**Data Cleaning – Team 32 (Non-Communicable Diseases)**

The initial dataset was sourced from the Institute for Health Metrics and Evaluation (IHME) Global Burden of Disease (GBD) database, encompassing annual state-level health outcome data across the United States from **2010 to 2019**. The dataset included observations for **over 300 non-communicable diseases** (NCDs), each measured by four key outcomes: **Deaths**, **DALYs (Disability-Adjusted Life Years)**, **Prevalence**, and **Incidence**.

**Preprocessing Steps**

1. **Filtering by Disease Category:**  
   We limited our analysis to diseases classified as *non-communicable* within the GBD framework. Communicable, maternal, neonatal, and nutritional diseases were excluded at this stage.
2. **Removal of Aggregate or Redundant Categories:**  
   Disease groupings such as “All Causes” or duplicated parent-level aggregates were excluded to prevent overrepresentation and ensure disease-level granularity.
3. **Missing Data Handling:**  
   Observations with missing values in the val, year, or cause\_name fields were removed. Additional filtering excluded rows with missing gender (sex\_name) or age group (age\_name) metadata, which were critical for subgroup analysis.
4. **Standardizing Column Types:**  
   Key variables, including location\_name, cause\_name, measure\_name, and expansion\_group, were converted to categorical types. All numeric values, such as health outcomes (val), were checked for outliers or anomalies and confirmed to be in interpretable units per capita.

**Disease Selection for Causal Analysis**

To narrow the scope of our causal analysis and focus on conditions with the highest policy relevance, we conducted a **preliminary statistical filtering process**:

**Two-Sample T-Test Procedure:**

For each disease and measure, we compared **pre-expansion (2010–2013)** vs **post-expansion (2014–2019)** means. This test was applied within the subset of states that eventually expanded Medicaid.

* **Inclusion Criteria:**
  + The difference in means was **statistically significant** at *p < 0.05* in at least one outcome measure.
  + The disease showed **meaningful prevalence and public health relevance**, based on policy interest and published literature.
  + Diseases with **extremely low population burden** (e.g., <10 cases across all years) were excluded to ensure interpretability and robustness.
* **Selection Outcome:** This process resulted in a final list of **10 non-communicable diseases**, which were used for all subsequent Difference-in-Differences (DiD) and Fixed Effects modeling. The selected diseases spanned chronic conditions (e.g., diabetes, kidney disease), behavioral health (e.g., substance and opioid use disorders), and neurological disorders (e.g., Alzheimer’s disease).

**Final Dataset Characteristics**

* **Time Range:** 2010–2019
* **Geographic Coverage:** 50 U.S. states
* **Expansion Groups:** Early, Mid, Late, Never
* **Measures:** Deaths, DALYs, Prevalence, Incidence
* **Sample Size:** ~120,000 observations after filtering
* **Treatment Definition:** A binary variable treat\_group indicating whether a state belonged to an expansion group (vs. Never) and a post variable indicating the post-expansion period.

**Selected 10 Statistically Significant Diseases for Analysis**

1. Substance use disorders
2. Drug use disorders
3. Opioid use disorders
4. Diabetes and kidney diseases
5. Chronic kidney disease
6. CKD due to diabetes mellitus type 2
7. Alzheimer's disease and other dementias
8. Alcohol use disorders
9. CKD due to hypertension
10. Uterine cancer