Is There a Relationship Between Health Insurance Coverage Rates and Opioid Overdose Deaths at the County Level in the United States?

Luke Athans, Yiting Bao, Matthew Holland, Sophie Jubert, Benjamin Zivan

I. Introduction

Prior to COVID-19, the US was already in the throes of another decades-long health epidemic: Opioid addiction and overdose deaths. Beginning in the 1990's with the advent of plentiful oxycodone (via the brand name OxyContin), the US found itself in what would become the first wave of a currently three-wave opioid epidemic [8][1] that has claimed hundreds of thousands of lives and has worsened during COVID-19 [1].

II. Background and Problem Statement

While there have been many efforts to determine predictors of Opioid Related Deaths (ORD's) both through directly related statistics such as Per Capita Pill Volume (PCPV) [7] and general environmental (e.g., socioeconomic) factors [14], these efforts have generally been focused on small geographic regions [14], and static choropleth maps are the most advanced visual aids.

We built on these earlier studies by making an interactive county-level map of the United States that shows trends in opioid prevalence between 2006-2020 and layered it with US Census Small Area Health Insurance Estimate (SAHIE) model data (only available between 2008-2019) to assess what relationship, if any, healthcare coverage may have as far as impacting ORD's. Since Early Medicaid has been shown to mitigate ORD rates via lower PCPV [7], we wanted to investigate health insurance coverage in general for a similar effect.

III. Literature Review

With respect to data visualization, the journal articles we found generally display one or more of the following visualizations of results: A numerical table [12, 16] with model prediction accuracy [3, 4, 9, 11, 13]; a line chart or scatterplot [2, 4, 6, 9, 10, 11, 13, 14, 16]; a bar chart [4, 5, 6, 9, 16]; static choropleth maps [1, 2, 5, 6, 7, 8, 9, 10, 11, 13]. We improved on each of these by creating an interactive visualization that updates in real time, which we have experienced in the homework assignments for this class.

The Washington Post, who is responsible for public hosting of the ARCOS dataset and created the API for accessing it, created one time-lapse choropleth visualization, but this is displayed as an animated GIF on the website as opposed to an interactive feature allowing dynamic user control, such as zoom or filter options like we employed.

We have found many instances of machine learning models applied to data from the opioid epidemic, such as clustering [16], ensemble methods [2, 14], binary classification [6], clustering [17], regression [1, 2, 3, 4, 6, 7, 11, 14, 16, 17], deep learning [3, 4], and NLP [13]. Among papers seeking better information about predictors of the opioid epidemic [1, 3, 4, 6, 7, 10, 11, 13, 14, 15, 16, 17], several use privileged datasets such as electronic health records [3, 4], geo-located social media posts [11], or hospital admission data [16, 17]. While these sorts of datasets can be important tools for fighting the opioid

[1] Bracketed numbers correspond to the numbers in the *Sources* section of this proposal.

crisis, they are not always freely available, so if similarly helpful conclusions can be drawn from public datasets, such as the three that we use in this project, it will be easier to build further projects from the results.

IV. Proposed Method

Intuition

In the following sections, we describe a novel database designed to merge previously unmerged datasets from the CDC, DEA (via the Washington Post), and Census Bureau, a dynamic user-controlled visualization of these merged datasets, and several algorithms that we used to search for relationships in the data that could in turn help future efforts to combat the US opioid epidemic.

Database

The original data sets were acquired from a combination of file downloads and API's. Due to limited documentation and functionality with the CDC API, we opted to build a custom database allowing us to easily aggregate and facilitate our analysis. The raw data was downloaded locally and analyzed for data irregularities such as missing categorical information, incomplete counts across selected variables, standardization of id's for joining (e.g., county codes, state codes), and calculation of summary statistics. A python script was written to automate the data cleaning process and load the data into a Google Big Query database on Google Cloud Platform. We uploaded the data from our three data sources into five distinct tables within Big Query: SAHIE, DEA, UCD (underlying cause of death), and MCD (multiple cause of death), and a decoded version of MCD. We later added CDC opioid volume data between 2006-2020 since DEA data for opioid volumes is only publicly available through 2014. We also noted that for ORD rates obtained through the CDC database that data for many counties was supressed due to confidentiality constraints and NCHS data use restrictions. Due to these data limitations, the choropleth maps showing crude rate data for UCD and MCD appear sparser than those shown for opioid volume or health insurance coverage rates.

Visualization

Our visualization is based on the D3 JavaScript library, which was chosen for both its versatility in visualization as well as the universality of JavaScript. We used it to build choropleth maps to update based on user choices for up to two variables of interest. The left dropdown allows the user to view a choropleth map of annual data at a county level for the following selections: Opioid prescription rates per 100 people, Uninsured percentage, and Deaths per 100,000 people. The right dropdown introduces another D3 map that displays the results of clustering based on different variables of interest (UCD or MCD). The cluster colors are displayed above the clustering map along with the number of cluster members beneath each cluster's respective color. Counties labeled N/A on the clustering maps are those where data was available. However, they were deemed outliers where no cluster could be assigned. The user can also select the year for which they are interested in the results via a slider underneath the maps. Tooltips were built to allow the user to see the specific rates for each county at a glance (the maps are zoomable if the user wants to browse particularly small counties). For the MCD and UCD cluster tooltips, users can see Death rates per 100,000 people for age and gender demographic categories and drug specific causes of death.

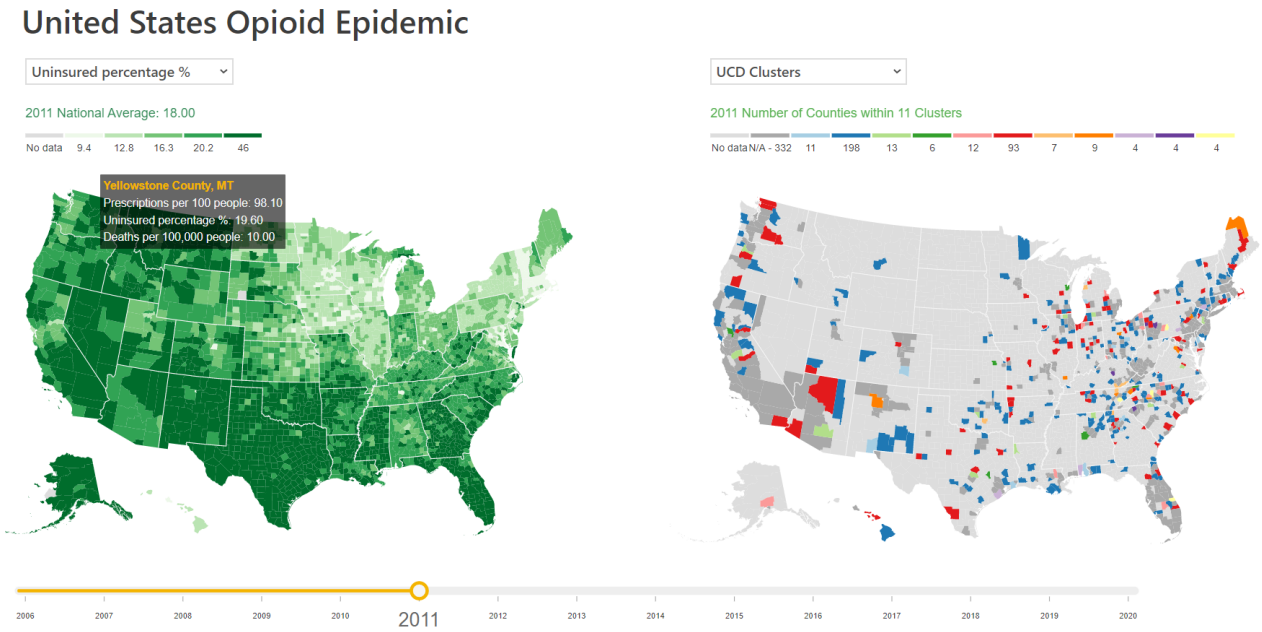

Figure 1: Our D3.js visualization with side-by-side maps comparing Prescriptions per 100 people and clustering for MCD data in the year 2020. Year can be adjusted at the bottom between 2006-2020.

Figure 2: Dropdown Menus for left and right maps, respectively

Figure 3: The details from the mouseovers on counties with available data from the maps on the visualization. The mouseover on the left is for prescriptions per 100 people, uninsured percentage, and deaths per 100,000 people. The center mouseover is for MCD clusters. The mouseover on the right is for UCD clusters.

And lastly, clicking on a county with data on the right-side map will open a new tab displaying a bar chart showing cumulative gender-based data for deaths by drug group. Using a python script to produce bar charts, we isolated each county's cumulative deaths and displayed these values based on the International Classification of Diseases (ICD) code associated with each entry [18]. This data was reported by gender with each bar representing a different drug category.

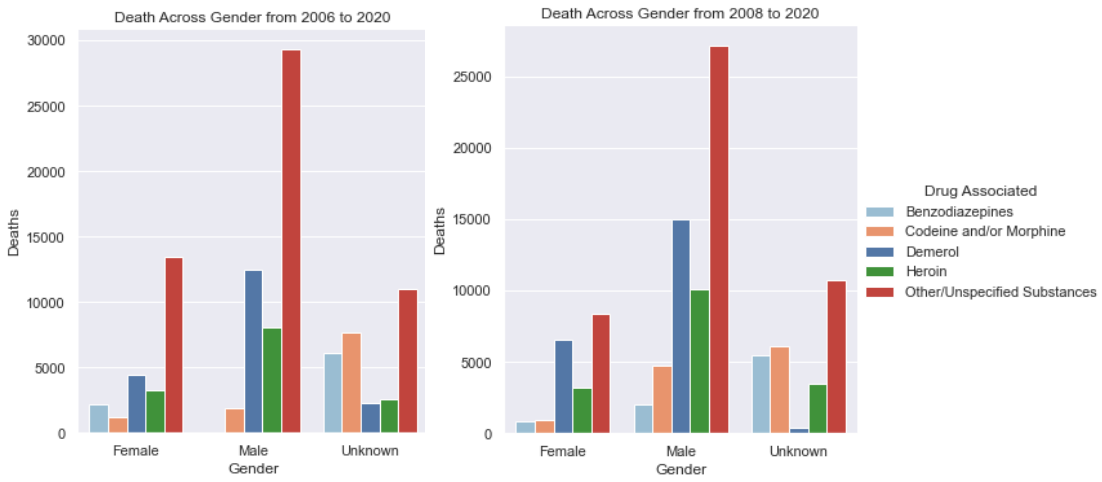

Figure 4: Deaths by gender for Cumberland County, NC (Left) and Berkeley County, WV (Right)

Algorithms and Modeling

We first ran regression to see whether there appeared to be any obvious correlations between features. After noting both gaps in the datasets (e.g. many counties not reporting data, or otherwise reporting lumped data, such as with the "Unspecified Substances" category in the screenshots above) as well as the seeming lack of any clear relationships between any predictors other than PCPV and ORD's, we changed our analysis to clustering to provide a broad view of the different facets of the opioid epidemic in the US.

After seeing inconclusive results from regression models, we decided to apply a clustering model to find similarities between different counties. We settled on a density-based spatial clustering algorithm (DBSCAN) because its clusters are flexible, relying not on absolute distance between cluster members, but on the relative distance between the next-closest member. This allowed us the greatest potential for discovering latent patterns of self-similarity in the data. We chose a minimum cluster size of 4 to prevent overfitting, but roughly half the counties with data for any given year failed to be clustered.

List of Innovations

1. Interactive D3.js visualization with extensive user control and a wide variety of data options.
2. Simultaneous linking of CDC, DEA, and SAHIE datasets to seek significant factors that are related to opioid related deaths. CDC and DEA datasets have been previously studied, but ever since the Affordable Care Act's debate and passage in the late 2000's, there has been increasing public interest in single-payer insurance, or government-paid healthcare, thus our interest in examining the SAHIE health coverage rates in relation to ORD rates.
3. Density-based spatial clustering model on UCD and MCD data across age, gender, and race to find self-similarities (if any) between how the opioid epidemic manifested between 2006-2020.

V. Experiments/Evaluation

Testbed/Questions

1. Does health insurance coverage correlate with ORD's on a county level?
2. Does the PCPV correlate with ORD's on a county level?
3. What commonalities, if any, exist between counties' MCD statistics?

Experiments and Observations

Regression models were created with Python and evaluated using an 80/20 train-test split. The quality of the model was quantified by examining the following statistics: R-squared, Adjusted R-squared, and mean squared error. The data was preprocessed by one-hot encoding categorical data and using StandardScaler to scale numerical data.

Our initial approach was to perform regression on a national level, but the correlations between predictors were quite weak, with R^2 ranging between 0 and 0.4, the highest of these between crude rate (ORD rate/population) and PCPV. When considering categorical variables related to race, the ORD rate appeared significantly higher in the Native American population, and lower with Asian/Pacific Islander.

Box-and-whisker plots for pills per capita grouped by race indicate that Asian/Pacific Islanders and Black/African Americans receive *fewer* opioids than other races. And while there appears to be no major difference in that same metric grouped by gender, from looking at pills per capita grouped by age, we see that the aggregated results appear to mimic most closely the statistics for the "Unknown" category. And while this sounds surprising, we also note that the "Unknown" category for each grouping accounts for the majority of datapoints.

Upon seeing these weak correlations, we redid the regression analysis at a county level, yielding generally stronger correlations, but again due to the overall paucity of the data (around 50% of counties are not reported in any given year), as well as the COVID epidemic, we concluded that the model had uncertain predictive utility at best.

Conclusions about the data are elusive from our clustering analysis as well, except to say that in any given year, roughly 50% of counties for which we have data are assigned to no clusters, which we believe implies that the US opioid epidemic is extremely unique insofar as the ability of health insurance, PPC, or types of opioids to predict ORD's with any certainty.

With respect to the cumulative deaths for the gender-based data, males appeared to be overwhelmingly more affected by drug overdose death than females. One shortcoming encountered was poorly detailed data collection involving drug overdose deaths. A large proportion of drug deaths are classified as "Other/Unspecified Substances".

Overall, the same issue that plagues the gender-based analysis of drug overdose deaths is a significant factor in analyzing data about the US opioid epidemic as a whole: Despite the seeming plethora of data

available, it is often redacted, incomplete, or poorly tabulated, thus confounding more targeted analyses.

Future Research

The challenges we faced during our research were all related to inconsistencies in data collection and reporting between different jurisdictions. Although the CDC provides the most complete data set available to report on ORDs in the United States, it contains mostly incomplete information due to both data privacy constraints and data availability. When looking at demographics such as race, gender or age, data was often incomplete, and the specific drugs contributing to underlying cause of overdose deaths remained unreported. Furthermore, death crude rates were shown to be on the rise year-over-year, signaling that the opioid epidemic is still widely affecting our national population.

These observations suggest that the true impact of the opioid epidemic remains vastly underreported, therefore making any research and mitigation efforts uncertain at best. In future research on this topic, we would employ machine learning techniques to provide better estimates of true ORD numbers across different demographic categories and attempt to provide better visibility into which specific drugs induce the highest overdose rates. Additionally, we would broaden predictors to include various economic indicators, such as per capita income, unemployment statistics, and investment figures. Our hope is that with better data availability of specific contributing factors related to opioid abuse that support efforts can be optimized through targeted outreach to the groups and communities with the highest risk indicators.

VI. Team Member Effort Statement

All group members had distinct tasks of equal importance, difficulty, and time commitment for this project. In addition, we left room for dynamic collaboration as time and interests allowed.

Sources

[1] Bharat, Chrianna, et al. "Big Data and Predictive Modelling for the Opioid Crisis: Existing Research and Future Potential." *The Lancet Digital Health*, vol. 3, no. 6, 2021, doi:10.1016/s2589-7500(21)00058-3. https://tinyurl.com/mwehyzvm

[2] Boslett, Andrew J., et al. "Using Contributing Causes of Death Improves Prediction of Opioid Involvement in Unclassified Drug Overdoses in US Death Records." *Addiction*, vol. 115, no. 7, 2020, pp. 1308–1317., doi:10.1111/add.14943. https://pubmed.ncbi.nlm.nih.gov/32106355/

[3] Dong, Xinyu, et al. "Machine Learning Based Opioid Overdose Prediction Using Electronic Health Records." *AMIA ... Annual Symposium Proceedings. AMIA Symposium*, American Medical Informatics Association, 4 Mar. 2020, www.ncbi.nlm.nih.gov/pmc/articles/PMC7153049/.

[4] Dong, Xinyu, et al. “Predicting Opioid Overdose Risk of Patients with Opioid Prescriptions Using Electronic Health Records Based on Temporal Deep Learning.” *SSRN Electronic Journal*, 2020, doi:10.2139/ssrn.3708326. https://pubmed.ncbi.nlm.nih.gov/33711546/

[5] Furst, John A., et al. “Pronounced Regional Disparities in United States Methadone Distribution.” *Annals of Pharmacotherapy*, vol. 56, no. 3, 2021, pp. 271–279., doi:10.1177/10600280211028262. https://pubmed.ncbi.nlm.nih.gov/34184584/

[6] Gavali, Sachin, et al. “Understanding the Factors Related to the Opioid Epidemic Using Machine Learning.” *2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 2021, doi:10.1109/bibm52615.2021.9669486. https://arxiv.org/abs/2108.07301

[7] Griffith, Kevin N., et al. ”Implications of county-level variation in U.S. opioid distribution.” *Drug and Alcohol Dependence*, Volume 219, 2021,108501, ISSN 0376-8716, https://doi.org/10.1016/j.drugalcdep.2020.108501.

[8] Kiang, Mathew V., et al. “Assessment of Changes in the Geographical Distribution of Opioid-Related Mortality Across the United States by Opioid Type, 1999-2016.” *JAMA Network Open*, vol. 2, no. 2, 2019, doi:10.1001/jamanetworkopen.2019.0040. https://tinyurl.com/23hrhu6n

[9] Madera, Joshua D., et al. “Declines and Pronounced State Disparities in Prescription Opioid Distribution in the United States.” 2021, doi:10.1101/2021.12.02.21266660. https://www.medrxiv.org/content/10.1101/2021.12.02.21266660v2

[10] Mattson, Christine L., et al. “Trends and Geographic Patterns in Drug and Synthetic Opioid Overdose Deaths — United States, 2013–2019.” *MMWR. Morbidity and Mortality Weekly Report*, vol. 70, no. 6, 2021, pp. 202–207., doi:10.15585/mmwr.mm7006a4. https://tinyurl.com/3peuaa2w

[11] Nguyen, Thuy, et al. “Comparison of Rural vs Urban Direct-to-Physician Commercial Promotion of Medications for Treating Opioid Use Disorder.” *JAMA Network Open*, vol. 2, no. 12, 2019, doi:10.1001/jamanetworkopen.2019.16520. https://tinyurl.com/36ybpexd

[12] Rose, Mark Edmund. “Are Prescription Opioids Driving the Opioid Crisis? Assumptions vs Facts.” *Pain Medicine*, vol. 19, no. 4, 2017, pp. 793–807., doi:10.1093/pm/pnx048. https://pubmed.ncbi.nlm.nih.gov/28402482/sa

[13] Sarker, Abeed, et al. “Machine Learning and Natural Language Processing for Geolocation-Centric Monitoring and Characterization of Opioid-Related Social Media Chatter.” *JAMA Network Open*, vol. 2, no. 11, 2019, doi:10.1001/jamanetworkopen.2019.14672. https://jamanetwork.com/journals/jamanetworkopen/fullarticle/2753983

[14] Schell, Robert C., et al. “Identifying Predictors of Opioid Overdose Death at a Neighborhood Level with Machine Learning.” *American Journal of Epidemiology,* 2021, doi:10.1093/aje/kwab279. https://academic.oup.com/aje/article/191/3/526/6433428

[15] Wilson, Nana, et al. “Drug and Opioid-Involved Overdose Deaths — United States, 2017–2018.” *MMWR. Morbidity and Mortality Weekly Report*, vol. 69, no. 11, 2020, pp. 290–297., doi:10.15585/mmwr.mm6911a4. https://www.cdc.gov/mmwr/volumes/69/wr/mm6911a4.htm

[16] Fulton, Lawrence et al. “Geospatial-Temporal and Demand Models for Opioid Admissions, Implications for Policy.” Journal of clinical medicine vol. 8,7 993. 8 Jul. 2019, doi:10.3390/jcm8070993. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6678995/

[17] (**Preprint of [16]**) Fulton, L.; Dong, Z.; Zhan, F.B.; Kruse, C.S.; Granados, P.S. Geospatial-Temporal and Demand Models for Opioid Admissions, Implications for Policy. *J. Clin. Med.* **2019**, *8*, 993.

[18] World Health Organization. (2016). International statistical classification of diseases and related health problems (10th ed.). https://icd.who.int/browse10/2016/en

Data Sources

[A] Ba Tran, Andrew et al. “The Opioid Files: Drilling into the DEA’s pain pill database.” *The Washington Post.* 2020, doi: https://www.washingtonpost.com/graphics/2019/investigations/dea-pain-pill-database/?itid=lk_inline_manual_18

[B] Centers for Disease Control and Prevention: CDC Wonder Database, “Multiple cause of death.” 2022, doi: https://wonder.cdc.gov/mcd.html

[C] United States Census Bureau, “Small Area Health Insurance Estimates (SAHIE) Program.” 2021, doi: https://www.census.gov/programs-surveys/sahie.html