Using Machine Learning to Understand the Georgia Opioid Crisis

Seungkwan Baek

Abstract— **This study aims to produce an analytical solution to the rising rates of opioid-related overdoses. A binary classifier was used to predict counties in Georgia that may be at-risk for a dangerous level of opioid overdoses in future years. This predictive classifier was then used to test if an at-risk county might be re-classified as safe if it had more health resources, such as hospitals and emergency medical dispatchers. We found that demographic, medical, and drug-related statistics and data can be used to create an accurate Random Forest binary classifier. Using this classifier, we observed that increasing health resources does, indeed, reduce the amount of risky counties predicted for the state of Georgia.**

I. INTRODUCTION

America is facing an increase in opioid addiction. More than 90 Americans die of opioid overdose every day [10]. This epidemic of opioid abuse (Opioid Crisis) causes an annual economic burden of $78.5B [10]. In recent years, US lawmakers invested $100 million to improve surveillance of opioid prescriptions, but these changes have not been enough to solve the crisis [6]. It is still a public health problem that affects the entire nation. We want to analyze factors that signal at-risk counties in Georgia and study if these counties have enough resources to curb the opioid mortality rate.

II. PROBLEM DEFINITION

This study identifies Georgia counties with the highest opioid-related deaths and determine strong indicators of opioid overdose. Then we will formulate and evaluate multiple models to forecast opioid abuse by county. The resulting predictive model can be used to determine if enough resources are in place in these counties at risk. If not, this model can suggest how to efficiently allocate health-related resources in hopes of preventing future vulnerability. If successful, this study can aid Georgia state officials to re-allocate funding to higher-risk counties for emergency medical preparedness and increase public awareness of opioid drug abuse. We can measure success by collecting the data for those counties after few years to observe if the resource reallocation had a significant effect on the overall rate of overdose-related mortality. Upon proven success, the methods used in this study can be extended to other states.

III. RELATED WORKS

The Opioid Crisis is a nation-wide concern, so there are many relevant works in visualizing high-risk areas for opioid-related deaths and in examining factors contributing to these fatalities. This study adds advanced statistical forecasting and analysis of health-related resources in affected areas to move towards a solution that allows for better targeting of high-risk areas in Georgia.

A. Identifying High Risk Regions

Most current studies surrounding the Opioid Crisis involve identifying areas in America with high rates of drug-related mortality. There are many proposed methods to find these high-risk regions, such as unsupervised machine learning to scrape the social media sites for non-medical opioid use [1], natural language processing to mine health records to identify cases of opioid misuse [3], or analyzing if current resources are sufficient enough to tackle the ongoing crisis [11]. Another large part of identifying high-risk regions is the analysis of geographic-based data, such as examining opioid sales rates per region [2], which can be done using methods such as density-based clustering [8]. These studies have pinpointed problem areas in America at the state and county level, but they are limited in that they do not incorporate any forecasting to the analysis.

B. Predicting Future Opioid Abuse

Many prediction models link drug-related mortality to a single predictor, such as linear regression trend analysis of opiate sales in North Carolina [2], logistic regression model of association between opioid abuse diagnosis and patient demographics/characteristic s [12], relationship analysis between overdose death and prescription dosage of drugs [13], success prediction of substance abuse treatment using machine learning models [7], or examination of the effect of socioeconomic factors on drug-related emergency room visits or deaths in a region [4].

These studies have produced some single-predictor models for predicting future opioid abuse in a region, but there are few studies that compare multiple predictors [5] and allow for incorporation of effects at different level [9]. This study will expand upon the proposed models and examine multiple predictors to find the most significant factors in predicting opioid-related deaths in Georgia, and utilize appropriate measures to quantify drug abuse [14]. Additionally, current studies are lacking analysis of the health-related resources in these high-risk areas. This study will incorporate these resources to determine if a predicted high-risk area is well-equipped to handle opioid abuse and to potentially save lives. Studies have identified resources that would likely prevent future opioid outbreaks in a region, but they have not incorporated these into a model [15]. This will be successful because the outcome of the study will provide clear, concrete actions that can be taken to mitigate the opioid crisis.

IV. Proposed Method

A. *Innovation*

Our methods were designed to extend the focus of this study beyond related works. Our analysis incorporates prediction of future at-risk counties using a predictive binary classifier, as opposed to implementing an exploration of counties already experiencing opioid abuse or creating a simple regression model to examine the effects of certain factors. The real innovation in our methods stems from creating a modular classifier that can take in user-defined changes to a countys health resources to examine the effect on opioid overdose. We aimed to provide a solution that can better allow budget planning to prevent further opioid abuse in Georgia.

B. *Algorithm*

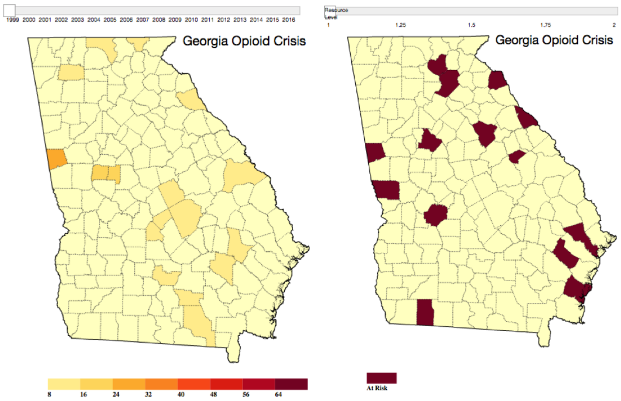
Implementation of our methods started with data collection. We scraped data from numerous demographic, health, and public safety-related resources dating back nearly ten years for each county (see Appendix for more detail). Features were selected based on intuition, internet research, and expert opinion. We spoke to Dr. Mansoor Khan, Vice-Dean of Texas A&Ms College of Pharmacy, about factors he believed were correlated with opioid abuse. This data was first cleaned via OpenRefine, normalized, and aggregated into a single dataset with predictors such as crime rates, number of hospitals, etc., and responses of overdose rates for each county. Then, we used percent change over one year, five years, and ten years to discretize time series data for our binary classification analysis. Then, we forecasted temporal data to allow us to test our model since we do not yet have data for future years. We used exponential smoothing to generate these forecasts.

The next step in our methods was to choose a machine learning classifier that could best predict at-risk counties in Georgia given a set of features. We defined at-risk to mean that the county exhibits overdose rate below the national overdose rate overdose rate now, but they will be above the national rate in 2017 and 2018. We bootstrapped samples from our data set to compare classifiers generated using GridSearchCV of Linear SVM, Logistic Regression, Stochastic Gradient Descent, and Random Forest. Random Forest outperformed compared to other classifiers; we used this to find at-risk counties and use this to classify hypothetical instances of these counties with increased health resources to see if their status changes from at-risk to safe.

C. *Visualization*

Our visualization was implemented using D3 choropleth capability. The first part of the visualization shows the overdose rates for all counties from 1999 to 2016. The second part of the visualization shows future at-risk counties in red, and updates the map as the user drags a bar with multiplicative changes to state-wide health resources. Figure 1 below shows our visualization.

Fig. 1: Visualization of the Georgia Opioid Crisis.



V. Experiments and Evaluation

Our experiments were designed to answer the following questions:

- Which counties are not in an opioid crisis now, but are at risk?
- Which model would help us deliver the best performance in prediction?
- Does changing health resources appear to have an effect on a county's opioid overdose risk?
- What is the optimal allocation of resources to minimize the number of counties in risk?

To answer the first two questions, we performed a grid search to find a binary classifier that best predicted if a countys opioid overdose rate was above our nationally-set threshold. The table below shows the results of our analysis.

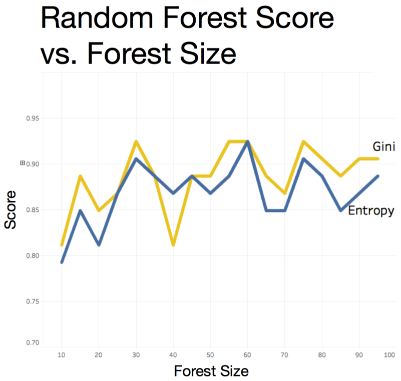
TABLE I: Grid Search Results to Select Binary Classifier

Model	**Accuracy**
Linear SVM	71%
Logistic Regression	67%
Stochastic Gradient Descent	60%
Random Forest	91%

The Random Forest classifier resulted in a higher accuracy, on average, than the other three types of classifiers. To fine-tune our model, we tested classification scores for classifiers created using random forests of varying sizes (number of decision trees).

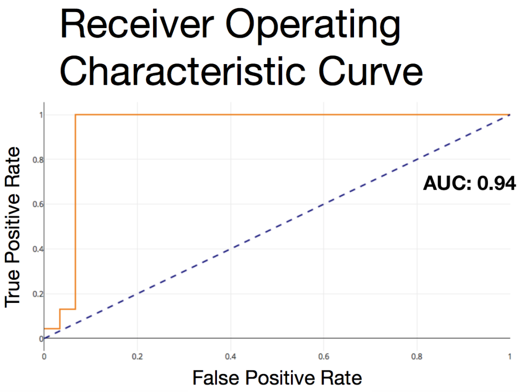
Figure 2 below shows the results of this experiment.

Fig. 2: Selecting Random Forest Size



A forest size of 60 decision trees resulted in the best classification score for our model using both Gini and Information Gain criteria. After selecting and fine-tuning our model, we tested it against our forecasted data from our exponential smoothing experiment. The Receiver Operating Characteristic Curve for this test is shown in Figure 3 below.

Fig. 3: Testing Classification Accuracy



Our classifier resulted in an Area-Under-the-Curve statistic of 0.94, suggesting that the ratio of true positives to false positives is close to 1. Thus, our model performed well against our test data. Further testing of our predictive model should be done when real data is available in the coming years.

The next question we set out to answer was whether or not changing health resources will lower the amount of opioid overdoses in any given county. To test this, we took sets of at-risk counties and changed the collective health resource features (number of hospitals, number of EMS dispatch units, etc.) by multiplicative and additive factors. We ran our classification 1000 times first for a normal level of resources and picked counties that were classified as ”at-risk” at least 40 times. We then ran our classification 1000 more times with an increased level of resources (either 1.5x, 2x, or so on), and counted counties that were classified as ”at-risk” at least 40 times and were considered to be ”at-risk” at a normal resource level. This prevented the odd cases where a county that wasn’t previous at-risk is classified as at-risk with more resources. The number of counties classified as ”at-risk” with increased resources was consistently less than the number of counties classified as ”at-risk” with normal resources.

VI. CONCLUSIONS

Our experiments showed that data reflecting demographics and health resources can be used to create a good binary classifier to predict whether or not a county in Georgia will have an opioid overdose rate above a dangerous threshold. Our results also showed that an increase in state-wide health resources should decrease the number of counties in Georgia that are at-risk for high rates of overdose. While these results were tested against a forecasted data set, if the results hold using real data gathered in the future, this study could prove to be valuable in combating the opioid crisis in states other than Georgia by suggesting more precise budget improvements for certain health resources.

APPENDIX

A. *Detailed Data Description*

Our data set was aggregated using the following sources:

Source	**Data**
Georgia Dept. of Public Health	-Overdoses 1999-2016 -Opioid Overdoses 1999-2016 -Heroin & Pain Reliever Overdoses 1999-2016 -Unknown Overdoses 1999-2016 -Opioid Overdoses Male vs. Female
Georgia Prescription Drug Abuse Prevention Initiative	-Georgia Prescription Drug Dropboxes
Association of American Medical Colleges (AAMC)	-List of Hospitals in Georgia
Georgia Free Rehab Centers	-List of rehab centers in Georgia
Georgia EMS Directory	-List of EMS services in Georgia
US Census	-Birth Rate -Death Rate -Population
CDC	-Opioid Prescription Rate per county
US Bureau of Labor Statistics	-Unemployment Rate
Georgia Bureau of Investigations	-Crime Rate by County

References

[1] Janani Kalyanam, Takeo Katsuki, Gert R.G. Lanckriet, Tim K. Mackey, Exploring trends of nonmedical use of prescription drugs and polydrug abuse in the Twittersphere using unsupervised machine learning, Addictive Behaviors, Volume 65, February 2017, Pages 289-295.

[2] F. Modarai, K. Mack, P. Hicks, S. Benoit, S. Park, C. Jones, S. Proescholdbell, A. Ising, L. Paulozzi, Relationship of opioid prescription sales and overdoses, North Carolina, In Drug and Alcohol Dependence, Volume 132, Issues 12, 2013, Pages 81-86.

[3] David S. Carrell, David Cronkite, Roy E. Palmer, Kathleen Saunders, David E. Gross, Elizabeth T. Masters, Timothy R. Hylan, Michael Von Korff, Using natural language processing to identify problem usage of prescription opioids, In International Journal of Medical Informatics, Volume 84, Issue 12, 2015, Pages 1057-1064.

[4] Alex Hollingsworth, Christopher J. Ruhm, Kosali Simon, Macroeconomic conditions and opioid abuse, In Journal of Health Economics, 2017.

[5] Christopher J. Ruhm, Geographic Variation in Opioid and Heroin Involved Drug Poisoning Mortality Rates, In American Journal of Preventive Medicine, 2017.

[6] Continued Rise in Opioid Overdose Deaths in 2015 Shows Urgent Need for Treatment. National Archives and Records Administration, National Archives and Records Administration, obamawhitehouse.archives.gov/the-press-office/2016/12/08/continued-rise-opioid-overdose-deaths-2015-shows-urgent-need-treatment.

[7] Acion, Laura, et al. Use of a Machine Learning Framework to Predict Substance Use Disorder Treatment Success. Plos One, vol. 12, no. 4, Oct. 2017, doi:10.1371/journal.pone.0175383.

[8] MacKinnon, David P., and Chondra M. Lockwood. Advances in Statistical Methods for Substance Abuse Prevention Research. Prevention science: the official journal of the Society for Prevention Research 4.3 (2003): 155171. Print.

[9] Imbens, Guido W., and Donald B. Rubin. Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction. Cambridge University Press, 2015.

[10] Abuse, National Institute on Drug. Opioid Crisis.NIDA, 1 June 2017, www.drugabuse.gov/drugs-abuse/opioids/opioid-crisis.

[11] Annals of the American Thoracic Society. The Critical Care Crisis of Opioid Overdoses in the United States — Annals of the American Thoracic Society — Articles in Press, www.atsjournals.org/doi/pdf/10.1513/AnnalsATS.201701-022OC.

[12] J. Bradford Rice, PhD, Alan G. White, PhD, Howard G. Birnbaum, PhD, Matt Schiller, BA, David A. Brown, PhD, MPH, Carl L. Roland, PharmD; A Model to Identify Patients at Risk for Prescription Opioid Abuse, Dependence, and Misuse, Pain Medicine, Volume 13, Issue 9, 1 September 2012, Pages 11621173.

[13] Bohnert ASB, Valenstein M, Bair MJ, Ganoczy D, McCarthy JF, Ilgen MA, Blow FC. Association Between Opioid Prescribing Patterns and Opioid Overdose-Related Deaths. JAMA. 2011;305(13):13151321.

[14] Secora A. M., Dormitzer C. M., Staffa J. A., and Dal Pan G. J. (2014) Measures to quantify the abuse of prescription opioids: a review of data sources and metrics, Pharmacoepidemiol Drug Saf, 23, pages 12271237.

[15] Hawk, Kathryn F., Federico E. Vaca, and Gail DOnofrio. Reducing Fatal Opioid Overdose: Prevention, Treatment and Harm Reduction Strategies . The Yale Journal of Biology and Medicine 88.3 (2015): 235245. Print.