Characterizing Allegheny County Opioid Overdoses with an Interactive Data Explorer and Synthetic Prediction Tool

Theresa Gebert
Department of Statistics and Data Science
Carnegie Mellon University
theresa@stat.cmu.edu

Shuli Jiang
Department of Computer Science
Carnegie Mellon University
shulij@andrew.cmu.edu

Jiaxian Sheng
Department of Computer Science
Carnegie Mellon University
jiaxians@andrew.cmu.edu

Abstract

There is an opioid epidemic in the United States. Pennsylvania's Allegheny County is among the worst, which is what motivates a deeper exploration of what characterizes this phenomenon, such as what risk factor are for people who ultimately overdose and die due to opioids. We show that some interesting trends and factors can be identified from openly available autopsy data, and demonstrate the power of building an interactive data exploration tool for policy makers. However, there is still a pressing need to incorporate further demographic factors. We show this by using synthetic Electronic Medical Record (EMR) data to simulate the predictive power of additional loosely correlated features. In addition, we give examples of useful feature extraction that enable model enhancement without sacrificing privacy.

1 Introduction

There is an opioid epidemic in the United States (REFERENCE). In order to best serve their communities, policy-makers need to know a variety of things about this epidemic, such as:

1. What exactly is happening right now: who is affected and how much?
2. Why and how did this happen: what factors influenced increased opioid usage?
3. What can be done to slow and stop this epidemic?
4. How can we prevent this from happening again in the future?

Data and statistical analysis should be able to help answer some of these questions, and machine learning may be able to create tools to aid prevention efforts. In this paper, we showcase the work we did during the 2-day 2018 HackAuton hackathon. Policy-makers frequently do not have the statistics or computer science background required to easily manipulate and explore data, or easily create and interpret models. That is why our first goal was to create an interactive data explorer, that provides a simple, intuitive user interface for the key stakeholders, and a lightweight, cheap backend in `RShiny` to enable easy codebase maintenance. Our second goal was to create a synthetic prediction tool: something that would show policy-makers the power of additional data and allow them to easily communicate this to others. In the last section of the paper, we also describe our desired future work and policy recommendations.

CMU HackAuton 2018, Pittsburgh, PA.

2 Background

Opioids are powerful painkillers, which have traditionally made them an attractive drug to treat patients with severe pain or to manage pain after surgery. Well-known opioids include heroin and morphine. When opioid drugs bind to opioid receptors in the brain, they can drive up dopamine levels and produce a state of euphoria and relaxation (REFERENCE). While the effects include euphoria, they also include drowsiness, nausea, confusion, constipation, sedation, tolerance, addiction, respiratory depression and arrest, unconsciousness, coma, and death. Opioids are highly addictive substances, which is why doctors recommend they only be used in the extreme cases. However, the number of prescriptions of opioids to take at home has also risen over the last decade, which makes it plausible that increased rates of prescription opioids is also affecting the rates of illegal opioid usage (REFERENCE). Opioids costs the United States billions of dollars every year in direct healthcare costs and lost productivity. Some of the states that are most affected are Massachusetts, Connecticut, and Pennsylvania.

The Allegheny County Medical Examiner's Office has made autopsy data from deceased overdose cases publicly available from the years 2008 thru 2017. The dataset contains the date and time of death and up to seven drugs identified as cause-of-death related to the overdose. Some additional demographic variables are included, such as age, sex[1], and race of the individual, as well as zip code of residence and zip code of the case (the hospital or medical examiner's location). It only includes cases considered closed and some features are missing or incorrect (e.g. 4-digit zip code). There are a total of 3483 cases.

3 Interactive Data Explorer

For all but the most simple datasets, the vast majority of time in model construction and analysis is spend on data exploration and visualization. Simply subsetting the data and extracting features from raw input can require formal computer science training. Not all offices have an in-house statistician or programmer; even if they do, they are frequently swamped with work; and hiring an external consultant is time-consuming and expensive. Giving policy-makers a way to easily interact with data and ask certain questions and get the answers themselves is crucial. As a result, we had three major goals in constructing our interactive data exploration tool:

1. interactive;
2. easy and intuitive to use;
3. lightweight and cheap to maintain.

3.1 Methods

Given our own time constraints, we had a great test case for satisfying our third engineering goal: lightweight and cheap to maintain. Our hypothesis was that if we can build this tool in a day, it will be easy for a high school intern at the Allegheny County Medical Examiner's Office to maintain, or an in-house IT employee swamped with work. We also used ourselves as guinea pigs in learning the tool, since none of us have experience with front-end development and design. Our tool of choice ended up being `RShiny`.

In less than one hour, we had `R` and the requisite packages installed and a working version of the `RShiny` tutorial. It only took us another hour to build our first interactive data visualization using the Allegheny County Overdose dataset. This highlights the usability of the product and its promise as a feasible, maintainable tool within a government organization (1).

3.2 Results

Our data exploration tool provides a few static graphs to highlight some known features of interest, such as top drugs used and geographic distribution. Our interactive graphs enable exploring more

[1] Sex data was recorded on a binary scale, which does not include individuals with uncertain or non-binary sex. While the paper may reference "both sexes," the authors intend this as a statement about the sex measurements included in the dataset, and not all possible sexes in the human population. In addition, the authors recognize that sex does not capture gender identity.

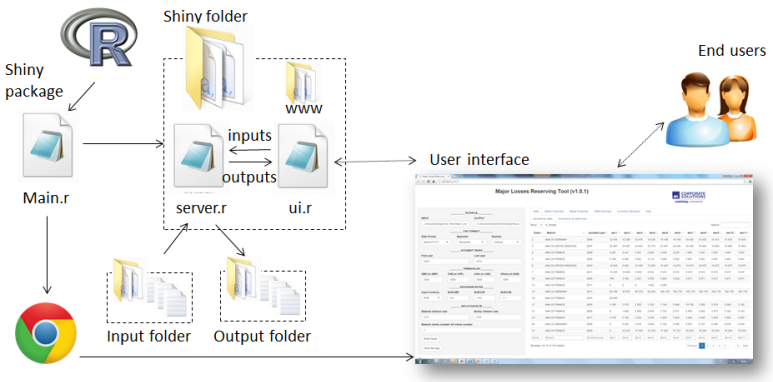

Figure 1: `RShiny` is a lightweight tool for building beautiful interactive web applications to visualize data. `R` calls the Shiny app in a main script with the Shiny package. Image courtesy of Little Actuary.

complex relationships, like change over time, and co-occurrence between drugs. From our exploratory analysis we discovered some simple characteristics:

- **opioid-related overdoses have been significantly increasing over time**; this is as expected, since this is what motivates this entire project (3).
- **overdoses are frequently related to many drugs**; the mean number of drugs involved in an overdose across all cases, all years, for both sexes, was 2.5. A density plot of the number of drugs involved shows that 2 drugs are most frequently involved (2).
- **a few drugs are involved in most of the cases**; this is what we might expect, but the data confirms that the top 8 drugs account for over 75% of cases (FIGURE).
- **cases are geographically concentrated in more populous area**; again, this is what we might expect, but what is more interesting is actually that, relative to population, the frequency of overdoses in rural and less populated areas is higher than in urban areas.

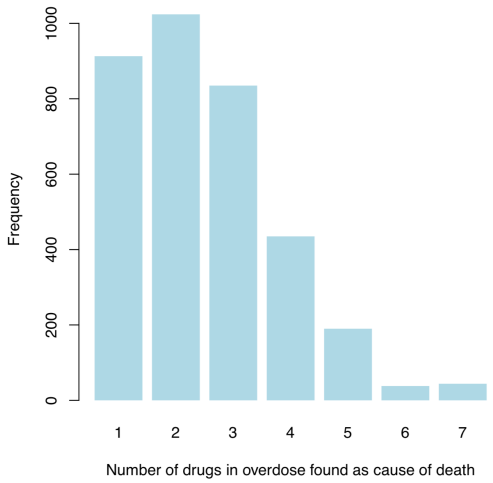

Figure 2: The frequencies of the number of drugs involved in an overdose aggregated across all cases and all years. Having at least 2 drugs involved in an overdose is extremely common.

From our exploratory analysis we also discovered some more nuanced phenomena:

- **Fentanyl is a major player in opioid use increases**; Fentanyl is a powerful synthetic opioid analgesic that is similar to morphine but is 50 to 100 times more potent (REFERENCE). Its

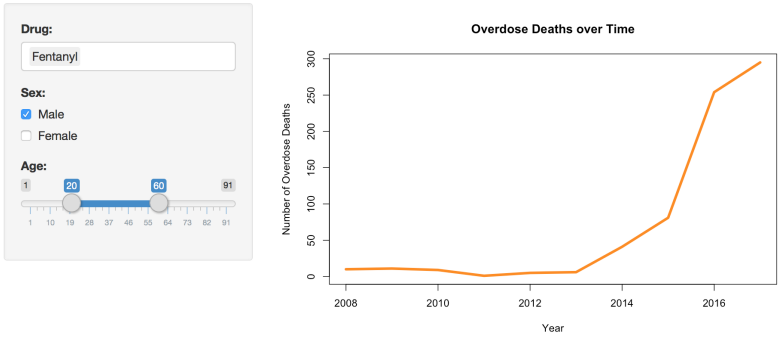

Figure 3: The first interactive graph we constructed for the tool. It enables the visualization of overdose deaths over time constrained to different variables: by drug, sex, and age. Drug and sex allow multiple selections. Age allows selection of an age range with variable length.

usage saw a major spike in 2015 across all races, age groups, and both sexes. Contrast this with acetaminophen, which has been in popular use for decades and usage associated with overdose is both overall low, and has not increased.

- **hard drugs tend not to be taken together**; while we may expect this behavior given anecdotal evidence and economic viability (illicit drugs are expensive), the data shows a clear trend here: the drugs that appear most frequently in the overdoses are also hard drugs (e.g. heroin, cocaine), and they are less likely to be taken together (4).
- **opioids, stimulants, and depressants tend not to be taken together**; again, while we may expect this behavior given anecdotal evidence, the data shows a clear trend here: classifying drugs within their major categories (opioids, stimulants, depressants) reveals that these drugs are more likely to be mixed within-category than between-category (DATA).

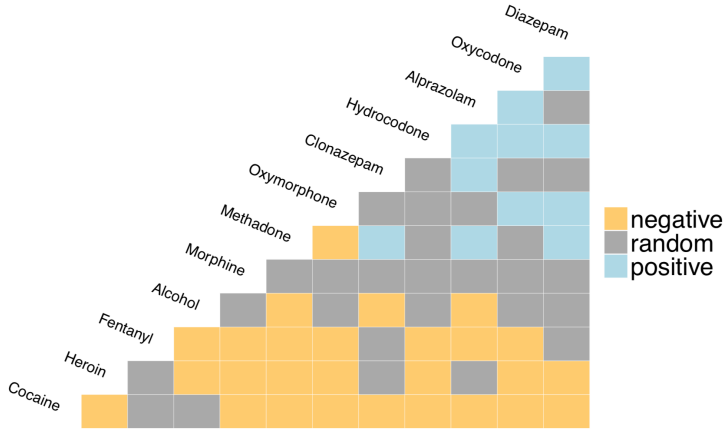

Figure 4: The cooccurrence of pairwise drugs for the top drugs used. Blue indicates those two drugs are frequently found together in an overdose, yellow indicates those two drugs tend not to be found together in an overdose.

The way that these results can inform policy decisions is discussed in the last section of this paper.

4 Synthetic Prediction Tool

The security and confidentiality of medical data is of high importance. This is why everything from conversational discretion by healthcare providers and the security of medical records is highly

regulated and strictly enforced. While patient privacy is becoming increasingly important, especially with regards to disclosure to insurance companies and employers, it can also slow and hamper progress in other ways. For example, the only patient information we have about the Allegheny County overdose cases is age, sex, and race. While these factors can be important factors in overdose risk, other factors may be far better predictors and improve our ability to identify individuals at risk of drug use, and at risk of dying from overdose. However, in order for such prediction models to be useful to social workers, policy-makers, and the patients themselves, they need to be easy to use and dynamically adapt to different patient characteristics. Finally, a major pitfall of traditional risk analysis tools is that they only provide point estimates, without giving any measure of uncertainty in their estimate. Therefore, we had three major goals in constructing our synthetic prediction tool:

1. dynamic;
2. easy and intuitive to use;
3. incorporates a measure of uncertainty.

4.1 Methods

In order to build our prediction tool, we needed to (1) generate synthetic data for a non-overdose population, (2) generate additional medical predictors, and (3) build a predictive model. Let $\mathcal{X}_o \in \mathbb{R}^{n_o,d}$ denote the n_o observations of d-dimensional features of the population that died from overdose. Let $\mathcal{X}_c \in \mathbb{R}^{n_c,d}$ denote the n_c observations of d-dimensional features of the population that does not die from overdose. Notice that our control population could be one of two types, both interesting: people in the general population, both users of opioids and non-users of opioids, or people in the opioid-using population. For simplicity, we generated synthetic data for the general population, but it would be very useful to know the risk factors for an individual to overdose given that they are already using drugs.

Our dataset is $\mathcal{X}_o$, so we do not need to generate that. Our first challenge is generating $\mathcal{X}_c \in \mathbb{R}^{n_c,d}$. We took advantage of a free synthetic Electronic Medical Record (EMR) database online in order to obtain the demographic information of the general population, under the assumption that the simulated data was representative of the general population (REFERENCE). Unfortunately, we already have evidence that it is not representative, since it did not include African-American males, for example, and a disproportionate number of Asian patients relative to the population-level statistics in the United States.

Our second challenge was to extract additional features, to show the prediction improvement in our model in the presence of additional covariates. In more precise terms, we transformed $\mathcal{X}_o \in \mathbb{R}^{n_o,d} \rightarrow \mathcal{X}_o \in \mathbb{R}^{n_o,d'}$, where $d' > d$. Here are some of the features we extracted from the medical records:

- **marital status**; whether an individual is single, married, divorced, or widowed.
- **socioeconomic status**; percentage of income below poverty line.
- **language**; primary language spoken at home.
- **sickliness**; time-discounted days spent in hospital.
- **disease history**; the number of occurrences in each category of disease as defined by ICD-10 (REFERENCE).

An additional challenge in generating these features was also *matching* them to the appropriate overdose cases. Given our time constraints, we used a very simple matching algorithm and have ideas for improvement described in the last section of the paper. The current implementation of our matching procedure works as follows: for each individual that overdosed, in a random order, randomly select a uniformly random individual from the synthetic patients that matches

1. exactly on gender; this seemed important to match on exactly, but resulted in dropping individuals with missing gender.
2. if race is white, black, or Asian, then match exactly, otherwise match randomly; specifically, for individuals marked as `Other` or `Unidentified` we randomly selected a race from the synthetic population.

3. within 3 years on age; this was more difficult to match on exactly, and ± 3 years seemed a reasonable first step.
4. and then remove that selected individual from the possible candidates; this is done to prevent duplicate records.

Finally, the last step was building our models. We built a simple model to predict the number of drugs used in the overdose using only the original data. Then, we build predictive models using the n_0 overdose cases with the d original feature set and with the $d' > d$ enhanced feature set to showcase the improvement of additional information.

4.2 Results

Our first model was a Poisson generalized linear model with the canonical link function to associate the outcome, number of drugs involved in overdose, with three features of interest: age, sex, and race. We did not include interaction terms due to the small size of our dataset relative to the number of features (sex and race are categorical variables and generate several indicator variables in the model). We did not run additional models in order to avoid multiple comparisons problem. Our model summary is shown in 4.2. We performed a Goodness of Fit test and reject the null hypothesis that the model does not explain the data well. The diagnostic residual and QQ plots also looked reasonable (5). We also performed a Deviance Test to assess overdispersion and concluded that the Poisson model fit the data well and there was no overdispersion.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	0.6871	0.0548	12.54	**0.0000**
as.factor(Sex)Male	-0.0353	0.0230	-1.54	0.1243
as.factor(Race)White	0.1390	0.0329	4.23	**0.0000**
as.factor(Race)Hispanic	-0.0946	0.2519	-0.38	0.7072
as.factor(Race)Other	-0.0823	0.3550	-0.23	0.8167
as.factor(Race)Asian	-0.2169	0.2907	-0.75	0.4555
as.factor(Race)Middle Eastern	0.5753	0.2902	1.98	**0.0475**
as.factor(Race)Unidentified	-0.8342	1.0005	-0.83	0.4044
as.factor(Race)Indian	-0.0258	0.7082	-0.04	0.9709
Age	0.0032	0.0009	3.66	**0.0003**

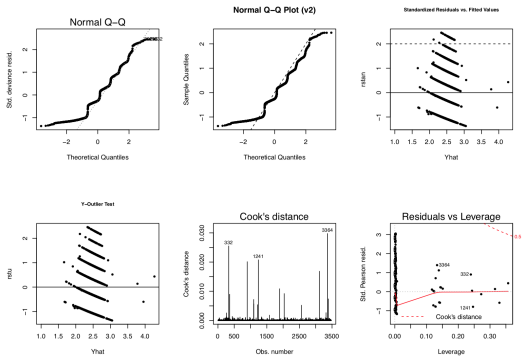

Figure 5: Diagnostic plots for the Poisson generalized linear model. The assumption of Normality of errors is not perfect, but reasonable. No apparent outliers are detected. Despite the odd appearance of the residual plots and the slight positive skew, this is actually a reasonable fit.

As we can see from the coefficients, age is slightly but significantly positively associated with number of drugs used, even when accounting for the other variables. In addition, being white is slightly but significantly positively associated the number of drugs used in the overdose.

Finally, we used the synthetic data to develop a predictive model of risk of overdose with and without our new features. As expected, the new features improved the model accuracy by 20%. In particular, we see that given our original, minimal covariates, model accuracy is little better than random chance

(baseline is 50% by construction). Meanwhile, with the features from the EMR data, we see a mean accuracy of the model for binary classification increases to 59% for neural network and 99% for random forest. We are still investigating the discrepancies between these models. These errors were calculated using 10-fold cross-validation. We also tried multi-class classification, which we predicted overdose vs. not overdose, as well as the drug that would be overdosed on, but our prediction errors were still quite poor and our models require further tuning.

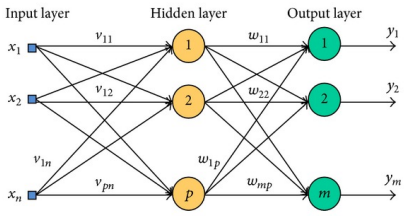

Figure 6: Example of simple neural network model, which can be used to predict outcomes, such as overdose vs. not overdose, given a set of features. Image courtesy of [2].

In particular, we applied random forest and neural network prediction algorithms to the dataset. We chose the neural network due to its known track-record of highly accurate predictions. We chose the random forest since it is also a efficient, effective algorithm, with the added benefit that it has more easily interpretable feature importance. For example, one can build the average decision tree from the random forest and show this to policy-makers. In our interactive prediction tool, we show how the relevant features can be used to calculate a score of risk overdose with a 95% confidence interval (7).

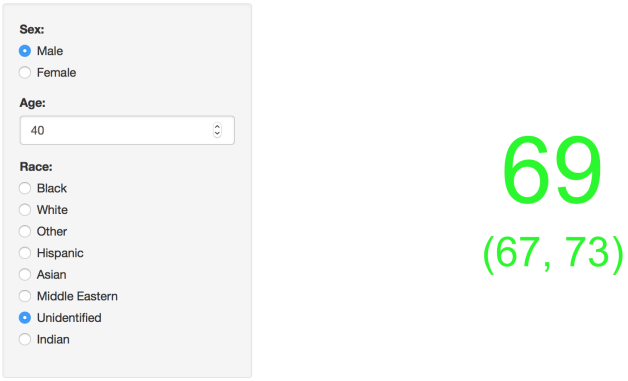

Figure 7: Example of an interactive prediction tool. Given the set of covariates of a patient, it recalculates the risk score of overdose with a 95% confidence interval.

5 Conclusion

In this section, we use the results from our interactive data explorer and synthetic prediction tool to recommend policy changes. We also describe the major flaws in our current approach and future work to improve our models and analysis.

5.1 Policy Recommendations

Given the trends we discovered using our interactive data explorer, the characteristics of overdoses themselves and the individuals who did so, as well as the substantial improvement in our model after adding just five features minimally correlated with the outcome of interest, we have a few policy recommendations to slow and, ultimately, stop this opioid epidemic.

1. **Collect more data and learn more about this problem.** First are foremost, more data is necessary in order to make more effective policy recommendations and assess individual risk

factors. While we have preliminary evidence to suggest age and certain races are positively associated with the number of drugs used, our synthetic data analysis shows how powerful additional medical information could be in improving these methods.

2. **Crack down on Fentanyl.** Again, more data is required to understand the more nuanced relationship here, but it appears increase in Fentanyl usage is a major driver of increased numbers of overdose deaths.
3. **Combinations are killers.** Most overdose deaths are related to combinations of drugs being used, usually 2 or 3, but up to 6 or 7. Further analysis is required to understand which combinations specifically may be more dangerous, and which health conditions may make certain combinations more risky.

5.2 Future Work

One major area of improvement is our matching algorithm in order to provide a more representative synthetic dataset and associate covariates with the overdose cases. In particular, we would like:

- matching on more exhaustive criteria, such as geographical location (zip code), specific diagnoses within the past hospital history, or the number of hospital visits;
- more flexible ways of drawing samples that allows criteria to match closely;
- statistically sound way of dealing with missing data or skewed samples in our synthetic distribution.

One major area of further investigation are the following questions, which we did not have time to address in this analysis.

- Does the significant effect of white and Middle Eastern race on the number of drugs used disappear once socioeconomic status is taken into account?
- Does the significant effect of age on the number of drugs used disappear once health is taken into account?
- What are the bigger killers: hard drugs combined with alcohol, or a dangerous variety of less potent drugs?
- What is the elbow point of drug usage: is there a particular drug or drug combination that leads to a significant spike in drug usage and overdose risk?
- How correlated do additional features need to be to the outcome of interest in order to yield model improvements?
- What masking techniques are most effective at providing high prediction accuracy improvements at low cost to privacy?

The opioid epidemic in Allegheny County, and in the United States more broadly, is an important problem. Better data exploration tools can increase policy maker understanding, and synthetic predictive tools can showcase the usefulness of access to additional data in this space.

Acknowledgements

We would like to thank the organizers of the 2018 HackAuton hackathon: Nick Gisolfi, Chirag Nagpal, and the rest of the volunteers at Auton Lab. We would also like to thank the generous sponsors who made this possible: The Carnegie Mellon University School of Computer Science, The Robotics Institute, The Machine Learning Department, and Heinz College; as well as our corporate sponsors: UPMC Enterprise, Philips, and RxThinking.

References

[1] Anna L. Buczak, Steven Babin, and Linda Moniz. Data-driven approach for creating synthetic electronic medical records. *BMC Medical Informatics and Decision Making*, 10(1):59, Oct 2010.

[2] Ching Lee Koo, Mei Jing Liew, Mohd Mohamad, and Abdul Hakim Mohamed Salleh. A review for detecting gene-gene interactions using machine learning methods in genetic epidemiology. 2013:432375, 10 2013.