Spatial Investigation and Exploration of Factors influencing HIV-related Mortality

United States, 2020

Hayley Hemme

4/17/23

Warning: package 'knitr' was built under R version 4.2.2

Warning: package 'readxl' was built under R version 4.2.2

Warning: package 'lubridate' was built under R version 4.2.2

Warning: package 'stringr' was built under R version 4.2.2

Warning: package 'tidyverse' was built under R version 4.2.2

Warning: package 'ggplot2' was built under R version 4.2.2

Warning: package 'tibble' was built under R version 4.2.2

Warning: package 'tidyr' was built under R version 4.2.2

Warning: package 'readr' was built under R version 4.2.2

Warning: package 'purrr' was built under R version 4.2.2

Warning: package 'dplyr' was built under R version 4.2.2

Warning: package 'forcats' was built under R version 4.2.2

Warning: package 'here' was built under R version 4.2.2

# 1. Summary/Abstract

*Write a summary of your project.*

# 2. Introduction

Since the emergence of the HIV epidemic in 1981 in the United States, more than 1.2 million people in the United States have been diagnosed with HIV, resulting in more than 700,000 HIV-related deaths [cite]. Medical advancements have led to effective antiretrovirals therapies that not only treat HIV infection and reduce adverse HIV-related health outcomes, such as progression to AIDS, but have also led to use of these medications to prevent infection before and after potential exposure for people who are HIV negative Even still, more than 30,000 people in the United States continue to be diagnosed each year [cite]. While initiatives such the Ending the Epidemic may provide xxx, this is followed by nearly one decade of steadily maintained HIV diagnoses rates, suggesting that existence of effective medical therapy alone will not end the HIV epidemic in the United States given the complex barriers existence among populations who are most vulnerable for HIV infection. The goal of this project is to gain a between understanding of factors associated with HIV-related mortality in the United States in

## 2.1 General Background Information

*Provide enough background on your topic that others can understand the why and how of your analysis*

## 2.2 Description of data and data source

The data used in this project was obtained from multiple publicly available source. Data pertaining to HIV deaths, new diagnoses, HIV-related outcomes, and social determinants of health relating to HIV were obtained from AIDSVu . Population data for on persons of race/ethnicities included in the analysis and total population was obtained from the [U.S. Census Bureau](https://data.census.gov/) . Percentage population make-up by sex was obtained from [KFF](https://www.kff.org/other/state-indicator/) . Percentage distribution of persons under 14 was obtained from Kids Count [Kids Count](https://datacenter.kidscount.org/). All data described was aggregated at the the state-level.

## 2.3 Questions/Hypotheses to be addressed

Question: Can the HIV death rate from 2020 in the United States be modeled by social determinants of health like poverty and/ or by HIV related outcomes (new diagnoses, prevalence PrEP use, etc?) Are these factors correlated with disparities in HIV-related mortality across different racial and ethnic groups? Can a spatial model explain variation in HIV death-rates at the state-level and/or regional-level after accounting for variables associated with sociodemographic, SDOH, and HIV-related outcomes?

H0 : There is not a spatial correlation HIV-related deaths and SDOH or HIV-related outcomes.

Ha : There is a spatial correlation between HIV-related deaths and SDOH or HIV-related outcomes.

# 3. Methods

I used a number of different methods throughout the various stages of this analysis. To ensure that data importation was simple and reproducible, I used library(here). The vast majority of data cleaning processes were achieved using functions within the tidyverse. Some of these processes included renaming, pivoting, selecting, and mutating variables. Additionally, most of the data visualization and maps were obtained using ggplot2, with a few exceptions being packages with functions that are not compatible with the tidyverse and simple model diagnostics that base R. During the data visualization process, it became apparent that the dependent variable would require transformation. Data transformation methods of increasing complexity were performed. Ultimately, a Box Cox transformation was able normalize out dependent variable. A small value was added to each observation to ensure that all were greater than 0. To help assess relationships between variables, I performed a correlation analysis, most using principle component analysis (PCA). Data analysis methods were performed in order of increasing complexity, beginning with simple linear regression, then Poisson regression, negative-binomial regression, and finally a conditional autoregessive model (CAR) to account for spatial dependence. Models were judged for goodness of fit using AIC.

## 3.1 Data acquisition

*As applicable, explain where and how you got the data. If you directly import the data from an online source, you can combine this section with the next.*

## 3.2 Data import and cleaning

All data used in this analysis were obtained from publicly available sources, as note above.

## 3.3 Statistical analysis

Spatial dependence assessed using Moran’s I.

# 4. Results

## 4.1 Exploratory/Descriptive analysis

|  |
| --- |
| Histogram of the distribution of the ratio for observed/ expected cases(raw) |

|  |
| --- |
| Histogram of the distribution of ratio of observed/ expected mortality after Box-Cox transformation |

*Use a combination of text/tables/figures to explore and describe your data. Show the most important descriptive results here. Additional ones should go in the supplement. Even more can be in the R and Quarto files that are part of your project.*

## 4.2 Basic statistical analysis

|  |  |
| --- | --- |
| Black | Multiple |
| Hispanic | White |

*To get some further insight into your data, if reasonable you could compute simple statistics (e.g. simple models with 1 predictor) to look for associations between your outcome(s) and each individual predictor variable. Though note that unless you pre-specified the outcome and main exposure, any “p<0.05 means statistical significance” interpretation is not valid.*

## 4.3 Full analysis

*Use one or several suitable statistical/machine learning methods to analyze your data and to produce meaningful figures, tables, etc. This might again be code that is best placed in one or several separate R scripts that need to be well documented. You want the code to produce figures and data ready for display as tables, and save those. Then you load them here.*

|  |  |
| --- | --- |
|  |  |
|  |  |
|  |  |

# 5. Discussion

## 5.1 Summary and Interpretation

*Summarize what you did, what you found and what it means.*

## 5.2 Strengths and Limitations

*Discuss what you perceive as strengths and limitations of your analysis.*

## 5.3 Conclusions

*What are the main take-home messages?*

*Include citations in your Rmd file using bibtex, the list of references will automatically be placed at the end*

Note that this cited reference will show up at the end of the document, the reference formatting is determined by the CSL file specified in the YAML header. Many more style files for almost any journal [are available](https://www.zotero.org/styles). You also specify the location of your bibtex reference file in the YAML. You can call your reference file anything you like, I just used the generic word references.bib but giving it a more descriptive name is probably better.

# 6. References