Spatial Investigation and Exploration of Factors influencing HIV-related Mortality

United States, 2020

Hayley Hemme

4/25/23

# 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. In the United States, these disparities are most apparent when comparing HIV-related outcomes between Black and White persons. 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.

The goal of this project is to gain an understanding of factors associated the disparity of HIV-related death rates between Black and White persons in the United States.

## 2.1 Question:

Can the ratio of Black-to-White HIV-related mortality rate (MRR) from 2020 in the United States be modeled by social determinants of health or by HIV-related outcomes (late diagnoses, viral suppression). Can a spatial model explain variation in MMR at the state-level and regional-level after accounting for variables associated SDOH or HIV-related outcomes?

H0 : There is no spatial correlation between Black-to-White HIV-related MRR and SDOH or HIV-related outcomes.

Ha : There is a spatial correlation between Black-to-White HIV-related MRR 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.

# 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

In order to assess spatial associations, this analysis was limited to the contiguous United States. Spatial object of the contiguous United States was obtained from library(maps) in R and a 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 Outcome Variables

The primary response variable in this analysis is the ratio of Black-to-White HIV-related death rate (MRR). To compute this variable, feature engineering was performed to create death rate (estimated) stratified by race (Black and White) using counts of HIV related deaths divided by the estimated population based on US Census data multiplied by 100,000 persons. The variable MRR was then computed for each state by dividing the estimate death rates for Black persons divided by the estimate death for White persons. When using methods that required count data,

Social Determinants of Health

Predictors used in this analysis were selected to serve as proxy for SDOH. A number of variables were initially considered, but ultimately five predictors were selected: the Gini coefficient was used measure income inequality; percent living in poverty was used to estimate economic stability; percent uninsured under 65 was used to measure health care access; and percent without internet subscription was used to estimate the neighborhood and built environment; and percent living with food insecurity.

## 4.3 Statistical Analysis

The goal of this analysis was to assess the spatial distribution of the Black-to-White HIV-related MRR. A number of predictors were analyzed using simple linear regression to predict MRR. Variables with a correlation coefficient of |-0.15| or more were selected as possible predictors. To address the possibility of multicollinearity between predictors, principal component analysis was then performed on these variable to aid in variable reduction.

Both non-spatial and spatial models were fitted to data as appropriate. Analysis of the full models began with multiple linear regressions model before moving on to models of increasing complexity, such as Poisson regression and the conditional autoregressive (CAR) model. Models that were flexible to continuous data used MRR as the primary outcome, while models that required count data used the number of deaths, Death Cases , and used log(Population) as the offset, and additionally included variables Race and Region in the models.

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

Moran’s *I* was computed to test the null hypothesis of spatial independence.

# 5. Analysis

## 5.1 Univariate analysis

Our primary outcome of interest was the Black-to-White HIV-related MRR, which was computed as described above. Table 1 shows the overall number of HIV-related deaths, population size older than 12 years, death rate and the MRR in the United States in 2020. The rate of HIV-related deaths among this Black persons was 7.2 times high than the rate of HIV-related deaths among White persons.

|  |
| --- |
| Table 1 |

|  |
| --- |
| Figure 1 |

Visual assessment (Figure 1) and Shapiro test were congruent that the distribution of Black-to-White HIV-related MRR in the US in 2020 followed a normal distribution (*p* = 0.63). No normalizing transformation was needed. Below is the histogram of the our primary outcome variable, MRR.

After verifying that the Black-to-White HIV-related MRR met assumptions of normality, I tested for spatial association of the Black-to-White MRR by computing Moran’s *I* (*I* = 0.1507). For uncorrelated data, the expected value of Moran’s *I* was E(I) = - 0.0213 with a variance var(I) = 0.0095. The data showed statistically significant evidence of spatial dependence (z =1.768; p = 0.039). Figure 2 shows a Moran’s scatter plot of spatially lagged Black-to-White HIV-related MRR and Black-to-White HIV-related MRR. The positive slope indicates that there is a degree of spatial clustering of the Black-to-White HIV-related MRR.

|  |
| --- |
| Figure 2 |

Figure 3 and 4 show the raw HIV-related death rate by Race and State.

|  |
| --- |
| Figure 3 |

|  |
| --- |
| Figure 4 |

Figure 5 compares the raw HIV-related death rate by Race and Region. These plots again show distinct disparities between rate of HIV-related death by race. Additionally, there is evidence of disparities in HIV-related death rate by region for Black persons.

|  |
| --- |
| Figure 5 |

Table 2 shows maps of selected predictors in the United States in 2020. We can see that the Gini Coefficient was generally higher in Southern half of the country, indicating a greater degree of income inequality. Percent Uninsured Under 65 Years Old was highest in Texas, with the South and non-coastal Western states also appearing to have higher rates of uninsured. Percent Without an Internet Subscription was generally low through the United States, however, Wyoming had the highest percent without Internet subscription.

Table 2.

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

## 5.2 Bivariate Analysis and Principal Component Analysis

To narrow down potential predictors to include in later models, I first started by performing simple linear regression models of MRR vs one of the possible predictors. Variables that had a correlation coefficient of |0.15| or more were then selected as possible variables for analysis. These possible predictors were then scaled and principal component analysis (PCA) was performed.

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

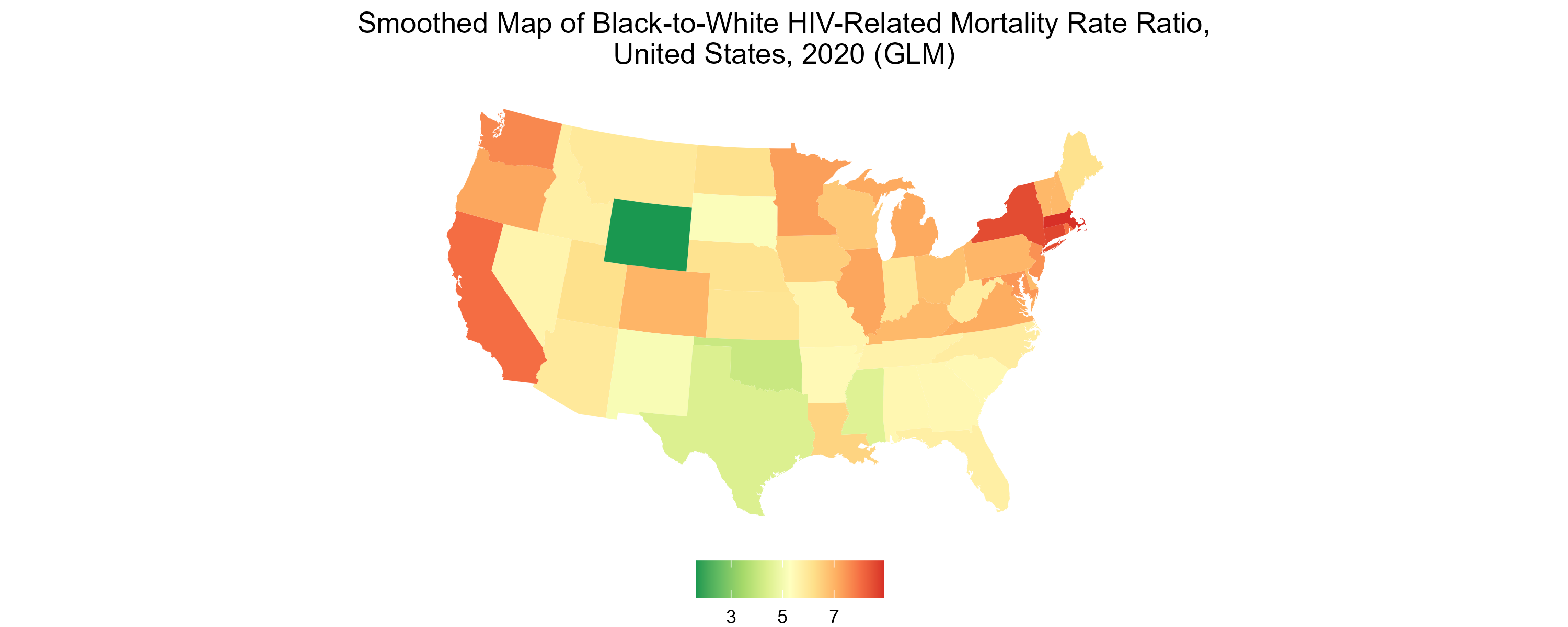
Table 2. shows the plot from the bivariate analysis of predictors that were scaled and included in the PCA. Table 3 shows the results of PCA.

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

Results of the PCA shows that the first three principal components contributed for x % of variation between predictors. The variables that best represent the components were Gini Coefficient , Percent Uninsured Under 65 Years Old, Percent Without Internet Subscription . Ultimately, these variables were included as predictors in the full model.

## 5.3 Full analysis

|  |
| --- |
|  |



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

The CAR model accounting for spatial variation only performed marginally better than the GLM model that did not account for spatial variation.

*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