Analysis and ranking of influential features in opioid abuse trends within the U.S.

Mario Iurlaro, Christopher Nelson, Thomas Swarbrick, Daniel Vela, Patrick A. Wall

1. INTRODUCTION

Prescription opioid abuse in the United States is an epidemic that has affected millions, with an estimation of drug abusers in recent years exceeding 15 million [Basak et al, 2019]. Determining the factors that create a greater risk of opioid overdose hospitalisation or death (overdose harm) is key to creating effective public health interventions to reduce the epidemic's effects. Our objective for this project was to build an interactive visual tool to identify emerging trends in the opioid epidemic, utilising available open-source data to analyse which demographic features could be used to predict opioid abuse within specific regions. We anticipate that the results may provide a new way for local public health departments and organisations to identify the unique circumstances that contribute to opioid abuse in their region, or determine which demographics in their jurisdictions are the most at-risk for prescription opioid abuse, to focus their interventions where they are most needed.

2. LITERATURE SURVEY

In recent years, an increasing amount of studies have used analytical approaches to potentially identify contributing factors, causal relationships and geographical variations of the opioid epidemic. However, the current practice is often restricted to small regional methods or limited in the scope of potential factors. Recent studies have used machine learning approaches to predict the risk of opioid use disorder on a population level [Hasan et al. 2021] or applied probabilistic models to find variation in prescription trends [Hu et al., 2015], but were restricted to Massachusetts or areas in Queensland (Australia), respectively. Areas of higher prescription rates have also been a central focus of research, in an attempt to identify areas of urgent action and causality factors. One approach used spatial cluster detection to monitor local patterns of prescription opioid abuse in New Mexico [Brownstein et al.]. A similar approach identified smaller hot-spot areas where the opioid prescription rate was significantly higher than the baseline measured as the rate of the whole state [Basak et al. 2019]. All these studies highlight the potential of using quantitative approaches to identify areas and predictive factors associated with the opioid abuse emergency, but also reveal limitations in the data.

Understanding the spatial-temporal effect of opioid abuse in the United States is a crucial aspect of this effort, as it might help highlight systemic differences that are at the basis of the increase in opioid usage. An argument is to be made that the misuse of opioids reflects the most substantial public health crisis that the US has faced [Maclean et al., 2020]. A multi-level analysis has shown as much as 10% of the variation in opioid intake can be explained by geographical data itself [Webster et al. 2009], even if using state-level data only. More recent studies have shown how there is a large geographical variation in opioid prescription rate and a general lack of consensus regarding opioid use to treat pain [McDonald et al. 2012; McDonald and Carlson, 2014].

Opioid abuse data can be hard to track as many health problems, including causes of death, may not seem opioid-related. Polysubstance abuse often occurs in those who are addicted to opioids. Studies have shown that the less widely used, the more likely a drug is to be used with other drugs [Compton et al., 2021]. Because of this, we could potentially use other types of drug overdoses as potential predictors in our analysis [Unick & Ciccarone, 2017; Unick et al., 2013].

There seem to be clear differences in opioid intake between urban and rural areas. An analysis of prescription opioid poisoning across counties in California indicated factors, such as higher pharmacy density in urban areas and low income with more manual labour industries in rural areas as potential causes for this variation [Cerda et al., 2017]. In the case of Purdue Pharma, they would target low-income miners in rural communities to target

Oxycontin which led to a major crisis [Whelan & Asbridge, 2013]. Availability of healthcare services, differences in demographics and unemployment rates have also been indicated as potential causes of this difference [Sun et al. 2022; Duan & Hand, 2021]. All these indicate the importance of including the rural/urban aspect in our analysis, while also clearly highlighting the limitation associated with single-state data or lack of accountability of confounding variables.

Public health departments and governments need to know where to best allocate resources to combat the ongoing prescription opioid epidemic in the US. Current research focuses on a single specific state or on a larger country scale. Our analysis investigated potential factors that contribute to the opioid epidemic at a state level using county-level data while building a visually interactive tool to visualise trends.

3. METHODS

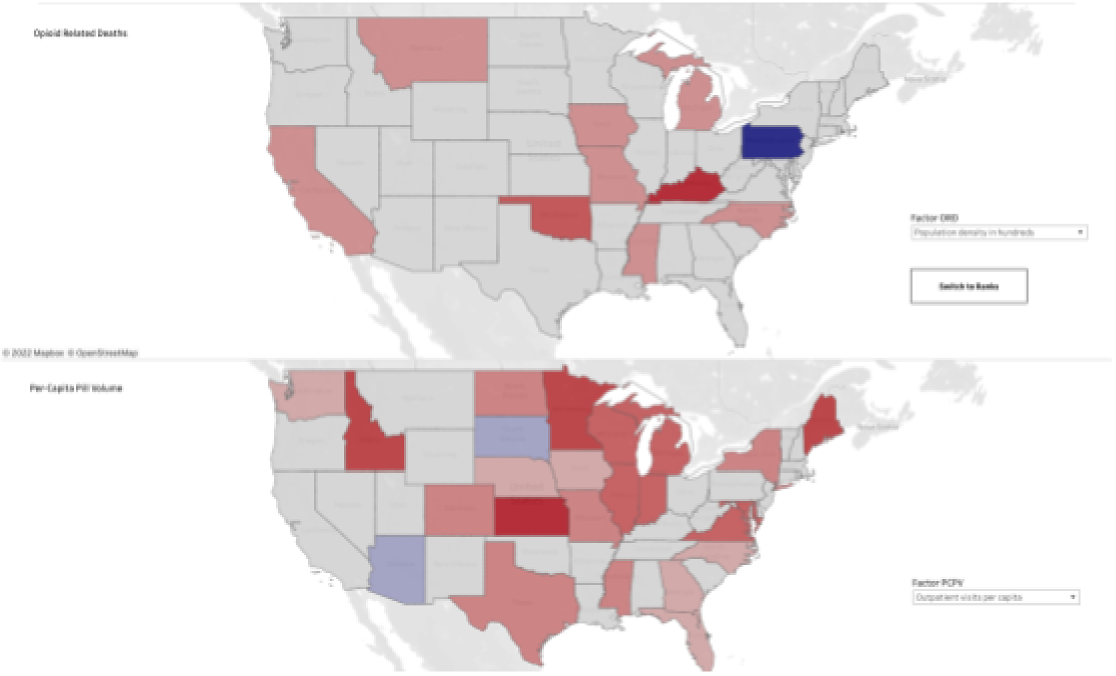
A county-level dataset [Griffith et al. 2021] with U.S. opioid prescription pill distributions, demographics and several other county-level variables from 2006-2013 was used in the analysis to determine the factors that contribute to increased levels of opioid use and abuse. The dataset consists of approximately 27,000 data points and 156 factors that leverage several U.S. federal data resources such as the Automation of Reports and Consolidated Orders System (ARCOS), the Center for Disease Control (CDC) Wide-ranging Online Data for Epidemiologic Research (WONDER) database and the Health Resources & Services Administration's (HRSA) Area Health Resource File (AHRF). It combines extracted pill shipments for oxycodone and hydrocodone to retail pharmacies in the U.S. from the database ARCOS, data on opioid-related deaths and cancer deaths from WONDER, and county-level characteristics such as annual data on demographics, healthcare workforce, rurality, unemployment rate, etc, from AHRF. The ARCOS data was aggregated to calculate county-level annual per capita pill volume (PCPV). Since WONDER suppresses data for counties with less than 10 deaths, the authors used imputation methods to estimate Opioid-Related Deaths (ORD). Other factors also had imputed data and the dataset contained both the imputed (*_IMP) and non-imputed (*_NOIMP) versions of these factors. The dataset also consisted of a few factors that were duplicates (e.g. N_BLACK and F04538). The duplicate columns and the non-imputed versions of the factors were removed before our analysis.

Based on our literature review, analysis to date has been completed on either individual states such as New Mexico, West Virginia or North Carolina [Brownstein et al. 2010, Basak et al. 2017] or on the entire country [Griffith et al. 2021]. To determine the features that create a greater risk of opioid abuse, we performed multiple linear regressions on county-level demographic data for each state. Our analysis consisted of 51 regression models (50 states + D.C.) using county-level data through the years 2006-2013 for two different response variables.

Given the large number of demographic features provided in the county-level dataset, it was difficult to determine feature relevance due to factors such as multicollinearity. To address this problem, we used a regularisation technique to determine which features provided the most predictive power and explained the geographic and temporal trends observed. Based on our review, this type of feature selection has not been done before on opioid prescription and mortality data at this scale.

The regression analysis used the L1 regularisation technique, LASSO, to determine feature relevance and overcome issues such as multicollinearity. The regularisation term, α, was determined using 5-fold cross-validation for each model to give the best fit to the data with the least amount of features. PCPV and ORD were selected as features to predict as we believed they provided the most insights into opioid abuse, with PCPV offering a leading indication of abuse and ORD offering lagging insights. For the models with PCPV as the response variable, we omitted ORD as a factor in the regression analysis. Our regression analysis results were pre-computed and exported as a ".csv" file for the interactive visual tool.

An interactive visual tool identifying opioid overdose trends was developed with the use of the results from our regression analysis. This tool will allow public health departments to identify factors which may contribute to opioid abuse in their local jurisdiction and design intervention techniques and policies focused on specific geographic areas and demographics. The interactive visual tool, shown in Figure 1, is two maps of the U.S. that can demonstrate the impact on opioid harm (ORD on top and PCPV on bottom) of any given factor from a dropdown list of factors located on the bottom right of the map.

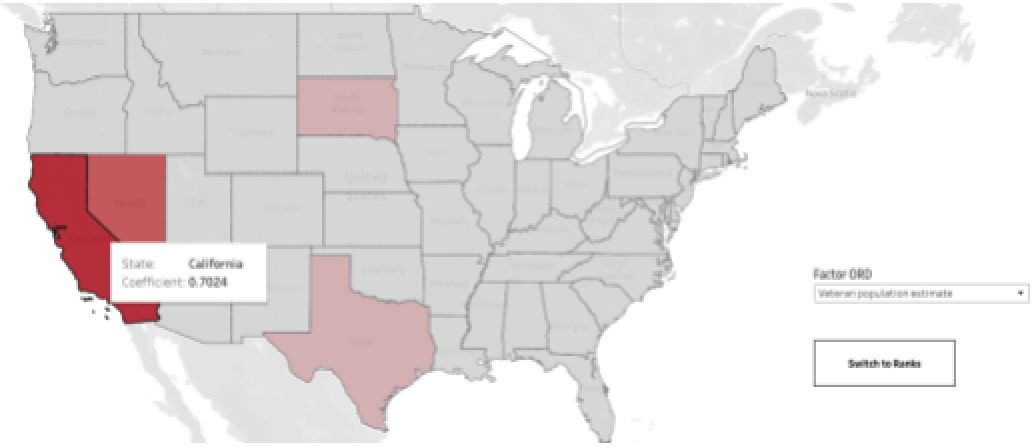


Factor ORD: Population density in hundreds

Factor PCPV: Outpatient visits per capita

Figure 1: The main screen of the interactive visual tool upon opening the application.

Once a factor is selected, the legend and colour-coding of each state will change to demonstrate which states have an opioid harm rate that is more impacted by that factor, based on the coefficient of the corresponding factor produced by the analysis. States with a greater positive coefficient are colored with a more saturated red colour, and states with a greater negative coefficient are colored with a more saturated blue colour. States which exclude the factor as part of their feature selection process are coloured grey. Hovering the mouse cursor over any given state produces a tooltip that names the state and specifies the coefficient corresponding to the selected factor as shown in Figure 2.



State: California
Coefficient: 0.7024

Factor ORD: Veteran population estimate

Figure 2: The tooltip function when hovering the mouse over a state.

Included in each map is a button to switch the map to display the colour-coded ranks of factors in each state. The rank of the factor for a state is the placement of that factor's coefficient in an ordered descending list of every coefficient in the equation. This feature allows the visualisation of groups of states holding a given factor in the same place of importance in determining opioid harm rate. This feature displays a similar colour-coded map and tooltip as displayed above, but displays the rank instead of the actual coefficient value.

4. EXPERIMENTS/ EVALUATION

A subset of the dataset was selected to validate the workflow and determine the demographic features that are most impactful for the counties within a state. To achieve this, an evaluation was performed on counties in West Virginia during the year 2013. West Virginia was selected due to its disproportionately high rates of opioid abuse compared to the rest of the country [Brownstein et al. 2010].

The analysis was performed to predict PCPV and ORD using combined county-level demographic data with each year (2006 - 2013). Given the significant difference in range between the features, standardisation was performed such that the transformed feature had a mean of 0 and a standard deviation of 1.

LASSO regression was used as a regularisation technique for variable selection and predictability of the model since the number of features far exceeded the number of observations. The LASSO regression included 5-fold cross-validation in order to determine the optimum value of regularisation which minimises squared error. The LASSO regression analysis determined 16 of the 156 were relevant for predicting PCPV and 14 for predicting ORD. Both models achieved an R2 well over 0.8 on both the test and training set, as shown in the figures below, indicating good predictive power. Shown is the PCPV model for reference.

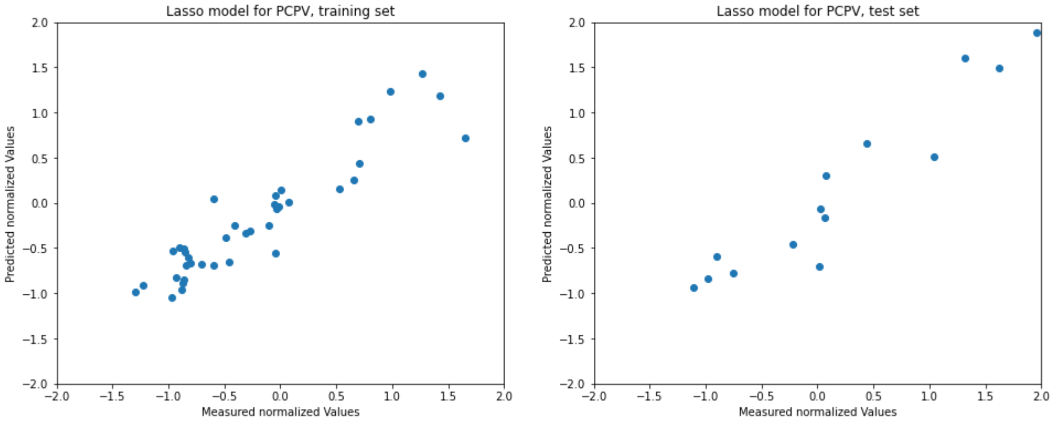
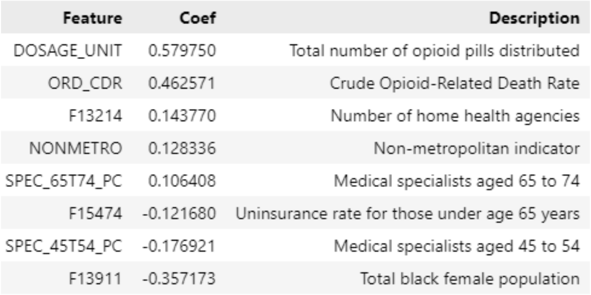


Figure 3. Predictive performance of the linear model with LASSO regularisation for predicting PCPV in West Virginia, 2013

The selected features with a coefficient magnitude greater than 0.1 are shown below for predicting PCPV:



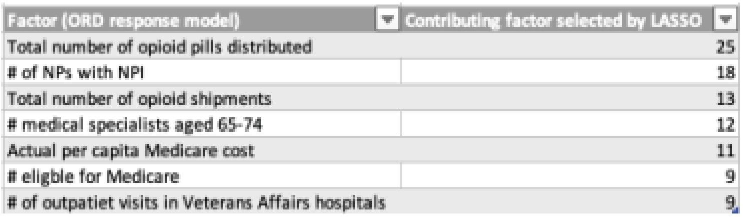
Feature	Coef	Description
DOSAGE_UNIT	0.579750	Total number of opioid pills distributed
ORD_CDR	0.462571	Crude Opioid-Related Death Rate
F13214	0.143770	Number of home health agencies
NONMETRO	0.128336	Non-metropolitan indicator
SPEC_65T74_PC	0.106408	Medical specialists aged 65 to 74
F15474	-0.121680	Uninsurance rate for those under age 65 years
SPEC_45T54_PC	-0.176921	Medical specialists aged 45 to 54
F13911	-0.357173	Total black female population

Figure 4. Significant features for predicting PCPV in West Virginia, 2013

The workflow described above provides accurate predictions which helped identify that the method with LASSO regression was adequate. ElasticNet was also evaluated as an alternative method but we found that this resulted in eliminating fewer features. Due to the L2-regularization term present in ElasticNet, the coefficient sizes were also shrunk closer together, resulting in underestimating the predictive ability of features. Based on this analysis, LASSO regression was found as a more adequate means to achieve our objectives.

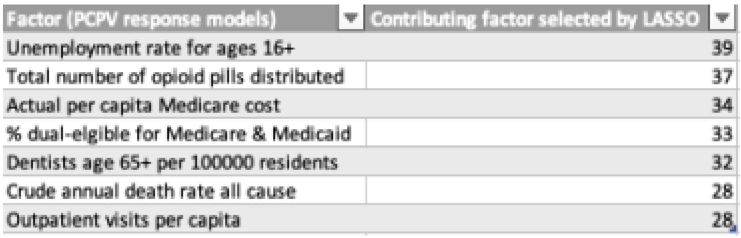
To evaluate our full model's predictive ability, the coefficient of determination (R2) was also calculated for our 51 models for each state for both response variables, ORD and PCPV. The explained variance from the LASSO selected variables for predicting ORD ranged from 0.61-0.99, where 50 models were above 0.80. The LASSO regression models for PCPV performed worse with an R2 value ranging from 0.53-0.98, where only 21 models were above 0.80.

The most common factors that were selected by LASSO regression as contributing factors to ORD and PCPV models are shown below in Figures 5 and 6. This highlights the most common issues that contribute to opioid deaths and opioid prescription numbers across states.



Factor (ORD response model)	Contributing factor selected by LASSO
Total number of opioid pills distributed	25
# of NPs with NPI	18
Total number of opioid shipments	13
# medical specialists aged 65-74	12
Actual per capita Medicare cost	11
# eligble for Medicare	9
# of outpatiet visits in Veterans Affairs hospitals	9

Figure 5: The most common predictors chosen by LASSO to predict ORD.



Factor (PCPV response models)	Contributing factor selected by LASSO
Unemployment rate for ages 16+	39
Total number of opioid pills distributed	37
Actual per capita Medicare cost	34
% dual-elgible for Medicare & Medicaid	33
Dentists age 65+ per 100000 residents	32
Crude annual death rate all cause	28
Outpatient visits per capita	28

Figure 6: The most common predictors chosen by LASSO to predict PCPV.

Whilst the results of the tool is mainly observational, the results from this analysis align with prior studies linking opioid availability and overdose [Ruhm, 2018]. Our results show that the feature which appears most commonly as a significant predictor of county ORDs is "*Total Number of opioid pills distributed*", with "*Total number of opioid shipments*" in third. This phenomenon is also highlighted in the interactive visual tool, shown in Figure 7, where the '*Total number of opioid pills distributed*' is marked as significant across many states in the US. The tooltip shows the magnitude of the coefficient and is coloured accordingly, showing a strong positive correlation between availability and ORDs.

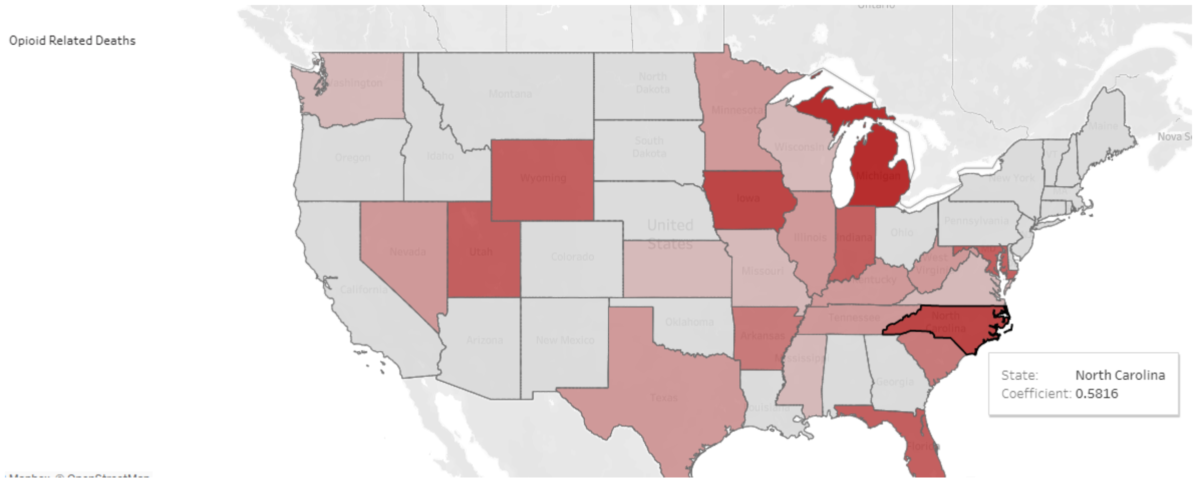


Figure 7. Coefficient magnitude for 'Total number of opioid pills distributed' as a predictive feature for ORDs

5. CONCLUSION/ DISCUSSION

Our analysis has helped identify factors that may contribute to the increasing rate of opioid overdose deaths and opioid prescriptions observed within the United States. Our goal was to identify the potential factors contributing to these geographic hot spots and to provide an interactive visual tool to allow public health representatives to implement strategies efficiently. This analysis is a beneficial first step to identifying a potential contributing factor to an area's increased opioid use.

The work outlined in this project has a number of limitations which should also be considered. The results from this work highlight correlation and may not be causal. The dataset used also suppressed death data for counties with fewer than 10 ORDs and imputation was used to overcome these missing values. The dataset used is also from 2006 - 2013 which may not still be reflective of the current climate of opioid abuse in the United States. Further work could be done on taking the methodology outlined in this project with more recent data once it becomes available. Finally, some predictive features may have been arbitrarily removed from each linear model as a result of using LASSO regression which is a known limitation of LASSO and many other feature selection techniques; this means that predictive features that are multicollinear with significant features could have been arbitrarily not selected, and therefore a pattern of what would have been a significant feature across multiple states may go unnoticed because other significant features in each of those states were multicollinear with the feature in question.

It is recommended to review the plethora of additional research identifying common contributing factors before allocating resources to one identified in this analysis to ensure they align. Incorrectly allocating resources would be a waste of time and money. The tool is best used as a starting point with the potential payout being extremely beneficial in the decline of opioid abuse.

All team members have contributed a similar amount of effort to the successful delivery of this project through completing the literature review, finding data sources and wrangling the data, running and interpreting the analysis, creating the visualisations and documenting the results.

6. REFERENCES

1. Basak, A., Cadena, J., Marathe, A., & Vullikanti, A. (2019). Detection of Spatiotemporal Prescription Opioid Hot Spots With Network Scan Statistics: Multistate Analysis. *JMIR Public Health Surveill*, *5*(2). 10.2196/12110
2. Brownstein, J., Green, T., Cassidy, T., & Butler, S. (n.d.). Geographic information systems and pharmacoepidemiology: using spatial cluster detection to monitor local patterns of prescription opioid abuse. *Pharmacoepidemiol Drug Saf.*, *19*(6), 627-37. 10.1002/pds.1939
3. Cerda, M., Gaidus, A., Keys, K. M., Ponicki, W., Martins, S., Galea, S., & Gruenewald, P. (2017). Prescription opioid poisoning across urban and rural areas: identifying vulnerable groups and geographic areas. *Addiction*, *112*(1), 103-112. https://doi.org/10.1111/add.13543
4. Cifuentes, M., Barbara S Webster, Verma, S., & Pransky, G. (2009). Geographic variation in opioid prescribing for acute, work-related, low back pain and associated factors: A multilevel analysis. *American Journal of Industrial Medicine*, *52*(2), 162-171. https://doi.org/10.1002/ajim.20655
5. Compton, W. M., Valentino, R. J., & DuPont, R. L. (2021). Polysubstance use in the U.S. opioid crisis. *Molecular Psychiatry*, *26*(1), 41-50. https://doi.org/10.1038/s41380-020-00949-
6. Duan, W. R., & Hand, D. J. (2021). Association between opioid overdose death rates and educational attainment – United States, 2010-2019. *Preventive Medicine*, *153*. https://doi.org/10.1016/j.ypmed.2021.106785.
7. Griffith, K.N., Y Feyman, SG Auty, EL Crable, TW Levengood. (2021). County-level data on U.S. opioid distributions, demographics, healthcare supply, and healthcare access. Data in Brief 35: e106779. https://doi.org/10.1016/j.dib.2021.106779
8. Hasan, M. M., Young, G. J., Patel, M. R., Modestino, A., Sanchez, L. D., & Noor-E-Alam, M. (2021). A machine learning framework to predict the risk of opioid use disorder. *Machine Learning with Applications*, *6*. 100144
9. Hu, X., Gallagher, M., Loveday, W., Connor, J.P., & Wiles, J. (2015). Detecting anomalies in controlled drug prescription data using probabilistic models. *Australasian Conference on Artificial Life and Computational Intelligence*, 337-349. https://doi.org/10.1007/978-3-319-14803-8_26
10. Maclean, J., Mallatt, J., Ruhm, C. J., & Simon, K. (2020). Economic Studies on the Opioid Crisis: A Review. *NBER Working Paper Series*. https://doi.org/10.3386/w28067
11. McDonald, D. C., Carlson, K., & Izrael, D. (2012). Geographic Variation in Opioid Prescribing in the U.S. *Journal of Pain Official Journal of the American Pain Society*, *13*(10), 988–996. https://doi.org/10.1016/j.jpain.2012.07.007
12. McDonald, D.C., & Carlson, K.E. (2014). The ecology of prescription opioid abuse in the USA: geographic variation in patients' use of multiple prescribers ("doctor shopping"). *Pharmacoepidemiology and Drug Safety*, *23*(12), 1258–1267. https://doi.org/10.1002/pds.3690
13. *Overdose Death Rates*. (2022, January 20). National Institute on Drug Abuse. Retrieved October 12, 2022, from https://nida.nih.gov/research-topics/trends-statistics/overdose-death-rates

14. Ruhm, C.J., (2018). Death and Despair or Drug Problems? *National Bureau of Economic Research,* accessed November 5, 2022 from https://www.nber.org/papers/w24188
15. Sun, F. (2022). Rurality and opioid prescribing rates in U.S. counties from 2006 to 2018: A spatiotemporal investigation. *Social Science & Medicine*, *296*(1), 114788–114788. https://doi.org/10.1016/j.socscimed.2022.114788
16. Unick, G. J., & Ciccarone, D. (2017). US regional and demographic differences in prescription opioid and heroin-related overdose hospitalizations. *International Journal of Drug Policy*, *46*, 112-119. https://doi.org/10.1016/j.drugpo.2017.06.003
17. Unick, G. J., Rosenblum, D., Mars, S., & Ciccarone, D. (2013). Intertwined epidemics: national demographic trends in hospitalizations for heroin- and opioid-related overdoses, 1993-2009. *PloS one*, *8*(2). https://doi.org/10.1371/journal.pone.0054496
18. *U.S. Overdose Deaths In 2021 Increased Half as Much as in 2020 - But Are Still Up 15%*. (2022, May 11). Centers for Disease Control and Prevention. https://www.cdc.gov/nchs/pressroom/nchs_press_releases/2022/202205.htm
19. Whelan, E., & Asbridge, M. (2013). The OxyContin crisis: Problematisation and responsibilisation strategies in addiction, pain, and general medicine journals. *International Journal of Drug Policy*, *24*(5), 402-411. https://doi.org/10.1016/j.drugpo.2013.01.007