Spatial Investigation and Exploration of Factors influencing HIV-related Mortality

United States, 2020

Hayley Hemme

4/21/23

# 1. Summary/Abstract

*Write a summary of your project.*

# 2. Introduction

HIV has a disproportionate impact on marginalized groups, resulting in significant disparities in rate of new HIV diagnoses and poor HIV-related health outcomes, such as late HIV diagnosis, progression to AIDS, and HIV-related death. Previous study have found direct links to indicators of social determinants of health and poorer health outcomes for persons with HIV

Other studies have found an association between different SDOH measures and COVID-19 mortality that varied across racial and ethnic groups and community types

## 2.1 Question:

Can the HIV death rate from 2020 in the United States be modeled by social determinants of health and/ or by HIV-related outcomes (late diagnoses, viral suppression) 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. Background Information

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 antiretroviral 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

# 4. Materials and Methods

## 4.1 Data Source

The data used in this project was obtained from multiple publicly available source.

### 4.1.1 AIDSVu

Data was obtained from AIDSVu on counts and rates of HIV-related mortality by state in 2020. Social determinants of health (SDOH), were also obtained from AIDSVu by state from 2020. SDOH included a number of variable represented as a percentage particularly pertinent to health persons with HIV. AIDSVu is a public resource presented by Emory University’s Rollins School of Public Health in partnership with Gilead Sciences, Inc. and the Center for AIDS Research at Emory University (CFAR) and contains HIV surveillance data and other pertinent data to monitoring HIV-related outcomes in the United States.

### 4.1.2 US Census Bureau

Data on 2020 demographic composition by race/ethnicity and age by state was obtained from the [U.S. Census Bureau](https://data.census.gov/) . A five-year estimate of percent with access to broadband internet by state (released in 2020) was also obtained from the US Census Bureau.

### 4.1.3 Kids Count

Data on the percentage distribution of children under 14 years old was obtained from [Kids Count](https://datacenter.kidscount.org/), an online resource that provides high-quality data on the well-being of children in the United States.

### 4.1.4 Spatial data

Spatial object of the contiguous United States was obtained from library(maps) in R. The list of neighboring contiguous United States was obtained from a public repository on GitHub. https://github.com/ubikuity/List-of-neighboring-states-for-each-US-state

## 4.2 Data processing

### 4.2.1 Cleaning

Data was cleaned to adjust variable format, merge dataframes, subset by race/ ethnicity, and for feature engineering. Several observations in response variables were suppressed due to small counts/ rates were represented by -9, and were re-coded to NA. To removed multicollinearity and to assess the relationship between predictors, principal component analysis (PCA). was used. Variables that had correlation greater that 0.3 in the first two principal components were deemed important. Scaled predictors of the the first three principal components were then used in various regression models multiple linear regression, Poisson regression, and negative binomial. Non-spatial and spatial methods were used.

### 4.2.2 Feature Engineering

The primary response variable in this analysis was HIV-related death rate. Feature engineering of death rate (estimated) was performed using counts of HIV related deaths divided by the estimated population based on US Census data multiplied by 100,000 persons. Initially, persons of multiple racial backgrounds were considered in the analysis, however, significant discrepancies between the death rates provided by AIDSVu and the calculated death rate based on counts and estimated population meant that I could not reliably interpret these data. The decision to used death rate (estimated) instead of death rate (AIDSVu) was made to help ensure that differences across various models that used rates versus counts was not ‘artifact’. Transformation of this variable was performed as necessary, and is discussed further in later sections.

### 4.2.3 Variables considered

Predictor variables used in this analysis were selected to serve as proxy for SDOH. The Gini coefficient was used measure income inequality; percent living in poverty was used to estimate economic stability; percent with high school education was used to estimate access to education; percent uninsured under 65 was used to measure health care access; and percent with access to broadband internet was used to estimate the neighborhood and built environment.

## 4.3 Model Selection

The relative goodness of fit of the various models used in this analysis was assessed using the Akaike information criterion (AIC).

# 5. Analysis

|  |
| --- |
|  |

## 5.1 Univariate analysis

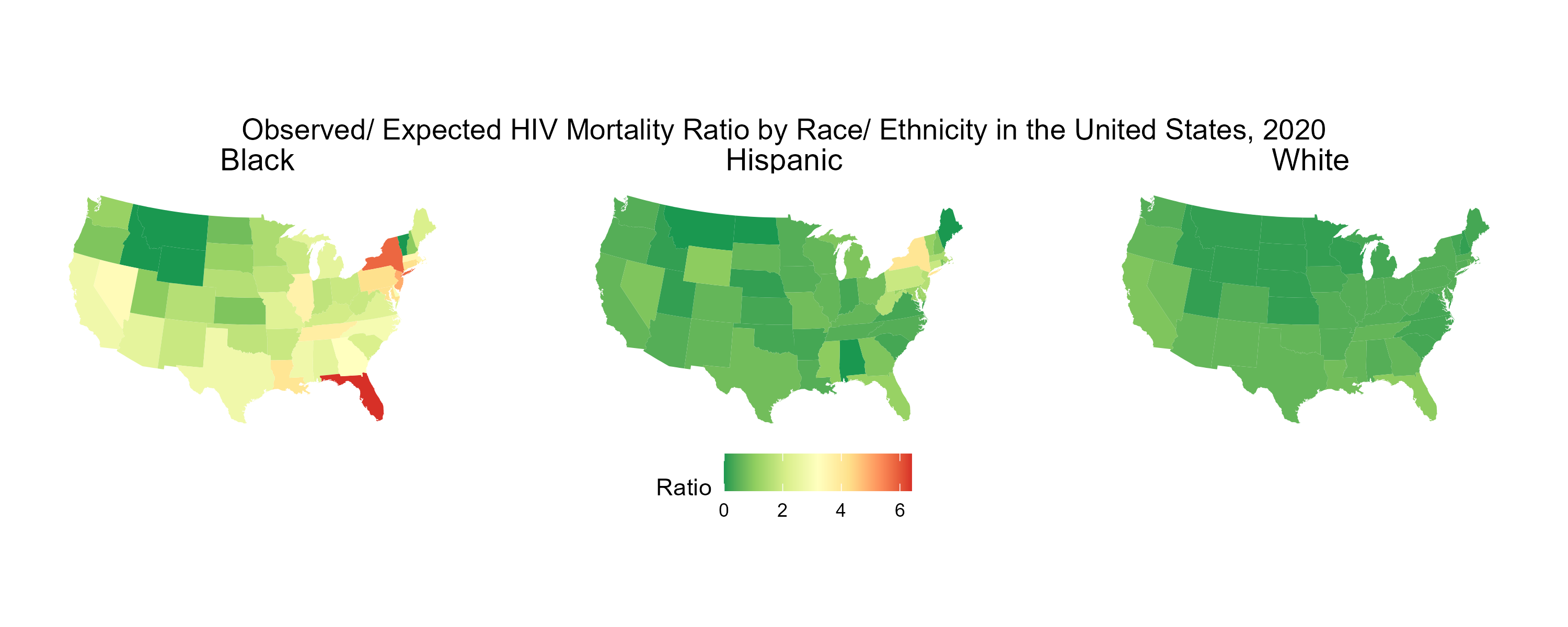
|  |
| --- |
|  |

The distribution of death rate was assessed for each race both visually and using the Shapiro test for normality. Below is the histogram of the distributed of the raw death rate (estimated).

|  |
| --- |
|  |

The death rate for Black persons did not require transformation and the death rate for Hispanic persons was normalized using a log + 1 transformation. Despite utilizing transformations of increasing complexity and power, including Box Cox and Yeo Johnson transformations, I was not able to normalize the distribution of death rate for White persons. As a result, I was unable to include White persons in the model for assessing for spatial correlation. While there are techniques to handle spatial data that is not normally distributed, such as geographically weighted Poisson regression and geographically weight negative binomial regression, my knowledge of the implementation of these techniques limited my ability to uses these methods in my analysis.

|  |
| --- |
|  |



|  |
| --- |
|  |

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

*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.*

## 5.2 Bivariate Analysis

|  |
| --- |
|  |

|  |
| --- |
|  |

|  |
| --- |
|  |

|  |
| --- |
|  |

|  |
| --- |
|  |

*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.*

## 5.3 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.*

# 6. Discussion

## 6.1 Summary and Interpretation

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

## 6.2 Strengths and Limitations

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

# 7. 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.

## Software

sessionInfo()

R version 4.2.1 (2022-06-23 ucrt)  
Platform: x86\_64-w64-mingw32/x64 (64-bit)  
Running under: Windows 10 x64 (build 22621)  
  
Matrix products: default  
  
locale:  
[1] LC\_COLLATE=English\_United States.utf8   
[2] LC\_CTYPE=English\_United States.utf8   
[3] LC\_MONETARY=English\_United States.utf8  
[4] LC\_NUMERIC=C   
[5] LC\_TIME=English\_United States.utf8   
  
attached base packages:  
[1] stats graphics grDevices utils datasets methods base   
  
other attached packages:  
 [1] kableExtra\_1.3.4 here\_1.0.1 forcats\_1.0.0 dplyr\_1.1.0   
 [5] purrr\_1.0.1 readr\_2.1.4 tidyr\_1.3.0 tibble\_3.2.0   
 [9] ggplot2\_3.4.1 tidyverse\_2.0.0 stringr\_1.5.0 lubridate\_1.9.2   
[13] readxl\_1.4.2 knitr\_1.42   
  
loaded via a namespace (and not attached):  
 [1] cellranger\_1.1.0 compiler\_4.2.1 pillar\_1.8.1 tools\_4.2.1   
 [5] digest\_0.6.31 viridisLite\_0.4.1 jsonlite\_1.8.4 evaluate\_0.20   
 [9] lifecycle\_1.0.3 gtable\_0.3.1 timechange\_0.2.0 pkgconfig\_2.0.3   
[13] rlang\_1.1.0 cli\_3.6.0 rstudioapi\_0.14 yaml\_2.3.7   
[17] xfun\_0.37 fastmap\_1.1.1 xml2\_1.3.3 httr\_1.4.5   
[21] withr\_2.5.0 systemfonts\_1.0.4 hms\_1.1.2 generics\_0.1.3   
[25] vctrs\_0.5.2 webshot\_0.5.4 rprojroot\_2.0.3 grid\_4.2.1   
[29] tidyselect\_1.2.0 svglite\_2.1.1 glue\_1.6.2 R6\_2.5.1   
[33] fansi\_1.0.4 rmarkdown\_2.20 tzdb\_0.3.0 magrittr\_2.0.3   
[37] ellipsis\_0.3.2 scales\_1.2.1 htmltools\_0.5.4 rvest\_1.0.3   
[41] colorspace\_2.1-0 utf8\_1.2.3 stringi\_1.7.12 munsell\_0.5.0

# 8. References