Heinz 95-845: Project Proposal

Mridul Gangwar MGANGWAR@ANDREW.CMU.EDU
Heinz College of Information Systems and Public Policy
Carnegie Mellon University, Pittsburgh, PA, United States

Lauren Rost LROST@ANDREW.CMU.EDU
Heinz College of Information Systems and Public Policy
Carnegie Mellon University, Pittsburgh, PA, United States

Nikita Setia NIKITAS@ANDREW.CMU.EDU
Heinz College of Information Systems and Public Policy
Carnegie Mellon University, Pittsburgh, PA, United States

1. Proposed Analysis and Likely Outcomes

This project will use a set of demographic features, usage of county-provided programs, and opioid prescription information to predict overdose deaths (opioid and non-opioid). Specifically, we propose to analyze the aforementioned data of Medicaid beneficiaries from 2009 to 2017 provided by Allegheny County Department of Human Services (DHS) to predict the risk of an individual dying due to an non-opioid or opioid-related overdose. An additional component will be to cluster these individuals to see whether sub-groups exist and how clusters relate to overdose deaths. The likely outcome of this analysis is the finding of at least one feature and subgroup that is highly deterministic in the prediction of overdose deaths, thereby enhancing the space of addiction and overdose death prevention.

2. Importance of Proposed Analysis

National drug overdose deaths have increased over the past two decades, from 16,849 in 1999 to 70,237 in 2017[1]. This trend is especially affected by the onset of the opioid epidemic in the late 1990s[2]. Approximately 67% of the overdose deaths in 2017 are due to involvement of any opioids and 24% are specifically due to prescription opioids [3]. The U.S. Department of Health and Human Services has a 5-point strategy to combat the opioid crisis [4]. These points include gathering better data and conducting pertinent research to prevent addiction, thereby preventing overdose deaths. Through this analysis, we hope to contribute to the HHS strategy to combat the opioid crisis.

3. Contribution to Existing Work

In 2005, the Opioid Risk Tool was developed to help flag patients at risk for opioid abuse and overdose[5]. However, due to the subjective nature of this tool and the spike of deaths in 2017, machine learning has been sought to provide a more objective and quantitative approach to estimate risk of opioid abuse and overdose.

Acion et al. used a super learning approach to predict the successful treatment for patients with substance use disorders[6]. Haller et al. utilized natural language processing on electronic health record data to assess risk and predict opioid abuse [7]. Crosier et al. predicted overdose frequency using random forests in order to uncover important features related to overdose frequency and course[8]. Lobo et al. identified sub-groups of Pennsylvania patients at greater risk for opioid abuse in a k-means clustering algorithm[9].

Our research will expand upon previous work seeking to understand opioid abuse and overdose, specifically as it relates to machine learning. We will implement a more in-depth approach to predicting opioid abuse and will be specific to Allegheny County, an area that is greatly impacted by the opioid crisis[10].

4. Data Description

The population (W) is 120,650 individuals who are enrolled in Medicaid and have obtained at least one opioid prescription in Allegheny County between 2009 to 2017. Once an individual enters the system, all demographic information, subsequent interaction with DHS, and prescription fills were recorded in separate datasets (demographic, program, and prescription).

The demographic data contains each person's race and gender. The program activity entails each individual's usage of DHS programs and criminal records over time. Programs include: Child, Youth and Family (CYF) program (as child or parent), mental health service, drug and alcohol service, and opiate prescription service. Criminal records entail an individual's count of criminal cases and incarceration status. The prescription dataset includes various features pertaining to each fill, such as dose, brand, etc. Further information can be accessed in our Github repository.

The covariates (V) are select features present in each of the datasets. Some of these features will need to be pre-processed or customized. Our binary outcomes of interest (Y) are whether the individual dies of overdose (any) and whether the event was an opioid-related overdose.

5. Evaluation Measures

To evaluate the performance of various machine learning models, we intend to create ROC curves and determine area under the curve. Given that the number of overdose deaths (especially opioid-specific) are very low, our dataset is imbalanced. As such, we intend on using other measures more sensitive to this imbalance, such as: specificity, sensitivity, f-1 score, positive predicted value (PPV), negative predictive value (NPV), and precision versus recall.

6. Study Design, Pre-Processing and Machine Learning Methodology

We expect the missing values to be either missing completely at random (MCAR) or missing not at random (MNAR). To address this, we will either create missingness as its own category or use a pertinent imputation technique. We may use oversampling to overcome the problem of the imbalanced dataset or some form of cost-sensitive classification. We plan to use the following machine learning algorithms: multivariate logistic regression (MLR), regularized MLR (lasso), random forest (RF), boosting, gradient boosting machine (GBM), and k-means clustering. We aim to use cross-validation and tune the hyperparameters to obtain the best possible results.

7. Potential Future Use of Pipeline/Study and Possible Limitations

We hope that our study and its analytic pipeline can be used by counties to analyze their data to predict and prevent overdose deaths and opioid addiction. The possible limitations of this study include our ignorance to Allegheny County data collection methodologies. This may potentially limit the generalizability of this study to other counties. Limitations affecting internal validity are the lack of additional meaningful features, such as: insurance claims, diagnosis, medical history and socio-economic status. We strive for our opioid pipeline to be sufficiently robust and verbose to allow for appropriate adaptation.

Endnotes

[1]National Institute On Drug Abuse (2019a)
[2]National Institute On Drug Abuse (2019b)
[3]National Institute On Drug Abuse (2019a)
[4]Health and Services (2019)
[5]Webster and Webster (2005)
[6]Acion (2017)
[7]Haller (2017)
[8]Sage Crosier (2017)
[9]Lobo (2017)
[10]County (2019)

References

Kelmansky D. Laan M. V. Sahker E. Jones D. Arndt S Acion, L. Use of a machine learning framework to predict substance use disorder treatment success, April 2017. URL https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0175383.

Allegheny County. Information about opioids and overdose prevention, 2019. URL https://alleghenycounty.us/Health-Department/Programs/Special-Initiatives/Overdose-Prevention/Information-About-Opioids.aspx.

C.M. Renier M. Juusola P. Hitz W. Steffen M.J. Asmus T. Craig J. Mardekian E.T. Masters T.E. Elliott Haller, I.V. Enhancing risk assessment in patients receiving chronic opioid analgesic therapy using natural language processing, October 2017. URL https://www.ncbi.nlm.nih.gov/pubmed/28034982.

Health and Human Services. The prescription drug and heroin overdose epidemic, March 2019. URL https://www.hhs.gov/opioids/.

H. Jalal C. Chang G. Cochran J. Donohue Lobo, C. Panel paper: Using unsupervised machine learning to identify potentially problematic opioid use in medicare, November 2017. URL https://appam.confex.com/appam/2017/webprogram/Paper23485.html.

NIDA National Institute On Drug Abuse. Overdose death rates, Jan 2019a. URL https://www.drugabuse.gov/related-topics/trends-statistics/overdose-death-rates.

NIDA National Institute On Drug Abuse. Opioid overdose crisis, Jan 2019b. URL https://www.drugabuse.gov/drugs-abuse/opioids/opioid-overdose-crisis.

J. Borodovsky P. Mateu-Gelabert H. Guarino Sage Crosier, B. Finding a needle in the haystack: Using machine-learning to predict overdose in opioid users, February 2017. URL Findinganeedleinthehaystack:Usingmachine-learningtopredictoverdoseinopioidusers.

L.R. Webster and R.M Webster. Predicting aberrant behaviors in opiod-treated patients: Preliminary validation of the opiod risk tool, December 2005. URL https://www.ncbi.nlm.nih.gov/pubmed/16336480.