# BMI706 - Project: Dataset and Tasks

*Group: 415*
Joanna Shen, Ziqian Liao

## 1. Identification of Dataset

We will be utilizing a publicly available dataset of the U.S. Chronic Disease Indicators (CDI), sourced from the Division of Population Health at the Centers for Disease Control and Prevention (CDC) [1]. This comprehensive dataset includes chronic disease data pertaining to key areas of public health practice as reported by U.S. states and territories via the Behavioral Risk Factor Surveillance System (BRFSS). The dataset consists of an extensive array of time-labeled chronic disease prevalence statistics and data on relevant risk factors. (e.g., asthma mortality rate and sale of cigarette packs) The abundance of entities and observations, along with the potential for identifying associations between variables, made the dataset an ideal subject for exploration and visual presentation.

## 2. Data Summary

The CDI dataset is a comprehensive collection of reported statistics spanning a time frame of 2001 to 2021, sourced from a total of 54 states and territories across the United States. Comprising 1,185,676 entries and 34 attributes, this dataset includes responses to 203 unique public health questions, with a majority of the responses stratified according to gender and race/ethnicity.

Here is a sample of the range of topics covered by this dataset, including but not limited to:

- Asthma prevalence and mortality rates
- Various types of cancer prevalence and mortality rates
- Melanoma prevalence and mortality rates
- Hospitalization rates for chronic obstructive pulmonary disease
- Alcohol use patterns across different age groups
- Obesity prevalence rates among different populations
- And more.

Given the extensive availability of quantitative and qualitative data labeled with both time and location, we will carefully select those elements that

demonstrate the potential for establishing associations or inferring causality and thus be of significance in advising public health practices.

## 3. Exploratory Analysis

Through exploratory analysis of this dataset, we aim to show temporal and spatial patterns that can be found in the statistics related to different kinds of chronic diseases such as obesity, arthritis, chronic lung disease, and so on. This will allow us to present visual information on the changes in trends associated with these types of chronic diseases, which are the leading causes of death and disability in the United States and leading drivers of the nation's healthcare costs. In the end, we hope to find out the most common characteristics among different types of chronic diseases.

## 4. Target Audience

Our main goal is to produce visualizations that can raise awareness of chronic diseases and correlating factors and reduce the burden on the healthcare system. This visualization targets the general population so as to provide them with a more comprehensive and detailed overview of how chronic diseases develop in the United States, healthcare providers so as to prepare them for upcoming treatment of chronic disease, and policy makers who focus on public health so that they can make evidence-based decisions in the future.

## 5. Visualization Tasks

The main tasks for this exploratory visualization are as follows:

- Temporal chronic liver disease cases and deaths per state
- Relating alcohol use, poverty, and physical activity to different kinds of cancer cases and deaths
- Relating cigarette use and tobacco-related laws to asthma deaths
- Relating obesity among high school students to whether they are allowed to purchase soda or fruit drinks their computer use, and other behavior
- And more.

## References

1. Centers for Disease Control and Prevention. (2021). U.S. Chronic Disease Indicators. CDC Division of Population Health. Retrieved February 16, 2023, from https://chronicdata.cdc.gov/Chronic-Disease-Indicators/U-S-Chronic-Disease-Indicators-CDI-/g4ie-h725