Data Analysis For Humane Incarceration

Ben Klingensmith

Abstract

This paper outlines my analysis of health and mortality in Pennsylvania jails using publicly available datasets from the Pennsylvania Department of Corrections, US Census, and Centers for Disease Control and Prevention. I explored jail data from the Pennsylvania DOC, showing a chronic overcrowding problem in several jails and highlighting concerns about data on mental health hospitalizations, population flow, and sentencing status. My primary focus was mortality; this research revealed that suicides are an enormous health problem in jails. Nearly forty percent of deaths in Pennsylvania jail from 2015 to 2018 were suicides. A novel analysis suggests that the jail suicide rate is more than twice that of similar people in the resident Pennsylvania population. Further, I cleaned, standardized and combined several datasets, so that my colleagues at CHRS can continue further investigations on this topic.

Introduction

A large number of Americans spend time in jail in any given year. According to the Bureau of Justice Statistics (BJS), which gathers data from across the country and synthesizes it into comprehensive datasets and reports, 745,200 people were incarcerated in jails at midyear 2017[1]. Two-thirds of those individuals were not convicted or sentenced at the time. That was the jail population on one day, but jails are for short-term incarceration. The total number of people going through the jail system every year is much higher. Nationally, jail incarceration rates have decreased over the past decade, from 259 per 100,000 residents in 2007 to 229 per 100,000 in 2017, though this is driven more by a growing national population rather than a shrinking incarcerated population. The national rate of jail incarceration varies wildly between groups, with men being jailed almost six times as often as women and the racial breakdown being: Black (616 per 100,000), American Indian/Alaska Native (366 per 100,000), White (187 per 100,000), Hispanic (185 per 100,000), Other (39 per 100,000) and Asian (26 per 100,000).

BJS is also the primary source for mortality data in correctional facilities. In 2016, people incarcerated

---

[1] Zhen Zheng, U.S. Bureau of Justice Statistics; Jail Inmates in 2017; February 2020 https://www.bjs.gov/index.cfm?ty=pbdetail&iid=6547

in jails died at a rate of 149 per 100,000 in 2016, with 40% of all deaths occurring within the first week of incarceration[2]. From 2000 to 2016, the number one cause of death was illness (51%) which was largely driven by heart disease (23%), followed by suicide (31%) and drug/alcohol intoxication (7%). Suicide rates have been rising over the 2007-2016 period, with a ten year low of 23.8% (2008) and a ten year high of 35.2% (2014) which translates to 29 and 51 suicides per 100,000 people. Much of the reporting and research on incarceration relies on this data, but these reports tend to be released several years after data collection, and summarize data at high levels. Mortality data hasn't been reported since 2016[3], which has made related research much more difficult. Their most recent report included summary statistics on mortality rates by cause of death, demographics, and state, as well as aggregated data records along those criteria[4]. The BJS reports aggregate their findings at the state and demographic levels, as is necessary due to the huge scope of their data. The underlying data is available by request through the National Archive of Criminal Justice Data for analysis onsite at a secure facility at the university of Michigan. My analysis focuses on Pennsylvania and uses data from the PA DOC that enables more up to date, granular comparisons between facilities and a focus on the needs of this state. Because jails are run at the county level, different facilities can have very different policies, and a facility-level analysis can help understand how the variation between counties and understand how different circumstances and practices affect the health of incarcerated people.

In 2017, the Rand corporation put out a detailed report of deaths in US correctional facilities titled "Caring for Those in Custody"[5]. This report breaks deaths down into five categories: illness or disease, drug and alcohol intoxication, accident, suicide, and homicide. It analyzes the state of deaths for each of these categories, using BJS data up to 2014, and summarizes expert opinions on how deaths can be reduced. The report found that jail facilities are lacking in both medical and mental health capacity, which contributes to the two most common causes of death in jail, illness and suicide. Other than increasing mental health funding, experts suggested that the best steps to reducing suicides in jails include better suicide risk assessment and the spread of mental health best practices. Regarding drug and alcohol related deaths, they recommended greater access to countermeasures such as naloxone and expanded use of medication-assisted treatment.

Our partner in this work is the Pennsylvania Prison Society (PPS)[6], a human rights organization founded in 1787, dedicated to representing the interests of prisoners and ensuring their humane treatment by the justice system. Furthermore, the Pennsylvania constitution gives PPS access to correctional facilities and

---

[2] Margaret E. Noonan, U.S. Bureau of Justice Statistics; Mortality in Local Jails, 2000-2016 – Statistical Tables; February 2020 https://www.bjs.gov/content/pub/pdf/mlj0016st.pdf

[3] https://www.bjs.gov/index.cfm?ty=tp&tid=19#data_collections

[4] Margaret E. Noonan, U.S. Bureau of Justice Statistics; Mortality in Local Jails, 2000-2016 – Statistical Tables; February 2020 https://www.bjs.gov/content/pub/pdf/mlj0016st.pdf

[5] Russo, Joe, Dulani Woods, John S. Shaffer, and Brian A. Jackson, Caring for Those in Custody: Identifying High-Priority Needs to Reduce Mortality in Correctional Facilities. Santa Monica, CA: RAND Corporation, 2017. https://www.rand.org/pubs/research_reports/RR1967.html.

[6] https://www.prisonsociety.org/

government officials that is unique among similar organizations. PPS's activities, which are conducted by volunteers, include monitoring prison conditions, investigative prison visits, mentoring, and transportation for family members. Currently PPS uses their access to investigate specific issues at the individual level, relying on reports from incarcerated people and their families to spot problems. Executive director, Claire Shubik-Richards, approached our group for support with research and data analysis to support their policy reform efforts. The Center for Human Rights Science (CHRS) at Carnegie Mellon has partnered with PPS to develop methods to assess prisons & jails through the analysis of public data, and the work I have done constitutes much of the early progress of this relationship.

The primary goal of this project has been to better understand the treatment of incarcerated people in Pennsylvania, focusing specifically on deaths in jails. Jails hold people who are awaiting trial or have short sentences (under 2 years in PA) and are operated at the county level, unlike prisons which are for longer sentences and are operated by the state. I focused my analysis on jails, creating a template for this kind of work, while another student from the CHRS, Zhenzhen Liu, has since begun similar work on prisons. I focused on deaths specifically because they are obviously important, well documented, and death rates are often indicators for other problems such as lacking medical care or oversight.

Data

The primary data source of my analysis is two annual jail data files that are publicly available through the Pennsylvania Department of Corrections (DOC) website[7], though I later supplemented this with data from the US Census, PA department of health, and CDC Wonder. The first dataset, County Statistics and General Information, was originally shared with me by PPS. It includes numerous relevant variables on topics such as incarcerated population demographics, staff and funding, facility programs and practices, and mental health. I found the second, Extraordinary Occurrence Statistics, by searching the website. It counts various outcomes of interest such as deaths, assaults, uses of force, and other emergencies, all of which are broken up further by type. All this data is reported by counties individually to the PA DOC, which aggregates it. The state gives counties guidelines on what and how to report but doesn't directly oversee any part of the process. As a result, there are a few instances of missing data, and it isn't clear if all facilities calculate variables in the same way. As of writing this, these files have been posted annually from 2014 to 2018 (General Information) and 2015 to 2019 (Extraordinary Occurrences).

To better understand how the PA jail population compared with the general population, I gathered PA demographic data from the US Census website[8]. Of particular use was the population count and racial

---

[7] https://www.cor.pa.gov/Facilities/CountyPrisons/Pages/Inspection-Schedule,-Statistics-And-General-Info.aspx

[8] U.S. Census Bureau; American Community Survey, Annual Estimates of the Resident Population: April 1, 2010 to July 1,

breakdown per county. To help visualize my results, I found a shapefile of Pennsylvania on ArcGIS Online[9], a website dedicated to GIS software. This allowed me to project statistics onto a map of Pennsylvania, making geographic trends easier to spot.

The most difficultly I had gathering data was finding detailed mortality data for Pennsylvania. Specifically, I wanted mortality data broken up by cause of death and various demographic variables, to better understand how mortality in jails compares to the general population while controlling for demographics. I originally looked to the PA Department of Health website[10], which does have mortality counts split by cause of death, race, age, and sex. However, not all causes of death were counted, with many being grouped into an 'other' category. Overdose deaths were not separately counted, which seemed to be a concerning oversight considering that it is one of the most common causes of death. This was particularly a problem for me as I was most interested in death among younger people, who are much more likely to be in jail or die from an overdose. After more searching, I found a source which broke up demographics and cause of death variables to a very granular level, the CDC WONDER Underlying Cause of Death database[11].

The CDC WONDER dataset uses the ICD-10[12] to code deaths, an extremely detailed, international system of medical classification. This, and the demographics, were at a greater level of granularity than I even needed. There was one limitation to the dataset however, which was imposed to protect individuals' privacy. Any data subset which contained less than ten deaths was suppressed, so I could not know the exact amount. This made for a complicated balancing act, where I had to find data at the most granular level I could while not creating categories with less than ten deaths. Ultimately, I ended up extracting data in the form of seventy different files, focusing on the variable splits most important to my analyses and aggregating wherever possible.

Methods

My work began with cleaning the PA DOC data, as it was spread across several excel files with different formats. My focus for this was to put the data in a usable format and produce simple and understandable code that can be reused or altered by others. This is both to allow replication of my analysis, as well as to support analyses of future data releases or additional variables. This included higher level decisions such as how to structure this multidimensional data, deal with changes in the structure of the data between years,

---

2018; generated using American FactFinder; http://factfinder.census.gov ; (Spring 2019).

[9] https://www.arcgis.com/home/item.html?id=04e3f70b4b7f401faafd431da9355ab4 Data sourced from the US Census Bureau, modified to only include Pennsylvania.

[10] These data were provided by the Pennsylvania Department of Health. The Department specifically disclaims responsibility for any analyses, interpretations, or conclusions. https://www.health.pa.gov/topics/HealthStatistics/VitalStatistics/DeathStatistics/Pages/death-statistics.aspx

[11] Centers for Disease Control and Prevention; Underlying Cause of Death 1999-2017; generated using WONDER http://wonder.cdc.gov , (February 2020)

[12] See, for example https://www.cdc.gov/nchs/icd/icd10cm.htm

and handle inconsistent reporting styles between counties.

When deciding how to structure the data, my largest obstacle was the Philadelphia facilities. All other counties have one or no jail facilities (a few export the need to neighboring county facilities) with the exception of Philadelphia, which has five due to its huge incarcerated population. The reporting of these five facilities is inconsistent with many variables, like demographics and deaths, having values for each facility while other variables, like funding and mental health statistics are aggregated across Philadelphia. This problem could be avoided by simply aggregating the five Philadelphia facilities across all variables, but I was hesitant to sacrifice so much data granularity. Philadelphia accounts for over one-fifth of the state's total incarcerated population for jails, with each Philadelphia facility holding more people than many of the smaller counties. So instead I collected the jail into two separate datasets, with one holding only the variables which are split across Philadelphia facilities and the other with all variables aggregated to the county level. I proceeded to analyze these, using the more granular dataset when possible to compare different facilities and using the other dataset to compare at the county level otherwise.

But as my research progressed, I received more information which led me to ultimately treat Philadelphia as a single facility. I learned from PPS that the Philadelphia facilities are geographically very close, essentially being different parts of a larger complex. Additionally, I noticed that the admission and discharge statistics for Philadelphia were very strange, with some facilities admitting many more people than were leaving and others doing the opposite. This suggested that between facility transfers were not being recorded as you would expect if the five facilities were reported separately. I saw a similarly high variance between Philadelphia facilities in some other variables, suggesting that Philadelphia reporting was inconsistent in treating itself as five small facilities or one large one. After consultation with PPS, I decided to treat Philadelphia as one facility in my final analyses.

Of the other obstacles I faced in preparing the data, the biggest was handling changes between years. Across the years that I looked at one Philadelphia facility closed, variables were added and reworded, a new tab about pre-incarceration populations was added, and the variables for which Philadelphia facilities were aggregated changed. This made it clear that I needed to make my code flexible and receptive to different input structures, to handle these changes as well as to anticipate future ones. To this end I unified the data structures to be consistent between years and allowed for new variables and observations to be added easily. While my code likely won't account for all future changes, I included thorough comments explaining my code to make future modifications easier.

Next I began exploratory data analysis, trying to understand the general structure and significance of the data. Across all 62 counties and 4 years I had data on, I focused on these areas: demographics, staffing and funding, deaths, uses of force, mental health commitments, admission/discharges, and pre-sentencing populations. My initial analysis was primarily graphical, using simple plots and choropleth maps to explore the geographic and relational landscape of different variables. I don't have much personal experience with jails or the landscape of Pennsylvania, so this was a good opportunity for me to learn the general context of the data, as well as to look for obvious outliers and trends. One common problem in all these analyses was the huge differential in population size between counties, with the smallest incarcerated population being less than 1% of the largest. The small facilities had much more volatile measures and very little significance when compared to all of Pennsylvania. I decided to group the smaller half of the counties together when comparing between facilities. There is no obvious cutoff point between big and small facilities, so after consulting with PPS I chose a cutoff that would create a group of counties that held fewer incarcerated people than the largest county, Philadelphia. The new group, which I labeled as 'small counties', contained 37 of the 62 counties and ~17% of the total incarcerated population. This allowed me to compare larger facilities with a more meaningful composite of smaller ones.

I also used this time to start thinking about what questions this data could answer, and what other datasets I might need to support my analyses. I found demographic data on the general population of Pennsylvania, and used it for comparison with the jail population, specifically regarding race. I explored the rest of the PA DOC website, the source of the jail data, to see what other information could be gathered. Regrettably, the data available on non-jail correctional facilities, including prisons, was much more limited, being mostly limited to demographics. This is the point where I decided to focus my analysis specifically on jails, because there were so many important measures reported for jails but not for prisons and other facilities.

As my analysis continued, I ran into problems with several variables which couldn't be addressed with the data alone. One county was a suspiciously high outlier on mental health commitments, making me question the validity of comparison with other facilities. The two primary datasets I was analyzing had different population measures, with one marked as average daily population and the other average monthly population. While the measures were roughly the same for each county, there were slight inconsistencies with no obvious pattern or explanation. When using variables from just one dataset I defaulted to using the population measure from that dataset, but when using both sources I favored the average daily population measure as it seemed more precise. I received input from the Claire Shubik-Richards, Executive Director of the PPS, that it is common for incarcerated people to be dealing with multiple charges at different stages of the correctional process. An incarcerated person could, for example, be serving time for a committed crime while also being in trial for another. This and other potential issues such as facility transfers brought into question the interpretability of the admissions, discharges, and pre-sentencing population variables. We approached the PA DOC with questions about all of these issues but received a limited response.

After communicating with the PPS about their priorities, I decided that the primary goal of my research would be to produce a comprehensive overview of mortality in PA jails. Death is the most extreme observed outcome, and it can also be indicative of other problems. Mental health problems, for example, could manifest as a higher suicide rate and underfunding could result in insufficient medical care. The initial investigation on deaths is included in a memo currently under review that will be released by the PPS for public consumption; further work, including the final section presented here, will later be fully detailed in an academic paper. The analysis has three main parts: a state overview, a cause of death analysis, and a comparison with the general population.

The state overview covers the jail population, incarceration rate, and death rate for each county. A significant complication, however, is that deaths are a very rare event and the PA DOC website only offers five years of mortality data. This causes the less populated county jails, some of which have less than one hundred people incarcerated on an average day, to have highly volatile mortality rates. Even heavily populated jails like Philadelphia's saw a significant variation from year to year. To better understand this, I estimated confidence intervals for the annual death rate of each facility by assuming deaths were independent and identically distributed binomial variables as shown in figure 1. While not a realistic assumption, it served to give a rough, conservative estimate of the significance of the observed rates. Because the intervals proved to be so wide, I decided to focus on death rates averaged across all five years of data and to focus primarily on large counties and state-wide statistics to minimize the randomness in my findings. For more detail, see figure 11 in the appendix with death rates per facility per year.

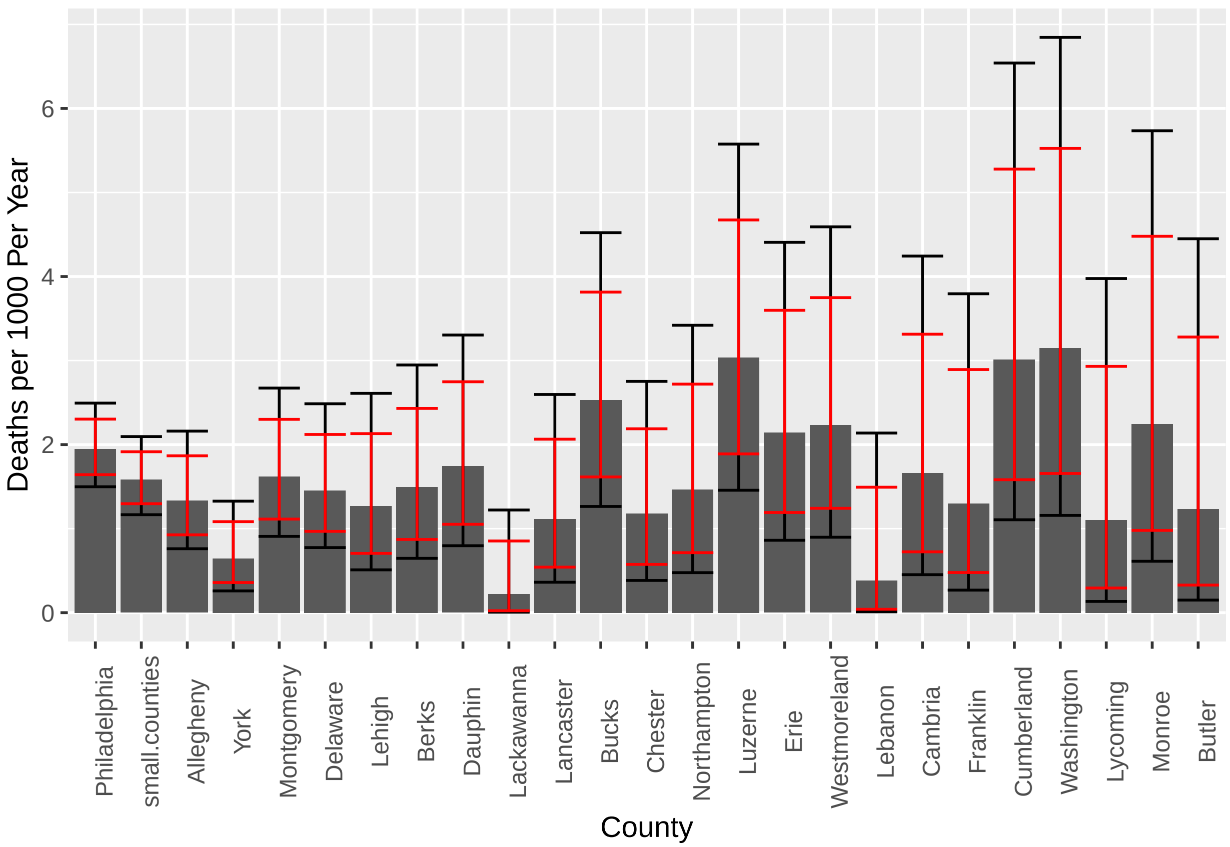

Figure 1: Shows death rates by county (averaged 2015-2019), with bars for the 80% and 95% confidence interval estimates. Note that many of the smallest counties are grouped together into the 'small.counties' bar. Note that most of the intervals are very wide, showing that the death rates are highly variable.

The cause of death section investigated the ten largest counties and the state overall, looking at the four reported causes of death: natural, suicide, accidental, and homicide. One major concern communicated to me by the PPS, my advisors, and the Chief Medical Examiner for Allegheny County, Karl Williams, is that deaths reported as natural might be the result of negligence or improper care, such as refusing to give an incarcerated person a necessary medication or ignoring the signs of a treatable illness. Homer Venters, the former chief medical office for New York's jails, termed the phrase "jail-attributable" death to refer to cases such as these[13]. Uncertainty about cause is of particular concern in smaller counties, where cause of death designations are made by a coroner who usually is not a doctor. To account for this, I needed to better understand what a 'normal' number of deaths would be for each cause of death designation. This was difficult, not only because jail is a very unique environment but also because incarcerated people differ in several ways from the general population. Dealing with this problem became the focus of the third section of this analysis.

---

[13] Venters, Homer. Life and Death in Rikers Island. Baltimore: Johns Hopkins University Press, 2019.

To predict what death rates a population similar to the incarcerated population would experience outside of jail, I adapted the concept of an excess death analysis. Specifically, I used mortality data from the general PA population and controlled for demographics to match the jail population. I obtained detailed mortality data from the CDC Wonder 'Underlying Cause of Death' database[14], though at the time the dataset limited me to only analyzing three years (2015-2017). Specifically, I controlled for age, sex, and race. Demographic data in the jail dataset was recorded on a single day, so I scaled it to match the average monthly population to make it more representative of the whole year. While other significant factors that are likely associated with both death and incarceration were not available, like income and pre-existing conditions, these controls are enough to give a roughly realistic point of comparison for jail mortality, aside from differences resulting from living in jail itself. However, limitations in data required that several assumptions needed to be made to control for these variables. These assumptions are:

1. Hispanic origin is considered an ethnicity by the Department of Health but a race by the county jails. I assume here that individuals marked having Hispanic ethnicity and race are identical with regards to mortality rates.
2. The jail dataset contained demographic data on all combinations of gender and race with age data recorded separately, so I don't know the gender/racial makeup of a specific age range. I assume here that the gender/racial makeup of all age ranges are identical, so for example: if a fifth of the total jail population is black then I assume that a fifth of the population in each age range is black.
3. I assume here that the jail population under 20 years old is best represented by the age range 15-19 in the general public. Similarly I assume that the jail population over 55 is best represented by the age 55-65 demographic.
4. Our jail demographic data is based on the population on a specific date, namely Jan. 31 of the next year. I assume here that the population at this date gives a reasonably unbiased representation of the average jail demographics.
5. Around 1.5% of the general population is marked as being of two or more races. Because the jail data doesn't record this category and the proportion of the general population is so small, I distributed this 1.5% proportionally among the single race categories. I assume here that doing so will not significantly misrepresent the Pennsylvania population on the topic of mortality.
6. The jail demographic data counts people at multiple level of granularity, such as counting the total number of women in a facility and also counting the number of women in each age group in a facility. It seems clear that the more granular variables should sum to the aggregate, i.e. the number of women

---

[14] Centers for Disease Control and Prevention; Underlying Cause of Death 1999-2017; generated using WONDER http://wonder.cdc.gov , (February 2020)

in a facility equals the sum of all women in each age group. However this is not always the case, with a quarter of facilities reporting gender counts which are either above or below the sum of that gender across all age groups. I assume here that the more granular data is accurate.

Because these assumptions are likely to introduce additional variance into death metrics which are already highly variable, the excess death analysis focused on state-level death rates rather than on specific counties. I plan to investigate whether these assumptions lead to significant bias through a sensitivity analysis, while working to put this into a paper in the future. Lastly, these assumptions are necessary for the construction of death rate categories that match the demographic categories of the jail data, but they are only sufficient for claiming that a population with the described characteristics would experience those death rates. They are not sufficient to make stronger claims such as that incarcerated people would experience these death rates outside of jail, because there are a number of confounders not addressed here which connect a person being in jail with their death rate.

One additional obstacle to this analysis was the suppression of small values in the CDC Wonder dataset. To protect individuals' privacy, the data would suppress any death count lower than ten, limiting the granularity of data that I could get. This forced me to carefully choose which variables to control for, so that the combination of categories wouldn't be so specific as to produce death counts below ten. I decided to use data aggregated across the whole state and all three years of data to alleviate this problem, though that restricted my ability to consider trends across counties or years. I was also forced to combine the racial groups American Indian or Alaska Native and Asian or Pacific Islander into a single 'other race' category, as they were all very small populations in Pennsylvania. Causes of death were listed with high specificity using the ICD-10 standard, so I grouped them into five categories: internal cause, overdose, accident, suicide, and other. Internal cause being the closest comparison to the natural death category in the jail data, and the other category including homicide as well as other rare or hard to categorize causes of death. No non-binary categories of gender were present in either dataset. Here is a list of the variables I controlled for, as well as the values these variables could take on:

Gender: Female, Male

Race: White, Black, Hispanic, Other

Age: 24-and-under, 25-34, 35-44, 45-54, 55+

Cause of Death: Internal, Overdose, Accident, Suicide, Other

Even with these concessions made regarding granularity, a small portion of the data was still suppressed. This was largely due to the race variable, for which the 'Hispanic' and 'other' groups were very small. I decided to impute the suppressed values rather than sacrifice any further granularity. Because I have access to data at multiple levels of granularity, I knew I could estimate the true value of the suppressed data with reasonable accuracy. However I was unable to find any common imputation practices which could leverage this kind of data, so I made a simple, novel imputation method for this. This method is definitely not optimal, but based on the small portion of the data it's being applied to I felt that simplicity was more important than perfect accuracy, as long as the method was not significantly biased. And although I created this method on my own, it's so simple that I expect it has been used in other forms many times, and I just failed to find it while searching.

The method I used to impute values was essentially to assume independence between some variables, and use that to mathematically calculate an estimator for the desired value. The first stage of imputation I did was to assume no interaction between race and gender, which allowed me to estimate most of the suppressed values. For the remaining suppressed values I proceeded to the second stage, which replicated the process but with the stronger assumption that race had no interaction with either age or gender. The exact calculations I made are shown in the Imputation Methods section of the appendix.

Across all combinations of demographic factors and cause of death groups in the general population, there were 200 different death rates in total being used with one for each combination of gender, race, age, and cause of death. For stage one imputation, 27 of those death rates were imputed, with those categories accounting for .06% of the total death rate of the comparison population. The second stage imputed 6 death rates, which accounted for .01% of the total death rate. Note that while more than 15% of the death rates were imputed, these accounted for a very small proportion of the total deaths because only the least common causes of death and demographics required imputation. To test the accuracy of this method, I estimated the death rates of all unsuppressed values using the same process and compared that to the true values. The only parameters of the model are my choice of which variables to assume independence between, so there's no risk of biased results from overfitting. The average absolute error rates for each demographic / cause of death category were 8.2% (stage one) and 20.9% (stage two), with the total death error rates being .05% and 1.07% respectively. Given these low error rates and the small portion of the dataset being imputed, I decided that my method was sufficient for this analysis. Once I had the death rates for each demographic combination, I simply calculated the number of deaths each facility would experience if they had the same death rates as in the general population.

Results

The Pennsylvania jail population and total populations are heavily concentrated in certain areas, with over 50% of incarcerated people living in the seven largest jails. This is shown in figure 2, where most of the state's jail and total populations live in the few dark regions such as Philadelphia and Allegheny counties. Special attention should be paid to Philadelphia, which contains over one fifth of the total reported jail population. Figure 3 shows incarceration rates, which show less variance between counties. Note that some counties with very high and low incarceration rates also have a low jail population, which could be the result of higher variance.

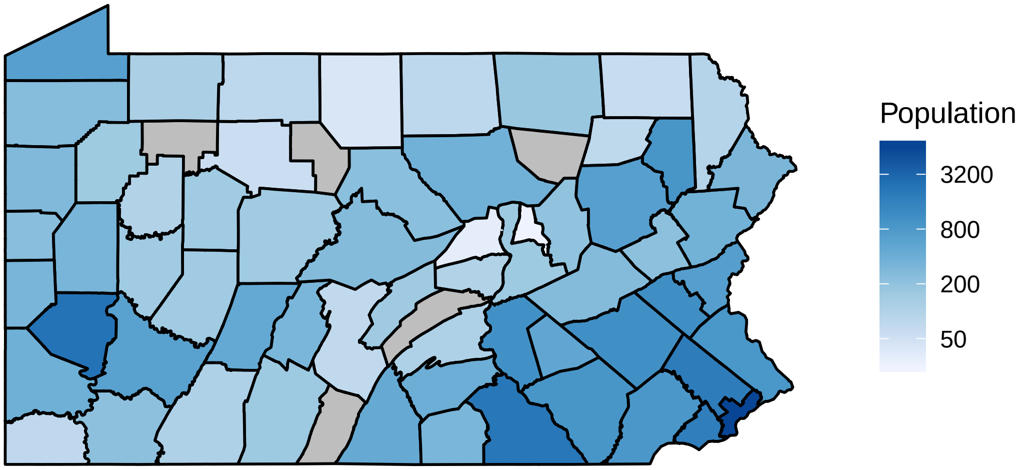

Figure 2: Shows incarcerated jail population within each county. Note that the color scale is logarithmic. 5 counties without their own jail facility are marked in gray.

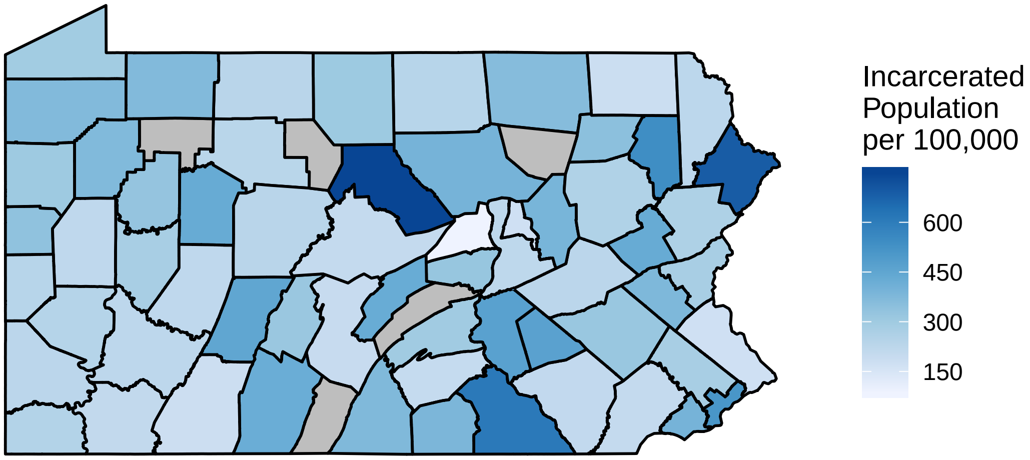

Figure 3: Jail incarceration rate per 100,000 people in the general population. Note that some several counties have less than 100,000 inhabitants, so the incarceration rate will exceed the real incarcerated population.

The incarcerated population in Pennsylvania jails is heavily skewed towards younger ages, though there are still a significant number of people above fifty five as shown in table 1. There's an even stronger disparity of gender, with the large majority of incarcerated people being male. While the racial categories used in the jail datasets are broad, they show that the large majority of people are either white or black, along with a smaller Hispanic group.

Age	Under 18	18-19	20-24	25-29	30-34	35-39	40-44	45-54	55+.
Male	134.4	1033.4	5189	5645.6	4503.8	3436.2	2398.6	3183.4	1450.8
Female	5	107.8	705.8	1019.8	894.2	637.6	412.6	528.8	160.2

Table 1: Shows the average population of jails in Pennsylvania, split by age and gender. All counts are averaged across all counties and five years of data (2014-2018). Demographic counts are recorded on only one day each year.

White	Black	Hispanic	Other
15653.8	12236.8	3752.4	522.4

Table 2: Shows the average population of jails in Pennsylvania, split by race. Racial groups are presented as in the data. All counts are averaged across all counties and five years of data (2014-2018). Demographic counts are recorded on only one day each year.

Using census data, we can see that jail incarceration rates vary by gender and race in Pennsylvania. As shown in figure 4 (left), men are more than five times as likely as women to be in jail. Racial disparities are also apparent in the jail population, as shown in Figure 4 (right). We can see that African Americans are more than 4 times as likely as whites to be in jail, and Hispanics are more than twice as likely as whites to be in jail.

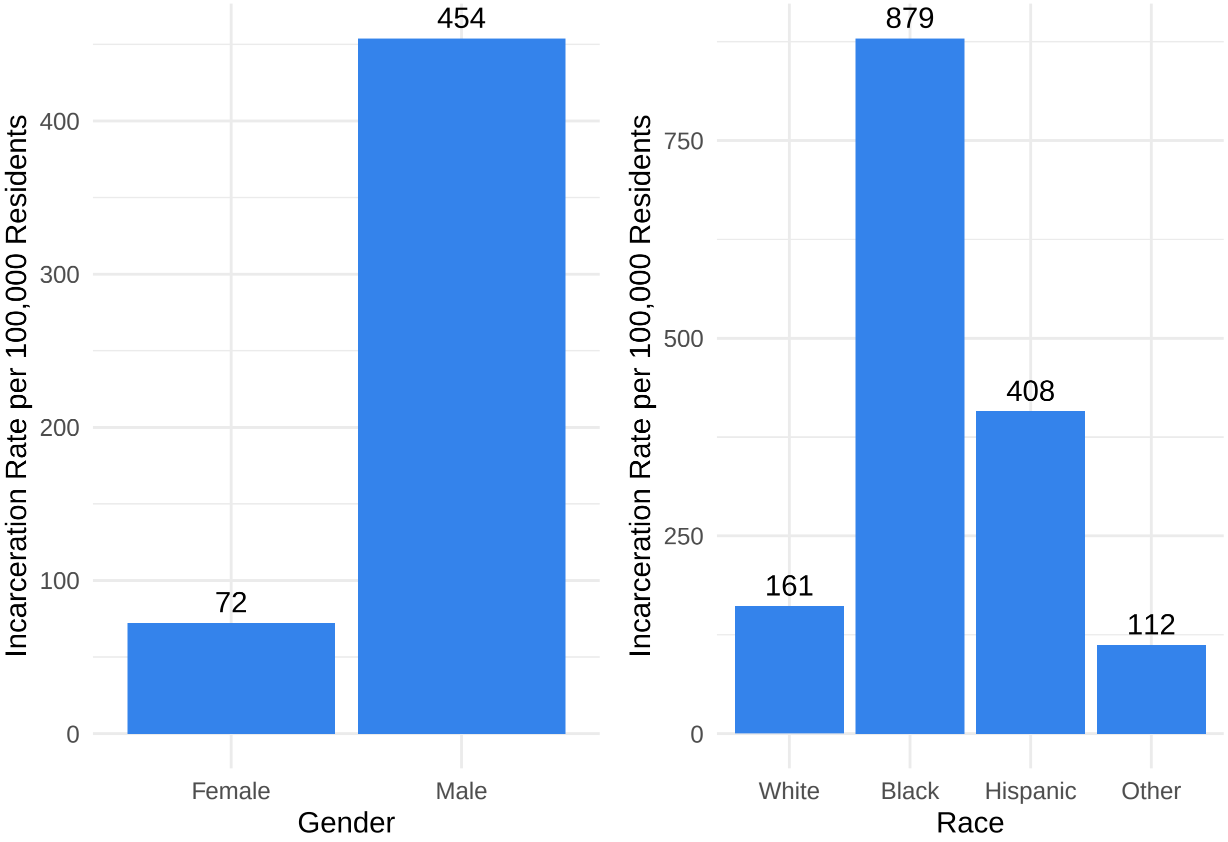

Figure 4: Jail incarceration rate per 100,000 Pennsylvania residents, split by gender (left). Jail incarceration rate per 100,000 Pennsylvania residents, split by race (right).

Exploring the Data

My initial analyses were focused on understanding the context and significance of different variables, as well as looking for interesting interactions. I first looked at demographic and staffing differences between jails. The most significant interaction I found was with population, which predicts variance in demographics and staffing as shown in figure 5. We can see that the size of a jail's incarcerated population is negatively correlated with the staff to incarcerated population ratio, suggesting that jails are able to operate more efficiently at scale and/or larger jails are understaffed. Larger jails are also clearly more racially diverse, as we would expect given that jails house primarily local populations and rural areas in Pennsylvania are predominantly white and sparsely populated.

I found similar trends in staff composition, with bigger jails having fewer part-time workers, slightly more treatment staff, and slightly less administrative and support staff. The proportion of security staff, which make-up 70% of all staff overall, varies wildly in the range of 45-90% but isn't correlated with population size.

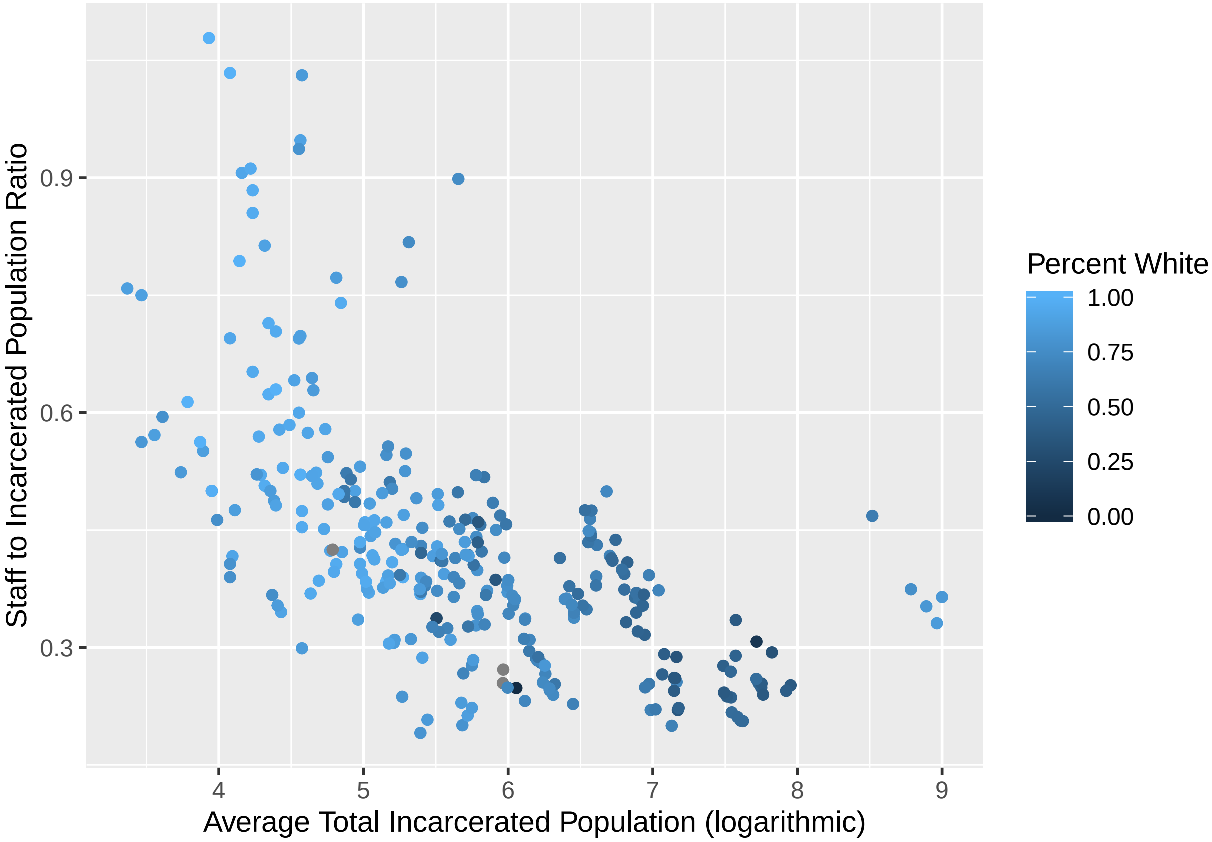

Figure 5: Ratio of jail staff to incarcerated population plotted against the logarithmic incarcerated population count and colored according to racial demographics of the incarcerated population. Each county projects five dots, one for each year of data 2014-2018. The population scale is logarithmic to account for the large differences between county populations. Note the clear negative correlation, indicating that heavily populated jails have fewer staff per person. Also note that more populated jails hold a higher percentage non-white people, and so racial diversity is negatively correlated with staff ratio.

Next I examined overcrowding and off-site housing. Across all counties the average bed capacity usage ranges from 36% to 119%, with 89.2% of facilities being between 50% and 100% usage. We can see in figure 6 that overcrowding is a common problem for several counties. The dataset only recorded population as an average for the year, but 11 counties incarcerated more people on the average day than they had beds to put them in. Considering that jail populations shift fairly quickly, this suggests that these jails were probably heavily overcrowded at several points throughout the year. Furthermore, these counties (except for Berks) were all regularly near or above full average capacity. It is not clear from the data what exactly the impact of overcrowding is or what procedures are in place to handle it, but it is definitely a cause for concern. Population counts were split between on-site and off-site populations, with 95% of the population being housed on-site. There was no obvious correlation between off-site housing and overcrowding, suggesting that it is not used to handle overcrowding. PPS believes that off-site housing is not likely to be worse than on-site, so I didn't pursue further.

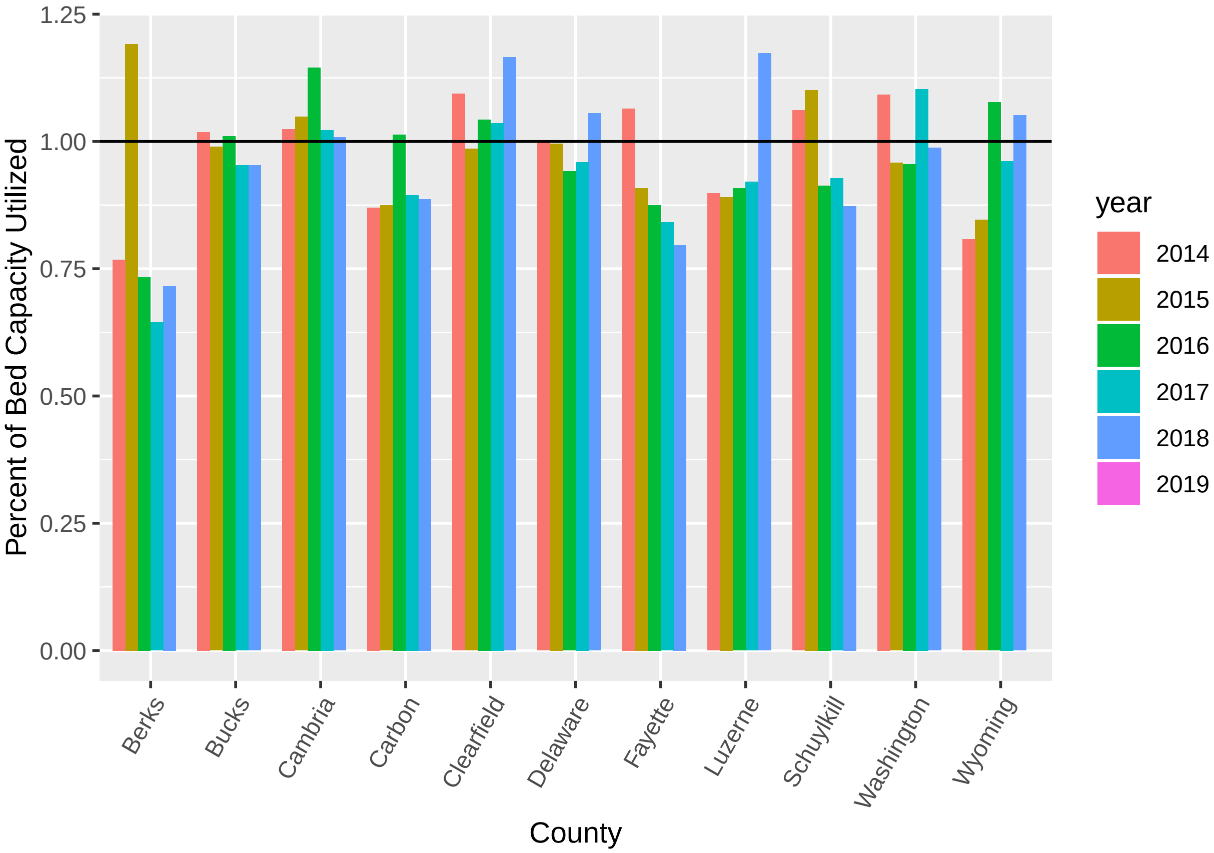

Figure 6: Average percentage of bed capacity used on site for each county, as calculated by dividing the total bed count for the year by the average on-site incarcerated population. Only shows counties which exceeded 100% average daily bed usage in at least one year.

Upon looking at mental health commitments, it was immediately apparent that there were concerns with the data. Most mental health commits were reported from Philadelphia, which reported per person commitment rates that were twenty times higher than the rest of the county for short-term commits and forty times higher for longer term commits. While Philadelphia was a huge outlier, other counties had much more stable numbers. It seems very strange that we would see a lone outlier of such magnitude, raising suspicion that the difference could be the result of different reporting practices, clerical error, or some other unforeseen circumstance. We intend to contact Philadelphia county to resolve this issue.

Death Rates

Mortality rates are obviously an extremely important topic when studying any population, but they can also be indicative of other important factors such as healthcare and security. Additionally, while those other factors can be difficult to quantify and study, deaths are easier to measure and interpret. Thus I decided to focus the rest of my analyses on death rates and different causes of death.

The yearly death rate of incarcerated people is displayed for each county in the figure 7. These rates can be difficult to interpret, especially for jails with a small population. We have data on the whole population of PA jails so there is no uncertainty from sampling, but death rates will fluctuate from year to year even if a facility is run exactly the same way. So to isolate the impact of variables we care about, jail conditions and policy, we have to account for this base level of randomness. With large datasets this randomness averages out, but it has much more impact when looking at a small county. Because deaths are rare, one or two additional deaths could double the death rate from one year to the next. For example: Potter County only had two deaths from 2015 to 2018, but also had the highest county death rate of 12.5 per thousand because their incarcerated population is only 40 people on average. In small counties like this, it is difficult to know whether they are especially dangerous or are just experiencing a spike in deaths due to chance.

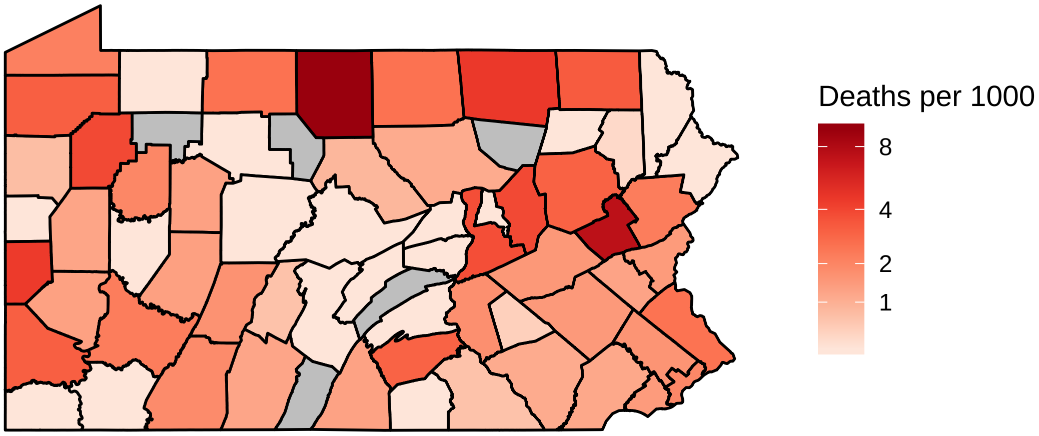

Figure 7: Shows deaths in jails per 1000 incarcerated people for each county. Note that the color scale is logarithmic. 5 counties without their own jail facility are marked in gray.

Looking at different causes of death in the incarcerated jail population in Figure 8, the most common causes were natural death (54%) and suicide (40%), whereas cases of accidental death (4%) and homicide (2%) were rare. The total death rate seen in this dataset (149 per 100,000) was higher than the national average jail death rate from 2000-2016 (139 per 100,000)[15], a difference which is more than explained by the suicide rate (59 per 100,000) being 37% higher than the past national average (43 per 100,000). Because we do not have national data past 2016, I cannot ascertain whether this difference indicates that Pennsylvania has a relatively high suicide rate or that suicide rates are rising nationwide.

Shown in figure 8 are the death rates of 10 counties with the largest jail populations, which together hold almost half the state's total jail population. Some facilities have significantly lower death rates, suggesting that others have room for improvement. Also note that Philadelphia, which houses over a fifth of the jail population, has a relatively high death rate. Most of those deaths were reported as natural, but that says little about the circumstances of the death. Cause of death designation is made with minimal oversight by the local coroner who is usually not a doctor, though larger counties like Philadelphia and Allegheny are more likely to have a qualified medical examiner. As a result, we cannot be certain that a death reported as 'natural' is not the result of negligence or improper care, such as refusing to give an incarcerated person a necessary medication or ignoring the signs of a treatable illness. Drug and alcohol related deaths, for example, are the third most common cause of death in jails nationally[16] and yet there's no indication of how common they are in Pennsylvania because there is no category for overdose and no explanation of where they fit. While we don't have any data to indicate how common these cases are, they fall under the accidental death category in the ICD-10 coding scheme so they may be reported that way here. However, this seems at odds with the very low accident rates (6%) we see in the data.

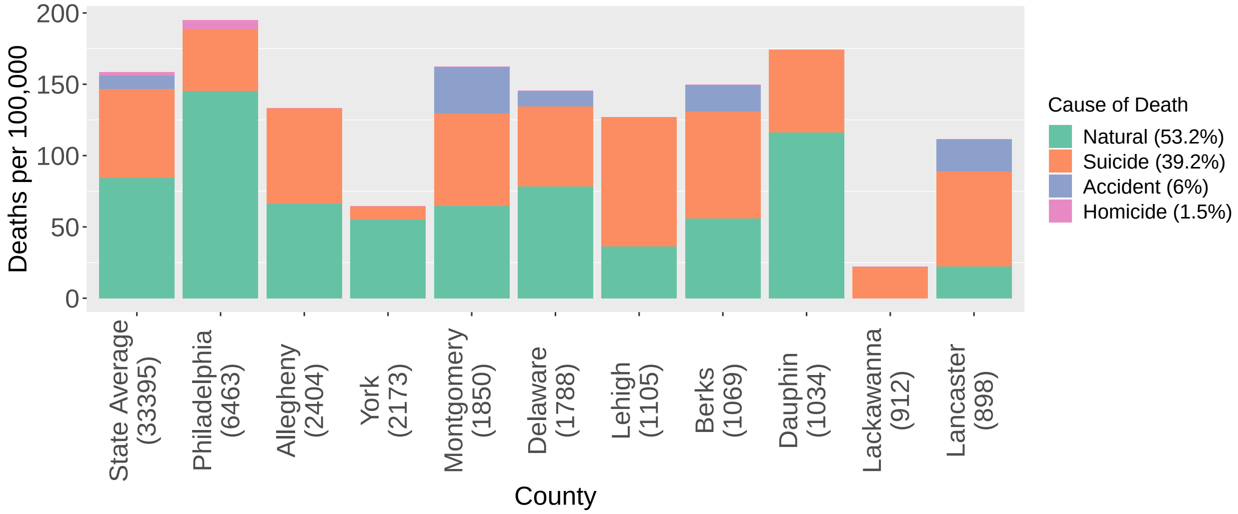

Figure 8: Deaths rates per 100,000 for the ten counties with the largest jail populations next to the state average, by reported cause of death. Next to each county name is the incarcerated population of that

---

[15] Margaret E. Noonan, U.S. Bureau of Justice Statistics; Mortality in Local Jails, 2000-2016 – Statistical Tables; February 2020 https://www.bjs.gov/content/pub/pdf/mlj0016st.pdf

[16] Margaret E. Noonan, U.S. Bureau of Justice Statistics; Mortality in Local Jails, 2000-2016 – Statistical Tables; February 2020 https://www.bjs.gov/content/pub/pdf/mlj0016st.pdf

county's jail. Note that all county jail populations are lower than 100,000, and so all death rates shown here are higher than the real death counts.

To better understand death rates in jails, I estimated the death rates of a similar population in the general population. Specifically, I calculated death rates for non-incarcerated Pennsylvanians of the same age, race, gender profile as the incarcerated population. By controlling for demographic factors, I estimate death counts for a hypothetical comparison population which I use as a reference point to better understand how incarceration is correlated with mortality. The purpose of this analysis is to estimate the amount of 'excess death' which results from jail incarceration by attempting to compare two populations which only differ by whether or not they are incarcerated. Ideally, we would like to know the death rate in the same individuals were they not incarcerated, or in an exchangeable population. Due to data limitations, my estimates of 'excess death' cannot account for other relevant variables such as income or mental health status, but it improves on a simple breakdown of jail deaths or age-adjusted rates.

Deaths in Pennsylvania jails have a very different distribution of causes than the comparison population, which can be seen in Figures 9 and 10. The death rate in the comparison population was 2.82 deaths per thousand, which is over twice the death rate in jails of 1.39 per thousand. It is encouraging to see that incarcerated people are safer than their non-incarcerated counterparts, though not surprising due to the inherent safeguards and limitations on risky behavior that come with living in a controlled facility, such as not getting into car accidents.

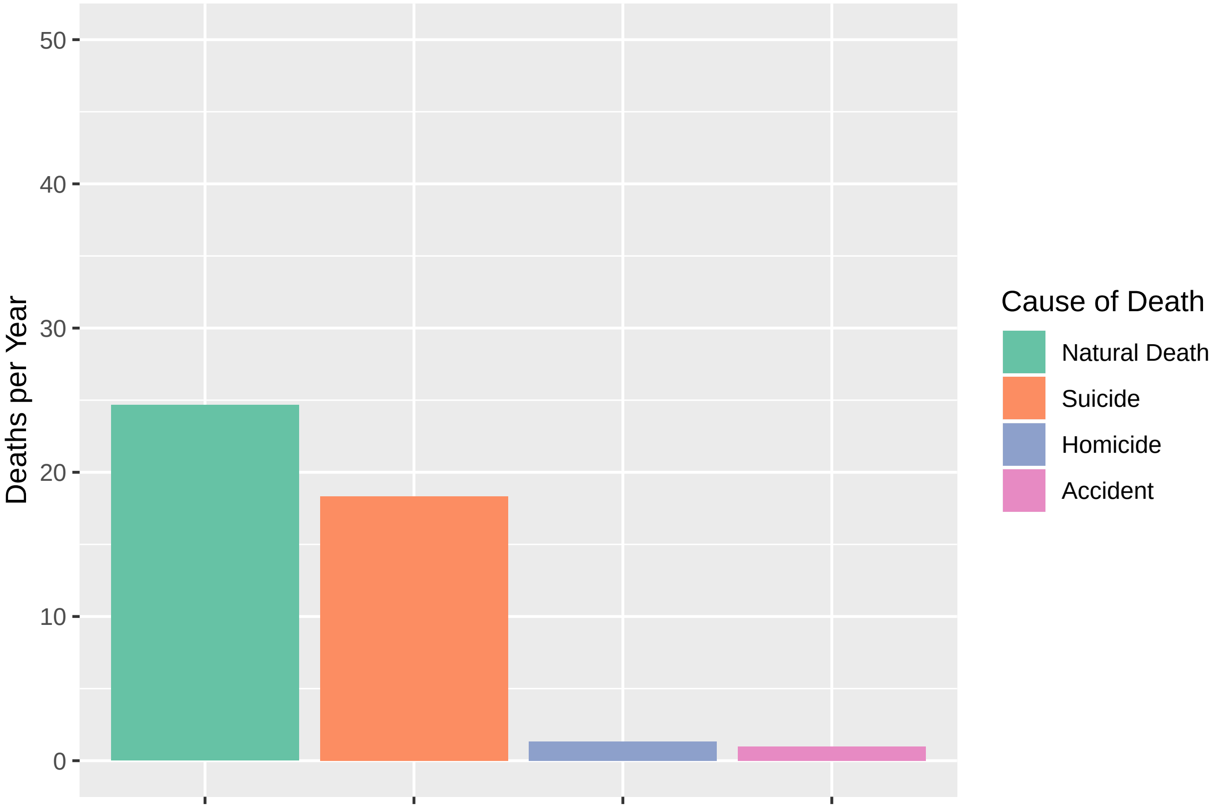

Figure 9: Average deaths per year in jails(2015-2017), broken up by cause of death.

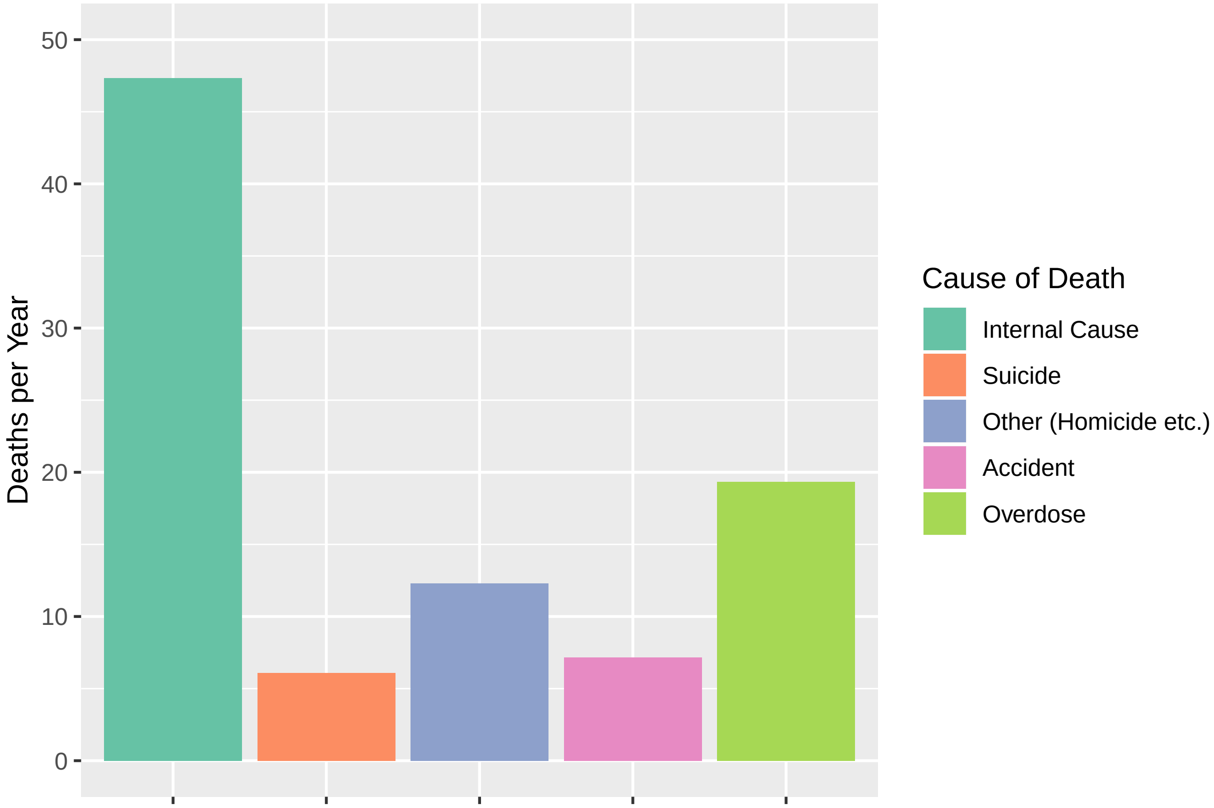

Figure 10: Estimated average deaths per year for the comparison population(2015-2017), broken up by cause of death. Note that the ordering and coloring is chosen for ease of comparison with Figure 9. 'Internal Cause' is the closest comparison to the 'Natural' deaths reported in jails, and the 'Other' category here includes homicides. Overdose deaths are in a new category here because they are unfortunately common and do not clearly fit into any of the listed categories of jail deaths.

Both accidental deaths and homicides are much lower in jails, as expected. However, the suicide rate for incarcerated people (.56 per thousand) was almost triple that of similar non-incarcerated people (.19 per thousand). Interpreting this difference is difficult, because there are reasons to expect the jail suicide rate to be both higher and lower than in the general population. People in jail are much more likely to be in a crisis or to suffer from mental health problems, both of which would increase the likelihood to attempt suicide. But like homicides, suicides should be difficult to carry out in a controlled facility. Given that jails seem to successfully prevent most homicides, it is strange that they have so much difficulty preventing suicides. The last cause of death in the general population is overdose, which was the most common other than internal causes. While this is typically included as a type of accident, I felt that overdoses were too common and distinctly different from other accidents to be counted together. It is not obvious what cause of death an overdose would be reported as in the jail dataset, but the high overdose rate in the comparison population demonstrates that drugs must have a significant impact on the lives of incarcerated people. According to the BJS, the national jail death rate from drug or alcohol intoxication was 15 per 100,000 in 2016, the third highest cause after illness and suicide[17] . These datasets have no ability to diagnose the prevalence of drugs in Pennsylvania jails, but even if they are not present there is still the problem of people suffering from withdrawal and drug related health issues. I can't assess the performance of jails addressing these issues with this data, but it is definitely an area of concern and warrants further investigation.

Conclusion

My research has uncovered several noteworthy findings about jails in Pennsylvania. Using jail population data I found that several jails are chronically overcrowded, likely impacting the health and wellness of incarcerated people. Additionally, jail incarceration rates varied significantly between counties, suggesting differences in crime, policing, and/or charging and judicial systems. My most significant and concerning finding was the high rate of suicides in Pennsylvania jails. High suicide rates are a problem in jails across the country, but this is the first analysis to look at the problem by facility and focus on Pennsylvania, which has a higher suicide rate than the national average. The mortality benchmarks I produced suggest that suicides in jails are much more common than we would expect from similar people in the general population. While the benchmarks are limited by data limitations, their results add a new analysis to the discussion and are significant enough to warrant further thought and research. It is clear that suicides are the number one cause of preventable deaths in jail, and there is a huge potential for health improvements by addressing this issue. In the Rand report "Caring for Those in Custody"[18], interviewed experts recommend that jail suicides could be reduced by increasing mental health funding, improving suicide risk assessment, and spreading mental health best practices between facilities.

Concerningly, limitations of the data stopped me from calculating the number of different people going through jails, as well as their lengths of stay and the proportion who were awaiting sentencing. These problems, as well as others like the inconsistency in mental health data, need to be resolved to more fully understand the state of jail incarceration. Another concern is that 'jail attributable deaths' might be reported as natural, but questions about this issue cannot be answered with these datasets alone. Through cooperation with the state of Pennsylvania, my fellows at the Center for Human Rights Science will attempt to address these issues in the future.

---

[17] Margaret E. Noonan, U.S. Bureau of Justice Statistics; Mortality in Local Jails, 2000-2016 – Statistical Tables; February 2020 https://www.bjs.gov/content/pub/pdf/mlj0016st.pdf

[18] Russo, Joe, Dulani Woods, John S. Shaffer, and Brian A. Jackson, Caring for Those in Custody: Identifying High-Priority Needs to Reduce Mortality in Correctional Facilities. Santa Monica, CA: RAND Corporation, 2017. https://www.rand.org/pubs/research_reports/RR1967.html.

My analysis constitutes a preliminary but significant look at these jail datasets, which have still more variables available for analysis. There is significant potential for more research with this data, which I hope CHRS and others will take advantage of in the future. Analyses like these are essential for the effective oversight of incarceration, so that government and researchers can work together to ensure the fair and humane treatment of incarcerated people.

Appendix

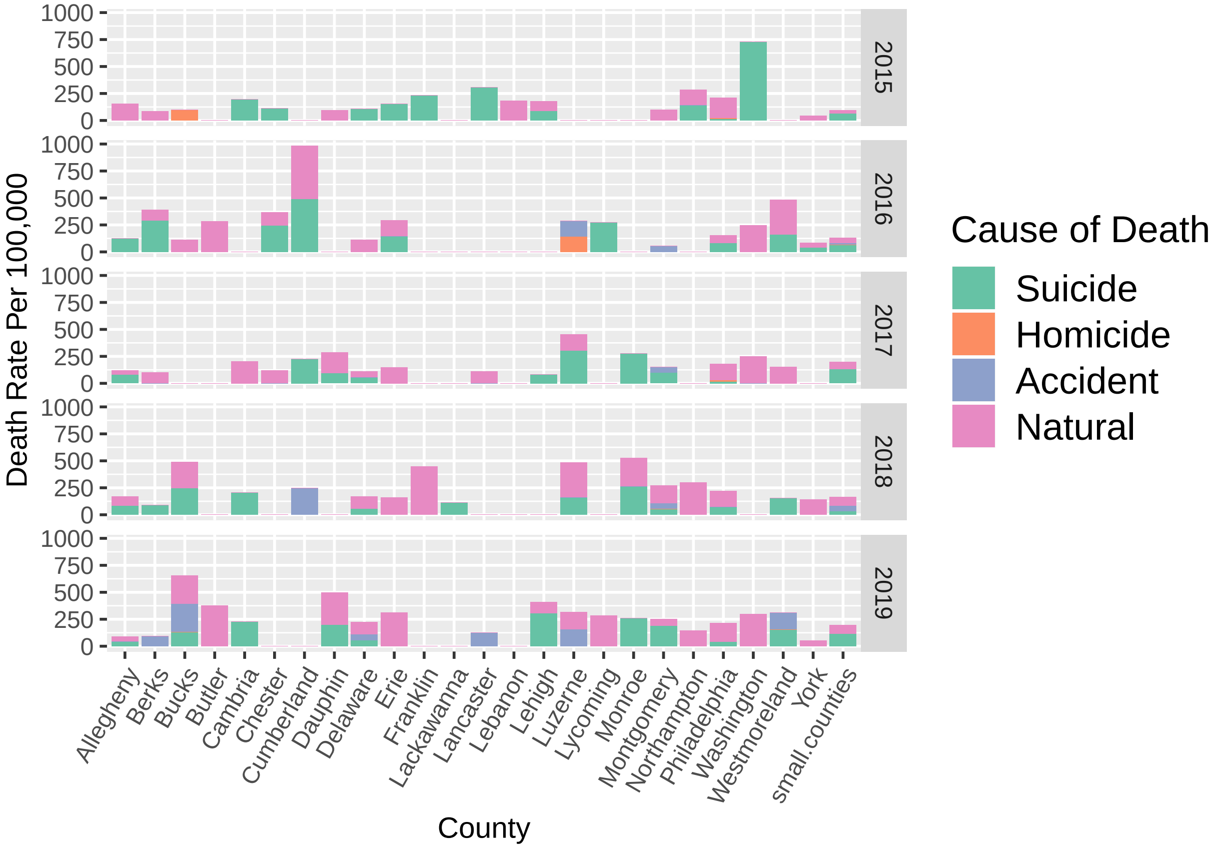

Figure 11: Shows death rates per 100,000 per facility per year for all five years of data 2015-2019, broken up by cause of death. Note that many small facilities are grouped together as 'small.counties'. Notice how death rates vary significantly between years, even within a single facility.

Imputation Methods

Stage 1

Let D_{carg} be the death rate of a population given a cause of death c, age range a, race r, and gender g. And let $M_{x|y}$ be the multiplicative impact of x on D_y, or in other words $M_{x|y} = \frac{D_{xy}}{D_y}$. So for example: $D_{cag} = D_{ca} * Mg|ca$.

For the first stage of imputation, we assume that gender and race are not correlated, so the impact of gender on a death rate is independent of the race and the impact of race is independent of the gender. Thus we have:

$$
\begin{aligned}
& D_{carg} && \text{The value we want to impute.} \\
&= D_{ca} * M_{r|ca} * M_{g|car} && \text{By defn. of M} \\
&= D_{ca} * M_{r|ca} * M_{g|ca} && \text{Assumption of independence} \\
&= D_{ca} * M_{r|ca} * M_{g|ca} * D_{ca}/D_{ca} && \\
&= D_{car} * D_{cag}/D_{ca} && \text{Simplify}
\end{aligned}
$$

Because we have sample estimates for D_{car}, D_{cag}, and D_{ca}, we can now estimate D_{carg}.

Stage 2

For the second stage of imputation we use the same notation but with a stronger assumption. We now assume that race is independent of both age and gender.

$$
\begin{aligned}
& D_{carg} && \text{The value we want to impute.} \\
&= D_{cag} * M_{r|cag} && \text{By defn. of M} \\
&= D_{cag} * M_{r|c} && \text{By assumption of independence} \\
&= D_{cag} * M_{r|c} * D_c/D_c && \\
&= D_{cag} * D_{rc}/D_c && \text{Simplify}
\end{aligned}
$$

Imputation Methods: Calculates the imputation formulae which are used to impute suppressed death rates from the CDC Wonder dataset.The death rates were then multiplied by the corresponding general population count (which was never suppressed) to get an estimate of the deaths in the corresponding group. The general death count estimates were used while testing the suppression methods, but only the death rate was used in the actual imputation process.