratory data
    - 2015-2016 data is what's available for now, will check for the update for 2017-2018 data release as project proceeds
    - https://wwwn.cdc.gov/nchs/nhanes/default.aspx
- Zipcode demographics
    - Collect demographic facts on zipcodes in Seattle
    - Data appear as web visualizations instead of table format. Might require developing a scraper or manually collect from the website.
    - https://censusreporter.org
- Cost of chronic disease per year
    - Cost for CVD and diabetes (controlled and uncontrolled)
    - For quantifying economic values of interventions. Ideal to collect cost for both patients and health insurance payers.
    - The data resource is not confirmed yet and requires additional research. Might be scattered in different research papers and CDC reports.

## Algorithms / solution technologies

1. From BRFSS and NHANES dataset, identify three categories of variables to keep because the dataset is large with lots of variables that might not be useful in our project.
    - Variables of demographics, such as age, gender, ethnicity, income, etc.
    - Variables of daily behaviors, including alcohol use, diet, sleep, exercise, and smoke.
    - Variables of chronic diseases, including diabetes and CVD.
2. Mash BRFSS and NHANES dataset to create a bigger and complete dataset because the examination and laboratory data in NHANES are instrumental in determining the health condition. BRFSS will be the base dataset since it has more features and records. The mashing algorithm will be based on inner joins and nearest neighbor algorithms.
3. After mashing the dataset, we will project the data on different zipcodes of Seattle. The data already has national weights, which represents how many persons does each record represents in the United States. Since zipcodes in Seattle have different demographic distributions compared to the United States, the weights will be different. We are going to generate zipcode weights for our dataset using linear optimization. The constraints will be distributions of demographic variables (probably also use prevalence rate of health conditions).

   For example, suppose we have two persons in our dataset, and two zipcodes, the following tables means person 1 represents 70% of 98105's population and represents 40% of 98103's population.

   | ID | 98105 | 98103 |
   |---|---|---|
   | 1 | 0.7 | 0.4 |
   | 2 | 0.3 | 0.6 |

4. Build predictive models on health condition variables using behavior variables. The target variable could be a binary indicator of whether the person has a specific disease, or a continuous measurement of blood sugar level. Therefore, both regression and classification models are needed. We will try different linear models and tree-based models for the best outcomes.
5. Apply behavioral interventions to the population, then see how the health condition variables will change using predictive models. Aggregate the prevalence rate and total cost on a specific zipcode using the weights produced in step 1.

   For example, an intervention would say "Reduce intake of carbohydrates by 20%", and we expect a decrease in the diabetes prevalence rate. Therefore, when we quantify the effects, we could say "annual healthcare cost of 98105 decreased by XXX because diabetes prevalence rate in this region decreases from XX% to XX%".

   Note here we are not looking at the individual level, because a specific person could not revert from diabetic to un-diabetic, even with the correct intervention. We are looking at the diabetes prevalence rate as a whole population.

   In this way, if we have several intervention options, we could apply them all, and identify the most effective intervention on each zipcode, and Seattle city level.

6. (Optional if have extra time) We could develop several segmentations over the population, using demographic and behavior variables. Different population segmentations share similar characteristics and often similar medical needs. For example, the population segmentation could be "goes to the gym often and mainly eat protein" or "sleep for 12 hours and take 6 shots of alcohol drink every day". We can then try out different interventions on different segmentations of the population for a more diverse and effective outcome. The clustering algorithm will be used for creating population segmentations.

## Risks

1. There is a risk that none of the behavior variables are strong predictors of health condition variables. That means the effects of interventions are minimal because the interventions are only performed on one or few behavior variables.
2. The data is mainly from public surveys, so the quality of the data is an issue. We need to check for invalid values, missing values, and outliers in data cleaning steps, and make some decisions on whether to drop features (columns) or drop records (rows) or impute the data.
3. Another issue from the survey is that the sample size is always limited. Some minority groups might not be well represented in the survey. The minority groups are the groups with low prevalence rates in the combination of various demographic variables. For example, we might only have one record represents the group "native American + 70-79 years old + <15000 annual income + high school education". As a result, we will either completely ignore these minority groups or have a very high bias on them in our solution.

## Challenge

1. One of the challenges we will most likely face is the difficulties during the EDA and data cleaning process. We are looking at different large datasets and will need to be careful with outliers and missing data.
2. Identifying the necessary variables from the different datasets might be challenging as well. We will ultimately need to decide what will variables will have a direct effect on the health condition.
3. Trying out different predictive models requires lots of experiments, which is time-consuming. We need a good design for the experiment to pick the right models.
4. Definitions of different intervention are relatively subjective, and it is challenging to convert behavioral interventions from descriptive text to numerical modification of variables.