Patterns of the American Overdose Crisis

Adam N. Wlostowski

2020/02/23

Table of Contents

[Take Home Messages 1](#_Toc33369055)

[Motivation 1](#_Toc33369056)

[The Data 2](#_Toc33369057)

[Data Structure 2](#_Toc33369058)

[Data Cleaning 4](#_Toc33369059)

[Investigation 5](#_Toc33369060)

[National trend, temporal 5](#_Toc33369061)

[State ranking by overdose rate (2017) 6](#_Toc33369062)

[State ranking by rate of change in overdose rate (2003 - 2017) 7](#_Toc33369063)

[Geographic patterns of overdose rates (2017) 8](#_Toc33369064)

[Identifying hot spots 9](#_Toc33369065)

[Disucssion 11](#_Toc33369066)

# Take Home Messages

1. **Overdose deaths in the United States have increase by 167% between 2003 and 2017.**
2. **Over 70,000 people died from an overdose in 2017.**
3. **Hot-spot analysis reveals the overdose crisis is worst in Appalachia, New Mexico, Mid-Atlantic, and New England.**

# Motivation

Recently, the Centers for Disease Control (CDC) [announced](https://www.cdc.gov/drugoverdose/index.html) that drug overdose was the leading cause of death among young people in the U.S. Over *70,000* Americans died from an overdose in 2017, alone. Surely, the opioid epidemic is one of the greatest public health crises of the 21st century.

In this post I investigate overdose death count data to better understand Americas’ drug overdose epidemic. Specifically, I use data from the Centers for Disease Control to show how overdose death rates changed between 2003 – 2017.

\***Objective**: Quantify patterns of overdose deaths in America.

**\*Rationale**: Patterns in the data are useful targets for predictive models. We can learn from predictive models. So, is pattern identification the first step of the learning process?

# The Data

The CDC publishes a wealth of data characterizing the health of American communities. The [Underlying Cause of Death](<https://wonder.cdc.gov/ucd-icd10.html>) database reports the demographics, timing, location and cause of death. In other words, CDC tracks the \*who\*, \*what\*, \*where\*, \*why\* and \*when\* of people dying in the United States. These data are made available through the [WONDERES](https://wonder.cdc.gov) database - an excellent (free) resource for characterize drug overdose death patterns in the United States.

However, the WONDERS database has some drawbacks. First, fatalities data are highly suppressed for privacy reasons. The CDC suppresses death counts of less than ten (0-9) for a specific demographic/location/time categorization. Becuase so many counties in America have small populations, it is common to end up with a lot of suppressed data. In turn, it can be [difficult to draw inferences at small spatial scales (e.g. counties)](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4103252/). In response, researchers have come up with clever ways to infill suppressed data and produce serially complete data products. One such data product is provided by CDC, called *[NCHS - Drug Poisoning Mortality by County: United States](https://www.cdc.gov/nchs/data-visualization/drug-poisoning-mortality/)*.

I chose to work with the NCHS data product, rather than working with raw data. I made this decision, in part, because of the troubles of working with highly suppressed data. For additional details on the NHCS data product was assembled, read the [Technical Note] (<https://www.cdc.gov/nchs/data-visualization/drug-poisoning-mortality/#techNotes>).

I downloaded data from a nifty [Tableau dashboard](https://www.cdc.gov/nchs/data-visualization/drug-poisoning-mortality/) created and hosted by CDC.

## Data Structure

OK, let’s load the data and see what it looks like!

# navigate to where the data are housed  
setwd("data/")  
  
# read in the data  
data <- read.csv("NCHS\_-\_Drug\_Poisoning\_Mortality\_by\_County\_\_United\_States.csv", header = T, fill = T)  
  
data %>% head()

## ï..FIPS Year State FIPS.State County Population  
## 1 1001 2003 Alabama 1 Autauga County, AL 46,800  
## 2 1001 2004 Alabama 1 Autauga County, AL 48,366  
## 3 1001 2005 Alabama 1 Autauga County, AL 49,676  
## 4 1001 2006 Alabama 1 Autauga County, AL 51,328  
## 5 1001 2007 Alabama 1 Autauga County, AL 52,405  
## 6 1001 2008 Alabama 1 Autauga County, AL 53,277  
## Model.based.Death.Rate Standard.Deviation Lower.Confidence.Limit  
## 1 6.752270 1.532372 4.394907  
## 2 6.948645 1.449189 4.644302  
## 3 6.587989 1.376595 4.399895  
## 4 8.454333 1.725717 5.693863  
## 5 8.758229 1.777117 5.905313  
## 6 8.898952 1.799995 6.006984  
## Upper.Confidence.Limit Urban.Rural.Category Census.Division  
## 1 10.37080 Medium Metro 6  
## 2 10.30284 Medium Metro 6  
## 3 9.77489 Medium Metro 6  
## 4 12.43308 Medium Metro 6  
## 5 12.84598 Medium Metro 6  
## 6 13.03717 Medium Metro 6

It looks like the data set contains 13 variables: For a complete list of variable descriptions go [here]( <https://data.cdc.gov/NCHS/NCHS-Drug-Poisoning-Mortality-by-County-United-Sta/p56q-jrxg>).

The variable we will be paying lots of attention to in this project is `data$Model.based.Death.Rate`. This variable is also known as the “crude rate” - the number of fatalities reported each calendar year per 100,000 people:

`data$Model.based.Death.Rate = (Deaths/Population) \* 100000`

## Data Cleaning

There are a few clean-up needs before we start working with the data. First, I want to change the fips code variable data$ï..FIPS to something a little less strange

colnames(data)[colnames(data) == "ï..FIPS"] <- "fips"

Next, we need to convert population from a factor to a numeric variable type

# convert from factor to character  
data$Population = as.character(data$Population)  
  
# remove those pesky commas!  
data$Population = gsub(",", "", data$Population)  
  
# convert from character to integer  
data$Population = as.integer(data$Population)  
  
# print the a snippet of the data to visually inspect  
data %>% head()

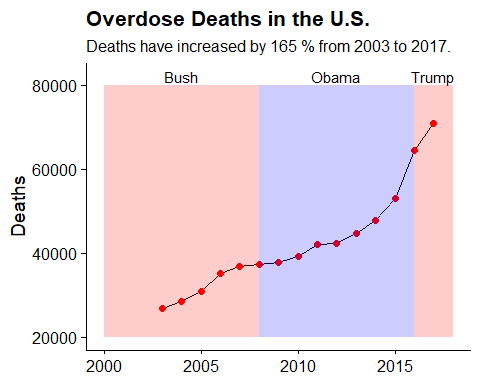
## fips Year State FIPS.State County Population  
## 1 1001 2003 Alabama 1 Autauga County, AL 46800  
## 2 1001 2004 Alabama 1 Autauga County, AL 48366  
## 3 1001 2005 Alabama 1 Autauga County, AL 49676  
## 4 1001 2006 Alabama 1 Autauga County, AL 51328  
## 5 1001 2007 Alabama 1 Autauga County, AL 52405  
## 6 1001 2008 Alabama 1 Autauga County, AL 53277  
## Model.based.Death.Rate Standard.Deviation Lower.Confidence.Limit  
## 1 6.752270 1.532372 4.394907  
## 2 6.948645 1.449189 4.644302  
## 3 6.587989 1.376595 4.399895  
## 4 8.454333 1.725717 5.693863  
## 5 8.758229 1.777117 5.905313  
## 6 8.898952 1.799995 6.006984  
## Upper.Confidence.Limit Urban.Rural.Category Census.Division  
## 1 10.37080 Medium Metro 6  
## 2 10.30284 Medium Metro 6  
## 3 9.77489 Medium Metro 6  
## 4 12.43308 Medium Metro 6  
## 5 12.84598 Medium Metro 6  
## 6 13.03717 Medium Metro 6

# Investigation

## National trend, temporal

Let’s take a look at how national overdose deaths are growing with time.

# calculate the number of fatalaties each year, nationwide.   
national\_deaths <- data %>%  
 mutate(Deaths = (data$Population/100000) \* data$Model.based.Death.Rate) %>%  
 group\_by(Year) %>%  
 summarize(Total.Deaths = sum(Deaths, na.rm = T))  
  
# calculate the percent increase from 1999 to 2018  
pct\_increase <- ((filter(national\_deaths, Year == 2017)$Total.Deaths - filter(national\_deaths, Year == 2003)$Total.Deaths)/filter(national\_deaths, Year == 2003)$Total.Deaths) \* 100  
  
# plot the number of deaths through time  
p <- ggplot(national\_deaths, aes(x = Year, y = Total.Deaths)) +  
 geom\_line() +  
 geom\_point(size = 2, color = "red") +  
 labs(x = NULL, y = "Deaths") +  
 ggtitle("Overdose Deaths in the U.S.",   
 subtitle = paste("Deaths have increased by",as.character(round(pct\_increase)),"% from 2003 to 2017.", sep = " ")) +  
   
 # add annotations: the start and end of presidencies   
 annotate(geom = "text", label = "Bush", x = 2004, y = 82000) +  
 annotate(geom = "rect", ymin = 20000, ymax = 80000, xmin = 2000, xmax = 2008, fill = "red", alpha = 0.2) +  
   
 annotate(geom = "text", label = "Obama", x = 2012, y = 82000) +  
 annotate(geom = "rect", ymin = 20000, ymax = 80000, xmin = 2008, xmax = 2016, fill = "blue", alpha = 0.2) +  
   
 annotate(geom = "text", label = "Trump", x = 2017, y = 82000) +  
 annotate(geom = "rect", ymin = 20000, ymax = 80000, xmin = 2016, xmax = 2018, fill = "red", alpha = 0.2)  
   
print(p)

 This figure shows the rise in overdose deaths between 2003 to 2017.

* The national death count increased 165% over this period.
* There were over 70,000 overdose deaths in 2017!

I overlaid U.S. presidential terms on top of the data. I don’t mean to imply that a certain president’s policies were responsible for the trend. Rather, presidential terms are a convenient way for me to frame the passing of time. We see that overdose deaths accelerated in the second term of the Obama administration. Death counts pushed even higher under the Trump administration.

## State ranking by overdose rate (2017)

Where is the opiod crisis the worst, as of 2017? Rank states by overdose crude rate to learn.

# calculate statewide per capita overdose rate in 2017, then rank  
state\_rank\_2018 <- filter(data, Year == 2017) %>%  
 group\_by(State) %>%  
 summarize(Rate = mean(Model.based.Death.Rate, na.rm =)) %>%  
 arrange(desc(Rate))  
  
# Show the top ten states   
state\_rank\_2018 %>% head(10)

## # A tibble: 10 x 2  
## State Rate  
## <fct> <dbl>  
## 1 District of Columbia 42.3  
## 2 West Virginia 40.0  
## 3 Kentucky 33.9  
## 4 Maryland 32.5  
## 5 Ohio 31.9  
## 6 Delaware 30.0  
## 7 Pennsylvania 29.6  
## 8 New Jersey 29.5  
## 9 Connecticut 28.9  
## 10 Massachusetts 28.8

The highest crude rate is observed in Washington D.C. Notice that all states in the top ten are east of the Mississippi River.

## State ranking by rate of change in overdose rate (2003 - 2017)

Where have overdose fatalities increased most rapidly between 2003 and 2017? To quantify this, I built linear models to quantify the annual growth rate of per capita fatalities in each state. Then I ranked states by growth rate (i.e. regression slope)

# calculate the time rate of change in per capita mortality rates in each state, rank   
state\_change <- data %>%  
 group\_by(State, Year) %>%  
 summarize(Rate = mean(Model.based.Death.Rate)) %>%  
 group\_by(State) %>%  
 do(lm\_trend = lm(Rate ~ Year, data = .)) %>%  
 mutate(slope = lm\_trend$coefficients[2]) %>%  
 ungroup() %>%  
 arrange(desc(slope))  
   
# Show the top ten states   
state\_change %>% head(10)

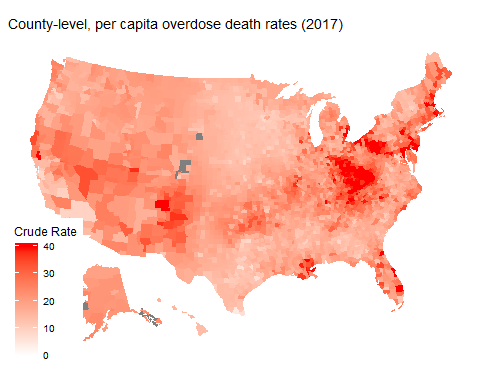
## # A tibble: 10 x 3  
## State lm\_trend slope  
## <fct> <list> <dbl>  
## 1 West Virginia <lm> 1.53  
## 2 District of Columbia <lm> 1.32  
## 3 Delaware <lm> 1.31  
## 4 Ohio <lm> 1.27  
## 5 Connecticut <lm> 1.22  
## 6 Maryland <lm> 1.19  
## 7 New Hampshire <lm> 1.18  
## 8 Kentucky <lm> 1.17  
## 9 Pennsylvania <lm> 1.17  
## 10 Massachusetts <lm> 1.13

The greatest increses in overdose deaths are seen in many of the same states with the greatest 2017 per capita fatalities. West Virginia has seen the most rapid growth in per capita overdose deaths.

## Geographic patterns of overdose rates (2017)

Let’s create a map that visualizes 2017 per capita overdose deaths by county.

library(usmap)  
library(ggplot2)  
  
# prepare the data  
county\_rates <- filter(data, Year == 2017) %>%  
 group\_by(fips, Year) %>%  
 summarize(Rate = mean(Model.based.Death.Rate, na.rm = T))  
  
plot\_usmap(data = county\_rates,   
 values = "Rate",   
 regions = c("states", "state", "counties", "county"),   
 # include = .northeast\_region,   
 color = NA, size = 0.1) +   
 scale\_fill\_continuous(low = "white", high = "red", name = "Crude Rate", limits = c(0,40), label = scales::comma, oob = scales::squish) +  
 ggtitle("County-level, per capita overdose death rates (2017)")

 Overdose death rates are greatest where the map is redder. Rates look to be highest in Appalachia, Western Pennsylvania, South Eastern Pennsylvania, New Hampshire, Connecticut, Oklahoma, New Orleans, New Mexico.

## Identifying hot spots

Can we objectively identify hot spots of drug overdose deaths? Rather than just eyeballing patterns in the map above, we can use spatial statistics to objectively identify hot spots. The Getis/Ord Gi\* Z-value is used to evaluate local correlation in geospatial data and objectively identify hot and cold spots. In this case, high positive Z-values indicate a local cluster of high death rates. On the other hand, very low values identify clusters of low death rates. The original publication behind the Getis/Ord Gi\* statistic can be found [here](https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1538-4632.1992.tb00261.x).

There are a few steps necessary to pull this analysis off. First, we need to get a shapefile of U.S. counties and spatially join our overdose data. Then, we identify neighboring counties (e.g. Boulder County touches Weld County - they are neighbors). Finally, we can use the LocalG function to calculate the Getis/Ord Gi\* Z-value for each county. We will rely heavily on the [spep](https://cran.r-project.org/web/packages/spdep/spdep.pdf) package to do much of the heavy lifting.

1. Load a SpatialPolygons oject of U.S. counties throught the tigris package

library(tigris)  
  
# load the data  
p <- counties()  
  
# create a new fips variable, as integer  
p$fips <- as.integer(p$GEOID)

1. Join the 2017 crude rate data to the shapefile by fips

rate\_join <- left\_join(as.data.frame(p),county\_rates, by = "fips")  
  
p$Rate <- rate\_join$Rate  
  
p <- p[!is.na(p$Rate),]

1. Identify neighboring counties with the the poly2nb funtion from the [spdep](C:\\Users\\awlostowski\\Documents\\Rblog\\awlostowski.github.io\\content\\(https:\\cran.r-project.org\\web\\packages\\spdep\\spdep.pdf)) package.

library(spdep)  
nb <- poly2nb(p, row.names = paste(p$STATEFP,p$COUNTYFP,sep = ""))

1. Supplement the neighbors list with spatial weights.

listw <- nb2listw(nb, style = "B",zero.policy = T) # use a binary coding, B

1. Calculate the local spatial statistic, Getis/Ord Gi\* Z-statistic on the county-level 2017 crude rate data.

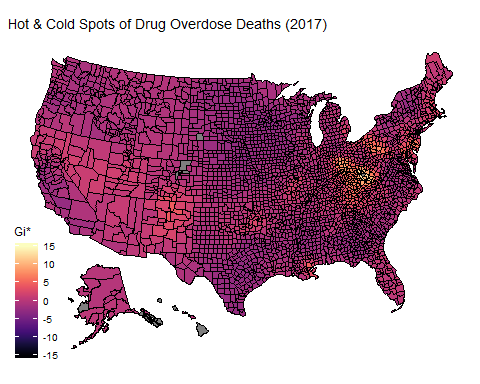
g <- localG(p$Rate, listw, zero.policy = T)  
  
# join county fips number  
gstar <- data.frame(g = as.numeric(g),  
 fips = paste(p$STATEFP,p$COUNTYFP,sep = ""))

1. Plot the variation in the Gi\* Z-value.

library(viridis)

## Loading required package: viridisLite

plot\_usmap(data = gstar,   
 values = "g",   
 regions = c("states", "state", "counties", "county"),   
 # include = .northeast\_region,   
 color = "black", size = 0.1) +   
 scale\_fill\_viridis(option = "magma", limits = c(-15,15),name = "Gi\*") +  
 ggtitle("Hot & Cold Spots of Drug Overdose Deaths (2017)")

 Warmer colors indicate hot spots - where per capita overdose death rates are high. Cooler colors indicate cold spotsWhat - where per capita overdose death rates are low. This plot is useful for comparing and contrasing the prominance of overdoses between different regions of the country.

# Disucssion

* \*\*Data considerations\*\*: I analyzed a data *product*, not raw data. Specifically, the [NHDC](https://www.cdc.gov/nchs/data-visualization/drug-poisoning-mortality/) data product is a derrivative of highly suppresed CDC death count data. This is a useful product, becuase otherwise it would be difficult to see small-scale spatial patterns in the data. However, I need to learn more about how the product was created, else there is a risk of being misled by biases intrinsic in the data product.
* \*\*What factors are behind geographic patterns?\*\*: Additional analysis may focus on using demographic, socioeconomic, and other data to explain emergent spatial patterns in overdose rates. Surely, a natural place to pick up on this topic later.
* \*\*This is my first post!\*\*: In order to get this published I learned a lot about using GitHub and Jekyll to create web pages. I have \*zero\* background in web development, so this was a pretty big lift for me. Hopefully, I have a repeatable workflow that will save me time in the future!
* Currently reading a book called American Overdose by Chris McGreal ([buy it on Amazon](https://www.amazon.com/American-Overdose-Opioid-Tragedy-Three/dp/1549176927)). Mr. McGreal hypothesize that overprescribing presription drugs is at the root of the problem. Is there a data set that tracks prescriptions for narcotics? If so, I could imagine analyzing those data to identify the location of pill mills. Then, use that to explain the patterns I described in this post.