CSE6242 Project Final Report

Fall 2019

Group Members: Adam Awad, Arthur Wang, Evan Marcantonio, Jay Kinzer

Table of Contents

Motivation 1
Problem Definition 2
Survey of Related Literature 2
  Geographic Factors 2
  Demographic and Economic Factors 3
  Other Factors 3
  General Areas for Improvement 3
Methodology 3
  Data Pipeline 4
  Predictive Modeling 6
  Interactive User Interface 6
    Investigation 6
    Prediction 8
  Innovations 9
Experiments and Evaluation 10
Conclusions and Discussion 10
Plan of Activities, Timeline, and Distribution of Effort 11
References 13

Motivation

In the United States, the opioid-related mortality rate has been increasing steadily over the past two decades. While this is a common problem across the US, it manifests and affects different parts of the country differently. The ability to analyze the mortality rate alongside contributing factors at a local-level yields additional insights into areas that are most-affected by this crisis.

Problem Definition

Changes in the opioid mortality rate can vary greatly in different areas of the country and amongst different subsets of the population. While relevant data and literature on the topic are available, they are typically static papers or diagrams with several common shortcomings:

i. **Geographic** – Analyses are typically generalized at a state / national level, or focused on a specific region (city, MSA, or single-state).

ii. **Contributing Factors** – Factors that may explain or contribute to the mortality rate are often omitted or evaluated in isolation. The type of area (urban vs. rural) as well as the person's age, gender, or ethnicity need to be considered together to understand the issue fully.

iii. **Temporal** – Analyses frequently use data from a snapshot in time and are always done retrospectively.

We are addressing these shortcomings by analyzing the rate of opioid-induced overdose deaths in the United States using a variety of data sources and have developed an interactive visualization to enable users to explore this data. Our analysis is presented at a county-level, giving a granular look at how rates vary geographically, and focuses on variations in different population subsets based on age, gender, and ethnic groups. Lastly, we provide forecasted future mortality rates; leveraging available county and demographic-level data as predictors.

Survey of Related Literature

Our problem definition builds on existing research on certain factors contributing to opioid-related mortality: geographic, demographic, and economic factors. Research in each area motivates our approach; below, we summarize existing research as well as areas we have identified as opportunities for improvement.

Geographic Factors

Much of the existing research examines variation in opioid-related mortality by geography. Specifically, research shows variation along the rural/urban continuum [5, 6, 11], as well as between regions [4], states [1, 10] and counties [9]. This research motivates our use of various geographic categories as variables in our predictive model and interactive visualization.

One shortcoming of some papers is that they only show variation between states or rural/urban areas [10, 11], masking variation between counties. In addition, two studies focused specifically on New York [2, 15]. Our approach will surface data at the most

granular level available in order to enable “deep dives”. We plan to retain the state and urban/rural geographic attributes in order to enable higher-level analysis as well.

Demographic and Economic Factors

Existing research also examines variation in mortality along lines of race, sex, age, and education [1, 2, 3, 5, 8, 10, 15]. There are also associations with income, poverty/economic distress, and unemployment [3, 4, 7]. This research indicates that such demographic and socioeconomic attributes will be useful variables for our predictive model; in addition, the data are freely obtainable and can be attributed to specific US counties.

While some papers [2] include detailed results of their statistical analysis, including coefficient sizes and p-values, others present figures that only describe variations [3, 10]. One improvement our project can bring is a visualization of which factors have the largest effect on mortality and in which direction. In addition, we can analyze demographic/socioeconomic factors at the county level, rather than a state or regional level.

Other Factors

Several papers also found an association between mortality rate and other factors not covered in the previous sections. For example, graduation rates [1]; the rate of opioid prescriptions [7, 12, 14]; the number of opioid-related emergency department visits [8]; physician availability [9]; incarceration and homelessness [3]; and Medicaid expansion [13] were all found to be associated with variation in mortality rates between geographic areas.

We were able to acquire county-level data for prescription rates, but incomplete or unavailable data for other factors prevents us from expanding on those papers’ methods and conclusions. Our project will improve on existing research regarding prescription rates and mortality by harmonizing that data with geographic, demographic, and economic data.

General Areas for Improvement

As described above, even papers that analyze opioid-related mortality along multiple dimensions do not effectively visualize how those variables may affect mortality in combination. This is partly due to the static nature of journal articles. Our interactive visualization will allow users to observe how mortality varies based on different combinations of variables. In addition, the analyses we cite are retrospective; our project will include predicted mortality rates.

Methodology

Our methodology can be divided into three components: the automated data pipeline; predictive modeling; and an interactive user interface.

Data Pipeline

The foundation of any analytical product is the underlying data that powers it. We acquired our core dataset by using the **Selenium** and **requests** Python libraries to automate web scraping. Specifically, our scraper performs the following actions:

- Drive a web browser to generate a session ID in the CDC's WONDER interface
- Set filter values to limit results to opioid-related deaths[1]
- Iterate through combinations of filter settings (variables) available in WONDER
- For each iteration, send a POST request with associated request form data to simulate the act of running a query using the web form provided by WONDER

Our iterative, filter-based approach was motivated by two key limits set by WONDER[2]:

- WONDER does not return results if the result set exceeds 75,000 records
- Death rates are suppressed (indicated as "Suppressed" in the results) for confidentiality reasons[3] if the number of deaths for a given row in the result set is less than 10

We worked around the result set limit by narrowing our queries to return one state's worth of data at a time. We worked around the suppressed data by running our queries in six-year chunks: 2000-2005, 2006-2011, and 2012-2017. Some counties' mortality rates were still suppressed after running queries using this method. For those counties, we imputed mortality rates by scaling the county's population to the range of 0-9 deaths. The total response dataset comprises ~10,000 records (~1 MB) across several hundred files.

We supplemented scraped data with datasets downloaded manually as flat files from government archives. These datasets include demographic and economic variables identified by related literature as being potentially significant. Because our scraper acquired data on the response variable in six-year chunks, we chose three representative years for each chunk: 2000, 2009, and 2017[4]. In total, these datasets comprise ~6.9 million records (~526 MB).

After collecting our datasets, we reshaped and harmonized them using FIPS codes as a common key. In addition, we transformed the response variables into mean rates, as

---

[1] We filter to the following Underlying Cause of Death codes: **X40-X44, X49, X60-X64, X83-X85, Y10-Y14, Y33-34**. In addition, we filter to deaths that include at least one of the following Multiple Cause of Death codes: **T40.0-T40.4, T40.6**.

[2] We observed that these limit only applies to the WONDER's publicly accessible web form; a future project researching the same issue could bypass the limit by sending a formal proposal to the CDC and paying a fee.

[3] https://wonder.cdc.gov/wonder/help/mcd.html#Assurance%20of%20Confidentiality

[4] The choice of "representative" years was partially based on data availability. For example, data related to the percentage of people in poverty was not available for 2001 through 2005. In order to remain consistent across datasets, we chose a year (2000) for which all ancillary datasets were available.

WONDER returns the **total** deaths for each selected time period. This provided us with a county-level view of opioid mortality as well as demographic and economic features over time.

For ease of access for users, we have packaged the data with our application. Although the size of the final data files are small (~2.5 MB), it contains a very large dataset comprising numerous distinct datasets from separate sources.

Below is a table listing each of our datasets and their respective sources.

Dataset	**Data Source**
CDC Wonder - Multiple Cause of Death data - county data	CDC WONDER (https://wonder.cdc.gov/mcd-icd10.html) Scraped using cdc_multi_cod_scraper.py
CDC Wonder - Multiple Cause of Death data - state data	CDC WONDER (https://wonder.cdc.gov/mcd-icd10.html) Downloaded manually
County Fips, Lat/Lon, and census Data (2018)	US Census Bureau https://www.census.gov/geographies/reference-files/time-series/geo/gazetteer-files.2018.html
State Populations by Year	US Census Bureau 2000: https://www2.census.gov/programs-surveys/popest/datasets/2000-2003/state/totals/ 2010-2018: https://www.census.gov/data/tables/time-series/demo/popest/2010s-state-total.html
County Education Data	US Department of Agriculture https://www.ers.usda.gov/data-products/county-level-data-sets/download-data/
County Poverty Rates	US Census Bureau, American FactFinder 2000: https://factfinder.census.gov/faces/tableservices/jsf/pages/productview.xhtml?pid=DEC_00_SF3_GCTP14.US05PR&prodType=table 2009 and 2017: https://factfinder.census.gov/faces/tableservices/jsf/pages/productview.xhtml?pid=ACS_09_5YR_GCT1701.US05PR&prodType=table
County Median Household Income Data	US Department of Agriculture https://www.ers.usda.gov/data-products/county-level-data-sets/download-data/
County Labor Data	Bureau of Labor Statistics https://www.bls.gov/lau/tables.htm
County Demographic Data	US Census Bureau 2000: https://www2.census.gov/programs-surveys/popest/datasets/2000-2003/counties/asrh/ 2010-2018: https://www.census.gov/newsroom/press-kits/2019/metro-county-pop-estimates.html

Predictive Modeling

In addition to gathering descriptive data, we trained a predictive model for our response variables (**mean opioid-related deaths per 100,000 people** and **age-adjusted mortality rate**):

- Read and pre-process the harmonized dataset using the **pandas** Python library
- Split data into training (80%) and test (20%) sets
- Fit models using **scikit-learn:**
    - Linear regression (using all variables)
    - Linear regression (using a subset of variables identified as significant in our literature survey)
    - Random forest (using all variables)
- Compare test RMSE of the three models

We found that the random forest performed better than the other models, with a test RMSE of 4.14 for mean death rate, compared to 5.74 and 5.48 for the regression models.

Next, we projected future values for our input variables using exponential smoothing (available in the **statsmodels** Python library). Using these projected values as inputs to our random forest, we generated predictions for our response variables representing what mortality rates might look like over the next six-year time period.

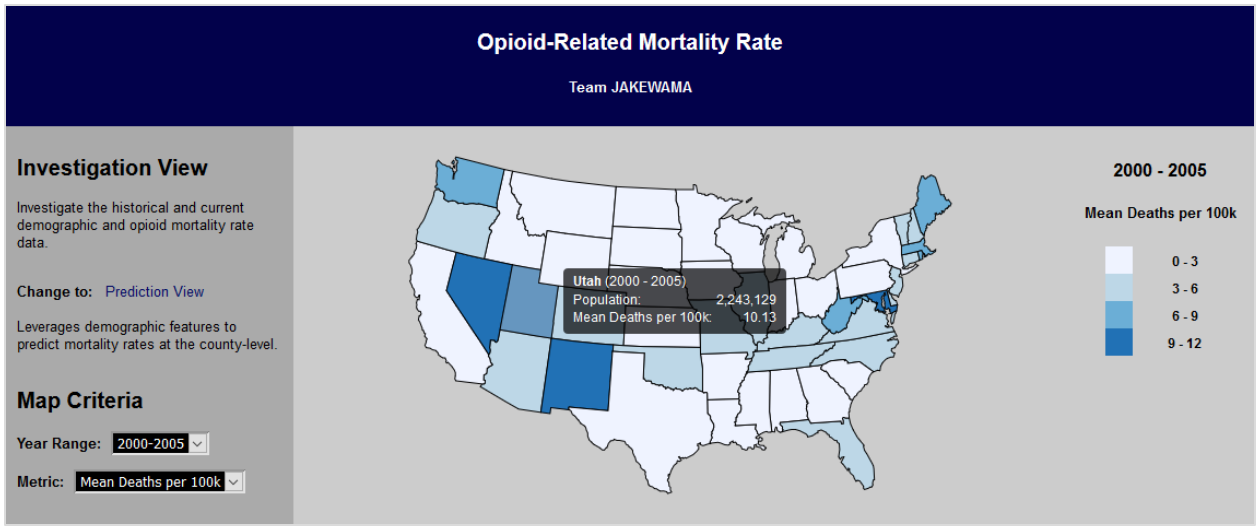
Interactive User Interface

We developed the user interface using **D3.js**. D3 provides functionality for mapping the geographies in our data (states and counties); altering visual characteristics based on data; and allowing users to interact with our visualizations. Our interface provides two primary modes: investigation and prediction.

Investigation

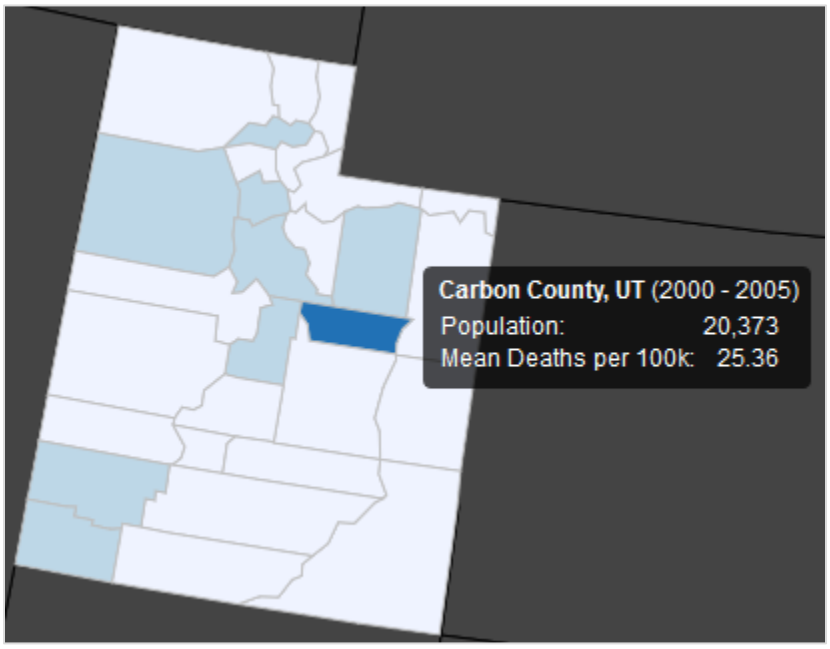
Starting from a nation-wide view of opioid-related mortality, users can make the following selections:

- Year range
- Mortality metric

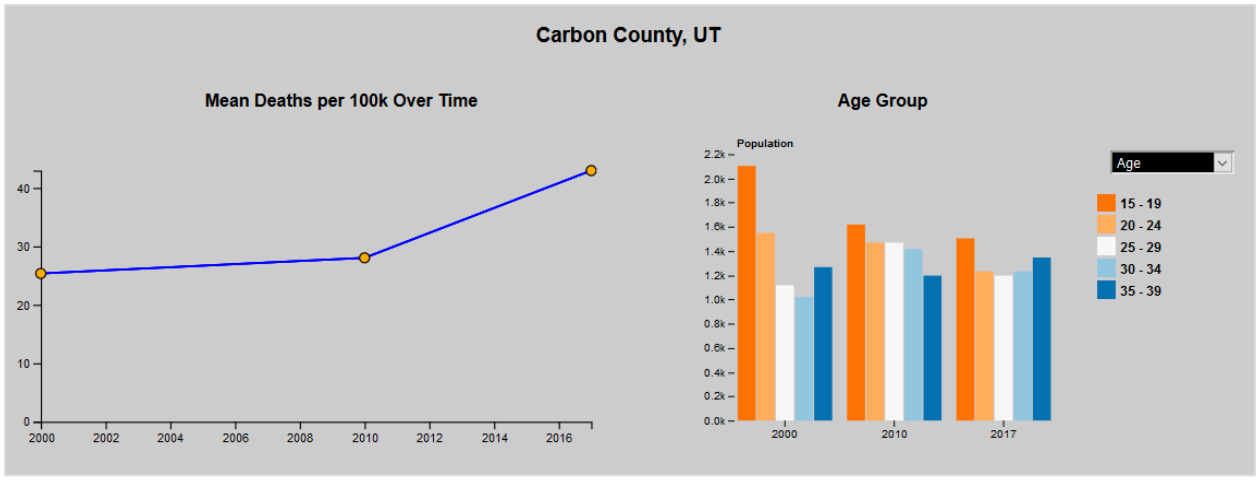


The user's selection dynamically updates the choropleth's colors and scale according to the selected data, allowing them to see differences between states as well as differences within a state over time.

The user can also zoom in on a specific state, showing them opioid-related mortality for each county in that state. The same selection controls are available in this zoomed-in view. From a zoomed-in state, the user can pan to neighboring states, remaining in the zoomed-in view.

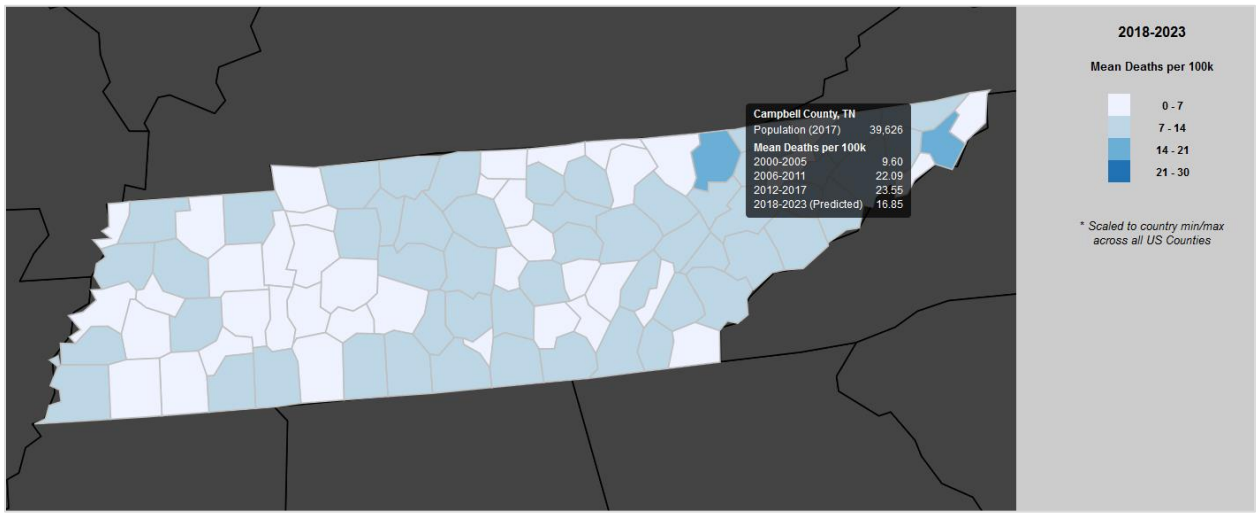


While in the zoomed-in view, the user can click a specific county to view that county's mortality over time as well as a breakdown of its demographic and economic features. This functionality gives them an intuitive way to provide context to observations on a specific county's mortality rates.

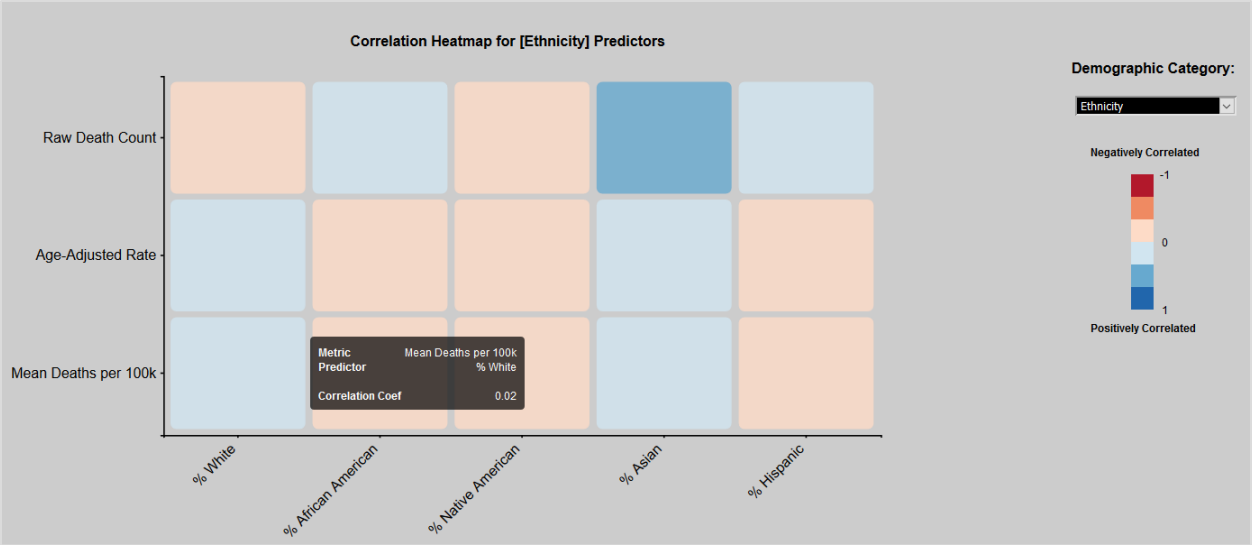


Prediction

In prediction mode, the choropleth provides much of the same functionality. However, there is no year range selector, as our predictions only cover a single time period (the next six-year chunk from 2018-2023).



In the zoomed-in view, we also replace the visualization of a county's demographic composition with a heatmap showing the correlation between each input variable with the selected response variable.



This functionality provides additional context to our predicted mortality metrics and could generate future research questions.

Innovations

Our project improves on existing work in four primary ways:

1. **Automation of data collection**
   Our primary dataset is normally available through a difficult-to-use web form or an API that provides difficult-to-use data. Using a Selenium web driver and Python requests, we have been able to eliminate significant manual effort in data collection. In addition, our pipeline enhances the response dataset with additional features.

2. **Harmonization of datasets at the county-level**
   Many datasets already exist, but do not capture details below the state-level. In addition, these datasets are not collected or cleansed in a way that makes it easy for a user to use them together.

3. **Prediction of the response variable**
   Existing data only describes observed death rates; our project allows users to get a picture of how opioid-related death rates might change in the future based on various county-level features.

4. **Use of interactive visualization**
   Using the D3 library improves on the mostly static visualizations used in scholarly papers and official reports.

Experiments and Evaluation

We asked a sample of **22** users to attempt to answer questions about the opioid epidemic from 2012-2017 – first, using WONDER system, then using our application:

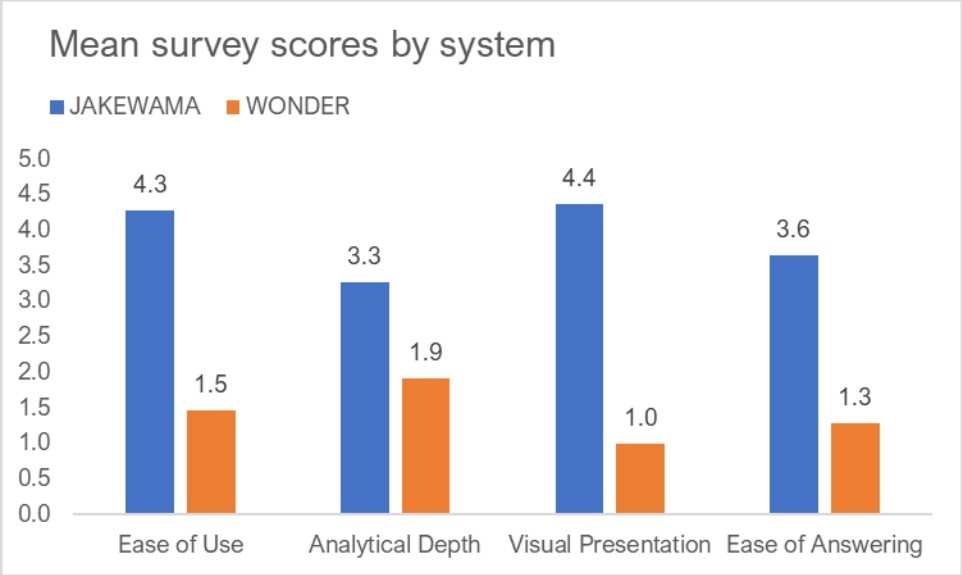
1. Which state had the highest mortality rate?
2. In that state, which counties had the highest mortality rates?
3. What makes those counties different from other counties?

Then, we asked users to rate both WONDER and our application on the following attributes:

1. Ease of use
2. Analytical depth
3. Visual presentation
4. How easy it was to answer the questions above

Users rated both systems on a scale of 1 (worst) to 5 (best).

On average, respondents rated our application higher than WONDER, with a mean score of 3.9 vs. WONDER's 1.4. The following chart displays the comparative results by attribute.



Based on our survey results, we conclude that our application outperforms existing applications and content – especially for communication of data-driven insights.

Conclusions and Discussion

We drew four primary conclusions from our analysis and evaluation:

1. **Confidentiality requirements and lack of data complicate analysis.** Although we were able to acquire response data for the majority of counties by

aggregating time periods or imputing values, we dropped approximately 1,300 county-year records due to missing feature data. As suggested in a footnote in this paper, a future project could perform a more complete analysis by submitting a formal proposal to the CDC with a fee.

2. **There is not a strong linear relationship between mortality rates and the various demographic variables we collected.** This is also supported by the literature; for example, Ghertner and Groves observed that mortality measures "had a high degree of spatial clustering" [4]. This suggests that a simple linear model applied to the entire dataset will not be very predictive. This observation holds even after we collected more demographic and economic variables as inputs; there was no combination of variables that displayed a strong linear relationship to mortality.

3. **However, there is a correlation between mortality rates and some of those demographic variables.** This was made clear when we drew a correlation heatmap during exploratory data analysis, which motivated us to include a heatmap visualization in the final application. We believe that this correlational analysis could drive future research on the relationship between opioid-related mortality and demographic variables.

4. **Providing an intuitive, accessible visualization is intrinsically valuable.** Based on the results of our experiments and evaluation, users found a visual, interactive tool to be more useful than a web form. Users were able to answer questions that they could not answer using existing resources, and they perceived it to be a better experience.

Plan of Activities, Timeline, and Distribution of Effort

The total duration of sub-tasks under task groups (highlighted in blue) may be greater than the task group's duration due to overlapping tasks.

Actual distribution of effort has been roughly equal between all team members.

Name	Duration	Start	Finish	Assignee	Complete?
Proposal Document	22 days	9/19/2019	10/11/2019		
Create plan of activities	3 days	9/19/2019	9/22/2019	Arthur	Y
Write a clear problem definition	0.5 days	9/26/2019	9/26/2019	All	Y
Write answers to Heilmeier questions	2 days	9/26/2019	9/28/2019	All	Y
Collect and summarize 12 papers	5 days	9/26/2019	10/1/2019	All	Y
Write proposal document	6 days	10/1/2019	10/7/2019	All	Y
Review proposal document	1 day	10/7/2019	10/8/2019	Evan	Y
Finalize proposal document	2 days	10/8/2019	10/10/2019	Evan	Y
Proposal Presentation	19 days	9/22/2019	10/11/2019		

Create slides	11 days	9/22/2019	10/3/2019		
Create slide for expected innovation	0.5 days	9/26/2019	9/27/2019	Jay	Y
Create slides for Heilmeier questions	1 day	9/28/2019	9/29/2019	Adam	Y
Create slides for literature survey	2 days	10/1/2019	10/3/2019	Arthur	Y
Create slides for plan of activities	0.5 days	9/22/2019	9/23/2019	Adam	Y
Compile slides into a single deck and edit	2 days	10/3/2019	10/5/2019	All	Y
Review proposal slides	1 day	10/5/2019	10/6/2019	Adam	Y
Finalize proposal slides	2 days	10/6/2019	10/8/2019	Jay	Y
Record presentation video	2 days	10/8/2019	10/10/2019	Evan	Y
Develop Project	60 days	9/19/2019	11/18/2019		
Collect data	8 days	10/11/2019	10/19/2019		
Collect initial dataset for exploration	2 days	10/11/2019	10/13/2019	Jay	Y
Develop data collection/ pipeline code	8 days	10/11/2019	10/19/2019	Jay	Y
Exploratory data analysis	5 days	10/13/2019	10/18/2019		
Identify features of interest	3 days	10/13/2019	10/16/2019	Evan	Y
Identify features that must be created	3 days	10/13/2019	10/16/2019	Evan	Y
Create prototype analysis/computation	5 days	10/13/2019	10/18/2019	Adam	Y
Create data dictionary/ readme	2 days	10/13/2019	10/15/2019	Jay	Y
Develop full analysis/computation	10 days	10/19/2019	10/29/2019	Adam	Y
Production code for analysis of full dataset	10 days	10/19/2019	10/29/2019	Adam, Evan	Y
Develop interactive user interface	28 days	10/11/2019	11/8/2019		
Design interface	2 days	10/11/2019	10/13/2019	Evan	Y
Review interface with team	1 day	10/13/2019	10/14/2019	All	Y
Develop interface	12 days	10/14/2019	10/26/2019	Evan, Arthur	Y
Develop data integration pipeline + interface	10 days	10/29/2019	11/8/2019	Jay, Arthur	Y
Iterations: test, fixes, enhancements	10 days	11/8/2019	11/18/2019	All	Y
Merge code to master	0 days	11/18/2019	11/18/2019	All	Y
Deploy project	0 days	11/18/2019	11/18/2019	All	Y
Progress Report	7 days	11/1/2019	11/8/2019		
Create document outline	0.5 days	11/1/2019	11/1/2019	Arthur	Y
Fill in progress to date	2 days	11/1/2019	11/3/2019		
Write introduction/ motivation section	0.5 days	11/1/2019	11/2/2019	Jay	Y
Write problem definition section	0.5 days	11/1/2019	11/2/2019	Evan	Y
Write survey section	1 day	11/1/2019	11/2/2019	Arthur	Y
Write proposed method section	2 days	11/1/2019	11/3/2019	Adam	Y
Write experiments/evaluation section	1 day	11/1/2019	11/2/2019	Jay	Y
Write conclusions and discussion section	0.5 days	11/1/2019	11/2/2019	Evan	Y
Write list of innovations	1 day	11/1/2019	11/2/2019	Adam	Y
Write plan of activities section + effort distribution	1 day	11/1/2019	11/2/2019	Arthur	Y
Compile progress report sections and edit	2 days	11/3/2019	11/5/2019	Arthur	Y
Review progress report	1 day	11/5/2019	11/6/2019	All	Y

Finalize progress report	2 days	11/6/2019	11/8/2019	Arthur	Y
Final Report	10 days	11/18/2019	11/29/2019		
Write first draft of report	3 days	11/18/2019	11/21/2019		
Update introduction/ motivation section	0.5 days	11/18/2019	11/18/2019	Jay	Y
Update problem definition section	0.5 days	11/18/2019	11/18/2019	Evan	Y
Update survey section	0.5 days	11/18/2019	11/18/2019	Arthur	Y
Update proposed method section	1 day	11/18/2019	11/19/2019	Adam	Y
Update experiments/evaluation section	2 days	11/18/2019	11/20/2019	Jay	Y
Update conclusions and discussion section	3 days	11/18/2019	11/21/2019	Evan	Y
Update plan of activities + effort distribution	1 day	11/18/2019	11/19/2019	Arthur	Y
Compile final report sections and edit	2 days	11/21/2019	11/23/2019	Arthur	Y
Review final report	2 days	11/23/2019	11/25/2019	All	Y
Finalize final report	2 days	11/25/2019	11/27/2019	Arthur	Y
Poster Presentation Video	10 days	11/18/2019	11/29/2019		
Create poster	2.5 days	11/18/2019	11/20/2019		
Create poster layout	0.5 days	11/18/2019	11/18/2019	Evan	Y
Create motivation/ introduction section	2 days	11/18/2019	11/20/2019	Adam	Y
Create approaches section	2 days	11/18/2019	11/20/2019	Jay	Y
Create data section	2 days	11/18/2019	11/20/2019	Arthur	Y
Create experiments and results section	2 days	11/18/2019	11/20/2019	Adam	Y
Review poster design	1 day	11/20/2019	11/21/2019	All	Y
Finalize poster design	3 days	11/21/2019	11/24/2019	Evan	Y

References

1. Spiller, H., Lorenz, D. J., Bailey, E. J., & Dart, R. C. (2009). Epidemiological trends in abuse and misuse of prescription opioids. *Journal of addictive diseases*, *28*(2), 130-136.

2. Schoenfeld, E. R., Leibowitz, G. S., Wang, Y., Chen, X., Hou, W., Rashidian, S., ... & Wang, F. (2019). Geographic, Temporal, and Sociodemographic Differences in Opioid Poisoning. *American journal of preventive medicine*.

3. Galea, S., & Vlahov, D. (2002). Social determinants and the health of drug users: socioeconomic status, homelessness, and incarceration. *Public health reports*, *117*(Suppl 1), S135.

4. Ghertner, R., & Groves, L. (2018). The opioid crisis and economic opportunity: geographic and economic trends. *ASPE Research Brief*, 1-22.

5. Rigg, K. K., Monnat, S. M., & Chavez, M. N. (2018). Opioid-related mortality in rural America: geographic heterogeneity and intervention strategies. *International Journal of Drug Policy*, *57*, 119-129.

6. Cicero, T. J., Surratt, H., Inciardi, J. A., & Munoz, A. (2007). Relationship between therapeutic use and abuse of opioid analgesics in rural, suburban, and urban locations in the United States. *Pharmacoepidemiology and drug safety*, *16*(8), 827-840.

7. Monnat, S. M. (2019). The contributions of socioeconomic and opioid supply factors to US drug mortality rates: Urban-rural and within-rural differences. *Journal of Rural Studies*, *68*, 319-335.

8. Hasegawa, K., Brown, D. F., Tsugawa, Y., & Camargo Jr, C. A. (2014, April). Epidemiology of emergency department visits for opioid overdose: a population-based study. In *Mayo Clinic Proceedings* (Vol. 89, No. 4, pp. 462-471). Elsevier.

9. Rosenblatt, R. A., Andrilla, C. H. A., Catlin, M., & Larson, E. H. (2015). Geographic and specialty distribution of US physicians trained to treat opioid use disorder. *The Annals of Family Medicine*, *13*(1), 23-26.

10. Seth, P., Scholl, L., Rudd, R. A., & Bacon, S. (2018). Overdose deaths involving opioids, cocaine, and psychostimulants—United States, 2015–2016. *Morbidity and Mortality Weekly Report*, *67*(12), 349.

11. Mosher, H., Zhou, Y., Thurman, A. L., Sarrazin, M. V., & Ohl, M. E. (2017). Trends in Hospitalization for Opioid Overdose among Rural Compared to Urban Residents of the United States, 2007-2014. *Journal of hospital medicine*, *12*(11), 925-929.

12. Wickramatilake, S., Zur, J., Mulvaney-Day, N., Klimo, M. C. V., Selmi, E., & Harwood, H. (2017). How states are tackling the opioid crisis. *Public Health Reports*, *132*(2), 171-179.

13. Cher, B. A., Morden, N. E., & Meara, E. (2019). Medicaid Expansion and Prescription Trends: Opioids, Addiction Therapies, and Other Drugs. *Medical care*, *57*(3), 208.

14. Cornaggia, K. R., Hund, J., Nguyen, G., & Ye, Z. (2019). Opioid Crisis Effects On Municipal Finance. *Available at SSRN 3448082*.

15. Cerdá, M., Ransome, Y., Keyes, K. M., Koenen, K. C., Tracy, M., Tardiff, K. J., ... & Galea, S. (2013). Prescription opioid mortality trends in New York City, 1990–2006: examining the emergence of an epidemic. *Drug and alcohol dependence*, *132*(1-2), 53-62.