Contagion Effect of Deaths from Suicides and Drug Overdoses in the US

(Capstone Faculty-Sponsored Project (FS#5)-- Fall 2019)

Team

Cristina Martinez-Finlay <cm2592@columbia.edu> - cm2592
Emma Nelece Smithayer <es3583@columbia.edu> - es3583
Namson Ngo-Le <nn2446@columbia.edu> - nn2446
Vineet Anthony Aguiar <vaa2114@columbia.edu> - vaa2114

Project Mentors

Jeffrey L. Shaman <jls106@cumc.columbia.edu>
Sasikiran Kandula <sk3542@cumc.columbia.edu>

Abstract

This paper discusses using data on Emergency Medical Service (EMS) calls in the US to estimate national mortality rates due to drug overdoses or suicide. It tests the hypothesis that EMS events with certain ICD-9 or ICD-10 codes as the cause of injury, or events with a certain type of medication administered to the patient, have a strong correlation with the relevant deaths reported in CDC WONDER dataset [2] in the same time period. EMS data is available much more quickly than final mortality data, so it may allow researchers and public health officials to identify and act on trends more quickly.

Opioid-related EMS events and opioid-related deaths reported in CDC WONDER are strongly correlated, even when facetted by demographics such as gender, age, race, and urbanicity of the population. Similarly high correlation was observed across Census Regions, except in the West. To account for a major drop in the events reported when the NEMSIS database version was upgraded in 2017, the ratio of opioid-related events to the total EMS events was used in our analysis instead of raw counts.

Among the models explored, the simple linear regression model predicts opioid-related deaths best. This encourages the prospect of further research as better quality data becomes available over time.

Introduction

Although mortality data in the US is available through systems such as CDC's WONDER [2], due to established reporting protocols and data quality checks, the data becomes publicly available 6 to 18 months after the fact. While the need for robust processes and checks are indisputable, more timely, albeit less accurate data could aid public health action, especially considering that mortality rates due to drug poisoning and suicide in the United States have been increasing in the past years. As events resulting from these two causes will likely generate EMS calls, the NEMSIS [1] systems, which records these events and makes the data available in a shorter period of time (approximately less than a week after the events occurred) might serve as a secondary source to estimate and track trends in mortality from these two causes.

Related work & your contributions

There are numerous studies of suicide and opioid overdose rates over time, but by necessity most focus on data that was already several years old at the time of publication, which is what is motivating our use of the NEMSIS data. One exception is this study from Australia that found that ambulance calls for *opioid overdoses and opioid deaths were correlated*, and that the EMS data could be used as an early indicator of trends in opioid deaths [3]

This study analyzed all of the different ICD codes used for drug overdoses in the emergency department. This is less directly relevant to our project, but it illustrates how ICD codes are often used inconsistently, making precise analysis more difficult [4]

Discussion of the data set and exploratory analysis

Data Sources:

- The *Emergency Medical Services (NEMSIS*) [1] dataset from the National Highway Traffic Safety Administration, collected in response to 911 emergency calls. These data are available in near real-time.
- The *Centers for Disease Control's WONDER* [2] dataset for mortality data. This is generally made public 6-18 months after required protocol and quality checks. The last available release is for the calendar year 2017.
- Provisional drug overdose death counts [5] which are based on death records received and processed by the National Center for Health Statistics (NCHS) which are available for up to April 2019.

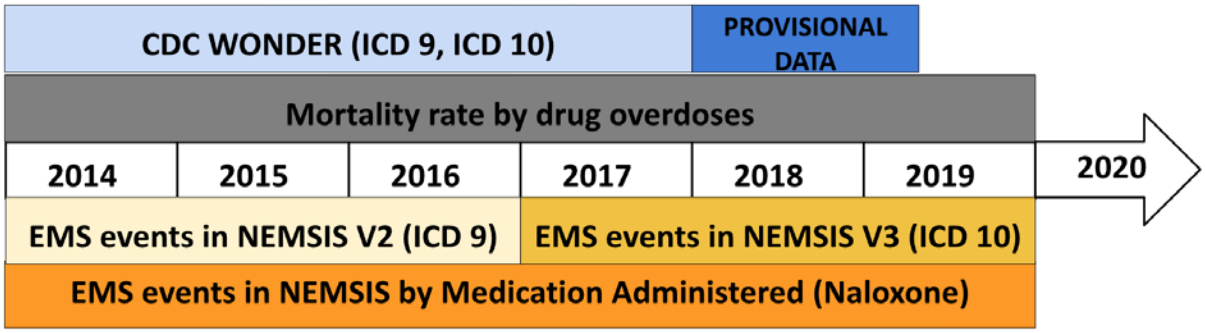


Fig 1. Diagram representing the data sources for different time periods

Identifying Opioid Overdoses and Suicides

All of our data sources include ICD codes, a system published by the World Health Organization for classifying medical symptoms, diagnosis, and causes. ICD is short for International Statistical Classification of Diseases and Related Health Problems. ICD-10 is the most recent version and contains far more codes than ICD-9. Most medical providers in the United States were required to switch to ICD-10 by October 2015. The NEMSIS v2 data uses ICD-9 codes, while NEMSIS v3 and the CDC Wonder data use ICD-10 codes. Even within the same ICD version, different codes may be used for the same medical condition, so choosing the codes to include is not an exact science. ICD-10 contains over 70,000 codes, many of them highly specific.

The NEMSIS datasets also have several different ICD codes fields. A single event may have a Primary Impression ICD, a Secondary Impression ICD, and an Injury ICD. The Primary Impression ICD field seemed the most comprehensive, so we focused on that.

To identify opioid overdoses in NEMSIS V3, we chose codes starting with F11 (“Opioid related disorders”) or T40 (“Poisoning by, adverse effect of and underdosing of narcotics and psychodysleptics”). Within each of those categories is a hierarchy of more specific codes (for example, F11.22: “Opioid dependence with intoxication”). However, it was more difficult to identify opioid overdoses and suicides in the NEMSIS V2 data. ICD-9 codes are less detailed than ICD-10, plus NEMSIS seems to only include a few dozen of the most general codes. For opioid overdoses, “977.90- poisoning/drug ingestion” is the most relevant code available, but that does not distinguish among types of drug overdoses or poisonings. Because the level of specificity was so different across the two versions, our plot of event counts per month showed a dramatic dropoff at the transition to NEMSIS V3.

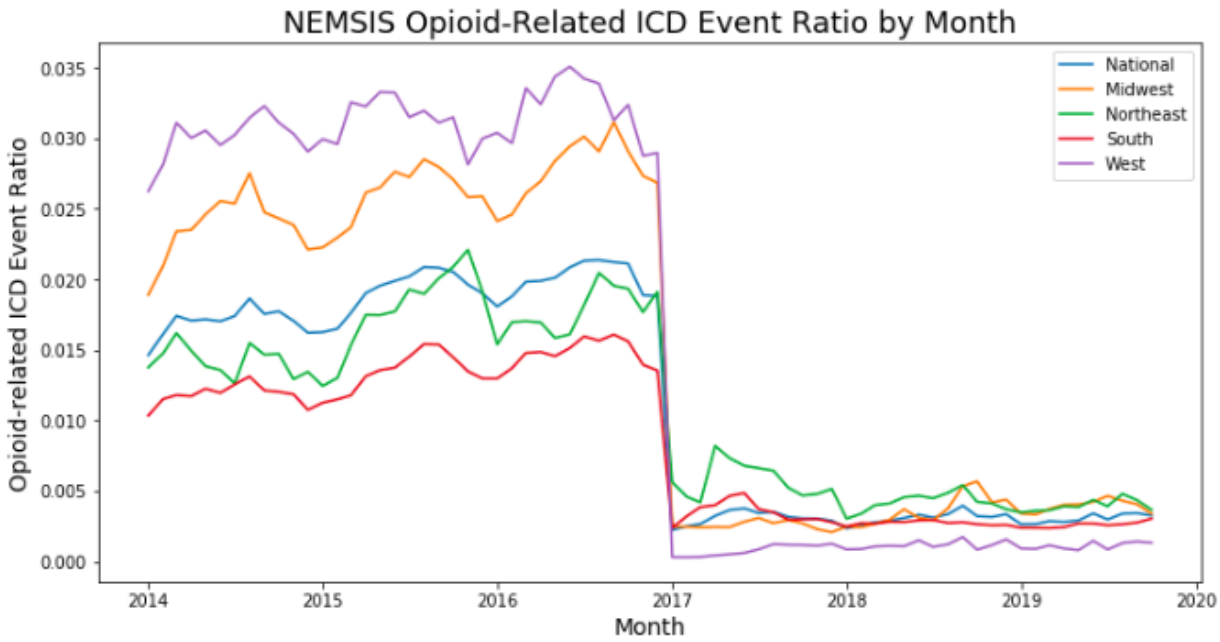


Fig 2: NEMSIS Opioid-Related ICD Event Ratio by month

We ultimately decided that the data was not comparable across the two versions. Instead, we decided to identify opioid-overdose events in both versions by filtering for events where “Medication Administered” included Naloxone (brand name: Narcan), a drug often used to reverse opioid overdoses. We also experimented with using the ICD10-filtered events in the V3 database, while dropping the V2 data, but found that the Naloxone event counts were more strongly correlated with mortality.

Originally, we also wanted to include suicides in our analysis. However, identifying those events in NEMSIS was even more difficult--no option allowed the identification of suicides specifically. “312.90- Behavioral/psychiatric disorder” seemed like the closest option, but we decided it was too broad to be useful.

Data Extraction, Cleaning and Augmentation

A significant amount of effort was dedicated to automating the extraction of data from the NEMSIS [1] systems. The user interface for interacting with the data required too many clicks to set up the right dimensions, the query result was returned in multiple pages that needed to be visited individually to be able to download the data and each download generated a new Excel file. All these manual interactions with the user interface had the potential for error resulting in either duplicate or missing data. Therefore we decided to automate the process using Selenium, a free (open source) automated tool for interacting with websites. We developed a Python script using the Selenium module and we were able to automate pasing MDX code to extract the data, instead of manually selecting dimensions and metrics, responding to popup messaging, moving

through all the pages and requesting the data download. We created a tool that was flexible and allowed us to get data at a very granular level avoiding unnecessary human intervention.

When combining data for incidents per month involving the administration of Naloxone reported on NEMSIS V2 and NEMSIS V3, we observed a significant drop in total reported incidents at the system conversion point and a steady, yet not complete, recovery of the total count more in line with what we expected to observe in Fig 3. We speculate that it could be due to a slow adoption of the new system nationally or regionally.

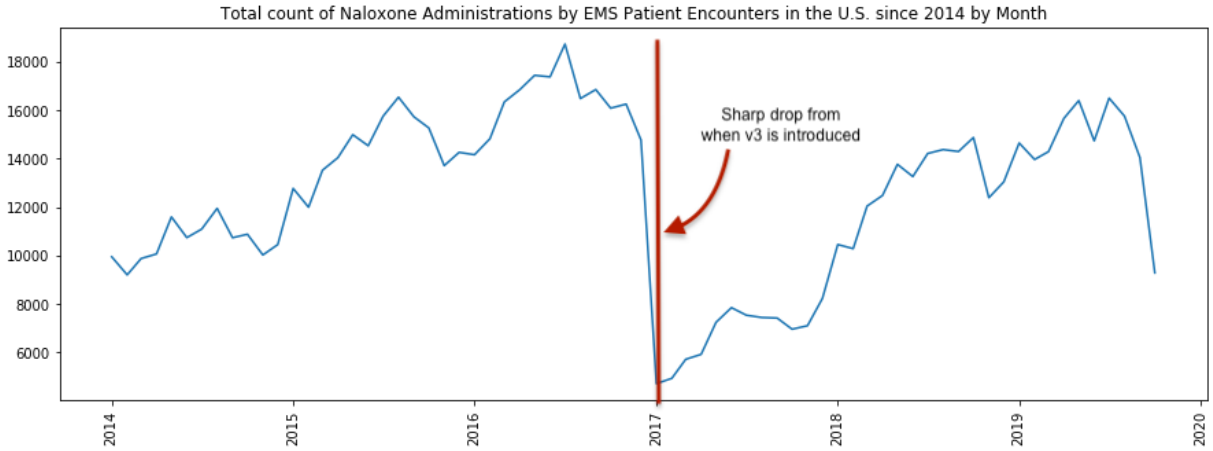


Fig 3. Count of Naloxone Administrations by EMS per 1,000 EMS Patient Encounters in the U.S. since 2014 by Month. (National EMS Information System, National EMS Database NEMSIS V2 and V3)

To mitigate the data loss as a result of observed effect, and inspired by some studies found on NEMSIS and WONDER [1], instead of using the counts of events, we will use the ratio between the reported incidents by medication administered with the total number of reported incidents. In Fig. 4 we show the time series graph using rates instead of totals and we can observe an improvement in the quality of the variable.

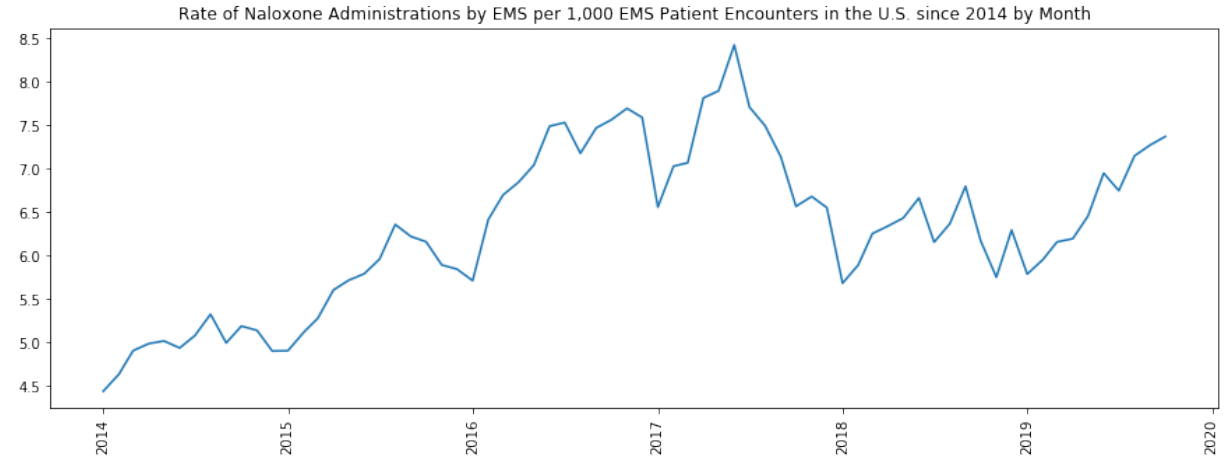
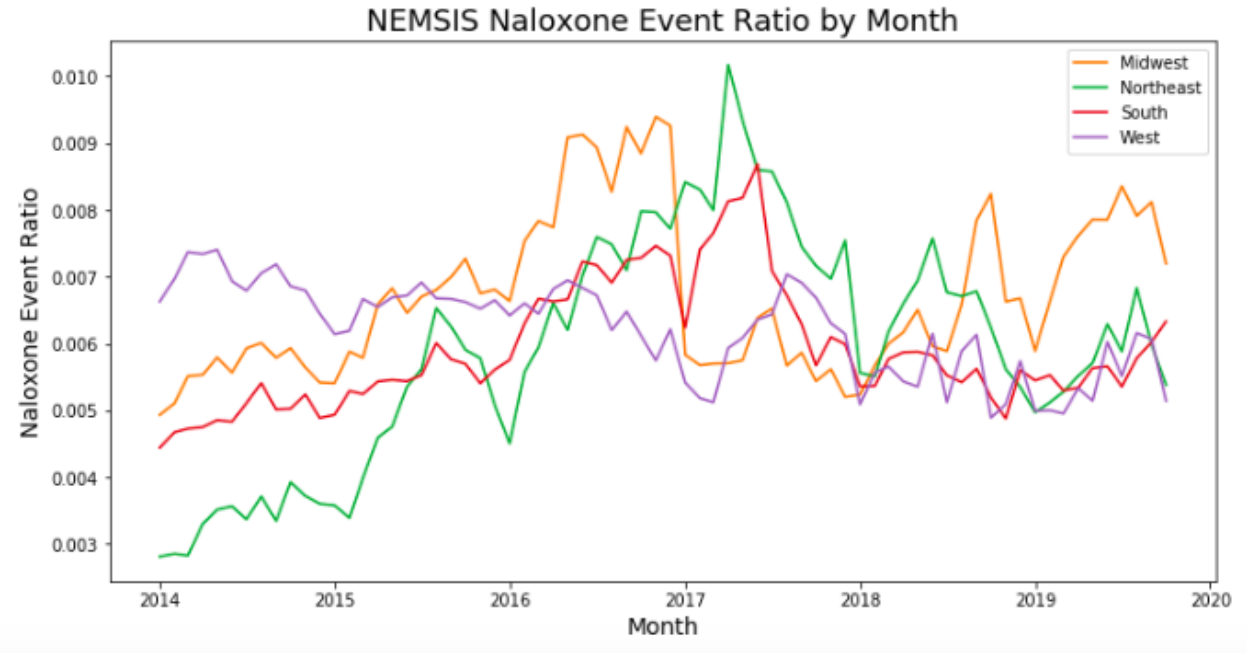


Fig 4. Ratios of Naloxone Administrations by EMS per 1,000 EMS Patient Encounters in the U.S. since 2014 by Month by Total events. (National EMS Information System, National EMS Database NEMSIS V2 and V3)

However, when we look at the same plot broken down by region, we see that the Midwest still shows a sharp dropoff at the beginning of 2017, when the NEMSIS V3 transition started. We suspect this reflects a reporting or technical issue rather than a true drop in Naloxone EMS events.



Data Augmentation with Provisional Dataset

Since CDC WONDER data is only available until December 2017, we explored extracting the data in provisional drug overdose death counts [2] which are based on death records received and processed by the National Center for Health Statistics (NCHS) which are available for up to April 2019.

While there are known differences between the CDC Wonder final data and the provisional data reported [5], our aim was to stitch data from both sets, spanning year 2014 to 2019 (April) in a format which would be useful for our analysis and model building.

However this process wasn't straightforward and we encountered numerous issues and had to perform some cleaning and normalization on the data. For instance, CDC WONDER records a *stock* value of number of drug deaths by month (Fig. 5) and the Provisional data records a *flow* value of number of drug deaths over the last 12 months (Fig. 6).

	A	B	C	D	E	F
1	State	State Co	Month	Year	Monthly (	Deaths
38	Alabama	1	Jan.	2017	1	91
39	Alabama	1	Feb.	2017	2	73
40	Alabama	1	Mar.	2017	3	92
41	Alabama	1	Apr.	2017	4	75
42	Alabama	1	May	2017	5	83

Fig 5. Death counts for the first five months of 2017 from CDC WONDER

Year	Month	Period	Indicator	Data Value	Percent	Percent Pendin	State Nar
2017	January	12 month-ending	Number of Drug Overdose Deaths	763	100	0.259428483	Alabama
2017	February	12 month-ending	Number of Drug Overdose Deaths	759	100	0.25610182	Alabama
2017	March	12 month-ending	Number of Drug Overdose Deaths	794	100	0.25160641	Alabama
2017	April	12 month-ending	Number of Drug Overdose Deaths	795	100	0.231687068	Alabama
2017	May	12 month-ending	Number of Drug Overdose Deaths	812	100	0.213695782	Alabama

Fig 6. Death counts for the first five months of 2017 from Provisional dataset

We extracted the monthly deaths for all the states, from 2018 onwards using the formula:

extracted monthly deaths in Jan '18 = provisional 12 mth data ending in Jan'18 - (provisional 12 mth data ending in Dec '17 - CDC data for Jan '17)

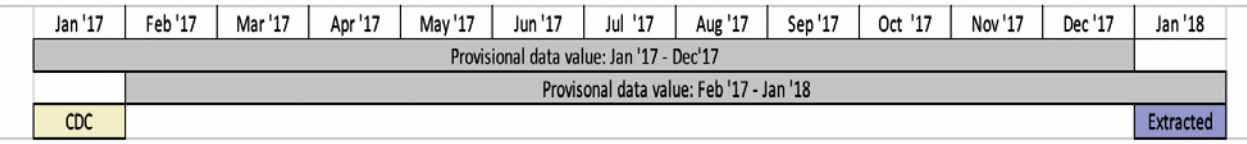


Fig 7. Temporal representation of the data extraction process

Additionally in the provisional data, we normalized date formats, merged New York City and New York state reported numbers into one and included a mapping of US Census Regions and Divisions in the dataset for further work.

NEMSIS vs. CDC Data

When we plot the CDC opioid-related death counts against the NEMSIS Naloxone event ratio by month, we see a clear relationship.

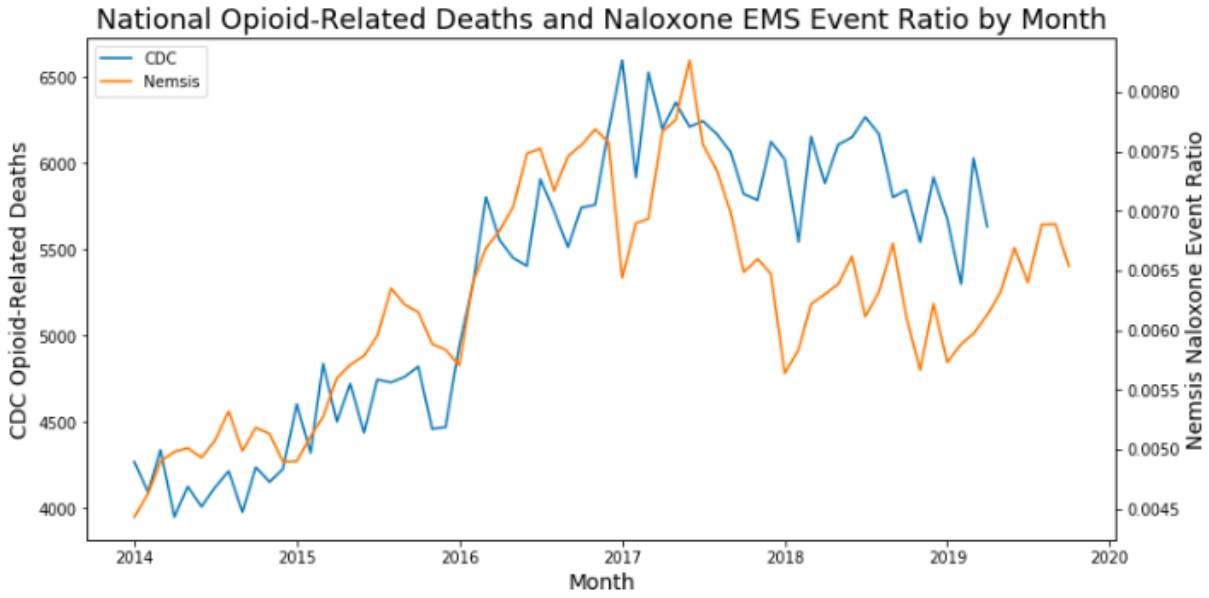


Fig 8) National Opioid-Related Deaths and Naloxone EMS Event Ration by Month

The plots for the Northeast and South Census regions show a similar strength of relationship, but in the West region there does not appear to be a correlation--opioid-related deaths trend upward while the Naloxone event ratio trends downward.

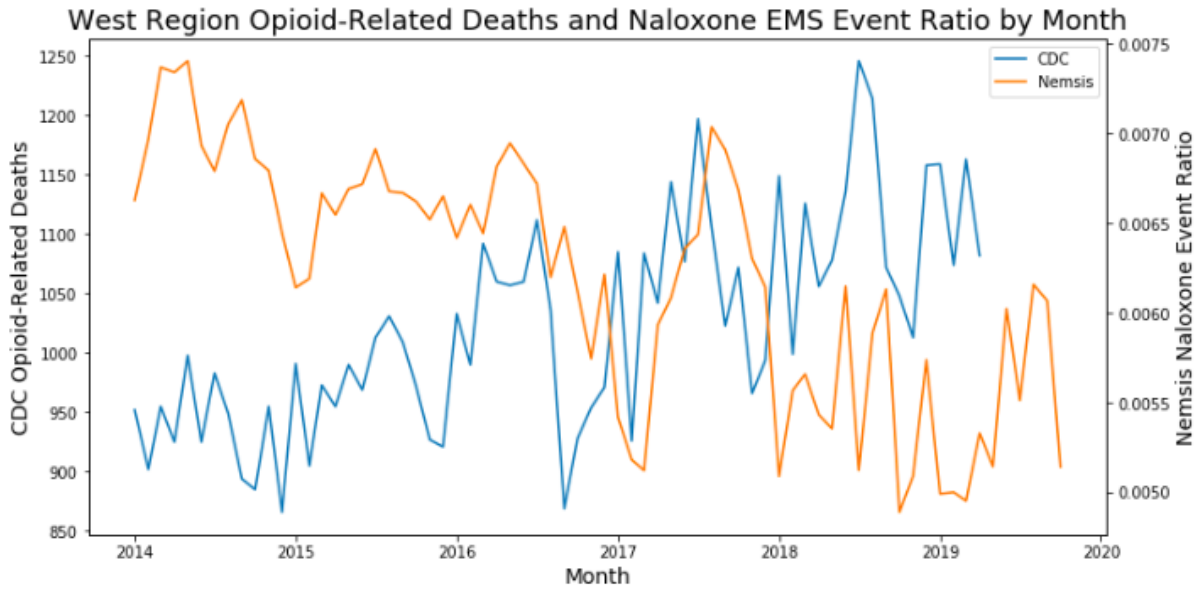


Fig 9) West Region Opioid-Related Deaths and Naloxone EMS Event Ratio by Month

In the Midwest, we see some correlation from 2014-2017, but a dramatic drop off in the NEMSIS data starting in 2017, as mentioned above. Since the drop coincides with the switch from NEMSIS V2 to NEMSIS V3, we suspect this represents a data collection/quality issue rather than real-world events.

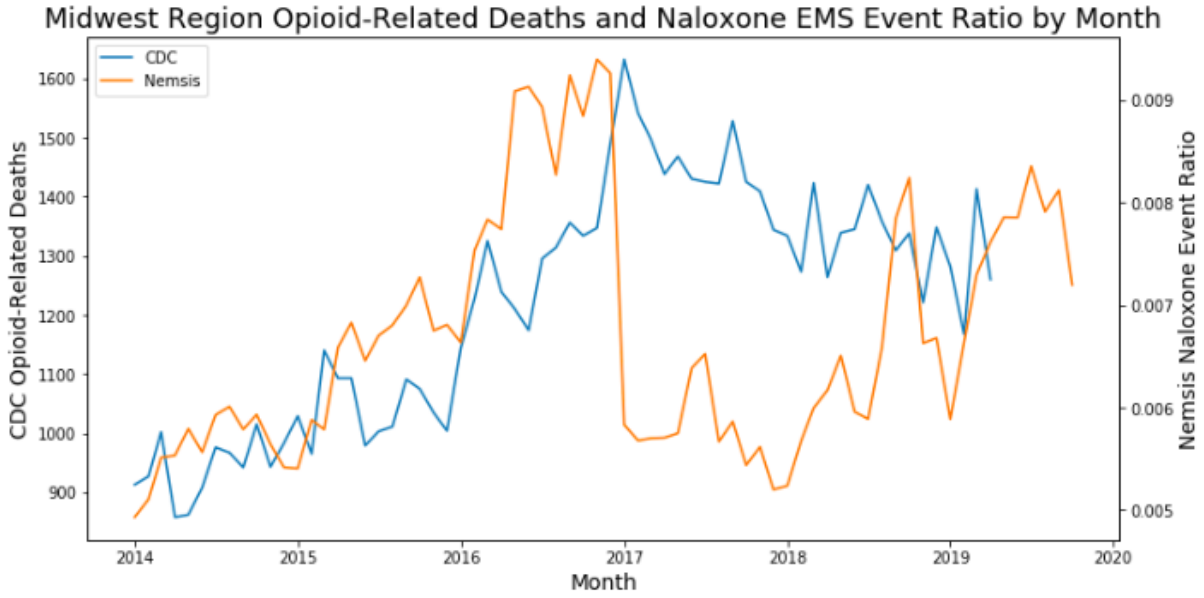


Fig 10) Midwest Region Opioid-Related Deaths and Naloxone EMS Event Ratio by Month

Our cross-correlation plots confirm that the CDC and NEMSIS numbers are strongly correlated in the Northeast and South but not in the Midwest or West. The correlation is strongest where

the lag is zero, i.e. when we compare data points for the same month, rather than one data source being a leading indicator of the other.

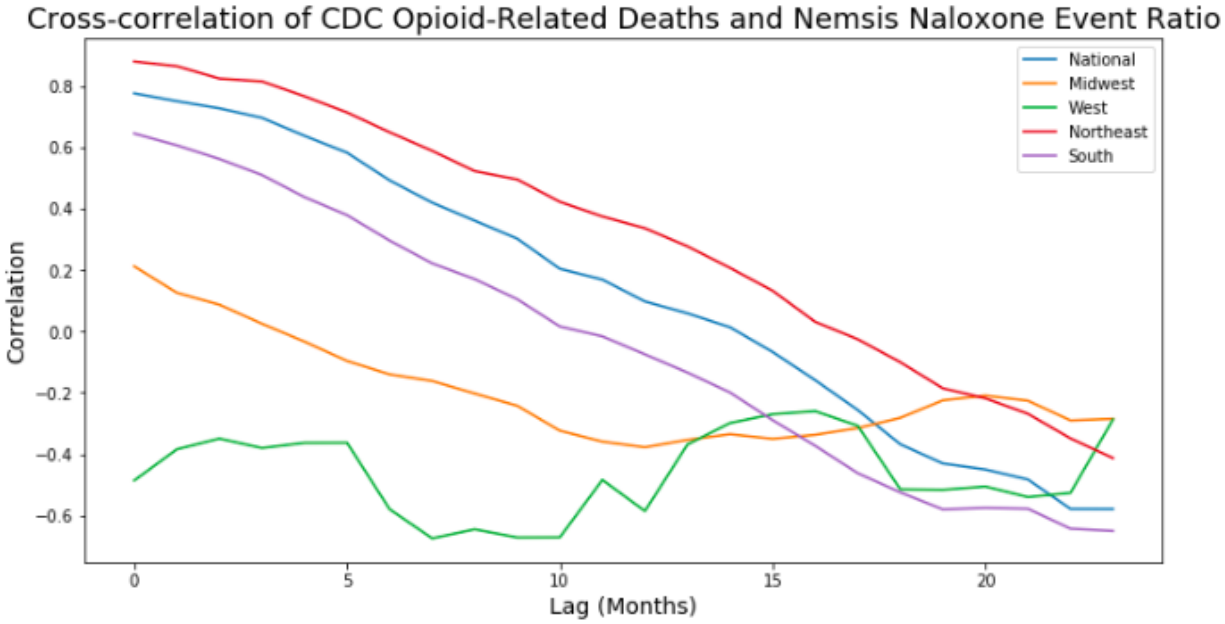


Fig 11) Cross-correlation of CDC Opioid-Related Deaths and Naloxone EMS Event Ratio

Since both the CDC and NEMSIS numbers trend upward between 2014 and 2017, we also tried detrending the data, which results in the following plot:

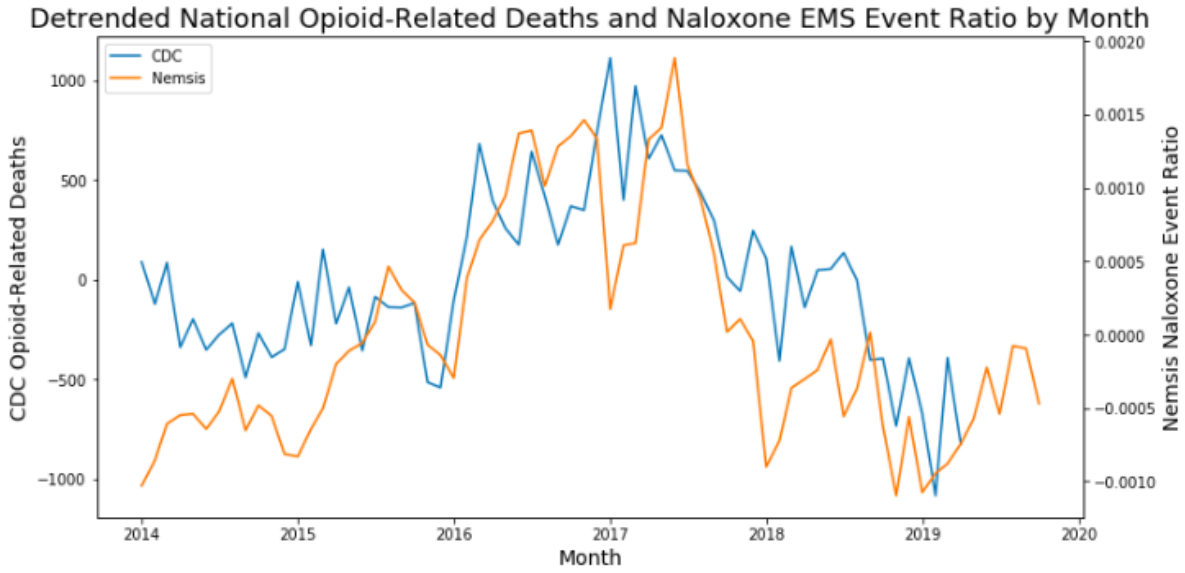


Fig 12) Detrended National Opioid-Related Deaths and Naloxone EMS Event Ratio by Month

As expected, this lowers the correlation, but there is still a meaningful relationship in the Northeast and South regions.

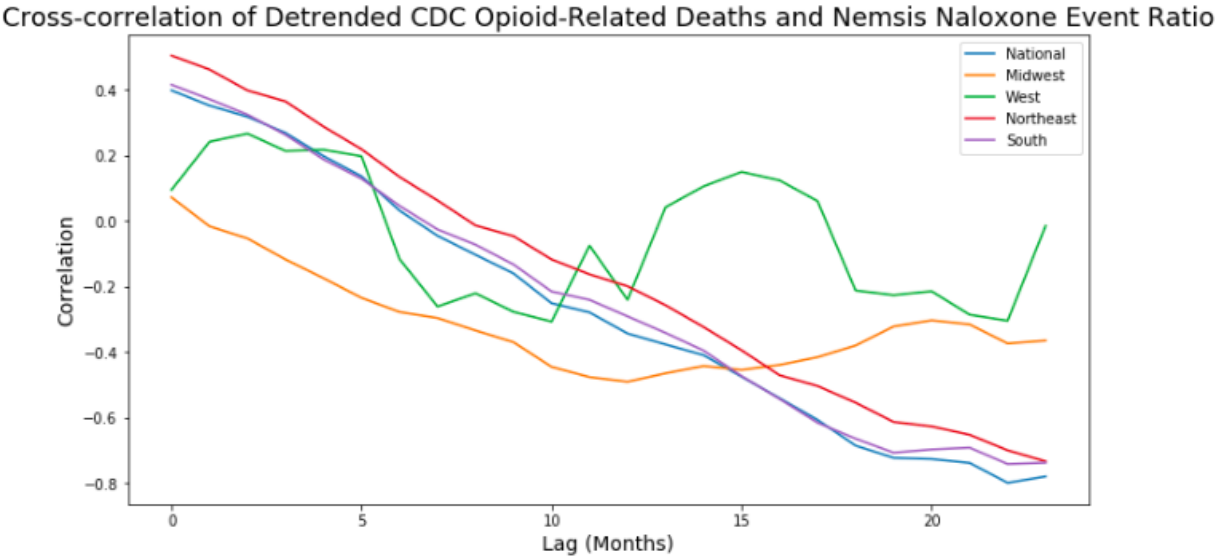


Fig 13) Cross-correlation of Detrended CDC Opioid-Related Deaths and Naloxone Event Ratio

When we performed a similar analysis with the ICD-filtered NEMSIS data, we see no correlation after detrending with the V2 data (that uses ICD-9). The V3 data shows more of a relationship, but is much noisier than the medication-filtered NEMSIS data.

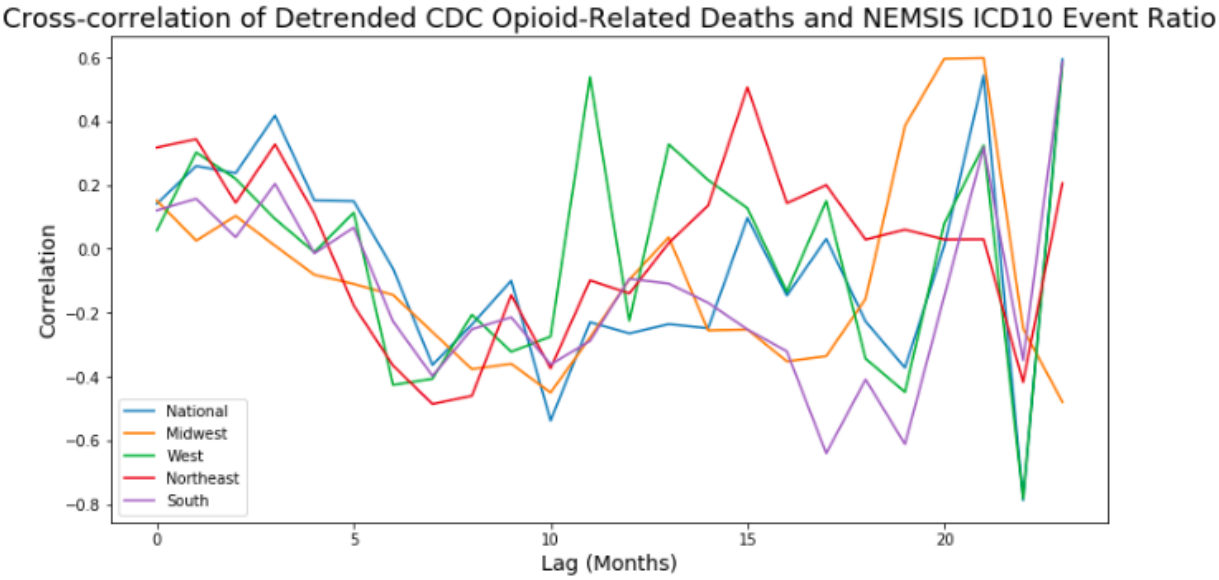


Fig 14) Cross-correlation of Detrended CDC Opioid-Related Deaths and ICD10 Event Ratio

Modeling approaches

We aimed to study different models for estimating mortality using the following *response* and *predictors*, as displayed below:

Approach	Response (monthly granularity)	Predictors (monthly granularity)
1	National Mortality Rates	National EMS data for Medications administered Or National EMS data for related ICD-10 codes
2	National Mortality Rates	Regional EMS data for Medications administered Or Regional EMS data for related ICD-10 codes
3	Regional Mortality Rates	National EMS data for Medications administered Or Regional EMS data for Medications administered
4	National Mortality Rates For a given month, we want to predict: $Y \sim X^{age1}, X^{age2}\ldots, X^{race1}, X^{race2}\ldots$	National/Regional EMS data for Medications administered faceted by age/gender/race Or National/Regional EMS data for ICD-10 codes faceted by age/gender/race

Results

Linear Regression

Our most successful model used Approach 3. Here, we trained a simple linear regression model with NEMSIS counts for a region or the country in a given month as predictors, and the corresponding CDC opioid-related death counts for that area in the same month as response.

One problem is that we cannot use our NEMSIS event ratios in this linear regression model when including data from multiple regions, but we know that the raw NEMSIS counts have severe quality issues. Instead, we estimated the true Naloxone event counts with the following logic:

Total events in NEMSIS in 2018 = 25 million
Population percentage by region:
Midwest: 21%
Northeast: 17%
South: 38%

West: 24%
Estimated count = NEMSIS Naloxone event ratio * region percentage * 25 million / 12 months
With this adjustment, we see a clear linear relationship.

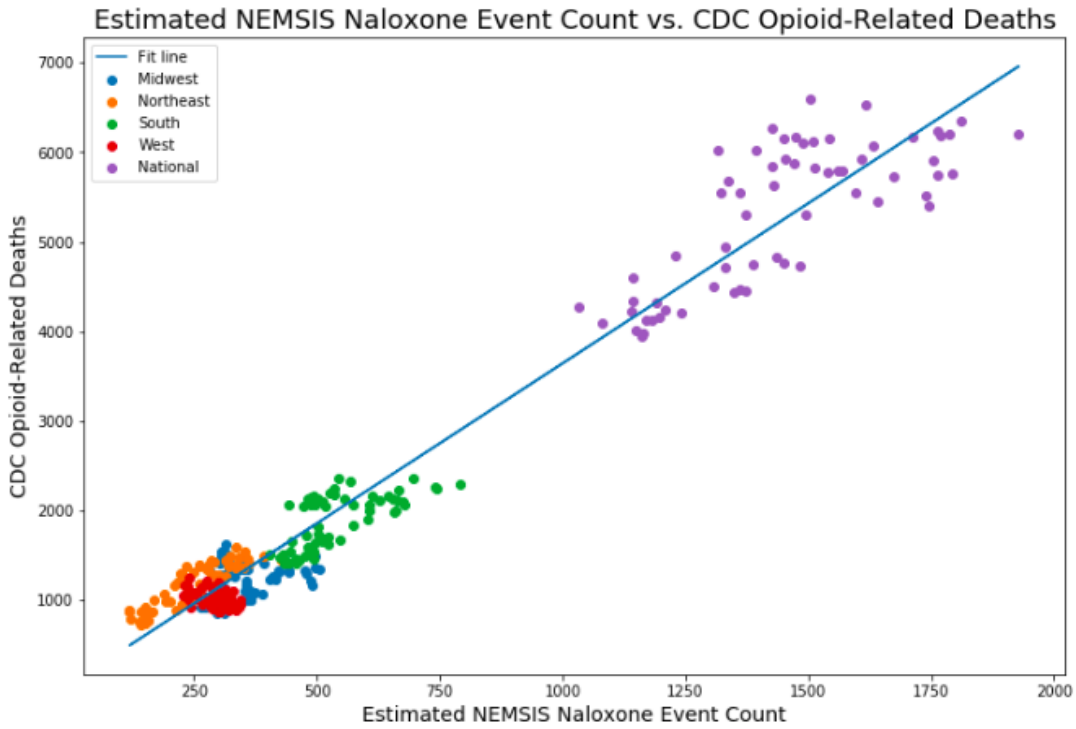


Fig 15) Estimated NEMSIS Naloxone Event Count vs CDC Opioid-Related Deaths

On a held-out test set of 20% of our data points, the mean squared error is 90528.07 (hard to reason about due to the different scale of our data depending on region) and the R2 score is 0.95. We used this model to predict death counts for the latest months in our NEMSIS data, for which CDC data does not yet exist, as shown below.

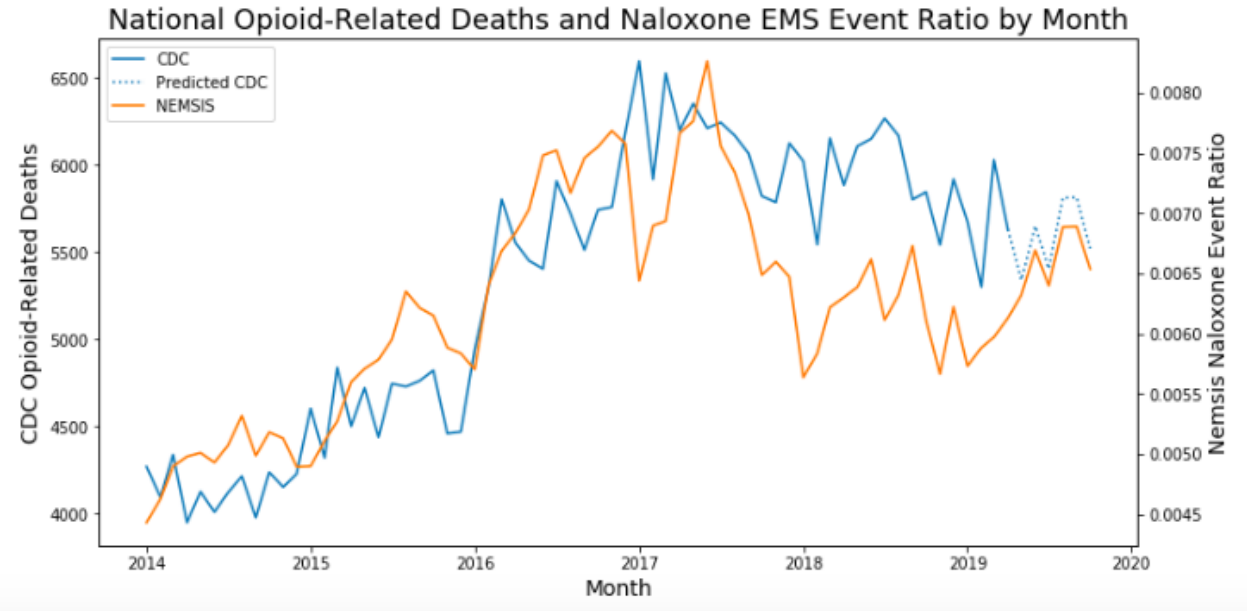


Fig 16) National Opioid-Related Deaths and Naloxone EMS Event Ratio by Month

ARIMA

We also looked into forecasting in order to model future values of national opioid-related deaths as a function of Naloxone EMS data by training a regression with ARIMA errors. Since this was done using R's auto-arima function, this enabled us to adjust for seasonality and differencing issues typical of time series data.

In this model, the data is split into training and test data sets, with the last six months being used for the test set. An ARIMA function is fitted to the training data, with national opioid-related deaths as the dependent variable and Naloxone EMS event ratios as the independent variable. Predicted values for the last 6 months are then generated using the auto.arima time series model and then plotted against the actual values of the last 6 months:

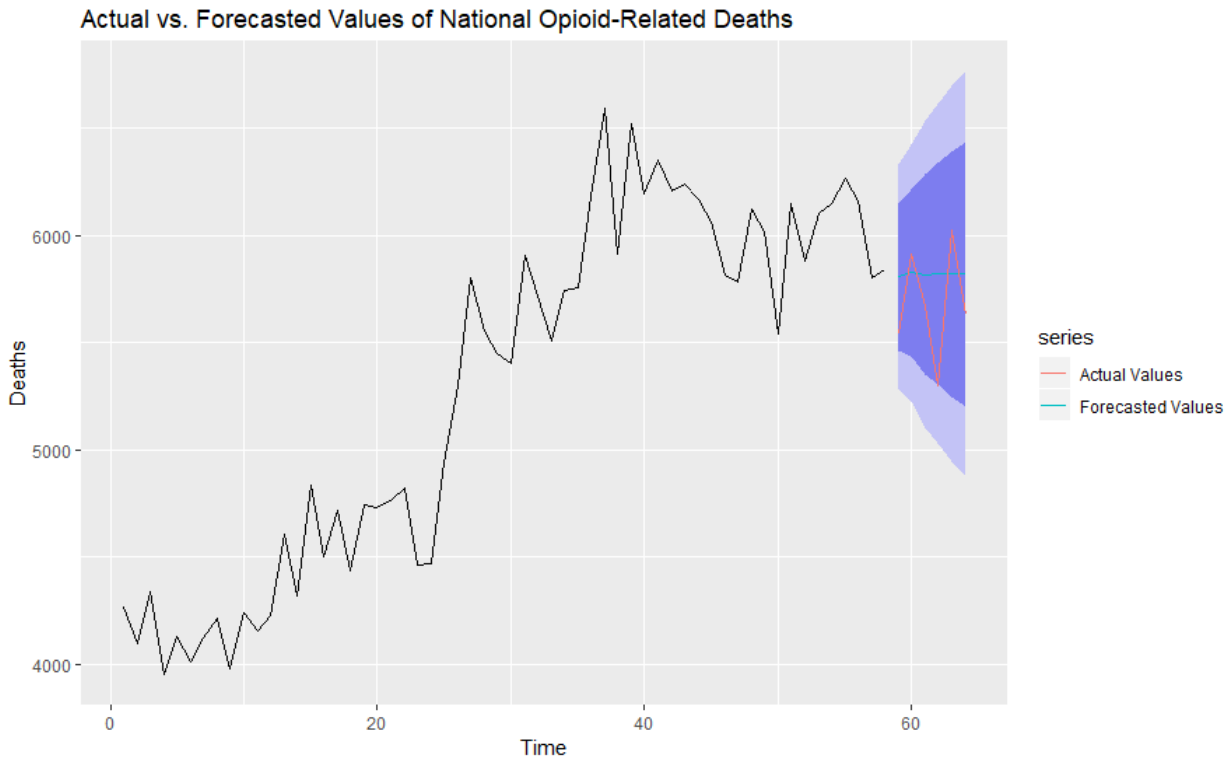


Fig 17) Actual vs. Forecasted Values of National Opioid-Related Deaths

We can see that the model produces very stable/invariant figures for the number of opioid-related deaths for the last 6 months, compared to the actual values which fluctuate rather wildly. While there is some regression to the mean to be expected when it comes to forecasting (i.e. the forecast will look smoother than the original time series), given the rather high variance of the actual number of opioid-related deaths the forecast could be more stochastic. Future work is needed on the best sort of forecasting model in order to generate predicted values of opioid-related deaths in the United States.

Conclusions and Limitations

The number of EMS events where Naloxone was administered correlates strongly with the number of opioid overdose deaths in that area at that time. Because EMS event data is available more quickly than mortality data, EMS event data can be used to estimate opioid overdose counts in a given month before official numbers are available.

Limitations of our research include:

- Using the ratio of Naloxone events to all events instead of raw counts means it's possible that a spike in non-opioid-related events could affect our data.
- In the Midwest, the data is clearly broken for some period of time after the NEMSIS v3 transition, so it's harder to evaluate whether our model works well there.

- In the West, it's less clear whether the NEMSIS events predict deaths. This may be a data quality issue or a true reflection of the nature of the opioid problems there.
- Not trained on a huge number of data points and the V2/V3 issues complicate things, so need to observe for longer to get a better understanding of prediction accuracy.

Future work

- Look at smaller geographic areas or demographic slices.
- Continue to tune and evaluate more models as more data becomes available, and develop a predictive model of deaths. Once there is sufficient data from after the NEMSIS V3 transition, it may make sense to drop the V2 data to avoid the noise caused by the transition.
- The ICD-9/ICD-10 discrepancies made using that data infeasible right now, but it may be possible in a few years, once more data exists from after the ICD-10 transition. This may mean that the approach could be extended to suicide or other public health concerns.

Team Member Contributions

Getting familiar with NEMSIS Cube / CDC WONDER system	Entire team
NEMSIS V3 - MDX scripting (Researching ORs in data)	Emma, Cristina
NEMSIS V2/V3 - Selenium scripts for Data extraction	Cristina
NEMSIS V3 - Initial ICD 10 codes set used for extraction	Vineet
NEMSIS V3 -> V2 code mappings + Naloxone study	Emma
NEMSIS V2/V3 - Page wise downloads + stitching pages together	Cristina
CDC WONDER data extraction	Namson
NEMSIS V3 Data exploration (Statistics, trends, uniqueness, facets)	Vineet, Cristina
CDC WONDER exploration	Namson
NEMSIS v2/v3 data extraction by medication administered	Cristina
Updating Selenium scripts - to automate our data extraction for future	Cristina
Investigating drops, differences and data quality in the Provisional data	Namson and Vineet
Provisional data extraction (normalizing with CDC WONDER) and adding extracting monthly counts from deltas - 2017 to April 2019	Vineet
Adding US Census Division and Region mapping to Provisional data	Namson
Study of NEMSIS and CDC correlation and cross correlations by months.	Emma
Data Techniques: Detrending, Estimating true Naloxone event counts	Emma
Linear regression model	Emma

Exploring other models including ARIMA	Namson
Fig 1 - 4	Cristina
Fig 5 - 7	Vineet
Fig 8 - 16	Emma
Fig 17	Namson
Final report	Entire team

References

[1] - https://nemsis.org/view-reports/public-reports/ems-data-cube/
[2] - https://wonder.cdc.gov/mcd-icd10.html
[3] https://ndarc.med.unsw.edu.au/resource/early-indicators-trends-opioid-overdose-deaths
[4] - https://bmchealthservres.biomedcentral.com/articles/10.1186/s12913-018-3756-8
[5] - https://www.cdc.gov/nchs/nvss/vsrr/drug-overdose-data.htm#differences_between_final_and_provisional_data