Kelly Cao and Kirsten Gilbert Report

Socioeconomic and geographic determinants of HIV Prevalence

Kelly Cao and Kirsten Gilbert

2024-12-09

# 1. Abstract

Human Immunodeficiency Virus (HIV) was first introduced to the human population around 1920 to 1940. (Hemother, 2016) It is an illness that compromises the efficiency of the immune system, which can impair the body’s response to other diseases or infections (U.S. Department of Health & Human Services, 2023). In the long term, it can cause the body to become more at risk of contracting heavy chronic illnesses such as cancer, cardiovascular disease, as well as neurological disease. HIV is an illness that requires quick and efficient treatment if it were to be left untreated, it can progress to even more severe stages that form into conditions that have a greater strain on the body (U.S. Department of Health & Human Services, 2023).

It also brings the next idea of the overall stigma around the Human Immunodeficiency Virus (HIV). The most common perception that is fostered around HIV is that it has disproportionate impacts on a certain demographic of people. The idea that those who are lower in socioeconomic status are having a hard time at receiving diagnosis and treatment for illnesses such as HIV. This idea can be correlated to a combination of historical and societal factors alongside years of the spread of misinformation, thus putting said demographic of people at a higher public health risk. Highlighting and progressing the current state of the health of society as a whole. Utilizing the year of 2022 as well as highlighting the United States, and the state of Georgia to specify just how said factors can affect the way individuals go about receiving treatment. What is gathered from said research and analysis is a reflection of how race and ethnicity is a socioeconomic determinant that carries heavy weight in the name of how it can prevent individuals from getting access to quality diagnosis services or treatment services.

Writings such as these are very important in the name of creating a catalyst in the discussions concerning not only HIV but other illnesses that find their way into the communities of people around the world, thus allowing for the challenge of public health epidemiological issues to be overcome.

# 2. Introduction

In the fields of healthcare and epidemiology, it is imperative to consider the multitude of factors that may influence how individuals access and receive treatment for various disease states. There can be numerous demographic factors that play a part in this, them being race, gender, sexual orientation, as well as income. An illness that has a notable reputation for having strong effects towards those who are lower in the socioeconomic hierarchy would be that of Human Immunodeficiency Virus(U.S. Department of Health & Human Services, 2023). For decades, HIV has been pervasive across many communities in the United States, often progressing to Acquired Immunodeficiency Syndrome (AIDS) if left untreated. Based on our previous readings, the hypothesis of this report is that socioeconomic factors have an effect on the prevalence and treatment of HIV. Adding to this, according to a study done in the UK concerning the topic of high HIV prevalence in areas where socioeconomic disparities are prevalent, also reflected the same results as that are intended for this writing (Burch et.al., 2016 ). In this report, socioeconomic factors will be used as a means of analysis, to determine which socioeconomic factors, if any at all, have a part in the prevalence of HIV. To further define, the means of which this will be done is through the comparison of prevalence rates in different areas that uphold different socioeconomic hierarchies. It is commonly known that individuals from minority communities are at an elevated risk of contracting HIV, and these same communities often experience delays in receiving treatment, contributing to worse health outcomes.

For the following project, the data collection was performed in the United States but the data analysis will be conducted specifically for the year of 2022 as well as a closer look at the state of Georgia. In the year of 2022, society was slowly awakening back up following the widespread pandemic of COVID-19. Along with this, COVID-19 also caused the already existing socioeconomic disparities to be highlighted within the state of Georgia. With a majority of people having some form of impact from the disease, it is no surprise that the high prevalence rates have been impacted through the heavy effects of socioeconomic inequalities. As of present day in 2024, Georgia has enacted an initiative called “Ending the HIV Epidemic”, with the intent of reducing HIV incidence by 90%. Recent literature shows that the main predictors that have played a part are racial/ethnic and geographic disparities being the main force behind the continual infection. (Piske, 2024) In 2017, CDC testing guidelines recommended HIV testing 2-4 times annually to demographics at risks. (Paschen-Wolff., M., et al, 2019) This goal continues to be a growing issue with the negative sentiments behind HIV, thus disincentivizing individuals from getting tested. With nearly 30% of people having never been tested for HIV, common barriers to testing are often being access to healthcare or perceived risks and costs. (Wise, J., et al, 2019)

The data set that will be analyzed is an open-sourced CSV file provided by the Center for Disease Control (CDC). By analysing open-sourced CDC data sets, the following conclusion should be tested and confirmed. Known as Behavioral Risk Factor Surveillance System (BRFSS), it presents data from 2011 to present regarding state-based surveillance on information about HIV prevalence, testing rates, and the numerous predictors that are associated with it.

Having been previously updated as early as September of 2024, the data is guaranteed to be up-to-date and relevant to the investigation. This is crucial in the name of advancing public health in a way that encourages change that is truly meaningful. Utilizing relevant data can also allow for the credibility of this report to be supported by a factual foundation that provides support in the name of being able to easily illustrate the presented claim, which only allows for an even greater impact. In the name of implementing ideas and possible solutions to the issue at hand, this report will also dwell into analyzing the current policies that are currently being enforced and comparing it to the reality that is being faced by those who are in the affected regions. For instance, Atlanta, Georgia, serves as a particularly pertinent case study due to its status as a hub of racial, economic, and cultural diversity. The city has long been a focal point for African American culture and continues to be an influential center for various socio-cultural movements. The diversity within the city means that a wide range of demographic groups are affected by HIV, and understanding how these factors interplay is essential for crafting effective public health strategies. (Saldana, C., et al., 2023) The data used during this analysis will also aid in answering the question of, what is the relationship between education level and HIV testing rates, along with other questions that will aid in the presentation of a conclusion. We will also be considering comparative data, by questioning exactly what demographic factors in Georgia are most strongly associated with HIV testing rates, and how these compare to national trends. After this, we will then be analyzing answers to conclusion questions such as can demographic and geographic factors predict HIV testing rates in the year of 2022 as well as the state of Georgia? Through a careful and thorough analysis, this report hopes to provide a clearer picture of the disparities in HIV treatment and offer concrete recommendations for closing the gap in healthcare access, ultimately improving health outcomes for all communities, particularly those at greatest risk. Igniting a dialogue that reaches everyone who is affected as well as those who aren’t affected. Thus allowing for evidence-based recommendations that can influence policymakers, healthcare providers, and epidemiologists, as well as new lessons being taught within a classroom. Thus addressing the complexities of the issue that surrounds the Human Immunodeficiency virus, and fostering an inclusive health strategy that isn’t affected by race, gender, sexual orientation, and income.

# 3. Method

The following steps were performed using R version 4.3.1, RStudio, and various R packages to perform data exploration and analysis. Github and git are used for version control.

## 3.1 Data Procurement

The following data used for analysis was procured as a dataset provided from the Center for Disease Control (CDC). It contains HIV-AIDS prevalence data procured from land-line and cell-phone use through state based surveillance systems known as the Behavioral Risk Factor Surveillance System (BRFSS). The data has been collected from 2011 to present and is continuously updated annually. The dataset is based on responses to the survey question: “Have you ever been tested for HIV?” The data represents the weighted percentage of respondents within various stratified sociodemographic groups who answered affirmatively to this question.

The data is downloaded from the CDC database as a CSV file with over 30,000 rows and 27 columns. Some notable columns include: HIV prevalence, location, stratification group, and stratification value. Stratification value represents the stratified sociodemographic groups, such as income levels, age groups, or even educational attainment. It represents the categories or levels within the stratification group. The weighted percentage of people in a stratified group is presented as a separate variable and is renamed as ‘prevalence’ for ease of use.

## 3.2 Data Processing

The data is loaded into RStudio as a csv file with base R functions. The dplyr and stringr packages are used to clean the data for further analysis. Out of the 27 possible columns, 13 columns were selected and filtered out for empty or NA rows. The column names were remapped and renamed for ease of future use and greater clarity. From that, every data point in which the ‘state\_name” == ‘Georgia’ is used and filtered further as a separate object in the RStudio environment. Separately, the cleaned data is processed into another csv file dubbed ‘HIV\_2023’ to reflect all data where ‘year’ == ‘2023’. This reflected all of the most recently available data. A summary of the cleaned US data, Georgia data, and 2023 data is reviewed and then saved as a separate CSV files under the ‘data’ directory.

## 3.3 Exploratory Data Analysis

When performing data exploration, R packages, such as here, dplyr and ggplot were commonly used to determine distribution of predictors for the HIV prevalence outcome. This was performed in two main parts: data manipulation and visualization through ggplot graphs.

## 3.4 Statistical Data Analysis

### 3.4.1 Linear Regression

When setting up the linear regression model, the HIV testing prevalence percentage column was chosen as the outcome. As for the predictors, the stratification values found in the different stratification group are chosen. From the linear regression model, the coefficient, standard error, t-value, and p-value were extracted for each determinant and ranked from highest to lowest.

### 3.4.2 ANOVA

Similar to the linear regression model, the ANOVA model was set up with the HIV prevalence as the outcome and the remaining determinants as the predictors. From the ANOVA model, the sum of square, degrees of freedom, mean square, f-statistics, and p-value were extracted and ranked in order of descending p-value.

### 3.4.3 LASSO

After performing ANOVA and Linear Regression, an attempt was made with LASSO. The LASSO model was created with the tidymodel package. The recipe, specification, and workflow were set with specific penalties to combat multi-collinearity. The data set was split to perform cross-validation. The following metrics were determined for the LASS regression model: RMSE, RSQ, and MAE. A list of the determinants were rank in descending coefficient.

# 4. Result

## 4.1 Exploratory Data Analysis

Following that, the distribution of the outcome was determined and graphed as a histogram graph to show the data distribution.Upon visualizing it, it becomes apparent that the data set is binomially distributed.

|  |
| --- |
| Figure 1. Distribution of Prevalence Rate Overall |

The outcome variable was then graphed as a heatmap in the US. Upon producing this, it suggests that the variance in the outcome variable is limited or much lower than expected.

|  |
| --- |
| Figure 2. Heatmap of the HIV Prevalence in the United States |

Using base R functions and dplyr r package, the data frame was manipulated and filtered to summarize by averaging the prevalence across the years for both the US and the state of Georgia. This was then plotted to create a time series plot, to clearly show the HIV prevalence from 2011 to present day, as shown in Figure 3. Something to note in figure three, is the sharp increase in HIV prevalence 2019 and 2022 in Georgia. Whether it is statistically significant is yet to be determined.

|  |
| --- |
| Figure 3. Time Series of HIV Prevalence between Georgia and the United States |

Figure 4 showcases the % distribution of prevalence stratified by the major determinants. For the US, it shows that across the board that the prevalence remains fairly consistant, but when look at Georgia’s data, Age group and Race/Ethnicity are the stand out group in terms of variance and range of data points.

|  |
| --- |
| Figure 4. Boxplot showing prevalence of HIV in the US and GA startified by major determinants |

Each determinant shows major change in the outcome over time with no consistency in terms of trend.

|  |
| --- |
| Figure 5. Time trend of HIV for Each Determinants/Stratification Group |

## 4.2 Statistical Data Analysis

From the EDA, it can be seen that the resulting ‘prevalence’ lacks variability in both the US and the Georgia data set. To offset this trait, the US data set was use for the following analysis.

### 4.2.1 Linear Regression

Based on Table 1, the linear regression results suggests that none of the predictors/determinants have any significant impact on the outcome variable (HIV prevalence). All of the listed p-values are well above the 0.05 and are near 1. Nearly all of the coefficients are near 0,further emphasizing this.

|  |
| --- |
| Table 1. Linear Regression Results |

### 4.2.2 ANOVA

In the ANOVA test, the sum of square attributed to the predictor variable is 43.8, which is much smaller to the residual variance at over 70,000. This suggests that the predictor variables have very little of the total variance in the outcome variable. The F-value is shown to be extremely small, and once again, the p-value is much too high to suggest any level of significance.

|  |
| --- |
| Table 2. ANOVA Results |

All of this may suggests that the predictors have little to no effect on the outcome. Alternatively, this could suggest that the chosen model or variables may not be adequate enough to reflect the factors that are driving this ‘prevalance’ percentage.

### 4.2.3 LASSO

In order to rule out multicollinearity, the LASSO test was performed to apply penalties and strip out determinants with near-zero coefficients. Despite this, the LASSO regression analysis showed limited success. With only 17.3% of the variance explained, an extremely high RMSE, and the intercept dominating the top predictor suggestions, these all suggest a weak or nearly non-existing relation between the chosen predictors and the outcome.

When ranking the coefficient from highest to lowest, only the intercept appears with a significantly high coefficient. All of the other determinants that were used as predictors had coefficients much closer to zero. This can be seen in table 3. Initially used to combat multicollinearity, the model reduced many of the highly correlated predictors’ coefficient to zero to avoid redundancy. However, even when reducing the penalty, the result remains virtually unchanged.

|  |
| --- |
| Table 3. LASSO Results |

# 5. Discussion/Conclusion

Analyzing HIV data sets doesn’t provide a clear connection between socioeconomic and geographic factors and the frequency of individuals being treated for HIV. Socioeconomic factors such as race/ethnicity, age group, income, as well as sexual orientation can’t be proved to play a really big role in causing high percentages of HIV spikes in specific groups within the year of 2022, along with an emphasis in the state of Georgia. This being presented through a data set that included a survey that was presented to those of a particular area as well as those who choose to participate. It can be difficult to isolate if there are any specific factors alone that are to blame for said health disparities, however it can be concluded that a combination of them together do not have the connection to said high prevalence rates. Multicollinearity can be reflected through how many socio economic factors that affect prevalence, with the data reflecting that income, educational status, gender, age, and race are all factors that play a part in high HIV prevalence. Minority communities such as those who have a heavy presence of people of color as well as those who are low-income are not put into a position of disadvantage as they are more at risk of contracting HIV.

Utilizing the analysis of the state of Georgia and putting it into comparison to the data of the entire US within the year of 2022, as well as pinpointing a specific area with the intent to target any socioeconomic inequalities that contribute to the higher prevalence rates that are being presented. When Georgia is put into comparison with the entirety of the United States it is seen that race/ethnicity and age are not the main factors that affect HIV prevalence as it was originally hypothesized. This is where it can be determined that geographical factors such as being located in the southeast has a play on socioeconomic factors, an example being not as much opportunity to get access to good healthcare. Multicollinearity is a prevalent issue in regards to HIV incidence relative to the large number of predictors associated with it. The data is shown reflecting that income, educational status, gender, age, and race are all factors that play a part in high HIV prevalence for the US. This is particularly true when viewing both the ANOVA and the linear regression results.

In minority communities, where income is a heavy determinant of health education alongside the lack of access to healthcare creating grounds for high infection rates, it is commonly believed that this can have a direct impact on how people are able to go and receive treatment for certain illnesses. HIV is an illness that is transmitted through contact with certain bodily fluids, whether it be through unprotected sexual contact, contact with infected blood, or even being passed down from mother to child. Along with this, also considering how areas within Georgia that can be considered metropolitan can also reflect a higher population that also reflect closer living quarters that cause for close contact with HIV to be less avoidable. With this it can be understood just how important creating an intersection that prevents the continuous cycle of socioeconomic factors having such a heavy impact on the way individuals go about receiving healthcare treatment is. Preventative measures for HIV to better fit areas are heavily impacted by the illness and creating health equity for everyone.

In the name of the limitations of the data that is being used to analyze the correlations between different factors, there was a lot of variation on the predictions and the outcomes, as well as models that were hindered in the name of the complexity of it. With the intention of utilizing a specific year as well as a specific area, it was difficult to pinpoint the correlation between socioeconomic status as well as high prevalence rates. This could be the result of having a broad data set alongside a narrow pool of data to utilize. Model misspecification is something that can have a strong impact on the way data is analyzed, as it could result in invalid inferences and poor reliability. Utilizing data that doesn’t uphold much variation could also aid in deriving accurate results. For future analysis, using data that is more randomized as well as models that could eliminate multicollinearity for the factors used would allow for more clear results. Using data sets that better reflect the specified intention of the overall writing. Randomizing the data would aid in the generalization of the findings thus causing the process of drawing a conclusion to be smoother. Using models that are more randomized would allow for results to be more easily interpreted, allowing for the process of drawing a conclusion to be smoother as well. Although the data failed to prove our hypothesis statement, it is still understood how socioeconomic factors have a part in health disparities, especially those concerning HIV.

# 6. References

German Advisory Committee Blood (Arbeitskreis Blut), Subgroup ‘Assessment of Pathogens Transmissible by Blood’ (2016). Human Immunodeficiency Virus (HIV). Transfusion medicine and hemotherapy : offizielles Organ der Deutschen Gesellschaft fur Transfusionsmedizin und Immunhamatologie, 43(3), 203–222. https://doi.org/10.1159/000445852

Piske, M., Nosyk, B., Smith, J. C., Yeung, B., Enns, B., Zang, X., Sullivan, P. S., Armstrong, W. S., Thompson, M. A., Daniel, G., & Del Rio, C. (2024). Ending the HIV Epidemic in Metropolitan Atlanta: a mixed-methods study to support the local HIV/AIDS response. Journal of the International AIDS Society, 27(7), e26322. https://doi.org/10.1002/jia2.26322

Centers for Disease Control and Prevention. (2024). BRFSS: Table of HIV-AIDS. U.S. Department of Health & Human Services. Retrieved November 16, 2024, from https://data.cdc.gov/Behavioral-Risk-Factors/BRFSS-Table-of-HIV-AIDS/475u-gzzh

HIV.gov. (2023). What are HIV and AIDS? U.S. Department of Health & Human Services. Retrieved December 1, 2024, from https://www.hiv.gov/hiv-basics/overview/about-hiv-and-aids/what-are-hiv-and-aids

Saldana, C., Philpott, D. C., Mauck, D. E., Hershow, R. B., Garlow, E., Gettings, J., Freeman, D., France, A. M., Johnson, E. N., Ajmal, A., Elimam, D., Reed, K., Sulka, A., Adame, J. F., Andía, J. F., Gutierrez, M., Padilla, M., Jimenez, N. G., Hayes, C., McClung, R. P., … Wortley, P. (2023). Public Health Response to Clusters of Rapid HIV Transmission Among Hispanic or Latino Gay, Bisexual, and Other Men Who Have Sex with Men - Metropolitan Atlanta, Georgia, 2021-2022. MMWR. Morbidity and mortality weekly report, 72(10), 261–264. https://doi.org/10.15585/mmwr.mm7210a3

Paschen-Wolff, M. M., Restar, A., Gandhi, A. D., Serafino, S., & Sandfort, T. (2019). A Systematic Review of Interventions that Promote Frequent HIV Testing. AIDS and behavior, 23(4), 860–874. https://doi.org/10.1007/s10461-019-02414-x

Wise, J. M., Ott, C., Azuero, A., Lanzi, R. G., Davies, S., Gardner, A., Vance, D. E., & Kempf, M. C. (2019). Barriers to HIV Testing: Patient and Provider Perspectives in the Deep South. AIDS and behavior, 23(4), 1062–1072. https://doi.org/10.1007/s10461-018-02385-5

Burch, L. S., Smith, C. J., Anderson, J., Sherr, L., Rodger, A. J., O’Connell, R., Geretti, A. M., Gilson, R., Fisher, M., Elford, J., Jones, M., Collins, S., Azad, Y., Phillips, A. N., Speakman, A., Johnson, M. A., & Lampe, F. C. (2016). Socioeconomic status and treatment outcomes for individuals with HIV on antiretroviral treatment in the UK: cross-sectional and longitudinal analyses. The Lancet. Public health, 1(1), e26–e36. https://doi.org/10.1016/S2468-2667(16)30002-0

# 7. Supplement

|  |
| --- |
| Supplement 1. Summary Statistics of each Variable in the data set |