Identification of At-Risk Groups for Opioid Addiction Through Web Data Analysis

A work in progress report submitted to epiDAMIK workshop to be held in conjuction with ACM SIGKDD 2018

K. Basu, S. Choudhuri, A. Sen
NetXT Lab, SCIDSE
Arizona State University
{kaustav.basu, schoud13, asen}@asu.edu

A. Majumdar
Galvanize, Inc.
aniket.majumdar@galvanize.com

D. Dey
Dept. of Statistics
University of Connecticut
dipak.dey@uconn.edu

ABSTRACT

The Opioid epidemic that claimed more than 63,600 lives in 2016, was declared as a public health emergency by the US government in October 2017. Although a few health insurance companies and commercial firms have examined this important issue from various available data sources, research findings from analysis of publicly available Opioid related web data is sparse. Accordingly, we have undertaken the important task of *identification of at-risk groups for Opioid addiction* through web data analysis, so that appropriate early intervention measures can be initiated by public health officials. We have collected Opioid incidences data for the states of Connecticut and Ohio for the time period of 2012 - 2018, and we are currently in the process of analyzing such data. In this paper, we present our preliminary findings and outline our plans for further research on this topic.

CCS CONCEPTS

• **Computing methodologies** → *Unsupervised learning*;

KEYWORDS

Opioid Addiction, Web Data, Risk Group Identification

1 INTRODUCTION

Opioids are drugs, prescribed by health professionals to relieve patients from pain. Unfortunately, these drugs often lead to addiction. This addiction has emerged as a full blown epidemic in the United States. In the last few years, there has been an alarming increase in Opioid related deaths, resulting in the loss of 63,600 lives in 2016 alone. In October 2017, the epidemic was declared as a public health emergency by the US government. Although a few health insurance companies and commercial firms have examined this important issue from various available data sources, research findings from analysis of publicly available Opioid related web data is sparse. Accordingly, we have undertaken the important task of *identification of at-risk groups for Opioid addiction* through web data analysis, so that appropriate early intervention measures can be initiated by public health officials. We have collected Opioid incidences data for the states of Connecticut and Ohio for the time period of 2012 - 2018, from various federal, state and local government databases, and we are currently in the process of analyzing such data. In this paper, we present our preliminary findings and lay out our plans for further research on this topic, which also includes finding the *pathways to Opioid addiction*. From the available literature, it appears that a major pathway to Opioid addiction is through drugs prescribed by medical professionals, to alleviate chronic pain of their patients. Our goal is to find out if this is the *only* pathway to Opioid addiction, or there exists other pathways, such as *peer pressure*, which is recognized as a pathway to alcohol and other non-Opioid drug addiction.

For our analysis, we have collected Opioid incidences data for the states of Connecticut and Ohio for the period 2012 - 2018. In particular, we have collected, (i) the Accidental Drug Related Deaths [1] dataset, for the state of Connecticut for 2012-2017, (ii) the USDA Economic Research Service dataset [2] for 2016-2017, (iii) the Centers for Medicare and Medicaid Services (CMMS) [3] dataset of 24 million Opioid related prescriptions written by 1 million unique prescribers in U.S. during 2014, (iv) A subset of CMMS dataset with 25,000 unique prescribers available on [4] and used by IBM researchers [5], (v) the Cincinnati Heroin Overdose dataset for 2015 - 2018 [6], and, (vi) Cincinnati neighborhood dataset for median income, median age and educational distribution [7, 8]. Brief descriptions of these datasets are provided in Section. 3.

In our effort to identify at-risk groups for Opioid addiction, we focus on the eight counties in Connecticut. Based on available data, our goal is to identify at-risk groups by taking *six* different factors - *location, race, gender, age, income and education level* into account. For our study, location is identified by one of the eight counties of Connecticut. We divide race into five categories (White, African American, Asian, Hispanic/Latino and Others), gender into three categories (Male, Female and Others) and age into four categories (Below 20, 20-39, 40-59 and 60 and above). Income level is divided into four categories (Less than $30k, $30k-$60k, $60k-$100k and above $100K), and education level is also divided into four categories (less than high school, graduated high school, some college, college graduate). Ideally, at the end of our analysis, we expect to identify

epiDAMIK @KDD '18, August 20, 2018, London, United Kingdom

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM.
https://doi.org/10.1145/nnnnnnn.nnnnnnn

at-risk groups at the following level of granularity and be able to make statements such as: *"White Male residents of Fairfield County in the age group of 20-39, within income bracket $30k-$60k, and educational level high school graduate"*, are the highest risk group in Connecticut.

In addition to identification of at-risk groups for Opioid addiction and pathways to addiction, we also examined the important contributing factors of the Opioid epidemic by attempting to answer the following questions,

- Q1: Is there a correlation between the prescribers, prescriptions and Opioid related deaths in U.S. states?
- Q2: Which prescribers are likely to prescribe more than 10 Opioid related prescriptions in a year?
- Q3: Is there a correlation between the income level, age, and the educational level and the Opioid related incidences, in a neighborhood?

From our analysis of historical data, we plan to identify the *characteristics* of risk groups and utilize it to develop a model to predict emerging risk groups in a state. In particular, we plan to develop a unsupervised learning techniques such as, clustering, for risk analysis of specific demographic groups. Using such a model, we plan to forecast the spatial effect (urban or rural) of the emerging risk groups.

In our ongoing effort, so far, we have analyzed (i) county level data of Connecticut, to rank various at-risk groups, according to their vulnerability to addiction, (ii) neighborhood level data of a specific city in Ohio (Cincinnati) to identify the impact of income level, age and educational level, on Opioid related incidences, (iii) national level data to identify the role of the individuals prescribing Opioid drugs on the spread of the epidemic.

2 RELATED WORK

In [9], the authors develop a model to identify patients at risk for prescription Opioid abuse, their dependence and misuse, using drug claims data. For exploration of pathways to prevention, the authors in [10] studied the effectiveness and risks of long term Opioid therapy for chronic pain. Machine learning techniques for surveillance of drug overdose was studied in [11]. Illicit sales of Opioid drugs, such as Fentanyl, through Twitter, was studied in [12]. Chary et. al. in [13] also analyzed Twitter data with a goal of identifying the location of the Opioid related Tweet. Data Science researchers from IBM Research and experts from IBM Watson Health have recently [14] undertaken studies in this domain. Their effort is directed towards the analysis of the relationship between factors surrounding an initial opioid prescription, and a subsequent diagnosis of addiction. The goal of this research is to identify causal factors that lead to addiction diagnosis, taking into account all the variables associated with the initial prescription, such as opioid class, quantity, and related medical procedures and diagnoses.

3 DATASETS FOR ANALYSIS

In order to identify the various at-risk groups, we first collected data from multiple sources and then munged the collected data to create additional datasets. The details of our data collection and data munging are provided in Sections 3.1 and 3.2.

3.1 Data Collection

Our collected data comprises of datasets DS1-DS7. In the following we describe each one of them.

3.1.1 DS1: It corresponds to the Accidental Drug Related Deaths 2012-2017, for the state of Connecticut [1]. This dataset comprises of 4083 unique incidences, across the state of Connecticut, during 2012-2017. Each record in DS1 includes Incidence Date, Gender, Race, Age, Residence/Death City/County of the individual involved, Location of Incidence in latitude/longitude and also the environment - hospital, residence, etc. Moreover, the records contains information related to description of injury, including drug use, substance abuse, multiple medications, etc. In addition, it provides the information about the immediate cause and specific type of Opioid and/or other drugs involved in the incidence.

3.1.2 DS2: The USDA Economic Research Service dataset [2] contains information related to poverty, population, employment/unemployment rates with median household income, and education, for the entire United States for the years of 2016 and 2017. As our focus is in the state of Connecticut, we extract county level information regarding poverty levels to educational levels, for Connecticut, from [2]. We refer to this extracted dataset as DS2.

3.1.3 DS3: The United States Census Bureau American Fact Finder dataset [15] contains information related to demographics, economics, education, etc. for the entire country. We extracted information pertaining to the state of Connecticut, by county, and refer to this extracted dataset as DS3.

3.1.4 DS4: It is the U.S. Opiate Prescriptions/Overdoses dataset available on [4]. This dataset comprises of 25000 unique prescribers, across the U.S., and the prescriptions written by them in 2014. This is a subset of the dataset maintained by the Centers for Medicare and Medicaid Services [3], that contains almost 24 million Opioid related prescriptions, written by 1 million unique health professionals (prescribers), in the U.S in 2014. Each record in DS4 includes *National Provider Identifier number, provider state, gender, credentials and the number of Opioid related drugs prescribed (among the set of 250 different drugs) by the provider.* In addition, it provides the information whether or not the provider prescribed more or less than 10 Opioid related prescriptions in 2014. It may be noted that determination of whether or not a prescriber has prescribed more than 10 prescriptions in 2014, is not done by summing up the number of drugs prescribed by the provider, as multiple drugs may be prescribed on a single prescription.

3.1.5 DS5: This dataset is also collected from [4]. It contains the population in each of the 50 states and also Opioid related deaths in that state.

3.1.6 DS6: It is the Cincinnati Heroin Overdose dataset available on [6]. This dataset is a subset of the Emergency Medical Services (EMS) dataset, where each record contains detailed information regarding an incident, such as location, time, EMS response type, neighborhood, and others, that required an EMS dispatch. This dataset contains information related to Heroin incidences from July 2015 to present time. As of April 18, 2018, there were 5568 such incidences. DS6 is a subset of EMS dataset in the sense that it contains

information only regarding Heroin incidences. It may be noted that heroin and opioid painkillers are extremely similar in terms of their chemical structure, mechanism of action and range of effects. Accordingly, for the purpose of this study, we use Heroin and other Opioid drug related data, in a similar fashion.

3.1.7 DS7: This dataset contains information regarding the median income, median age and educational distribution of various neighborhoods of Cincinnati. Information about the median income, median age and educational distribution were mined from three separate websites [7, 8].

3.2 Data Processing and Munging

We process and munge data from our collected datasets DS1-DS7, to create "secondary" datasets DS8-DS15 for the purpose of identification of at-risk groups, in the state of Connecticut. In the following, we describe our munging process:

3.2.1 DS8: It is a subset of DS1, restricted by the year 2016, to make it consistent with DS2.

3.2.2 DS9: It comprises of records of DS8, sorted by the counties, and filtered by gender, race and age.

3.2.3 DS10: It is created from DS9 and consists of 8 rows (corresponding to eight counties) and 13 columns (corresponding to number of incidences in each county; gender - Male, Female, Other; race - White, African American, Hispanic/Latino, Asian, other; age - below 20, 20-39, 40-59, above 60).

3.2.4 DS11: It is created from DS10 by considering all possible combinations of County, Gender, Race and Age. Since we have 8 different counties, 3 different genders and 5 different races and 4 different age levels, we will have 480 rows (8 * 3 * 5 * 4) and 5 columns (corresponding to a county, gender, race, age and the number of incidences for that specific county, gender, race and age).

3.2.5 DS12: It is created by sorting DS11 in descending order of the number of incidences.

3.2.6 DS13: This dataset is created by processing information available in DS4 and DS5. From DS4, we create a temporary dataset DS4A that contains information regarding the total number of prescribers and prescriptions written in each of the 50 states. DS4A was *joined* with DS5, to create DS13, that contains information regarding the total number of prescribers, prescriptions and Opioid related deaths in each of the 50 states.

3.2.7 DS14: This dataset was created by processing information available in DS6, and it contains information related to the number of Opioid related incidences in each of the 50 neighborhoods of Cincinnati.

3.2.8 DS15: This dataset was created by processing information available in datasets DS7 and DS14 and it contains information related to the median income, median age, *median education* and the number of Opioid related incidences in each of the 50 neighborhoods of Cincinnati. It may be noted that DS7 provides information related to the distribution of educational level of each of the neighborhoods. We define median education level of a neighborhood as the number of years, 50% of the residents of the neighborhood spend in school. In [8], the educational level is divided into 10 different categories from c_1, ..., c_{10} where c_1 corresponds to *None* and c_{10} corresponds to *Doctorate*. The categories c_1, ..., c_{10} correspond to n_1, ..., n_{10} years of education, with *None* implying 0 years of education and *Doctorate* implying 22 years of education. The precise definition of median education level of a neighborhood is as follows. The median educational level of a neighborhood is n_k years, if k is the smallest integer, such that $\sum_{i=1}^{k} x_i \geq 50$, where x_1, ..., x_{10} represents the percentage of neighborhood population that has educational levels corresponding to c_1, ..., c_{10}.

4 PROPOSED WORK

In order to identify at-risk groups with a high level of accuracy, one obviously needs to have access to relevant data. Data pertaining to Opioid related prescriptions, incidences, such as, calls to Emergency Medical Services (EMS) and deaths, are owned by multiple stakeholders, such as the insurance companies, hospitals, EMS providers and drug stores. Public health organizations at the federal, state and local level often collect such data, anonymize and aggregate them and make it available on the web. For our analysis, we have used such data made available by Centers of Medical and Medicaid Services [3], Health and Human Services department of Connecticut [1] and EMS responses for the city of Cincinnati [6]. However, such data often do not contain information at the individual level, e.g., it does not provide medical history of an individual as to how many or how often the individual was taking Opioid drugs. Moreover, it does not contain socio-economic and educational background of the individual. Data related to the medical history of an individual is available to the insurance companies, hospitals and drug stores. We did not have access to such data. However, we are making an effort to collect such data from these sources.

Once we acquire such data, we plan to develop a mathematical model to estimate the association between Opioid abuse and medical history and demographic characteristics of individuals. Given the demographic information and medical history of an individual, the response variable of the model, will assign the individual to a risk group. We also plan to develop unsupervised learning paradigms, as the number of risk groups may vary over time. We are currently developing mixture models, using the Expectation-Maximization algorithm.

5 PRELIMINARY WORK RESULTS

In this section, we present preliminary results for identification of at-risk groups, as well as answers to Q1-Q3, presented earlier. In the following, we briefly discuss these results

5.1 At-Risk Group Identification

As noted earlier, ideally, we would have liked to identify at-risk groups at the following level of granularity and able to make statements of the form: "White Male *residents of* Fairfield County *in the* age group of 20-39*, within* income bracket $30k-$60k*, and* educational level high school graduate*"*, are the highest risk group in Connecticut. However, the datasets available to us currently does not provide any information related to income and education level of the individual involved in the Opioid related incidence. Accordingly, we are unable to identify at-risk groups at a level of

granularity involving six factors (Race, Gender, County, Age, Income and Education). Instead, we identified at-risk groups involving four factors (Race, Gender, County, Age).

It may be recalled from Section. 3.2 that DS12 was created by sorting DS11 in descending order of the number of incidences, where DS11 contained all possible combinations of County, Gender, Race and Age. Thus, DS12 contains 480 rows (corresponding to 8 different counties, 3 different genders and 5 different races and 4 different age levels) and 5 columns (corresponding to county, gender, race, age and the number of incidences for that specific county, gender, race and age). Each of the 480 rows correspond to a *risk group* identified by county, gender, race and age. We define a risk group to be the *highest risk group* if the number of Opioid related incidences for this group is the highest among all risk groups. Based on our analysis, we present the five highest risk groups, in Connecticut, in Table. 1.

County	Race	Gender	Age Group	No. of Incidences
Hartford	White	Male	20-39	66
New Haven	White	Male	40-59	64
Hartford	White	Male	40-59	61
New Haven	White	Male	20-39	53
Fairfield	White	Male	20-39	44

Table 1: Five Highest Risk Groups in Connecticut

From Table. 1, we find that White Males as a demographic group, is the highest risk group in Connecticut. Moreover, two counties - Hartford and New Haven, are among the worst affected, by the Opioid epidemic. From a different dataset DS3, we can identify income and educational level characteristics of the highest risk groups (White, Male, Hartford/New Haven County). The income level for White males in Hartford, Connecticut was $63,200 in 2016 and 91.7% of them were at least High School graduates. The corresponding numbers for New Haven, Connecticut were $48,985 in 2016 and 91.3% respectively.

5.2 Results of Data Analysis for Q1-Q3

We used DS4, to answer the first two questions. Question3 was answered using DS15. We briefly summarize our findings below,

5.2.1 Data Analysis for Q1. We used partial correlation to analyze the relationship between the prescribers/number of prescriptions and the number of Opioid related deaths. From Table. 2, we can infer that there is a moderate positive correlation between the number of prescribers and prescriptions with Opioid related deaths.

	Number of Prescribers	Number of Prescriptions
Opiate Deaths (Partial Correlation)	0.4664	0.3619

Table 2: Partial Correlation Coefficients between Opiate Deaths and Prescribers and Prescriptions

5.2.2 Data Analysis for Q2. We used several machine learning algorithms to predict prescribers who are likely to prescribe high number of Opioid prescriptions, by analyzing the trend of prescribing *non Opioid* drugs. IBM ran some initial machine learning algorithms and attained accuracies ranging from 60% to 84%. Using XGBoost and CatBoost, we had accuracy scores of 81.8% and 84.7%. The CatBoost model provided a feature importance array, which identified "specialty" as the most important feature. The prescribers with specialty "Addictive Medicine", prescribed the highest average of annual Opioid drugs. We also examined the average annual Opioid prescription rates by state and observed a trend of higher prescription rates in the southern states. Furthermore, we implemented a Mulit-Layer Perceptron and a Random Forest on the dataset. We did not consider the specialty of the prescribers in these two models. We achieved a training accuracy of 95.6% and a testing accuracy of 89.7%, using the MLP, and a testing accuracy of 89% using the Random Forest model.

5.2.3 Data Analyses for Q3. We used the Cincinnati Heroin Overdose dataset, along with the Cincinnati neighborhood dataset to identify the relationship (using partial correlation) between the income levels, age and educational levels of a neighborhood, and the number of Opioid related incidences. The results are tabulated in Table. 3. From the table, we can infer that the Opioid addiction affects the entire spectrum of income levels and age, and is not restricted to a particular level. In addition, we found that, with an increase in the educational level of a neighborhood, there is a decrease in the Opioid related deaths, albeit slightly.

	Median Income	Median Age	Median Education
Opiate Deaths (Partial Correlation)	-0.0576	-0.0789	-0.1516

Table 3: Partial Correlation Coefficients between Opiate Deaths and Median Income/Age/Education

REFERENCES

[1] City of Connecticut, "Accidental Drug Related Deaths", https://data.ct.gov/Health-and-Human-Services/Accidental-Drug-Related-Deaths-2012-2017/ecj5-r2i9
[2] USDA, Economic Research Service, "https://www.ers.usda.gov/data-products/county-level-data-sets/", 2016
[3] Centers for Medicare and Medicaid Services, "https://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/Medicare-Provider-Charge-Data/Part-D-Prescriber.html.
[4] Kaggle Dataset: "U.S. Opiate Prescriptions/Overdoses", https://www.kaggle.com/apryor6/us-opiate-prescriptions.
[5] IBM Opioid Github: "https://github.com/IBM/predict-opioid-prescribers", 2017.
[6] City of Cincinnati, "Heroin Overdoses", https://insights.cincinnati-oh.gov/stories/s/Heroin/dm3s-ep3u/.
[7] City-Data, "http://www.city-data.com/city/Cincinnati-Ohio.html".
[8] Statistical Atlas, "https://statisticalatlas.com/place/Ohio/Cincinnati/Overview".
[9] J. B. Rice, A. G. White, H. G. Birnbaum, M. Schiller, D. A. Brown, and C. L. Roland. "A model to identify patients at risk for prescription opioid abuse, dependence, and misuse." Pain Medicine 13, no. 9 (2012): 1162-1173.
[10] R. Chou, J.A. Turner, E. B. Devine, R. N. Hansen, S. D. Sullivan, I. Blazina, T. Dana, C. Bougatsos, and R. A. Deyo. "The effectiveness and risks of long-term opioid therapy for chronic pain", Annals of internal medicine 162, no. 4 (2015): 276-286.
[11] D. B. Neill, W. Herlands. "Machine Learning for Drug Overdose Surveillance." Journal of Technology in Human Services (2018): 1-7.
[12] T.K. Mackey, J. Kalyanam, T. Katsuki, G. Lanckriet, "Twitter-Based Detection of Illegal Online Sale of Prescription Opioid", American J. of Public Health, 2017.
[13] M. Chary, N. Genes, C. Giraud-Carrier, C. Hanson, L.S. Nelson, A.F. Manini, "Epidemiology from Tweets: Estimating Misuse of Prescription Opioids in the USA from Social Media", Journal of Medical Toxicology, 13, 278-286, 2017.
[14] D. Wei, "Combating the Opioid Epidemic with Machine Learning", https://www.ibm.com/blogs/research/2017/08/combating-the-opioid-epidemic-with-machine-learning/, 2017.
[15] United States Census Bureau Fact Finder, "https://factfinder.census.gov", 2016.