STEM Fellowship Big Data Challenge 2017

Developing a Syndemic Framework for Understanding HIV Prevalence

Ann Seongju Yun[1], Maliha Islam[2], and Samee Mahbub[3]

[1,2,3]Faculty of Applied Science and Engineering, University of Toronto

May 27, 2017

Abstract

This paper aims to analyze the HIV prevalence of different countries based on various markers such as economic health and environmental status. While there exists studies showing how these factors affect and correlate with HIV prevalence, they are commonly done independently. Very rarely has it been considered how these factors correlate with themselves and in turn how these inter-correlations affect the HIV prevalence. Datasets relating to these markers was collected from organizations such as World Health Organization (WHO), World Bank and UNESCO and then analyzed using software tools from R such as K-means clustering and linear regression, and visualized using R's libraries. Statistical tools were implemented to interpret these datasets and to accept or reject the hypothesis that were originally constructed (or if the correlation is not strong enough, left inconclusive). Afterwards, these individual factors were then correlated to each other and a syndemic framework is constructed in order to link all these factors to each other in an attempt to understand how they affect the HIV prevalence of a given country, collectively. It was seen that economic factors such as GDP and Gini coefficient were strongly correlated to HIV prevalence, specifically GDP showed a negative correlation while Gini coefficient showed positive correlation. Other trends were also found in markers such as Climate Change Performance Index (CCPI)where it was found that most countries had mid to low values of CCPI. Furthermore, data based on various response methods to HIV prevention such as antiretroviral therapy and HIV testing and counselling were shown to have contradictory correlations. HIV testing and counselling was found to have a positive correlation with HIV prevalence (very contradictory, since a negative relation was expected). Lastly, a syndemic framework was constructed using correlational research between the factors themselves, to understand how different factors interact synergistically to increase the prevalence of HIV and to more accurately predict HIV prevalence trends.

Keywords
Human Immunodeficiency Virus, syndemic, Gross Domestic Product, Gini coefficient, literacy rate, Healthcare Access and Quality Index, Climate Change Performance Index

1 Introduction

The prevalence of HIV/AIDS in a given country is dependent on many factors, including, but not limited to, economic health, education and environmental status. The effects of each factor separately have been well documented. However, it has rarely been considered how these factors affect not only HIV prevalence, but also each other, in a complex interconnected framework, known as a syndemic. A syndemic is described as "a set of mutually enhancing health problems that, working together in a context of deleterious social and physical conditions increase vulnerability, significantly affect the overall disease status of a population" [1]. According to the article "Interactions Between HIV/AIDS and the Environment: Towards a Syndemic Framework published in the American Journal of Public Health" in 2013, a syndemic does indeed exist between the HIV epidemic and the environment, which is connected to many other factors [2]. It predicts how global factors such as climate change and macroeconomics can eventually lead to effects such as lack of education and poor health care [2]. These secondary effects can then lead to changes in HIV prevalence, but more importantly, these factors can also affect each other which can have a combined effect on HIV prevalence [2]. This paper will attempt to validate this syndemic by choosing certain metrics that can be used to gauge these factors, and by analyzing them while taking into consideration the inter-correlation between the factors themselves.

To gauge the macroeconomics and microeconomics of a country, the metrics Gross Domestic Product (GDP) per capita and the Gini coefficient was used. These two metrics were chosen in order to measure the economic health of a country from two different aspects. GDP per capita measures the monetary value of goods and services produced per person in a country [3], while the Gini coefficient measures the degree of inequality in the distribution of family income in a country [4]. It is hypothesized that the HIV prevalence should share a negative relationship with GDP, since a country with more economic resources is expected to be able to fight the HIV pandemic. Furthermore, it is conjectured that the Gini coefficient should share a positive relationship with HIV prevalence, because the less uniformly the economic resources are distributed (thus the higher the Gini coefficient), the more likely it is for people to have fewer resources to prevent HIV.

In an attempt to quantify the climate change of a given country, the Climate Change Performance Index (CCPI) is used. This metric is used to measure how much effort a country has put to reduce the effects of climate change, which can be used to measure how the country is affected by climate change. It is hypothesized that the HIV prevalence should be negatively correlated with the CCPI. The reasoning behind this is, if a country has a high CCPI, they are less likely to face the chain effects of climate change, which can have dire consequences such as poverty and lack of education, thus increasing the chances of HIV infection.

However, climate change, like economics, is global in nature, so it would be desirable to have metrics that can gauge the secondary effects of climate change such as access to education, lack of access to proper healthcare. For these factors, the adult literacy rate, Healthcare Access and Quality Index (HAQI) and various response method metrics (such as percentage of the population that have access to antiretroviral therapy, counselling and testing for HIV/AIDS) are used. These indexes are used to quantify the ease at which a country has access to these facilities (e.g. education, healthcare, therapy). Thus, it can be inferred that they should share some sort of negative relationship with the country's HIV prevalence, due to the fact that the higher the fraction of the population has access to facilities, the less prone they are to be infected by HIV.

Data based on these metrics will be collected based on country. They will be taken for the year 2015 (or if this is not possibly the closest year to 2015), in an attempt to avoid major market shocks and events occurring before the decade (such as the financial crisis of 2007-2008), so that the error due to these events are minimized and the HIV prevalence can be observed only due to the factors being analyzed.

Based on these findings, the paper will also attempt to verify and possibly reconstruct the syndemic framework based on the results and use them to suggest solutions in the form of public health decisions and policies, and potential ideas on how to better tackle the HIV/AIDS pandemic in the future.

2 Materials & Methods

2.1 Software Tools

RStudio Version 1.0.143 and R Version 3.4.0 was used to plot and analyze the datasets. R packages used are:

- readr
- ggplot2
- shiny

2.2 Sources

2015 data for total and adult HIV affected population for each country was collected from World Health Organization [5]. This data was divided by the data for total population (2015), collected from The World Bank [6], to find the percentage of total and adult population affected by HIV for each country.
Sources for all other datasets used are listed below:

- Climate Change Performance Index (CCPI), 2015 – Germanwatch and Climate Action Network Europe [7]
- Gross Domestic Product (GDP) per capita, 2015 – The World Bank [9]
- Gini coefficient, 2013* – United Nations Development Programme [8]
- Adult literacy rate, 2015 - – UNESCO Institute for Statistics [10]

- Healthcare Access and Quality Index, 2015 – GBD 2015 Healthcare Access and Quality Collaborators [11]
- Responses (HIV testing and counselling, ART coverage, prevention of motion-to-child transmission), 2013-2015* – World Health Organization [12]

*Datasets in 2013 and 2014 were used as they were the most current data found for these factors. It is assumed that they remained approximately constant over the course of one-two years and thus, these were analyzed with datasets from 2015.

2.3 Acquiring Data

The dataset was acquired from various sources (see Sources) to be used for analysis. Since this paper focuses providing conclusive statements about HIV prevalence worldwide, the various factors used to investigate were based on country. Factors ranged from population, HIV affected population, as well as other markers used for the analysis. Afterwards, the data was organized combining different datasets so that they could be analyzed.

2.4 Analysis

Two main statistical tools were used to analyze the different natures of the datasets. For datasets where the data appeared to show some uniform trend, a regression model was used. If the data appeared to follow a certain function, the data would be linearized by applying an operator on the dataset, and a linear regression model was used to find the relationship between the variables. A linear regression is preferred over a Gaussian regression model since the exact relationship can only be known using a linear model. Wellness of fit of this regression model is gauged by obtaining the R-squared value. For this paper, an R-squared value on the **magnitude of 0.1** or higher would be sufficiently large enough to conclude a good fit.

If a regression model is not suitable, then the dataset is clustered. The sample sizes used in the various datasets analyzed in this paper were in the range of 70-200. Due to this large sample size, it was deemed appropriate to use a K-means clustering algorithm to cluster the data. The Euclidean distance is used to measure similarity between data points. 5 clusters with randomly generated centroids were initialized and the algorithm was run iteratively until the cluster centroids didn't change significantly and those clusters were then presented and analyzed. The quality of the clusters were determined by measuring the standard deviation. The standard deviation of Cluster j C_j with cluster mean μ_j and n_j elements where x_{ij} represents the i^{th} element in C_j is:

$$SD(C_j) = \sqrt{\frac{\sum_i (x_{ij} - \mu_j)^2}{n_j}}$$

This would be compared to the standard deviation of the entire dataset. For this paper, if the cluster standard deviation is 70% or less of the entire dataset standard deviation, this is concluded as a "good" cluster.

3 Results

3.1 Climate Change Performance Index on HIV Prevalence

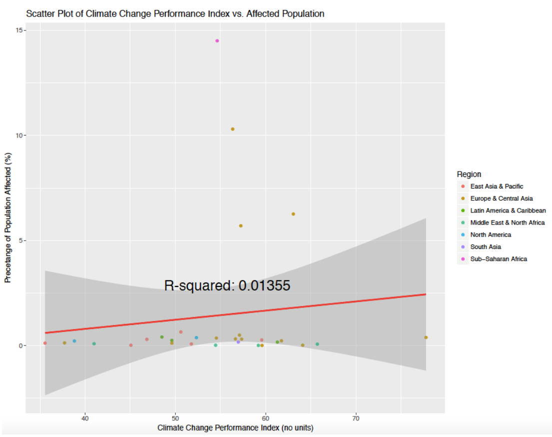


Figure 1: *Scatter plot of CCPI vs. percentage of population affected, sorted by region, fitted with a linear model*

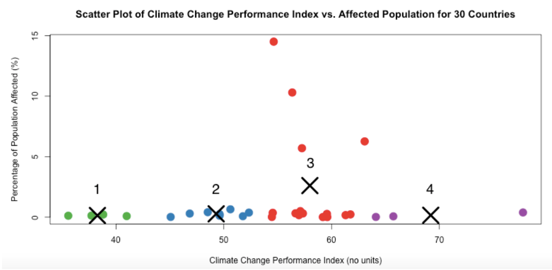


Figure 2: *Scatter plot of CCPI vs. percentage of population affected, k-mean clustered into four clusters.*

The linear curve in Figure 1 shows a very low, therefore statistically insignificant, R-squared value of 0.01355. Since the linear model does not explain the variability of the response data around its mean, k-means clustering was used to further analyze the effect of climate change on HIV prevalence. Four clusters were formed as shown in Figure 2. Trends and notable observations between countries within separate clusters were then made.

Countries in cluster 1 has CCPI in the range of 35.57 – 40.99 and affected population of 0.09 – 0.22 %. The GDP is similar and within the range of 180,000 – 1,500,000 USD and adult literacy rate within 74.99 – 79.72%. For this cluster, correlation is found between low CCPI and low GDP, as well as between low CCPI and moderate literacy rate. Cluster 2 consists of countries with CCPI in the range of 45.07 – 81.3 and affected population of 0.02 – 0.65%. Half of the countries in cluster 2 are in East Asia & Pacific region. The adult literacy rates of all countries are similar within the range of 92.99 – 99.77%. The GDP varies widely from 10,000 – 18,000,000 USD. Countries in cluster 3 has a CCPI in the range of 54.46 – 63.07 and affected population from 0.01 – 14.5%, Germany's affected population of 0.01% being the lowest and South Africa's 14.5% being the highest of the entire dataset. 9 out of the 15 countries in this cluster lie in Europe & Central Asia. The GDP and Gini coefficient in this cluster also vary widely from 16,000 to 3,400,000 USD and 25.6 – 61.1%, respectively. Cluster 4 has countries with CCPI in the range of 64.11 – 77.76 and affected population of 0.02 – 0.39%. These countries have similar GDP and Gini coefficient in the ranges of 100,000 – 2,400,000 USD and 41.5 – 45.7%, respectively.

3.2 Gross Domestic Product and Gini Coefficient on HIV Prevalence

To investigate the effect of GDP on HIV prevalence, the two variables were plotted against each other. Judging from the general shape, the data was adjusted and a favourable method of analysing the data was chosen.

3.2.1 Variables for GDP Plot

- percent2015 (Y): This is the percentage of the total population of a country whom are infected with HIV (%) as of 2015.
- gdp2015 (X): This is the GDP of a given country given in million USD dollars (as of 2015).

A clear trend is that most countries with a high percentage of HIV infected population are

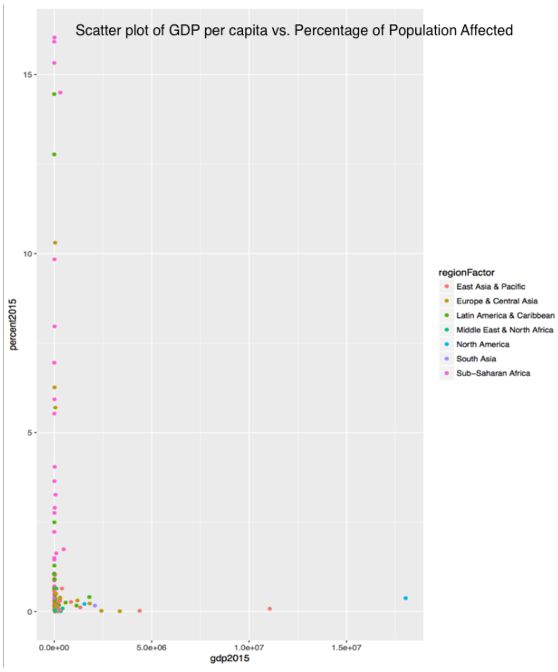


Figure 3: *Scatter plot of GDP per capita vs. percentage of population affected, sorted by region*

from the Sub-Saharan African region. Furthermore, we can see these countries are also in the low GDP region as they are close to the y-axis. The shape of graph suggests an inverse relationship. It looks similar to graphs in the form of:

$$Y = \frac{c}{X^n}$$

where 'n' is some positive number and 'c' is a constant.

To find out the value 'n', apply the natural logarithm operator to both sides:

$$ln(Y) = -nln(X) + c'$$

Where c' is another arbitrary constant. If ln(Y) is plotted against ln(X), a linear relationship should be obtained, and the slope of the curve should be equal to '-n'.

Thus, ln(Y) is plotted against ln(X) and a simple linear regression model is applied (Figure 4).

The linear model has a slope value of -0.40394. This will evaluate a value of 'n' to be 0.40394. (The fact that n has a positive value does confirm our hypothesis that there is in fact an inverse relationship)

The R-squared value is quite low here (while the p-value is not as low as wanted, although this calculation was done assuming each country was identically independently distributed, which may not be valid), indicating that this is quite a weak trend.

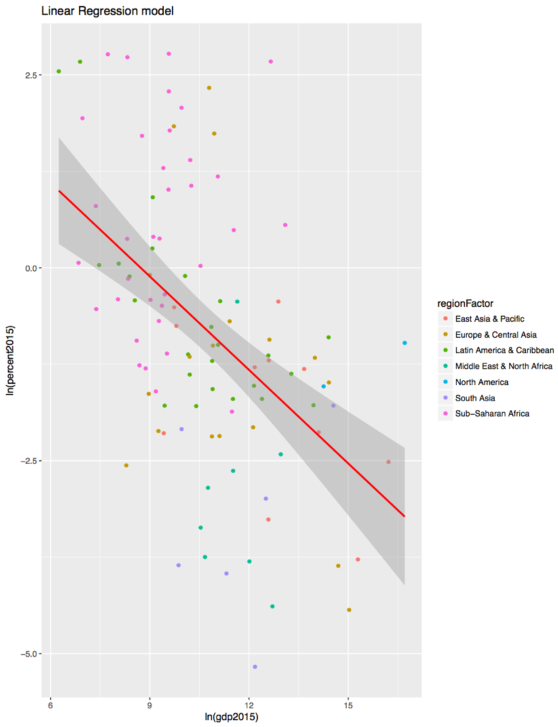


Figure 4: *ln(Y) plotted against ln(X) fitted with a simple linear regression model*

Thus, the relationship between Y and X follow approximately (albeit weakly) the relationship:

$$Y = X^{-0.40394}$$

This shows that clearly, there is a trend that GDP of a given country does in fact follow an inverse relationship, but due to that fact this is a weak trend, one must look at the research done on this field as well to further justify this hypothesis.

The research used to further investigate this will be reduced to only consider African countries. Judging from Figure 3, the color coding legend shows that the countries that are the greatest victims to the HIV epidemic are countries from the Sub-Saharan African region. Thus, if concluding statements about how the GDP of these countries are affected could be made, it is assumed that this can be generalized to other countries as well.

Another metric to gauge the health of the econometrics of a country is the Gini coefficient, which shows the income inequality within a country. Naturally, similar to the logic used in the case of GDP, we would naturally hypothesise that there would also be an positive relationship between the Gini coefficient of a country and its HIV prevalence.

3.2.2 Variables for Gini Coefficient Plot

- Gini (X): Represents the Gini coefficient shown as a percentage of a given country (%)
- Percent (Y): Represents the HIV prevalence of a given country shown as percentage of the total population (%)

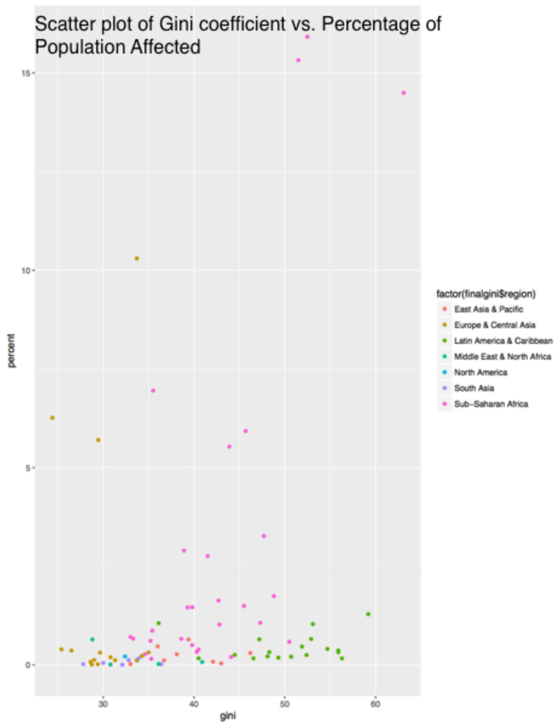


Figure 5: *Scatter plot of Gini coefficient vs. percentage of population affected, sorted by region*

from the Latin America & Caribbean region (colored in green) seem to be clustered together. Furthermore, it can also be seen that most of the countries with a high HIV prevalence originate from the Sub-Saharan African region (colored in pink).

A Gaussian regression model is attempted to fit the data in Figure 6, and the R-squared value will be able to indicate a goodness of fit.

From the adjusted R-squared value (0.04976) of the fit, this is deemed too low. Thus, other statistical tools must be considered here to derive meaning from this plot. Clustering can be another method of analysis, and this will then be implemented here. The K-means method will be used since the sample size for this dataset (82 countries) is quite large, and so it is preferred over hierarchical clustering.

A K-means clustering algorithm is then implemented using several different initial conditions and iterated many times until the clusters have converged. The algorithm initially started off

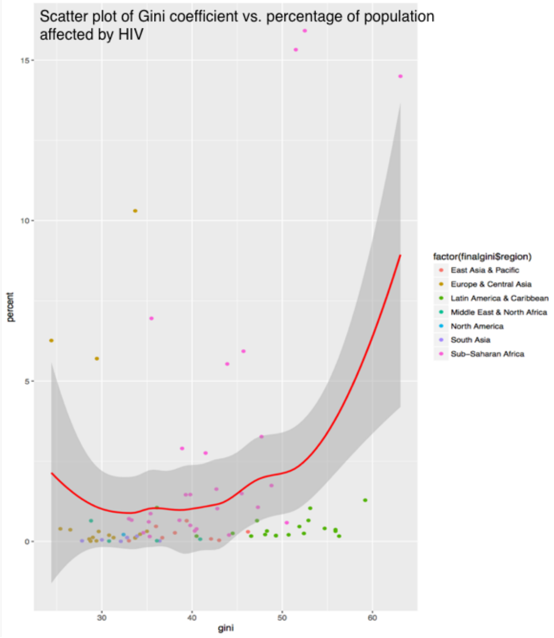


Figure 6: *Scatter plot of Gini coefficient vs. percentage of population affected, sorted by region, fitted with a Gaussian curve*

using 5 clusters, but due to the nature of the dataset used, upon convergence, only 3 clusters remained, and they are shown using the shapes of the points (circles, squares, triangles and triangles) and are outlines in red as shown in Figure 7.

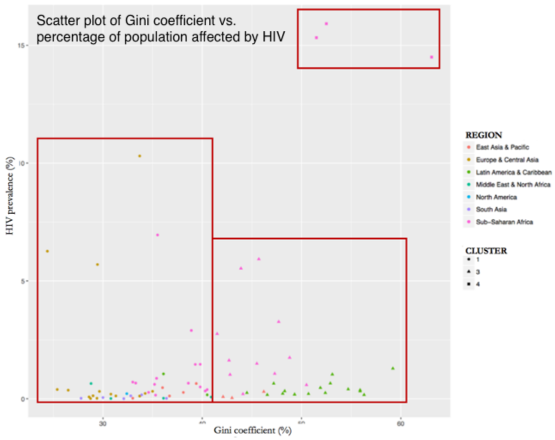


Figure 7: *Scatter plot of Gini coefficient vs. percentage of population affected by HIV sorted by region and clustered into three clusters using K-means clustering*

3.2.3 Cluster Properties

From the cluster data, it can be seen that the standard deviation of the elements in each cluster are all significantly lower than the standard deviation of the entire dataset. This indicates

Cluster	Mean Gini coefficient (%)	Mean HIV prevalence (%)	Number of elements	Standard-Deviation of cluster
1 (Circle)	33.85	0.972	48	4.734
3 (Triangle)	48.64	1.057	31	4.877
4 (Rectangle)	55.70	15.24	3	5.281
Entire Dataset	40.24	1.52	82	9.410

that the clusters are well formed and elements within the cluster are close together, while the clusters as far apart from each other.

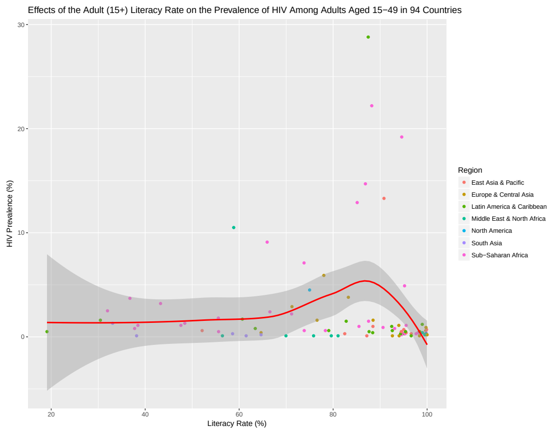
Also, it should be noted that Cluster 1 and 3 are quite similar with respect to their mean Gini coefficient, HIV prevalence, number of elements as well as their standard deviation, however Cluster 3 seems to be comprised mostly of countries originating from Sub-Saharan Africa as well as some from Latin America and Caribbean, where as Cluster 1 seems to contain countries from many regions quite uniformly.

It should also be noted that Cluster 4 contains three countries all originating from Sub-Saharan Africa, which have the highest mean Gini coefficient as well as the highest HIV prevalence. Thus, it can be seen that the cluster with the highest Gini coefficient have the highest HIV prevalence.

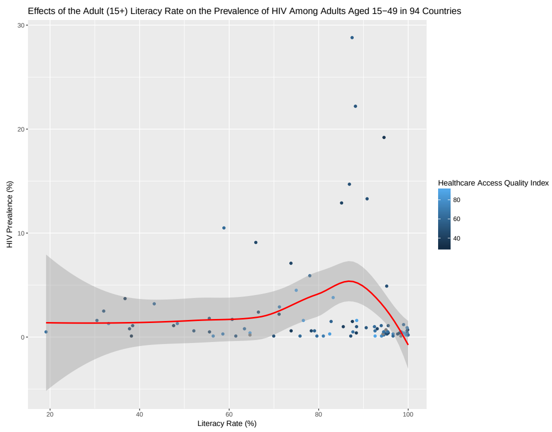
3.3 Literacy Rate on HIV Prevalence

It appears from the Gaussian regression model of Figure 8 that once literacy rate reaches 88% or higher, the prevalence of HIV significantly decreases as literacy rate increases. Also in the graph, outliers that have both high literacy rate and HIV prevalence percentage were identified to be Sub-Saharan African countries (see Figure 8a. To find potential factors that these countries have in common, HAQI was added as the third variable in Figure 8b. This plot shows that these countries have high HAQI, which means that the citizens can easily access high-quality health care services. Another notable result from Figure 8 is that the 95% confident interval, represented by the shaded grey region, gets smaller as it approaches the region with high literacy rate and low HIV prevalence. Although the convergence of the data points is visible, R-squared value of 8.295E-06 of the dataset signifies that the correlation cannot be considered accurate to draw conclusions. Consequently, the effects of higher literacy rate on HIV prevalence was investigated, and the results are shown in Figure 9.

The sample size of the data was reduced to from 94 to 45 countries as some countries did not have two or more years of data available on literacy rate or HIV prevalence to calculate the changes in the two variables. The linear regression on the graph shows inversely-proportional



(a) Region as the third variable



(b) HAQI as the third variable

Figure 8: *Scatter plot of adult literacy rate versus HIV prevalence of adults*

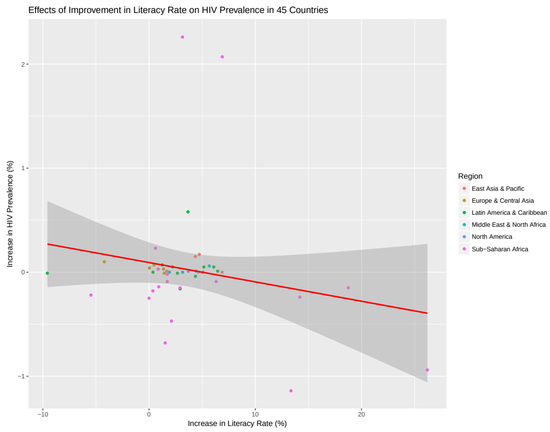


Figure 9: *Increase in literacy rate and the change in HIV Prevalence*

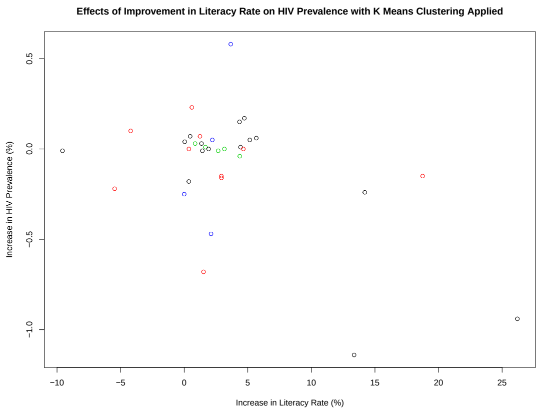


Figure 10: *Clustered version of Figure* ?? *based on Gini coefficient and Gender Parity Index*

relationship between educating more people and HIV affected population. The slope of the regression line is -2.056, which means that 1% increase in literacy rate will decrease HIV prevalence by 2%. The dataset has small R-squared value of 0.04 as well as p value of 0.1891 indicating weak correlations between the two variables. K-means clustering was applied to Figure 9 to further study the trends and the result is shown in Figure 10.

K was chosen to be 4 through trials and errors. Two factors that are likely to have influences on the number of educated people in a country, such as Gender Parity Index (GPI) in literacy and Gini Coefficients were taken into consideration for clustering. Since there are no clear trends in clusters, the relationships between Gini Coefficients and literacy rate and between GPI and literacy rate cannot be determined. Connection between these variables will be further investigated in Discussion section with research findings.

3.4 Healthcare Access on HIV Prevalence

The linear model in Figure 11 shows a negative correlation between HAQI and HIV population. Since this model has a very weak R-squared value of 0.04236, trends were observed based on region. From Figure 11, it is seen that Sub-Saharan African countries have the lowest HAQI and highest affected population while Europe & Central Asian countries have the highest HAQI and low to moderate affected population. The top four countries with the highest affected population are Botswana, Lesotho, Swaziland and South Africa (decreasing order). These countries all lie in Sub-Saharan Africa and have low to moderate HAQI of 51.1, 35.7, 41.9 and 52.0, respectively. Germany, a Europe & Central Asian country, has one of the lowest affected population and a very high HAQI of 86.4.

The effects of HAQI on HIV prevalence was further investigated using the two economic factors: GDP and Gini coefficient. From Fig-

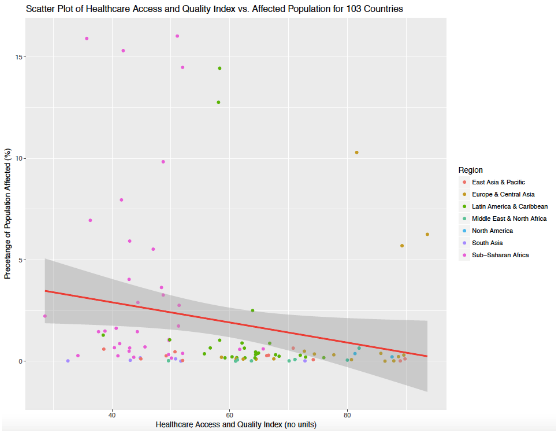
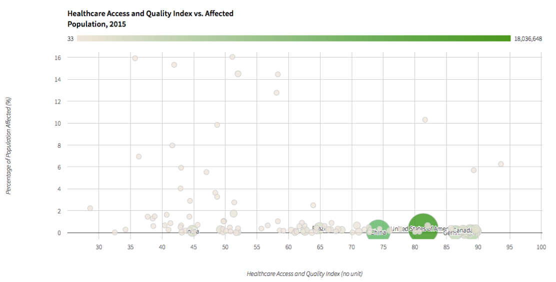


Figure 11: *Scatter plot of HAQI vs. affected population sorted by region, fitted in a linear model*

ure 12a, a positive correlation is found between countries GDP and HAQI. United States of American and China have the highest GDP, very high HAQI and very low affected population. Sub-Saharan African countries have the lowest GDP, lowest HAQI and very high affected population. Gini coefficient, on the other hand, appears to have a negative correlation with HAQI as seen in Figure 12b. In general, Sub-Saharan African countries with high affected population have very high Gini coefficient, while Europe & Central Asian countries with low affected population have very low Gini coefficient.



(a) *HAQI vs. affected population, sized and coloured by GDP*

(b) *HAQI vs. affected population, sized and coloured by Gini coefficient*

Data of three response methods to HIV outbreak are also observed:

1. Antiretroviral received by pregnant women for preventing mother-to-child transmission
2. Antiretroviral therapy received by general population
3. HIV testing and counselling

3.4.1 Variables for plot of ART received by pregnant women

- X: Estimated percentage of pregnant women living with HIV who received antiretroviral given per country (%)
- Y: Percentage HIV prevalence of the given country in 2015 (%)

There was no obvious trend seen in Figure 13. Using a linear regression model directly on the dataset resulted in a R-squared value of 0.0048, which is rendered as very weak a trend. Thus, a K-means clustering algorithm was implemented to cluster the data and then to observe the clusters. 4 clusters were formed and they are shown in Figure 13 using the shapes of the points as well as the red boxes (regions are color coded).

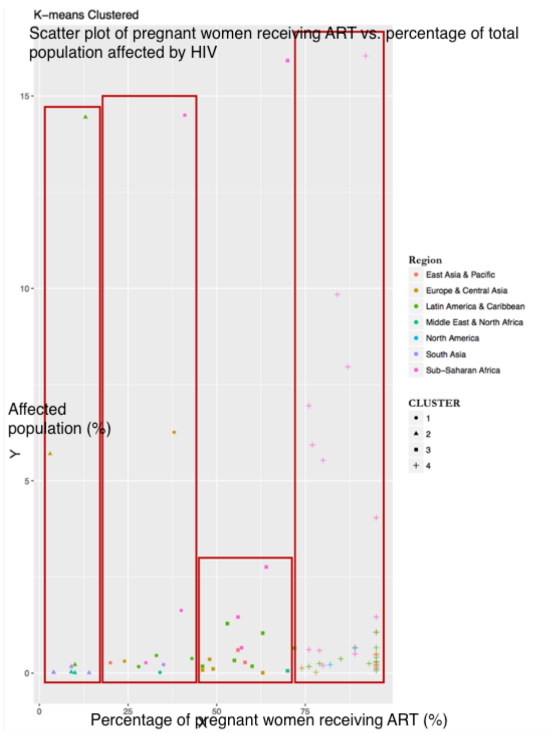


Figure 13: *Scatter plot of pregnant women receiving ART vs. percentage of total population affected, sorted by region and clustered into four clusters using K-means clustering*

Cluster	Mean of X	Mean of Y	Number of Elements	Standard Deviation
1	33.27	2.22	11	8.11
2	1.12	0.00	9	9.72
3	58.00	1.53	17	8.79
4	88.14	1.94	35	8.37
Entire Dataset	61.7	1.64	72	15.60

3.4.2 Cluster Properties

From the standard deviations of the clusters formed in relation to the that of the entire dataset, the clusters are not deemed to be "good" clusters, thus there can be no conclusions made about the dataset.

3.4.3 Variables for plot of general population receiving ART

- X: Percentage of population receiving antiretroviral therapy per country (%)
- Y: Percentage HIV prevalence of the given country in 2015 (%)

Once the dataset has been plotted (Figure 14), attempting a linear regression model produced an R-squared value of 0.0042, thus deeming this too poor of a fit to even consider it being a weak correlation. Thus, a K-means algorithm was implemented and produced 4 clusters as shown in Figure 14

.

3.4.4 Cluster Properties

. From observing the standard deviations of the clusters with respect to that of the entire dataset, only one cluster, Cluster 2 is a "good" cluster, since the standard deviations of the other clusters are much too high. Furthermore, it is seen that 61 of the 76 countries fall into this one cluster centered 0.04% HIV prevalence along with 0.02% population receiving antiviral therapy. This indicates that this response method has a weak or no effect on the HIV prevalence since there is no clear or even weak trend to show this.

3.4.5 Variables for testing and counselling facilities plot

- X: People aged 15 years and over who received HIV testing and counselling per 1000 adult population, given per country (dimensionless)
- Y: HIV prevalence as a percentage over the population of people aged 15 years and older, given per country (dimensionless)

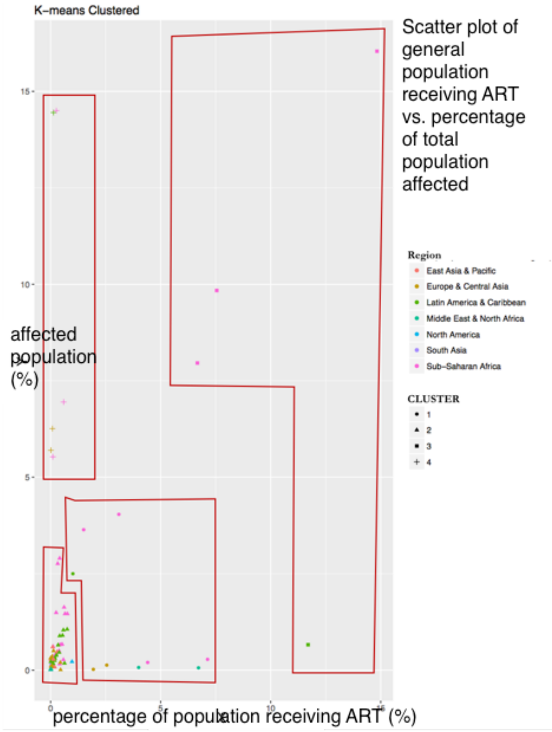


Figure 14: *Scatter plot of percentage of general population receiving ART vs. percentage of population affected, sorted by region and clustered into four clusters using k-means clustering*

Cluster	Mean of X	Mean of Y	Number of Elements	Standard Deviation
1	6.17	0.89	6	3.41
2	0.02	0.04	61	1.10
3	9.69	11.28	3	4.05
4	0.20	8.98	6	3.97
Entire Dataset	1.15	1.64	76	4.20

Plotting the data, a mild trend is seen (Figure 15), so a linear regression model is applied, producing an R-squared value of 0.382, thus it is concluded that this is a good fit.

This is highly contradictory data, since it claims that variables X and Y follow a positive relationship. This indicates that if percentage of people aged 15 years and older do in fact receive testing and counselling, this increases the HIV prevalence. However, the R-squared value of 0.382 suggests that this trend might actually be weak, despite the authors' definition of a high R-squared value.

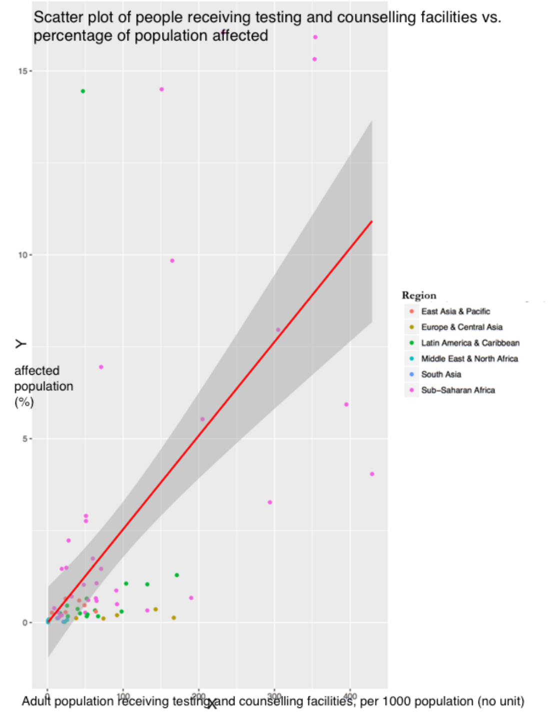


Figure 15: *Scatter plot of people receiving testing and counselling facilities vs. percentage of population affected, sorted by region, fitted in a linear model*

4 Discussion

4.1 Linking Global Crises: Climate Change and HIV Epidemic

Global climate change is causing significant variability and changes in climate factors such as temperature and precipitation, leading to extreme weather events like droughts, floods and heat waves [2]. Such events may lead to increased poverty, migration and livelihood instability, thereby increasing vulnerability and impeding the coping ability of people [2]. Extreme weather also leads to land degradation which decreases agricultural land productivity which in turn leads to malnutrition and health problems such as weakened immune systems [2]. Environmental changes amplify factors favourable for disease transmission, including infections (e.g., malaria and diarrheal disease) that affect people living with HIV/AIDS [2]. Coinfections significantly increases the amount of circulating virus in a HIV patient's blood [2]. Climate change can also decrease plant diversity, including the loss of species that have medicinal or anti-HIV properties [2]. Thus, global climate change disrupts environmental and ecosystem stability which indirectly, but significantly, increases HIV transmission and progression. As such, it was hypothesized prior to the experiment that the correlation between HIV and climate change is positive and so a higher climate change will lead to a higher population affected by HIV.

To establish this hypothesis, the Climate Change Performance Index (CCPI) was identified as a possible factor influencing the global HIV epidemic. CCPI evaluates the climate protection performance of countries that are collectively responsible for more than 90% of global energy-related CO2 emissions [7]. The index categories are weighted as following: 30% for emissions level, 30% for recent emissions development, 10% for renewable energy, 10% for efficiency, and 20% for national and international climate policy assessments [7]. Thus, a higher CCPI indicates that a country is performing well in terms of protecting itself from climate changes and vice versa. It was observed from the results that most countries with very poor to moderate CCPI have the highest affected population, while most countries with very good CCPI have very low affected population. This does not reject the hypothesis that climate change has a proportional impact on HIV distribution. Countries with moderate to good CCPI also have the highest GDP. Countries in Europe & Central Asia have relatively high CCPI. No correlation was found between CCPI and adult literacy rates.

4.2 Economic Factors and HIV

It was hypothesized that the HIV prevalence would be negatively related to the GDP per capita of a given country, due to the fact that the HIV disease will cause damage to the country's capital, thus decreasing the GDP. Overall, this hypothesis was confirmed. This may be because countries with high GDP have high economic resources which can be used to invest into resources such as improving quality of healthcare and education. So, efforts should be taken to increase a country's GDP in order to decrease its HIV prevalence. Research has shown that there is significant evidence showing HIV related deaths correlate to low GDP per capita for countries in South Africa [26]. It has shown this by using a linear regression model to model GDP has a linear combination of the country's arable land, life expectancy, labor force as well as other factors. So perhaps improving these areas could also lead to an increase in GDP, thus leading to a decrease in HIV prevalence.

It can be seen that indeed the countries with the highest HIV prevalence tend to contain the highest Gini coefficient. Furthermore, the countries with the lowest HIV prevalence tend to contain the lowest Gini coefficient, thus indicating a uniform spread of a countries economic wealth could potentially be used to decrease the HIV prevalence of the country. So, it can be advised that these countries should make efforts to bridge the gap between the rich and the poor and this action should be able to decrease the percentage of people whom have HIV. This is made more apparent by studies done on countries in Africa that the low-income sectors are very disadvantaged due to lack of proper education as well as HIV-related services, particularly poor women [27].

4.3 The Role of Education on the Fight Against HIV

Although the definition of literacy is evolving, in this experiment, it was defined to be "ability to read, write and do arithmetic" [17]. It was hypothesized prior to the experiment that high literacy rate will lead to low HIV pandemic. This is because many issues related to HIV are communicated through printed materials at national and international level, and only literate people would be able to understand them [13]. Therefore, they remain aware of crucial information such as what HIV is, how it gets transmitted and how to prevent it, which can help to reduce new HIV outbreaks. This hypothesis is supported by Global Campaign for Education advocacy report written in 2004, which states that "if all children received a complete primary education, around 700,000 cases of HIV in young adults could be prevented each year" [18].

The Results section of the report proves that the relationship between education level and HIV, in reality, is more complex. For example, Figure 1 shows that countries with high literacy rate need not have low HIV prevalence rate. In addition, as discussed earlier, sub-Saharan countries have higher prevalence compared to other regions in the world regardless of their high literacy rate. This finding confirms numerous research papers stating that sub-Saharan Africa has the highest number of HIV affected population [13]. Statistic shows there were 2.8 million new HIV patients from sub-Saharan Africa in 2006, but only 25% of the affected population receive antiretroviral therapy (ART) to control the virus and slow down the progression of the disease [14]. High HAQI values of the sub-Saharan African countries that are greatly affected by HIV seem to reject the observation regarding ART treatment.

Another potential reason why literacy rate does not strongly correlate to HIV infection rate lies on the information delivery method. As internet allows information to be shared across the world, international health organizations have put efforts to inform people about diseases through their websites. This might be an effective way to enhance HIV literacy in developed countries, because people have easy access to the Internet. In contrary, only 35% of people in developing countries had Internet access in 2015. Sub-Saharan countries had the lowest percentage of people who have internet access, some countries with less than 2% [15]. This indicates that the information is not reachable to people in those countries, even though they have the ability to educate and protect themselves from the disease.

Linear regression of figure Z is an indication that educating people is an effective way to prevent HIV. It has been found that people with more schooling are likely to change their behaviors after information about risks of HIV have been delivered [16], for instance, choosing to use condoms or to delay having sex [17]. When GPI and Gini coefficient were added to the analysis of literacy improvement and HIV infection rate, it was expected that data points with high HIV rates would have high GPI and Gini values. This hypothesis was based on earlier researches that showed that countries with a bigger gap in male and female education have higher HIV prevalence rate.

4.4 Access to Healthcare and Treatment for HIV

Countries with higher access to quality healthcare and treatment for HIV are hypothesized to have lower HIV prevalence. This is because increasing access to healthcare can improve quality of life and life expectancy of people with HIV by reducing illness and risk factors [23]. Testing and care facilities and treatment can also prevent HIV and reduce its progression [23].

Healthcare Access and Quality Index (HAQI) measures the national level of personal healthcare access and quality by measuring mortality rates from causes that would otherwise not be fatal if medical care is effective [2]. Highly standardised cause of death and risk factor estimates collected from Global Burden of Diseases, Injuries, and Risk Factors Study (GBD) was used to quantify this index for countries [2]. It is claimed that access to high-quality health care significantly improves health outcomes, including both communicable (e.g. HIV/AIDS, tuberculosis, malaria) and non-communicable diseases (e.g. diabetes and chronic kidney diseases) [2]. As a result, HAQI was used a factor affecting HIV prevalence.

From the results, the hypothesis that there is a positive correlation between HAQI and HIV affected population was accepted. Inequalities in health: access to treatment for HIV/AIDS, an article published in 2007, claims that inequalities in access to health care between rich and poor countries has an effect of HIV infection since resource-limited countries cannot treat AIDS [24]. This suggests that inequality plays an important role in the HAQI of a country. Gini coefficient was shown to have an inverse relationship with HAQI, suggesting that countries with higher Gini coefficient, meaning higher income inequality, have lower HAQI. These countries are also those that have high population of HIV patients. Similarly, GDP was shown to have a positive correlation with HAQI, suggesting that countries with high economic growth and development also have better access to quality healthcare, thus lower HIV population.

The impact of three response methods for the HIV outbreak were also considered as factors affecting HIV distribution. It was hypothesized that all responses– testing and counselling, antiretroviral therapy for the general population and for pregnant women for preventing mother-to-child transmission– have negative correlations with HIV. Antiretroviral therapy (ART) uses antiretroviral (ARV) drugs to suppress HIV virus by slowing the rate at which HIV multiplies, thereby stopping the progression and transmission of the disease [25]. ART has proven to significantly reduce death rates and infections when a potent ARV regimen is used, especially in early stages of the disease [25]. However, no correlation was found between the response methods and HIV prevalence. Countries with high affected population are also found to have the highest response facilities. One way to explain this is that the countries with high HIV prevalence invest more into response methods, but the treatment is ineffective and thus the prevalence does not decrease. This may indicate that this response method should be further investigated and other external factors must be considered in the effort to prevent and treat HIV.

It is a global responsibility to reduce inequality in order to provide all individuals with equal opportunity for a better quality of life. Taking efforts to reduce economic and social inequalities on a national and international level will improve the level of healthcare access and quality to those who need it, thereby decreasing HIV prevalence. Moreover, if HIV testing and counselling is successful, then a higher fraction of people receiving the facilities will decrease the HIV prevalence. It is also important to continue ART and ensure that people affected by HIV receive it as soon as possible after diagnosis to see a clear negative trend between ART and HIV prevalence.

The HIV Syndemic

Figure 16 shows the impact of global scale, secondary and local factors that are contributing to the global HIV epidemic.

Conclusions

In conclusion, while some of the analysis that was done left the hypothesis inconclusive, most of the data had correlations with the various factors that were investigated.

For the response methods, the data relating to pregnant women taking antiretroviral therapy was left inconclusive since the clusters formed had high standard deviation. However, the data corresponding to people receiving antiretroviral therapy formed good clusters and showed that most countries do not offer this form of treatment in the first place. This may indicate that this antiretroviral therapy should be offered in more countries until a definitive answer on their performance can be obtained. It should also be noted that this form of therapy is most effective during the early stages of the disease. Thus, it would be wise to educate more people about these facilities and where to access them. Lastly, HIV testing and counselling data was shown to have a strong positive correlation with HIV prevalence, which is very contradictory to what was hypothesized. One way of explaining this trend could be the fact that perhaps countries with high HIV prevalence are in fact receiving testing and counselling, but this method is not yielding the results expected (which would be to decrease the HIV prevalence). As a result, it is advised to perhaps investigate this response method more closely in order to get a clearer understanding of why this trend exists.

Trends found in economic and healthcare metrics were found to be consistent with the initial hypothesis formed. GDP and HAQI were found to show a negative correlation with HIV prevalence, so it is still advised to improve a country's economic health and provide more access to healthcare facilities to decrease it's HIV prevalence.

Furthermore, it was also found that there exists a negative relationship between adult literacy rate and the HIV prevalence of a given country. This also indeed confirms our hypothesis about this relation. Thus, it is advised to increase a countries Literacy Rate by making education more readily available to more people.

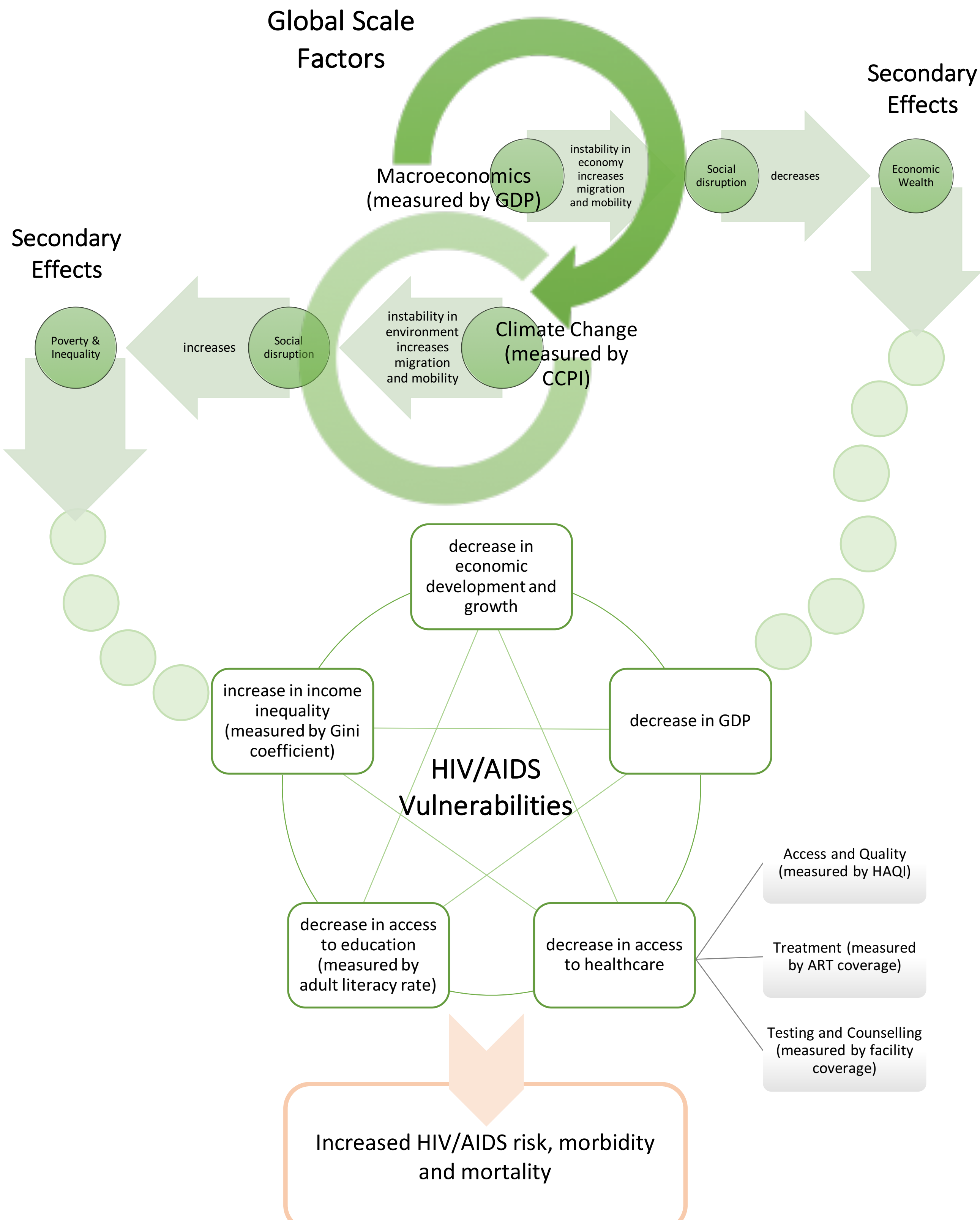


Figure 16 The syndemic framework developed from the results of this paper

(Possible) Mathematical method for analyzing HIV prevalence using the syndemic framework:

We represent the various factors contributing to HIV prevalence as functions.
GDP(t) – GDP at some year t
GC(t) – Gini Coefficient at some year t
LR(t) – Literacy Rate at year t
CCPI(t) – Climate Change Performance Index at year t
HAQI(t) -- Healthcare Access and Quality Index at year t

We can expect that any one of these factors will change because of these factors. In mathematical terms:
Take as an example GDP(t), we can express its change over time as $\frac{\partial}{\partial t}\mathbf{GDP(t)}$
We expect this change to be dependent on other factors as well. Simply put:

$$\frac{\partial}{\partial t}\mathbf{GDP(t) = f(GDP(t), GC(t), LR(t), CCPI(t), HAQI(t))}$$

where f is an arbitrary function
For simplicity we can assume that this function is a linear combination of the factors, so a linear regression model can be used to evaluate it.

$$\textbf{So } \frac{\partial}{\partial t}\mathbf{GDP(t) = a_1GDP(t) + a_2GC(t) + a_3LR(t) + a_4CCPI(t) + a_5HAQI(t)}$$

Where a_1, a_2, a_3, a_4, a_5 are constants (not functions of time, for simplicity)
We can represent these all of these functions as a vector **X** = **[GDP(t), GC(t), LR(t), CCPI(t), HAQI(t)]**
We can now represent how these factors change by taking the time derivative of this vector, shown as:

$$\frac{\partial}{\partial t}\boldsymbol{X(t)} = \boldsymbol{AX(t)}$$

Where **A** is a matrix with constant coefficients
The coefficients of **A** may be unknown, but using the syndemic framework we can make some assumptions on what their values may take on. If their values are known, the system of ODE's can be solved analytically, providing a solution to how these factors evolve over time, taking into account the inter-correlation effects . However this may prove to be tough due to the complex nature of the syndemic framework.
In this case the derivative of **X(t)** can be approximated using a finite difference method:

$$\frac{\partial}{\partial t}\boldsymbol{X(t)} \approx \frac{\boldsymbol{X(t+\Delta t) - X(t)}}{\boldsymbol{\Delta t}}$$

Where $\boldsymbol{\Delta t}$ is small (in this case perhaps a year so $\boldsymbol{\Delta t} = 1$)
Thus, rearranging will provide us the solution:

$$\boldsymbol{X(t+\Delta t) = X(t)(1 + \Delta t A)}$$

Where **1** is the identity matrix. This method can then be used to measure the value of **X(t)** at incremental time steps.
The **X(t)** obtained from this method takes into account the inter-correlation effects of the different factors, these factors can now be used to estimate HIV prevalence.
Note: This method has not been implemented due to its complex nature, it is simpley experimental

It should be noted that this education should be provided to people of all age groups because it was found that specially in African countries, school is the only source for this type of education.

Also, countries with low to moderate CCPI were found to have high HIV prevalence. Thus, countries should increase their efforts to improve their CCPI by decreasing CO2 emission levels and switching to renewable energies in order to maintain a sustainable status. This will significantly reduce the prevalence of HIV by producing a chain reaction in secondary factors, such as poverty, education and inequality, nationally and on a global scale, such that these factors work together in an optimistic manner to end the global HIV epidemic.

Finally, the syndemic framework that was used to chose the factors chosen was mostly validated based on the findings. Furthermore, a new syndemic framework was also constructed based entirely on the results, and provided an experimental mathematical model to use the information derived from the framework to potentially more accurately predict HIV prevalence, taking into account the inter-correlation effects of the factors

References

[1] Singer, Merrill. "Pathogen-Pathogen Interaction: A Syndemic Model of Complex Biosocial Processes in Disease." Virulence 1.1 (2010): 10–18. PMC. Web. 27 May 2017.

[2] Talman, Anna, Susan Bolton, and Judd L. Walson. "Interactions Between HIV/AIDS and the Environment: Toward a Syndemic Framework." American Journal of Public Health 103.2 (2013): 253–261. PMC. Web. 27 May 2017.

[3] "Finance & Development". Finance & Development | F&D. N.p., 2017. Web. 27 May 2017.

[4] "The World Factbook — Central Intelligence Agency". Cia.gov. N.p., 2017. Web. 27 May 2017.

[5] "GHO | By Category | Data On The Size Of The HIV/AIDS Epidemic". Apps.who.int. N.p., 2017. Web. 27 May 2017.

[6] "Population, Total | Data". Data.worldbank.org. N.p., 2017. Web. 27 May 2017.

[7] Jan Burck, Franziska Marten, and Christoph Bals. "The Climate Change Performance Index Results 2015". Think Tank & Research. N.p., 2017. Web. 27 May 2017.

[8] "Income Gini Coefficient | Human Development Reports". Hdr.Undp.Org, 2017, http://hdr.undp.org/en/content/income-gini-coefficient.

[9] "Gross Domestic Product 2015". World Bank. N.p., 2017. Web. 27 May 2017.

[10] "Education : Literacy Rate". Data.Uis.Unesco.Org, 2017, http://data.uis.unesco.org/Index.aspx?queryid=166.

[11] arber, Ryan M et al. "Healthcare Access And Quality Index Based On Mortality From Causes Amenable To Personal Health Care In 195 Countries And Territories, 1990–2015: A Novel Analysis From The Global Burden Of Disease Study 2015". 2017,.

[12] "GHO | By Category | Data On The HIV/AIDS Response". Apps.Who.Int, 2017, http://apps.who.int/gho/data/node.main.624?lang=en.

[13] Making The Connections: Why Literacy Matters For HIV Prevention. 1st ed. UNESCO Institute for Lifelong Learning, 2007. Print.

[14] "Treatment And Care". World Health Organization. N.p., 2017. Web. 26 May 2017.

[15] UN News - Billions Of People In Developing World Still Without Internet Access, New UN Report Finds". UN News Service Section. N.p., 2015. Web. 26 May 2017.

[16] Addressing Gender Relations In HIV Prevention Through Literacy. 1st ed. Carolyn Medel-Anonuevo, 2005. Print.

[17] HIV/AIDS And Literacy: An Essential Component In Education For All. 1st ed. United Nations Educational, Scientific and Cultural Organization, 2005. Print.

[18] Global Campaign for Education (2004). Learning To Survive: How Education For All Would Save Millions Of Young People From HIV/AIDS. Global Campaign for Education. www.campaignforeducation.org.

[19] "Income Gini Coefficient | Human Development Reports". Hdr.Undp.Org, 2017, http://hdr.undp.org/en/content/income-gini-coefficient.

[20] "Education : Literacy Rate". Data.Uis.Unesco.Org, 2017, http://data.uis.unesco.org/Index.aspx?queryid=166.

[21] Barber, Ryan M et al. "Healthcare Access And Quality Index Based On Mortality From Causes Amenable To Personal Health Care In 195 Countries And Territories, 1990–2015: A Novel Analysis From The Global Burden Of Disease Study 2015". 2017,.

[22] "GHO | By Category | Data On The HIV/AIDS Response". Apps.Who.Int, 2017, http://apps.who.int/gho/data/node.main.624?lang=en.

[23] "Access To Health Care". Catie.Ca, 2017, http://www.catie.ca/en/hiv-canada/7/7-1.

[24] S, Giuliano. "Inequalities In Health: Access To Treatment For HIV/AIDS. - Pubmed - NCBI". Ncbi.Nlm.Nih.Gov, 2017, https://www.ncbi.nlm.nih.gov/pubmed/18209264.

[25] "Treatment And Care". World Health Organization, 2017, http://www.who.int/hiv/topics/treatment/en/.

[26] "The Relationship Between HIV Infection Rates And GDP Per Capita In African Countries". Georgia Institute Of Technology, 2017, https://smartech.gatech.edu/bitstream/handle/1853/53303/hivandgdpecon3161.pdf.

[27] Wabiri, Njeri, and Negussie Taffa. "Socio-Economic Inequality And HIV In South Africa". 2017,.