Cole Hoffman

BIOS 7719 – Information Visualization

Final Project Report

Introduction

Chronic pain affects nearly 100 million adults in the United States, with severity influenced by factors such as age, gender, and health history (Dayer et al., 2018). Before the 1980s, research on opioid prescriptions for chronic pain was limited to studies that examined acute pain or disease-specific causes (Dayer et al., 2018). However, at this time, medical professionals began advocating for opioids to manage chronic pain, prompting pharmaceutical companies to increase production and marketing of opioids. From 1996 to 2001, Purdue Pharmaceuticals held conferences to train medical professionals on their long-lasting opioid, Oxycontin, and incentivized sales representatives to boost drug sales (Dayer et al., 2018). Such actions have increased the prescriptions of opioids for non-cancer pain management tenfold from 1996 to 2003, ultimately giving rise to the opioid epidemic. As federal regulations tightened on prescription opioids, many users turned to the black market for substances like heroin, synthetic opioids, and more recently fentanyl, further exacerbating this crisis. Understanding trends in historic opioid use is crucial for developing impactful public health policy, and treatment programs are vital in addressing the opioid epidemic that continues to ravage the United States.

Many Americans are deeply affected by the rippling effects of this crisis, and the data used in this project can begin to answer some of these dire research questions. These include discovering spatial and temporal patterns in drug and opioid use over the past decade, and whether previously enacted policies aimed to curb drug use had any discernible effect among the general population. Therefore, the intended audience for these data include medical professionals (i.e., physicians, pharmacists, nurses), public health researchers (i.e., biostatisticians and epidemiologists), policymakers, and the public. As a result, I decided to develop an interactive visualization tool that allows researchers and general users access to these data and manipulate key features so they can use the data to answer their desired questions.

Data Abstraction

The primary dataset for this project is titled "Provisional Drug Overdose Death Counts" and is available for public use on the Center for Disease and Control's (CDC) website (see ***References*** for the link). The data are sourced from the National Vital Statistics System (NVSS), which collects death certificate data from state and local vital statistics offices. The dataset consists of provisional drug overdose death counts across all 50 U.S. states and the District of Columbia. Provisional death counts, particularly for drug overdoses, often underestimate the final mortality figure due to the time lag in cause of death reporting. These provisional counts include foreign residents and are based on death certificates processed by the CDC's National Center for Health Statistics as of a monthly cutoff date (usually the first Sunday of each month). Lag times for cause of death typically range from 4-6 months and impact data completeness. States are evaluated based on data quality metrics including percentage of deaths pending investigation, data completeness, and drug specificity. For states that meet these thresholds, statistical methods are used to adjust provisional counts using multiplication factors (e.g., 1.1 for 90% completeness) to better approximate true final death counts. Drug overdose deaths may be recorded under multiple ICD-10 codes, and finals counts include only U.S. residents. These standards align with those outlined by the World Health Organization (WHO).

The primary dataset consists of monthly provisional and final drug overdose death counts for all U.S. states and D.C. from January 2015 to December 2024. The raw data are organized in wide tabular format comprising 69,738 items (rows) and 12 unique attributes (columns). Before conducting any statistical analyses, several important data cleaning measures were taken. First, the core data were sorted chronologically and reshaped from wide to long format. This transformation combined observed and predicted death counts into a single 'count' attribute, and a new 'type' attribute was created to distinguish between the two measurements. Second, the long format dataset was filtered to only include the following six relevant attributes for the development of the interactive visualization tool:

- *'state'*: a **categorical**, **nominal** variable containing the abbreviated 2-letter state name for all 50 U.S. states (including DC).
- *'month'*: a **categorical**, **ordinal** variable with **sequential ordering** to denote the month of the reported death counts/records.
- *'year'*: a **categorical**, **ordinal** variable representing the calendar year (2015-2024) for reported death counts.
- *'indicator'*: a **categorical** variable containing the underlying cause-of-death codes from the Tenth Revision of the International Statistical Classification of Diseases and Related Health Problems (i.e., ICD-10). These codes encompass 9 distinct drug classes (with ICD-10 codes) and 3 summary measures:
    - Heroin (T40.1)

    - Natural & semi-synthetic opioids (T40.2)
    - Methadone (T40.3)
    - Synthetic opioids, excl. methadone (T40.4)
    - Natural & semi-synthetic opioids, incl. methadone (T40.2, T40.3)
    - Cocaine (T40.5)
    - Natural, semi-synthetic, & synthetic opioids, incl. methadone (T40.2-T40.4)
    - Opioids (T40.0-T40.4,T40.6)
    - Psychostimulants with abuse potential (T43.6)
    - Percent of drug overdose deaths with at least one drug specified
    - Absolute number of drug overdose deaths
    - Number of total deaths reported (drug overdose or otherwise)
- *'type'*: a **categorical** variable indicating whether the death count was observed or predicted.
- *'count'*: a **numerical** count variable denoting the number of deaths.

At this point, I created subsets of tabular data based on the desired visualization idiom (see ***Visual Designs*** section for more). Line and stacked bar plots were constructed using state-level subsets which were derived by removing redundant indicators and aggregating certain drug classes into broader categories to avoid duplication and visual clutter. Furthermore, the three summary metrics were filtered out, leaving the following six indicators:

- Heroin (T40.1)
- Natural & semi-synthetic opioids (T40.2)
- Methadone (T40.3)
- Synthetic opioids, excl. methadone (T40.4)
- Opioids (T40.0-T40.4,T40.6)
- Stimulants (i.e., Cocaine and Psychostimulants with abuse potential) (T40.5, T43.6)

The national-level dataset for the U.S. cartogram included the same six drug classes, but also retained the three summary metrics. However, these data only considered the observed death counts, whereas the state-level data included both observed and predicted death counts.

Task Abstraction

The provisional death count dataset from the CDC provides a rich foundation for exploring temporal and geographic patterns in drug overdose mortality, with implications for public health response, surveillance, and policy. At a high level, users can **consume** the data to carry out **discovery** tasks. Analysts and policymakers can **explore** broad trends in drug

overdose deaths over time and **identify** any emerging patterns or outliers. These actions can lead to an effective assessment of the impact of the ongoing opioid epidemic, such as the surge in synthetic opioid (e.g., fentanyl) fatalities associated with increased illicit distribution. At a mid-level, users can perform **targeted lookups**, such as **querying** a specific state or drug class to monitor death counts, evaluate local spikes, or **compare** values against prior expectations. At a low level, the data allows for **comparative** analyses; for example, users can compare or contrast **trends** across states, within states over time, or among different drug classes. These data encourage the development of granular hypotheses regarding geographic disparities or the timing of mortality shifts.

Visualizations derived from this dataset, including line graphs and stacked bar charts, help translate complex data into more interpretable formats for diverse user groups. These graphics serve both exploratory and explanatory purposes at both the state and national level.

Visual Designs

All visualizations were developed using R Version 4.3.3. Showcasing these idioms involved building an interactive web application using the 'shiny' package in R, and visualizations were constructed primarily using the 'plotly' R package. Visualization designs were constructed by leveraging the CDC's provisional death counts dataset to explore temporal and spatial trends in drug overdose mortality.

At the top level, the interactive web app presents a cartogram of the United States (Figure 1), with each state shaded a different saturation level of red color to represent the total number of drug overdose deaths. This map serves as an interactive starting point for geographic exploration where users can hover over states to reveal tooltip information such as total overdose deaths for the selected date range, as well as the percentage of those deaths with at least one identified drug class.

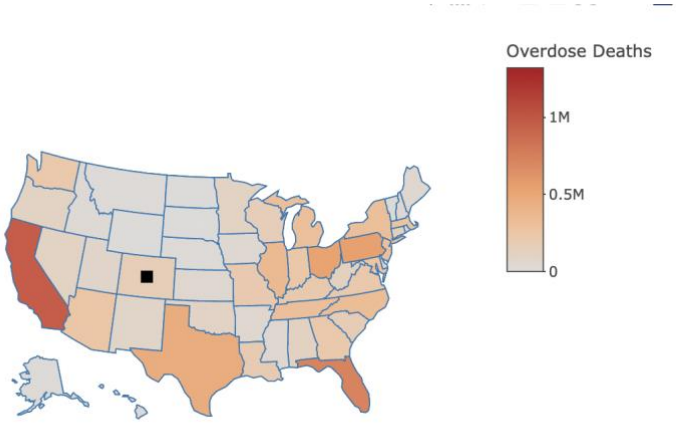
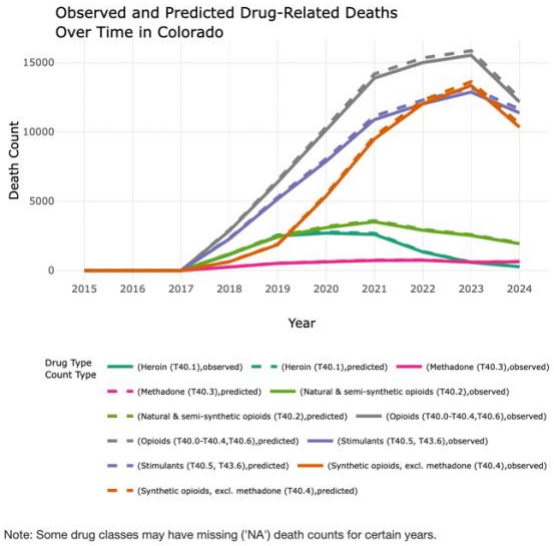


Figure 1: Cartogram of the United States in the R Shiny app

Selecting a state on the map dynamically reveals two interactive linked idioms below the U.S. map: a line plot and a stacked bar plot (Figures 2 and 3, respectively). The line plot displays the observed and predicted death counts over time, stratified by drug class. A solid line distinguishes the observed deaths, whereas a dashed line signifies predicted deaths. The stacked bar plot only shows observed death counts, with the height of each colored segment denoting the contribution of a specific drug class to the overall total death count. For both plots, the y-axis is a numeric measure of total death count, while the x-axis discretely measures time by year (2015-2024). Color encodes drug class in both plots, allowing for consistency across the linked views.



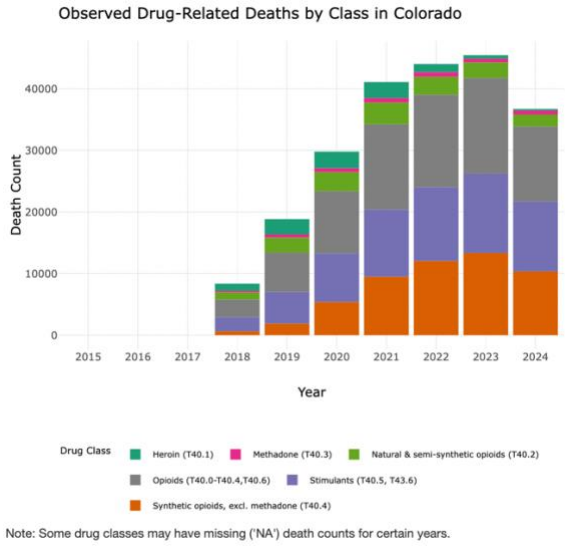


Figure 2 (left): Interactively linked line plot showing the observed and predicted death counts; Figure 3 (right) stacked bar plot showing the observed counts. Both plots feature death counts stratified by drug class.

Tooltips are implemented throughout to enhance user interactivity and bolster overall user experience, as well as provide detailed, on-demand information. In the line plot, hovering over the lines at discrete time points reveals the drug class, death count type (observed vs. predicted), and the total number of deaths. In the stacked bar plot, tooltips show the year, drug class, and exact death count corresponding to the highlighted bar segment.

To further enhance data exploration and the user's interactive experience, a menu on the top left provides additional interactive controls. Users can select a desired state from a dropdown if geographically identifying the state on the U.S. map proves to be difficult. Other controls include filtering by drug class, providing a cleaner view of death counts in the linked plots, as well as a year-range slider that enables users to zoom in on shorter time frames. These interactions allow users to filter data precisely and uncover patterns that may have been initially obscured in the broader 10-year range view.

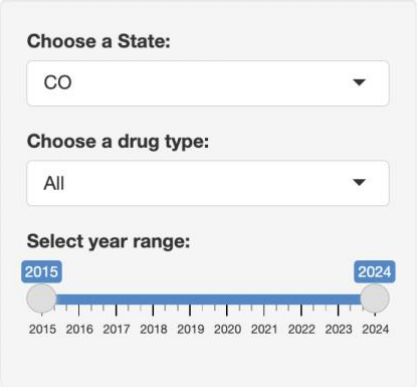


Figure 4: Interactive controls menu located in the top left corner of the R shiny app.

These visual idioms were chosen to conveniently support both spatial and temporal aspects of the data. The cartogram offers a geographical overview and highlights potential regional disparities, while the state-level linked idioms allow for deeper exploration of temporal trends over time. Coordinated views pave the way for seamless comparisons to be made between the selected state and national patterns. The use of both observed and predicted death counts in the line plot provides insights for both general users and public health researchers to retrospectively evaluate whether the CDC's algorithm for calculating predicted death counts closely aligns with the actual death count for a given state and year. Omitting predicted death counts from the stacked bar plot bypasses occlusion, reduces visual clutter and improves clarity. Overall, the app follows a structured overview-zoom-filter framework to effectively handle data complexity and interpretability while catering to a wide range of users and a diverse set of tasks.

Results: Implementation and Use Case

Upon launching the app, the user is brought to the first of three tabs labeled "**Summary of Dataset**." This tab summarizes the source of the data (CDC), how they were compiled, and provides basic information about the dataset's dimensions (see ***Data Abstraction*** section of this paper). It also identifies the app's intended end users, which include both public health researchers and members of the general public. The second tab, "**Dataset Information**," describes the CDC's rigorous data collection and processing methods, which ensure that predicted and observed death counts are accurately separated from the provisional death counts. For more details, users are referred to the linked dataset on the CDC's website or the ***Data Abstraction*** section above. The third tab, "**Plots**," features the three visualization idioms that make up the R Shiny app (see ***Visual Design*** section of the paper).

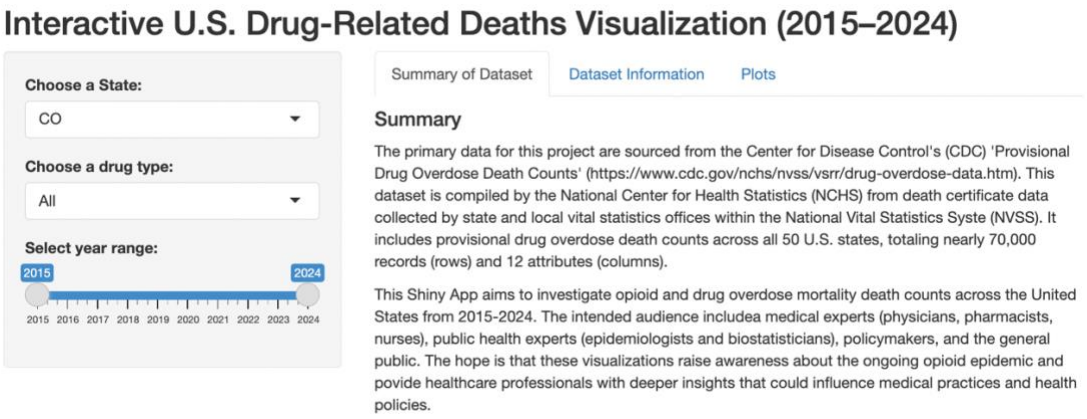


Figure 5: General R shiny app dashboard that is featured upon running the app's R code.

To illustrate the potential uses of this app, consider two potential users: a public health researcher and a member of the general public.

Example 1: Public Health Researcher

Suppose an epidemiologist is interesting in evaluating whether the establishment of several new drug rehabilitation centers in Washington state has had an impact on opioid-related deaths. These centers opened in 2019, and since the dataset spans from 2015-2024, it provides a suitable time window for investigation. The researcher can select Washington either by clicking on the state in the U.S. cartogram or by choosing "**WA**" from the dropdown menu. Upon selection, the linked line and stacked bar plots below update to reflect Washington state-specific data. The line plot illustrates a steep rise in drug overdose deaths in Washington from 2019 to 2023, followed by a decrease in 2024. While no causal relationship can be drawn from this visualization alone, the data serve as a valuable

hypothesis-generating tool. The researcher could use this exploratory insight to form new research questions and design follow-up analyses.

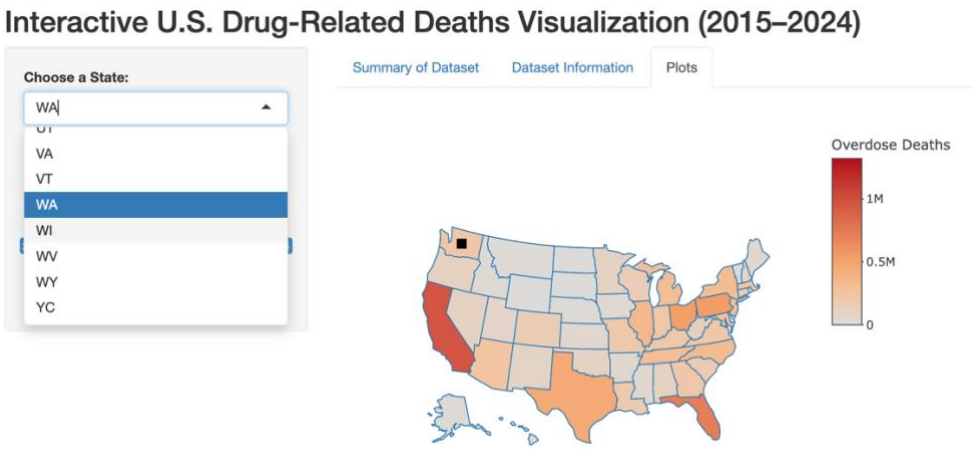


Figure 6: Selecting the state of Washington, either by clicking directly on the U.S. map or by navigating to the upper left dropdown menu and selecting from there.

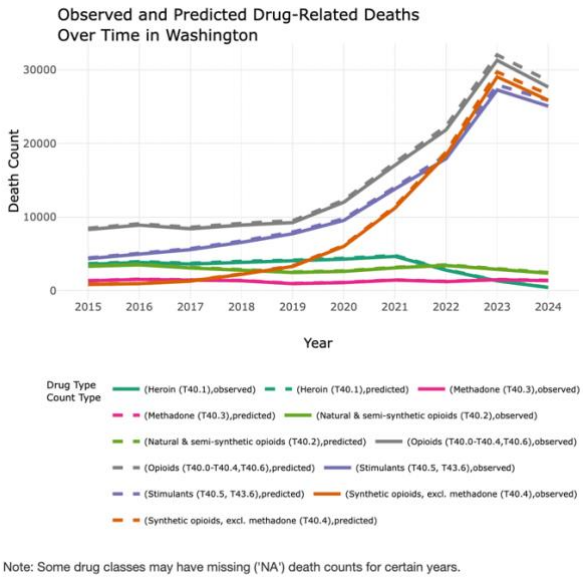


Figure 7: Line plot illustrating the observed and predicted drug overdose death counts for Washington state.

Example 2: General Public User

Now suppose a general user wants to learn more about the severity of synthetic opioid use in Colorado. They can click on Colorado in the cartogram or select "**CO**" from the dropdown menu. Then, by selecting "**Synthetic Opioids, excl. methadone (T40.4)**"from the drug type dropdown, all other drug classes are filtered out, effectively reducing visual clutter. By default, the time window sets to the full 10 year range (2015-2024), permitting the user to observe how synthetic opioid death counts have evolved over time. Upon

viewing the line plot, the user will likely notice lower death counts in the earlier years, followed by a sharp increase in deaths related to synthetic opioids in the early 2020s. This observation highlights two key insights: the lack of state data collection during the early years of observation, and the widespread increase in illicit synthetic opioid distribution.

The user can hover over specific time points in the line plot to reveal details like drug class, death county type (observed vs. predicted), and the exact number of deaths. Alternatively, the stacked bar plot provides a clearer picture of the increase over time, with bar height representing the magnitude of observed deaths.

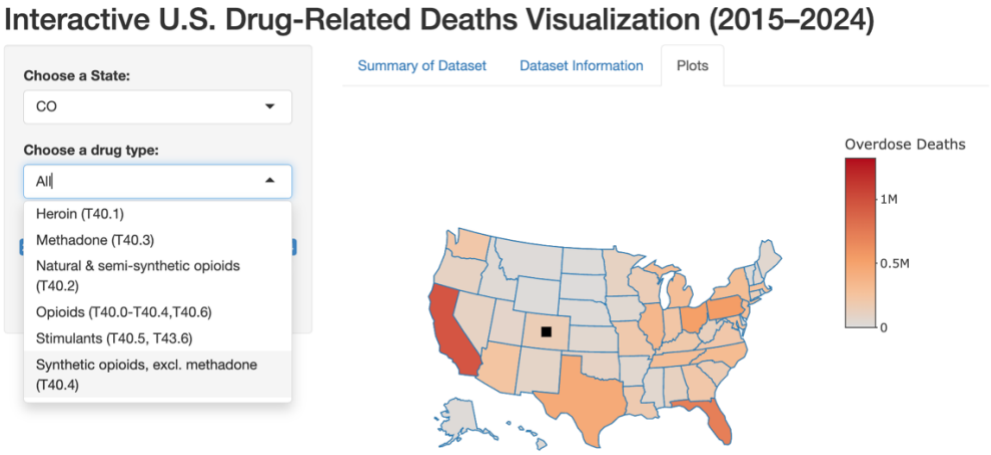


Figure 8: Selection process for specifying drug class of interest.

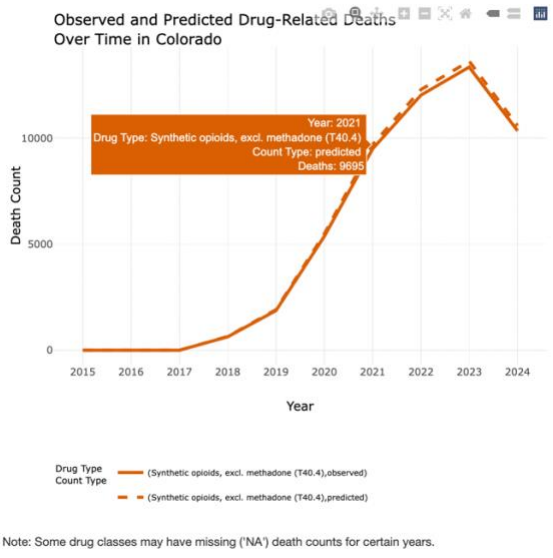


Figure 9: A filtered line plot showing only the drug overdose death counts for synthetic opioids for Colorado.

In both use cases, the epidemiologist and the general user used the app to explore how drug overdose mortality has changed in their respective states. They both employed lookup

techniques to locate their home states and used the interactively linked visualizations to compare death counts across time and geographical location. These use cases show how the app enables a diverse user base to **consume**, **explore**, and interpret the data for both research and raising general awareness.

Discussion

Although the CDC's provisional drug overdose death counts data are hypothesis-generating and offer a valuable insight into the issues of the ongoing opioid epidemic, users should be wary of interpretation. The visualizations support exploration and pattern recognition, but they cannot establish any causality. Observed associations may be confounded by unmeasured variables, and without further statistical analyses, trends must be interpreted as exploratory rather than conclusive.

Another limitation relates to how these data are displayed geographically. The U.S. map shows absolute overdose death counts, which may mislead users about the real impact of the opioid epidemic across states. For instance, California appears deeply saturated at all time points due to its large population (over 40 million people), making it the highest in absolute death counts. On the other hand, states like Montana and Wyoming appear lightly affected, not necessarily because overdose death rates are lower, but because their populations are much smaller. Without normalizing death counts (i.e., deaths per 100,000 people), users might misinterpret which states were most affected by the opioid epidemic.

Another important consideration involves how drug-related deaths were reported. Death certificates may list multiple substances, leading to polysubstance reporting. This means that one death might be counted under more than one drug class, thereby inflating death counts when investigating specific drug classes. This overlap emphasizes the importance of also reviewing national-level summaries, where methods for accounting such duplication could be more refined.

Lastly, the classification of drug types presents some challenges. Although nine drug classes are represented in this app, some of these categories are redundant and might overlap with others, potentially introducing ambiguity for users. To mitigate this, the visual design includes dropdown filters that permit users to isolate specific drug classes, allowing for a more tailored data exploration process.

In the future, the app could be improved upon by incorporating some individual-level covariates such as sex, age group, race/ethnicity, and geographical subregions (e.g. county-level statistics or urban vs. rural classification). These variables would provide more granular insights into drug overdose mortality trends and might help uncover underlying

health disparities that are obscured by state and national level aggregated data. For example, certain demographic groups may be disproportionately affected by synthetic opioid deaths, or simply devoid of accessible health care due to living in more rural environments. Enriching the dataset in this way would support the development of informed interventions aimed at reducing inequities in drug overdose mortality.

References

Dayer, L. E., Painter, J. T., McCain, K., King, J., Cullen, J., & Foster, H. R. (2018). A recent history of opioid use in the US: Three decades of change. *Substance Use & Misuse*, *54*(2), 331–339. https://doi.org/10.1080/10826084.2018.1517175

Centers for Disease Control and Prevention. (2025, March 12). *Products - vital statistics rapid release - provisional drug overdose data*. Centers for Disease Control and Prevention. https://www.cdc.gov/nchs/nvss/vsrr/drug-overdose-data.htm