Contents

Modeling Approach Determination and Justification

Mixed effects negative binomial regression was determined to be the best approach for undertaking this study because county-level overdose death rates are "zero-heavy" in nature. Given that a 1) majority of observations fell close to zero, 2) the potential values ranged into the low 100s, and 3) the assumptions of Poisson distribution were not met (the mean and variance not being equal) it was determined that a negative binomial distribution would best fit the outcome. (We note that although the negative binomial generally performs better than zero-inflated Poisson in the context of over dispersed data, we did attempt to implement zero-inflated Poisson, but, due to R implementation and increase in model complexity, this option was not practical due to concerns of run time and consistent model convergence). The random intercept for county allowed the model to leverage the longitudinal nature of the data to capture county-specific overdose death rate trends by accounting for variability not explained by the fixed effects. Each county will be at a different epidemic stage at a given point in time and the random intercept for counties allows to account for different "starting points" in the relationship between overdose deaths and predictors. While it could be expected that counties in the same state are more likely to experience similar epidemics (i.e. have similar intercepts), including a random intercept for state majorly affected the number of covariates that the model could take and did not improve model performance. We therefore did not include it. In addition, neighboring effects were accounted for by incorporating a gravity variable accounting for distance to counties weighted by their respective overdose death rates. The random slope for year allowed the model to capture varying rates of overdose deaths change over time. As the epidemic evolves, and a higher proportion of the population is affected or it reaches saturation, the effect of predictors on overdose death rates might change. The random slope for year allows the model to account for these likely changes through time. An offset term was incorporated into the model to account for the log-size of the population susceptible (i.e. carrying capacity) to fatal overdose. We describe this further in the next section.

Calculating the Carrying Capacity

In our negative binomial model, we include an offset term which "informs" the model of how many people are in a given county. In a practical sense, we may understand that a county with a smaller population will, on average, have a smaller total overdose count than a larger county, all other things held equal. The offset term provided to the model allows the model to adjust its predictions based upon the log of the population size.

We hypothesize that not all people in a county are at risk of fatal overdose each year, and that it would be much more effective to use the "susceptible" population as the offset term – where "susceptible" is defined as at-risk of overdosing. Two counties may have identical populations but there may be factors which influence differences in the total "susceptible" population size. Thus, it would improve model performance if we could effectively identify the *carrying capacity* (i.e. "susceptible" population) of each county-year observation.

Of importance, we hypothesized that as a given county experiences more overdose deaths, its carrying capacity will shrink. Our first goal in the implementation of carrying capacity was to better capture this shrinking. We further hypothesized that over time this carrying capacity would replenish (i.e., an overdose outbreak in 2010 would be unlikely to diminish carrying capacity in 2018). To account for this, we decided to make the carrying capacity 5% of the 2010 population minus the total number of overdoses the prior 3 years (or 1 year for 2011 and 2 years for 2012).For example in 2012 the carrying capacity was equal to 5% of the 2010 population minus the number of overdose deaths that occurred in 2010 and 2011. In 2017, it was equal to 5% of the 2010 population minus the number of overdose deaths that happened in 2014, 2015 and 2016. We chose 5% as the Substance Use and Mental Health Services Administration (SAMHSA) estimated that 3.6% of adults used illicit drugs in the prior month (excluding cannabis) in 2010.[1] We rounded this value up to 5% to capture that over the course of a year 3.6% of the population is an underestimate. We then subtracted the prior 3 years of overdose to reflect that an outbreak would, in the short-term, diminish the overall "susceptible population". In addition, we bounded the carrying capacity such that its minimum possible value was 50 to ensure that all counties had some risk of experiencing overdose deaths.

First, by setting the carrying capacity to 5% of the population, we are maintaining population-proportional estimates of the "susceptible" population. Second, by minimizing the carrying capacity to 5% of the population, the impact of overdose deaths can more readily be captured. For example, if a county with 1,000,000 people had 5,000 overdoses the prior three years, those 5,000 deaths would represent only 0.5% of the population and thus would minimally impact the offset term. By setting the carrying capacity to 5% (in this case, 50,000), we now see that the number of overdose deaths more substantially alters the carrying capacity (in this case, 5,000 deaths then reduce the carrying capacity by 10% from 50,000 to 45,000). Therefore, our approach maintains the information of the overall population size while also giving the model improved capability to capture the impact of overdose deaths on the "susceptible" population. Sumetsky et al., in a similar study, provided a more computationally intensive approach for identifying county-level carrying capacity.[2] We opted for a simpler approach to ensure the model runs smoothly (given it is applied to over 3000 counties with random intercepts and slopes) and to prevent the introduction of

unintended uncertainty, but we do advocate that improving carrying capacity calculation holds the potential to greatly improve the performance of both statistical and mathematical model of overdose. Of note, the choice of socio-demographical indicators or characteristics used to identifying county-level carrying capacity must carefully consider the use of immutable characteristics such as race. For example, the Sumetsky et al. paper does, at least partially, base carrying capacity on the proportion of white people in the county. Given that we hypothesize that the relationship between race and overdose is confounded and mediated by structural racism, we feel that such a design choice will result in a model which results in biased predictions that imply increased risk (and therefore need for resources to prevent them) in whiter counties (we provide further discussion on this issue in the section below). This, though, is not a criticism of the statistical methods employed in that work – but it is intended to highlight the complexity and ramifications of biased estimates of carrying capacity, and how, as a result, they may result in predictive technologies which replicate the dynamics of structural racism.

Variables Considered and Selected for Analysis

We identified four main categories of information we wished to capture, namely: **socio-economic vulnerability** (community susceptibility to drug use disorders), **healthcare access** (including to treatment of opioid use disorders), **drug markets**, to capture the drug's likely geographic presence and **geospatial effects**, to capture likely geographic spread. Variables were chosen a priori based on literature addressing overdose rates in the US.[3–5] Because we were seeking to identify outbreaks, we endeavored to include time-variant predictors (such as fentanyl seizure data) as they should better capture annual changes that may influence the likelihood of an outbreak occurring. However, we also incorporated time-invariant[1] predictors that have been consistently found to be associated with overdose death rates (such as urbanicity).

Healthcare access: In order to assess healthcare access, we included the number of active buprenorphine waivered physicians in each county. Evidence has indicated that access to opioid assisted treatment modalities (methadone and buprenorphine) is associated with a decreased risk of fatal overdose.[6] Over the studied period, providing opioid agonist treatment could only be done by practitioners approved by SAMHSA, which shares information on the number of new waivers provided each year and the number of waivers which lapse each year. From this, we calculated the number of active buprenorphine providers in each county by year. We also included the presence of an urgent care facility as a proxy for access to emergency healthcare – the Director of National Institute on Drug Abuse, Dr. Nora Volkow, has discussed the important role that urgent care facilities play in diminishing the fatal risk of overdose.[7] We would have ideally used IQVA data on naloxone pharmacy dispensing, as used by Guy et al.,[8,9] but these are not publicly available and the costs are relatively high. Similarly, including data on responses to overdoses by emergency medical services (EMS) from NEMSIS would have provided valuable information.[10] Unfortunately, due to confidentiality requirements, these are only shared publicly at national, Census Region and Census Division level (nine in total), which did not provide enough granularity for our analysis.

Socio-economic factors: We included a range of key socio-economic variables that have been consistently identified in other studies as being associated with overdose mortality, including the poverty rate,[11] household income,[11] unemployment rate,[11] and education level[11] (operationalized through high school graduation rate). Housing expenditure has also been found to be associated with overdose mortality and we incorporated this through including both the proportion of homeowner and renter households that spend over 35% of their income on mortgage or rent respectively. Finally, in order to capture drastic shifts in job markets due to factory closures or industry changes, which have also been found to be associated with overdose mortality,[12] we created two variables using data from the county business patterns (CBP) database. To calculate the "change in employment capacity" added the number of companies across all industries weighted by their personnel size category each year and subtracted past year from current year (yielding positive or negative values). A similar process was followed to estimate changes in payroll in a given year, but using payroll category as opposed to personnel size category.

We explicitly chose not to include identity-based demographic indicators, such as race, given the concern that such application may inappropriately diminish predicted risk in counties based on immutable characteristics, as discussed by Robinson et al.[13] Robinson et al. argue that while there is often a detectable relationship between race and a given health outcome, that machine learning researchers must ask if the "true" relationship is between race and the outcome or if the effect of race is mediated or confounded by structural racism. If it is structural racism which explains the relationship between race and the health outcome of interest, then including race as a simple covariate (which inherently hypothesizes a direct relationship between race and the outcome) may result in biased predictions which may replicate the dynamics of structural racism. It is thus necessary to ensure that the inclusion of race as a covariate will not replicate the dynamics of structural racism (or other similar immutable identity-based demographic indicators).

This risk of replicating structural racism cannot be understated, especially given that the opioid crisis has been characterized as primarily afflicting white people in the US. A model which accounts for race would likely output results suggesting that counties with greater proportion of white individuals are at higher risk of overdose (which may even improve model performance based on our provided indicators). While such a finding is informative in the context of an explanatory model, it fails to capture the underlying reasons for this concentration of risk. In the case of race, throughout the early stages of the opioid crisis, physicians were more likely to prescribe pain medications to white individuals than those of other races, especially Black individuals.[14] This over-prescribing is argued to be the basis of the concentration of opioid-related overdose deaths in white populations. A model which accounts for race as a fixed effect would fail to capture the full effect of changes in opioid prescribing patterns and would predict that counties with a greater proportion of white people are in greater need of resources and interventions to respond to future overdose death. Thus, a model which includes race may replicate the social dynamics which already result in inequitable distribution of health resources.

We do not hypothesize that a relationship between race and overdose exists which is not either confounded or mediated by structural racism. As such, to protect against the bias of concentrating risk within counties based on race, we do not include it as a fixed effect in our model.

Drug markets: Capturing information about drug markets is important because there cannot be an overdose outbreak in the absence of the drug. To assess the presence of opioids and fentanyl, two variables were included: the opioid prescription rate, using data from nearly 50,000 non-hospital-based pharmacies; the state-level count of seized drugs that were found to have fentanyl in them and These two variables captured both the presence of prescription opioids (which have driven up the number of people dependent on opioids and, thus, at higher risk of transitioning to heroin and fentanyl) as well as the presence of fentanyl at the state-level. Ideally, we would utilize IQVA data on opioid prescription pharmacy dispensing that includes detail on morphine milligram equivalents (MME)[8,9] but as mentioned, these are not publicly available. We would ideally also utilize county-level fentanyl seizure data, but these are not available. We also included the jail population size in each county, as this is partly driven by drug-related crimes, and high incarceration rates (among other punitive criminal justice practices) have been shown to be positively associated with overdose deaths.[11,15]

Geospatial factors: Finally, to capture likely geographic spread of fentanyl, we generated a gravity variable to represent distance to neighboring counties experiencing overdose outbreaks by measuring distance to contiguous counties weighted by their overdose death rate. In addition, we used a six-category variable ranking county-level urbanicity (as done by Van Handel et al.[3]): with a value of one indicating the most urban and a value of six indicating most rural.

Assessing Predictor Collinearity

We sought to ensure that the model was not poorly influenced by predictor collinearity. To assess collinearity of variables we fit a linear model with subsequent year overdose death rate as the outcome and each predictor under consideration as the set of predictors. We included data points for all years under study (i.e. predictors from 2010 to 2017 paired with outcomes from 2011 to 2018). We then calculated the variance inflation factor (VIF) for each variable (presented in **Table S1**). A rule of thumb is that if the VIF for a given variable is 10 or greater, then that variable should be removed and collinearity should be re-assessed.[16] As can be seen in **Table S1**, no variables approached this threshold and, thus, all were included in the final model.

Table S1. Variance inflation factor (VIF) for each predictor. VIF greater than 10 indicate that due to multicollinearity the predictor should be removed from the model. We kept all predictors in the model.

Variable	**VIF**
Buprenorphine waivered physicians	2.25
Urgent Care Presence	1.67
Opioid Prescription Rate	1.36
Log Fentanyl Seizure Data	1.54
Log Jail Population Size	2.61
High School Graduation Rate	2.21
Poverty Rate	3.80
Unemployment Rate	1.70
Employee capacity Difference	1.62
Payroll Difference	2.47

Log Median Household Income	4.25
Proportion of Homeowner Households That Spend At Least 35% of Income on Mortgage	1.35
Proportion of Renter Household That Spend At Least 35% of Income on Rent	1.69
Log Overdose Gravity	1.56
Urbanicity	2.30

Post-Hoc Analysis Methodology to Identify Fixed Effects with Greatest Predictive Utility

As is noted in the primary text, it is of interest to understand the contribution of fixed effects to the predictive accuracy of the model. When making predictions, it is also uncertain what the best set of fixed effects will be given that the model cannot be evaluated until after the predicted events occur. We employ a bootstrapped forward variable selection strategy similar to that described by Beyene et al to identify the fixed effects with the greatest predictive utility.[17] For this post-hoc analysis, we focus only on predicting overdose death rate for the year 2018 and the metric we are seeking to optimize is the proportion of counties correctly predicted in the top decile.

The procedure for a single bootstrap iteration is as follows. First, we generate a bootstrap sample by sampling county observations with repetition (i.e., when we sample a single county, we include all observations for that county). Specifically, there were 3,106 counties in our sample. The bootstrap sample was taken by first randomly sampling counties, such that we end up with 3,106 counties (with some repeating and some not appearing at all). The bootstrap sample dataset was then created by combining all observations for the selected sample counties (if a county was sampled more than once than its observations are added to the sample dataset multiple times). Second, we run the modeling procedure on the bootstrap sample as described above with the random effects and offset only model and calculate the proportion correctly predicted in the top decile. Next, we repeat the prediction process, adding each fixed effect individually to the model and calculate the proportion correctly predicted in the top decile. We then assess which fixed effect inclusion led to the best improvement in top decile identification. The optimal fixed effect is then added to the model and the process is repeated, adding the remaining variables to the updated model one at a time and choosing the next fixed effect to add by the same procedure. The process is terminated when either every variable has been added to the model or when no variable maintains or improves predictive performance (i.e. if the optimal fixed effects maintains performance, we add it and repeat the process). In order to ensure that variables that may provide similar contribution are equally considered, potential fixed effects are fed into the model in random order.

As is noted in the primary text, in total, we ran 100 bootstrap iterations. We then counted the number of times each variable was included in optimal model and we display, as the result, the proportion of the time each variable was included in the final model. Fixed effects that are chosen more frequently are considered to have greater predictive value than fixed effects chosen less frequently.

Model diagnostics

In order to better characterize the performance of the negative binomial modeling approach, we present the diagnostic plots in **Figure S1**. Scatterplots indicate that across all years that the predicted overdose death rates were, on average, less than their corresponding observed overdose death rate (as indicated by linear regression slope of less than 1). It appears that over time this slope approaches 1, providing initial indication that with more training data, the model's predictive performance improves. The histograms display the mean average error (MAE) for all counties with an observed overdose death rate rounded down to the nearest whole number (i.e. the bar corresponding to 5 on the x-axis is the MAE for all counties whose observed overdose death rate was greater than or equal to 5 and less than 6). Histograms indicate that across every year, the greater the observed overdose death rate, the greater the MAE – this is fairly intuitive, in that one would expect that model precision would be greatest for counties whose death rate is consistently close to 0, whereas predicting higher overdose death rates is subject to far greater variability. Further, wider errors for higher overdose death rates may have less impact on predictive utility – for example, if we predict a county to have 115 deaths per 100,000, but the rate ends up being 80 per 100,000, this error (35) is quite large, but, from a practical sense the model correctly predicted a very high death rate for this county. A converse example would be a county whose observed death rate was 12 deaths per 100,000. An error of just 10 (far smaller than 35) is consequential – predicting 2 deaths per 100,000 would wrongly indicate that this county is at minimal risk. Thus, the smaller errors for smaller observed overdose death rates indicates that the model is performing in a useful manner. The Bland-Altman plots in **Figure S2** confirm the findings of scatterplots and histograms. The rightward fanning of the plots indicates that the prediction error increases as the magnitude of the observed-predicted sum of an observation increases. Further, the downward "lean" of these plots (in all years except 2018) indicates that the model tended to under-predict overdose death rates for counties with higher observed overdose death rates.

Figure S1. Scatterplot (left) and histogram (right) model diagnostic plots for each year.

The scatterplots map the predicted overdose death rate for each year paired with its corresponding observed overdose death rate. The dotted blue line is a best-fit linear function – a slope less than 1 (as seen for all years) indicates that observed overdose death rates were, generally, greater than their corresponding predicted values. It appears that over time this slope approaches 1, indicating that the model's predictive capacity improves over time. Histograms display the average error for all counties with an observed overdose death rate (values were all rounded down to the nearest whole number). As is displayed in the histograms – the magnitude of errors corresponded directly with the observed overdose death rate, where the model was the least precise at predicting counties with the highest observed overdose death rates.

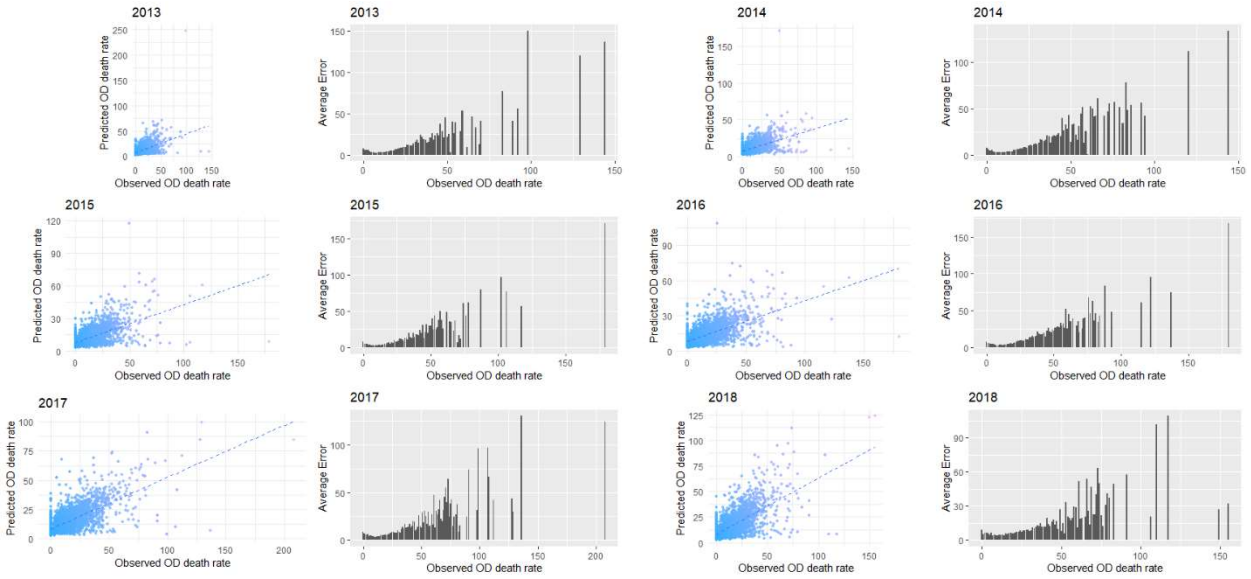
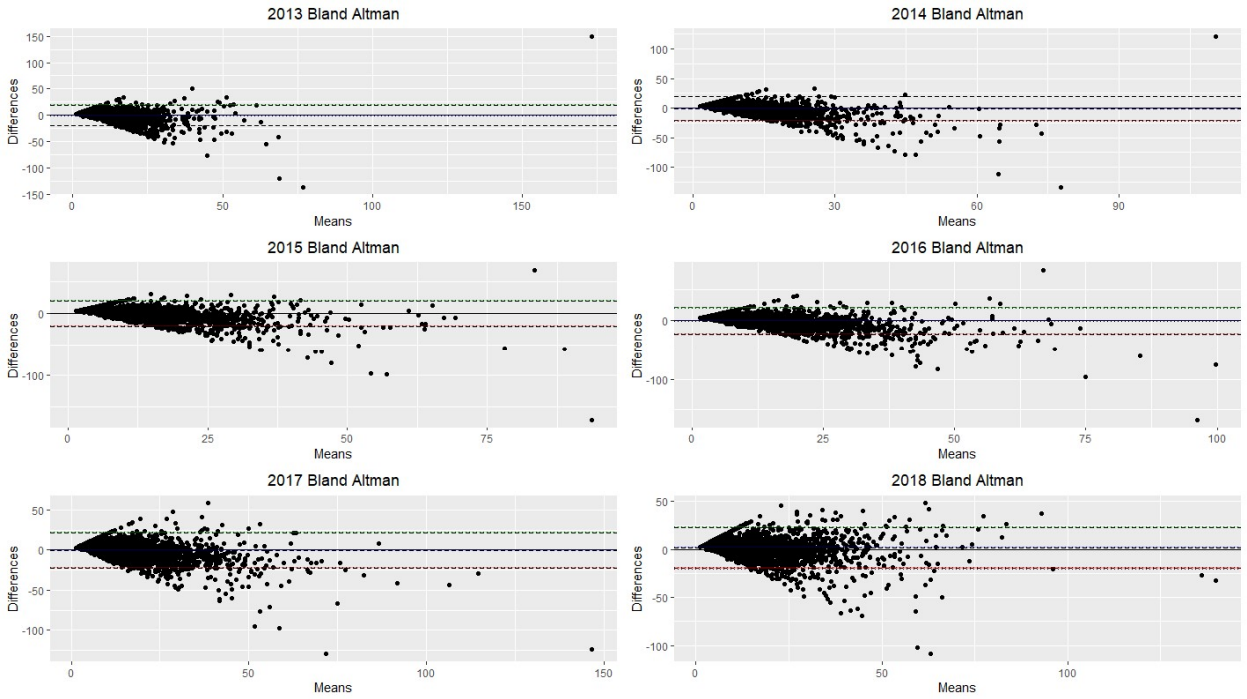


Figure S2. Bland Altman Plots. These plots offer a variation on the scatterplots and histograms above. The BA plot is a scatterplot mapping the difference between the observed and predicted observation for a given county (Y axis) and the mean of the two (X-axis). A point on the 0 line of the Y-axis indicates that the prediction was the same as the observed value for that given county. The further to the right on a plot, the greater the magnitude of the sum of the observed and predicted. Here we see for each year a fan shape formed by the scatterplot. This fan shape indicates that as the magnitude of the predicted-observed sum increases, so generally too does the error. This is consistent with the histograms above. Further, the fans appear to skew downward indicating that model had a slight tendency toward under-predicting the overdose death rate, especially when the magnitude of the observed was greater.



Post Hoc Sensitivity Analysis: Regional Analysis

It was of interest to investigate how the modelling approaches overall performance would be impacted if it was applied to smaller regions separately, as opposed to all counties within the contiguous US in a single model. Given the geographic heterogeneity of the fentanyl epidemic, in which the eastern portion of the country was impacted throughout the study time period to a much greater extent than the western portion of the country, it would seem reasonable to hypothesize that running our modeing and prediction procedure on each half of the US separately may improve overall predictive performance. As such, as a post hoc sensitivity analysis, we categorized counties into two groups: east of the Mississippi River and west of the Mississippi River. This natural, geological cut-off is also ideal in that there are a similar number of counties within both groups.

As such, we implemented our modeling approach in both the eastern and western US separately and then aggregated the predictions to compare the sensitivity analysis performance to the primary analysis. The sensitivity analysis was more precise five out of six years for both MAE (all years except 2013) and RMSE (all years except 2018), noting that MAE was never more than 0.12 better and RMSE was never more than 0.17 better than the primary analysis (see **Table S2**). Further, the Spearman's rho indicates near-identical (accounting for rounding) performance in terms of rank-ordering counties.

Table S2. Mean average error, root mean squared error and Spearman's ρ for the both the sensitivity analysis and the primary analysis for each year from 2013 to 2018.

	East West Sensitivity Analysis			Negative Binomial – With Fixed Effects		
	MAE	RMSE	Spearman's ρ	MAE	RMSE	Spearman's ρ

2013	6.55	9.96	0.57	6.58	10.04	0.57
2014	6.69	10.42	0.58	6.70	10.42	0.58
2015	6.73	10.33	0.62	6.74	10.34	0.62
2016	7.59	11.45	0.64	7.66	11.55	0.64
2017	7.40	11.05	0.67	7.52	11.22	0.67
2018	7.69	11.11	0.65	7.73	10.95	0.65

The sensitivity analysis identified more counties in the top decile in four out of six years (see **Table S3**). At best, the sensitivity analysis identified 6 (out of 310) additional counties. As well, across the final three years of the study period, the sensitivity analysis identified between 3 to 5 more counties newly entering the top decile than the primary analysis.

Table S3. Number of total and new counties in the top decile of overdose death rates correctly predicted by the sensitivity analysis and primary analysis each year. In the top portion of the table (**Top Decile)**, the number of counties in the top decile (n = 310) that were accurately predicted as such are presented for each year and each approach. In the bottom half, the number of counties that newly entered the top decile (i.e. were not in the top decile the year before) that were accurately predicted as such are presented for each year and each approach.

	East West Sensitivity Analysis	**Negative Binomial**
	Top Decile	**Top Decile**
2013	126/310 (40.6%)	129/310 (41.6%)
2014	147/310 (47.4%)	145/310 (46.8%)
2015	157/310 (50.6%)	158/310 (51.0%)
2016	160/310 (51.6%)	154/310 (49.7%)
2017	180/310 (58.1%)	176/310 (56.8%)
2018	174/310 (56.1%)	171/310 (55.2%)
	Newly in Top Decile	**Newly in Top Decile**
2014	46/175 (26.3%)	46/175 (26.3%)
2015	40/170 (23.5%)	40/170 (23.5%)
2016	42/165 (25.5%)	37/165 (22.4%)
2017	41/149 (27.5%)	38/149 (25.5%)
2018	36/149 (24.2%)	33/149 (22.1%)

Overall, the findings of our sensitivity analysis indicate that a more targeted stratification of the sample may lead to improvements in model performance, though we note that, based on the metrics presented, the predictions of the sensitivity analysis do not appear to be substantially different from the primary analysis.

References

1. Substance Abuse and Mental Health Services Administration. *Results from the 2010 National Survey on Drug Use and Health: Summary of National Findings*.; 2011. https://www.samhsa.gov/data/sites/default/files/NSDUHNationalFindingsResults2010-web/2k10ResultsRev/NSDUHresultsRev2010.pdf.

2. Sumetsky N, Mair C, Wheeler-Martin K, et al. Predicting the Future Course of Opioid Overdose Mortality. *Epidemiology*. 2020;Publish Ah. doi:10.1097/EDE.0000000000001264

3. Van Handel MM, Rose CE, Hallisey EJ, et al. County-Level Vulnerability Assessment for Rapid Dissemination of HIV or HCV Infections Among Persons Who Inject Drugs, United States. *JAIDS J Acquir Immune Defic Syndr*. 2016;73(3):323-331. doi:10.1097/QAI.0000000000001098

4. Rossen LM, Khan D, Warner M. Trends and Geographic Patterns in Drug-Poisoning Death Rates in the U.S., 1999–2009. *Am J Prev Med*. 2013;45(6):e19-e25. doi:10.1016/j.amepre.2013.07.012

5. Monnat SM, Peters DJ, Berg MT, Hochstetler A. Using Census Data to Understand County-Level Differences in Overall Drug Mortality and Opioid-Related Mortality by Opioid Type. *Am J Public Health*. 2019;109(8):1084-1091. doi:10.2105/AJPH.2019.305136

6. Larochelle MR, Bernson D, Land T, et al. Medication for Opioid Use Disorder After Nonfatal Opioid Overdose and Association With Mortality. *Ann Intern Med*. 2018;169(3):137. doi:10.7326/M17-3107

7. Volkow N. Emergency Departments Can Help Prevent Opioid Overdoses. NIDA Website. https://www.drugabuse.gov/about-nida/noras-blog/2019/08/emergency-departments-can-help-prevent-opioid-overdoses. Published 2019.

8. Schuchat A, Houry D, Guy Jr GP. New Data on Opioid Use and Prescribing in the United States. *JAMA*. 2017;318(5):425-426. doi:10.1001/jama.2017.8913

9. Guy GP, Haegerich TM, Evans ME, Losby JL, Young R, Jones CM. Vital Signs: Pharmacy-Based Naloxone Dispensing — United States, 2012–2018. *MMWR Morb Mortal Wkly Rep*. 2019;68(31):679-686. doi:10.15585/mmwr.mm6831e1

10. Friedman J, Beletsky L, Schriger DL. Overdose-Related Cardiac Arrests Observed by Emergency Medical Services During the US COVID-19 Epidemic. *JAMA Psychiatry*. December 2020. doi:10.1001/jamapsychiatry.2020.4218

11. Altekruse SF, Cosgrove CM, Altekruse WC, Jenkins RA, Blanco C. Socioeconomic risk factors for fatal opioid overdoses in the United States: Findings from the Mortality Disparities in American Communities Study (MDAC). Genberg BL, ed. *PLoS One*. 2020;15(1):e0227966. doi:10.1371/journal.pone.0227966

12. Venkataramani AS, Bair EF, O'Brien RL, Tsai AC. Association Between Automotive Assembly Plant Closures and Opioid Overdose Mortality in the United States. *JAMA Intern Med*. 2020;180(2):254. doi:10.1001/jamainternmed.2019.5686

13. Robinson WR, Renson A, Naimi AI. Teaching yourself about structural racism will improve your machine learning. *Biostatistics*. November 2019. doi:10.1093/biostatistics/kxz040

14. Om A. The opioid crisis in black and white: the role of race in our nation's recent drug epidemic. *J Public Health (Bangkok)*. 2018;40(4):e614-e615. doi:10.1093/pubmed/fdy103

15. Kajeepeta S, Rutherford CG, Keyes KM, El-Sayed AM, Prins SJ. County Jail Incarceration Rates and County Mortality Rates in the United States, 1987–2016. *Am J Public Health*. 2020;110(S1):S109-S115. doi:10.2105/AJPH.2019.305413

16. Miles J. Tolerance and Variance Inflation Factor. In: *Wiley StatsRef: Statistics Reference Online*. Chichester, UK: John Wiley & Sons, Ltd; 2014. doi:10.1002/9781118445112.stat06593

17. Beyene J, Atenafu EG, Hamid JS, To T, Sung L. Determining relative importance of variables in developing and validating predictive models. *BMC Med Res Methodol*. 2009;9(1):64. doi:10.1186/1471-2288-9-64

R Code

```
##
## Load the Needed Libraries
##

library(lme4)
library(MASS)

##
## Load the Data
## The data is not publicly available because the outcome was generated using
## Restricted data from the CDC. To learn more about the outcome data
## and what this restriction entails, please see the accompanying manuscript
## or head over to www.wonder.cdc.gov where the unrestricted mortality data is available
##

data <- read.csv("Analysis_Data.csv", header = T)


##
## Prior to running analysis we calculated the county-level carrying capacity
## See the manuscript for specific details
##
data$carrying_capacity <- NA

for(FIPS in unique(data$FIPS)){

  ## Carrying Capacity is initially set to 5% of the county population in 2010
  baseline <- data$population[data$FIPS == FIPS & data$year == 2010]*0.05

  for(year in 2010:2017){

    ## next we substract the number of overdose deaths from the prior three years
    ## noting that for 2011 we only subtract the prior year
    ## and that for 2012 only the prior two years

    if(year == 2010){

      data$carrying_capacity[data$FIPS == FIPS & data$year == year] <- baseline

    }else if(year == 2011){
      data$carrying_capacity[data$FIPS == FIPS & data$year == year] <- baseline -
data$overdose_deaths[data$FIPS == FIPS & data$year == year - 1]
    }else if(year == 2012){
```

```
    data$carrying_capacity[data$FIPS == FIPS & data$year == year] <- baseline - data$overdose_deaths[data$FIPS == FIPS & data$year == year - 1] - data$overdose_deaths[data$FIPS == FIPS & data$year == year - 2]
   }else{
    data$carrying_capacity[data$FIPS == FIPS & data$year == year] <- baseline - data$overdose_deaths[data$FIPS == FIPS & data$year == year - 1] - data$overdose_deaths[data$FIPS == FIPS & data$year == year - 2] - data$overdose_deaths[data$FIPS == FIPS & data$year == year - 3]
   }

  }

}

## we then limit the carrying capacity such that its minimum value is 50
data$carrying_capacity[which(data$carrying_capacity < 50)] <- 50

##
## Next we define our regression equation
##

func <- next_year_overdose_deaths ~  (year|FIPS) + ## random effect for county (FIPS) with a random slope for year
  offset(log(carrying_capacity)) + ## offset term for the log of the carrying capacity
  log(NFLIS) + log(jail_pop) + employee_diff + payroll_diff +
  (log(overdose_gravity_add)) + (opioid_prescriptions_per_100) +
  (buprenorphine_provider_waivers) + urgent_care +
  proportion_high_school_or_greater +  proportion_poverty +
  unemployment_rate + median_household_income +
  proportion_homeowners_35perc_income + proportion_renters_35perc_income +
  urbanicity


## First we create the results table where analysis for this half of the country will be stored
## It contains the names of the counties and their FIPS codes
results <- data[data$year == 2010, c("FIPS","county_name")]

##
## We begin by predicting overdose death rates for the year 2013 (i.e. 201X)
## We provide detailed code for this year and note that the analysis for remaining years is identical
## As noted in the manuscript, in order to predict overdose deaths from the year 201X
## We take predictor data from the years 2010 - 201(X-2) (paired with outcome data from 2011 - 201(X-1))
## We train the model on this dataset
## Then we take predictors for the year 201(X-1) and feed them into the model to generate predictions for 201X
```

```
## These values are the total predicted number of overdose deaths so then we population adjust them to get
## our final predicted overdose death rates for each county
##

## First we need the data to train our model
## We select all data corresponding to years 2010 through 2011 (i.e. 2010 - 201(X-2))
train_data <- data[data$year >= 2010 & data$year < 2012 & !is.na(data$next_year_synthetic_opioid_death_rate),]

## Next we need our test/prediction data
## We select the data corresponding to year 2012 (i.e. 201(X-1))
test_data <- data[data$year == 2012,]

## Next we train the model on our data using the function specified
## We set nAGQ = 0 in order to simplify the optimization routine which dramatically improves runtime
lme_model <- glmer.nb(func, train_data, nAGQ = 0)


## After training the model we need to use the model in order to predict the death counts for 2013 (201X)
## We use the predict function, supplying the trained model, the test/prediction data
## Making sure to set type = "response" to get the approrpriate metric
## We store the predicted value in the test_data data frame

test_data$prediction <- predict(lme_model,newdata=test_data, type = "response")

## Finally we need to store the results
## The following code handles this

## First, we create the columns for the observed and predicted SynthOD deaths rates in 2013 (i.e. 201X)
results$observed_2013 <- NA
results$predicted_2013 <- NA

## Then we create a loop to go through every county in the results and we fill in the observed and predicted values
## Note that the model predicts the crude count of deaths so we need to adjust the value by the population size
## In order to get the death rates
for(i in 1:nrow(results)){

  ## Extract the county FIPS code -- this is used as a key to match up data appropriately
  county <- results$FIPS[i]
  ## Extract the population size of the county for the given year
```

```
  population <- test_data$population[test_data$FIPS == county]

  ## Get both the observed and predicted death counts for the year 2013 (i.e. 201X)
  observed <- test_data$next_year_overdose_deaths[test_data$FIPS == county]
  predicted <- test_data$prediction[test_data$FIPS == county]

  ## We then create the rate by dividing the deaths counts by the population size and then
  ## by multiplying this value by 100,000 (thus we have a death rate as X per 100,000)
  ## These values are stored in the appropriate location in the results table
  results$observed_2013[i] <- (observed/population)*100000
  results$predicted_2013[i] <- (predicted/population)*100000

}

##
## We then repeat this process for each year from 2014 to 2018
## For parsimony we do not include comments on the identical code
## 2014
##

train_data <- data[data$year >= 2010 & data$year < 2013 &
!is.na(data$next_year_synthetic_opioid_death_rate),]
test_data <- data[data$year == 2013,]

lme_model <- glmer.nb(func, train_data, nAGQ = 0)

test_data$prediction <- predict(lme_model,newdata=test_data, type = "response")

results$observed_2014 <- NA
results$predicted_2014 <- NA
for(i in 1:nrow(results)){
  county <- results$FIPS[i]
  population <- test_data$population[test_data$FIPS == county]

  observed <- test_data$next_year_overdose_deaths[test_data$FIPS == county]
  predicted <- test_data$prediction[test_data$FIPS == county]

  results$observed_2014[i] <- (observed/population)*100000
  results$predicted_2014[i] <- (predicted/population)*100000
}


##
## 2015
##
```

```
train_data <- data[data$year >= 2010 & data$year < 2014 & !is.na(data$next_year_synthetic_opioid_death_rate),]
test_data <- data[data$year == 2014,]

lme_model <- glmer.nb(func, train_data, nAGQ = 0)

test_data$prediction <- predict(lme_model,newdata=test_data, type = "response")

results$observed_2015 <- NA
results$predicted_2015 <- NA
for(i in 1:nrow(results)){

  county <- results$FIPS[i]
  population <- test_data$population[test_data$FIPS == county]

  observed <- test_data$next_year_overdose_deaths[test_data$FIPS == county]
  predicted <- test_data$prediction[test_data$FIPS == county]

  results$observed_2015[i] <- (observed/population)*100000
  results$predicted_2015[i] <- (predicted/population)*100000
}

##
## 2016
##

train_data <- data[data$year >= 2010 & data$year < 2015 & !is.na(data$next_year_synthetic_opioid_death_rate) ,]
test_data <- data[data$year == 2015 ,]

lme_model <- glmer.nb(func, train_data, nAGQ = 0)

test_data$prediction <- predict(lme_model,newdata=test_data, type = "response")

results$observed_2016 <- NA
results$predicted_2016 <- NA
for(i in 1:nrow(results)){

  county <- results$FIPS[i]
  population <- test_data$population[test_data$FIPS == county]


  observed <- test_data$next_year_overdose_deaths[test_data$FIPS == county]
  predicted <- test_data$prediction[test_data$FIPS == county]

  results$observed_2016[i] <- (observed/population)*100000
```

```
  results$predicted_2016[i] <- (predicted/population)*100000
}

##
## 2017
##

train_data <- data[data$year >= 2010 & data$year < 2016 &
!is.na(data$next_year_synthetic_opioid_death_rate),]
test_data <- data[data$year == 2016,]

lme_model <- glmer.nb(func, train_data, nAGQ = 0)

test_data$prediction <- predict(lme_model,newdata=test_data, type = "response")

results$observed_2017 <- NA
results$predicted_2017 <- NA

for(i in 1:nrow(results)){

  county <- results$FIPS[i]
  population <- test_data$population[test_data$FIPS == county]


  observed <- test_data$next_year_overdose_deaths[test_data$FIPS == county]
  predicted <- test_data$prediction[test_data$FIPS == county]

  results$observed_2017[i] <- (observed/population)*100000
  results$predicted_2017[i] <- (predicted/population)*100000

}

##
## 2018
##

train_data <- data[data$year >= 2010 & data$year < 2017 &
!is.na(data$next_year_synthetic_opioid_death_rate),]
test_data <- data[data$year == 2017,]

lme_model <- glmer.nb(func, train_data, nAGQ = 0)

test_data$prediction <- predict(lme_model,newdata=test_data, type = "response")

results$observed_2018 <- NA
results$predicted_2018 <- NA
```

```
for(i in 1:nrow(results)){

  county <- results$FIPS[i]
  population <- test_data$population[test_data$FIPS == county]


  observed <- test_data$next_year_overdose_deaths[test_data$FIPS == county]
  predicted <- test_data$prediction[test_data$FIPS == county]

  results$observed_2018[i] <- (observed/population)*100000
  results$predicted_2018[i] <- (predicted/population)*100000
}

##
## The prediction algorithm is now complete!
## So last step is to save the results
##

write.csv(results, "final_results.csv")
```