UNIVERSITÉ
DE GENÈVE

Sociobehavioral characteristics and HIV: a longitudinal study in 29 Sub-Saharan African countries

Master's thesis presented
by
Jeffrey Post (18-342-576)
to
The Global Studies Institute

in partial fulfillment of the requirements
for the degree of
Master of Science
in the subject of
Global Health

University of Geneva
Geneva, Switzerland

15 August 2021

Thesis advisor: Aziza Merzouki, PhD Jeffrey Post (18-342-576)

Sociobehavioral characteristics and HIV: a longitudinal study in 29 Sub-Saharan African countries

Abstract

Levels of HIV epidemics vary widely across Sub-Saharan African (SSA) countries. Being an exceedingly diverse continent, sociobehavioral heterogeneity is also substantial. This reality complicates the design of effective interventions that are crucial in a context of changing epidemiology and transition from pandemic to low level endemic epidemics. In this study, we investigated how sociobehavioral heterogeneity across SSA could account for the different levels seen in their respective HIV epidemics.

We analyzed historical (2000-2018) survey data aggregated at the national-level from the DHS for 29 SSA countries. We used Principal Component Analysis to reduce dimensionality and visualized the surveys on a 2D reduced space. We identified the main drivers of sociobehavioral change in the context of progressing HIV epidemics given by HIV incidence and effective contact rates.

The most important characteristics that explained the change on the sociobehavioral space are ART coverage, HIV testing, increase in accepting attitudes towards People Living with HIV/AIDS (PLWHA), and increasing knowledge about AIDS. Three groups of countries transpire within which

HIV incidence is similar and dissimilar across, but effective contact rates are similar across.

Our findings suggest that the initial conditions of nascent epidemics, likely determined by the sociobehavioral factors, are critical in determining the long-term progression and levels of HIV epidemics. Our methods can help design targeted interventions in the pursuit of epidemic control.

Contents

Listing of figures

Listing of tables

0

Introduction

The burden of HIV in SSA is disproportionately large, accounting for nearly two out of three of the 6000 daily new infections that occur globally, and accounted for a staggering 74% of the 1.5 million AIDS related deaths in 2013 (Jones et al., 2014). Levels of HIV epidemics vary widely across the continent and within countries, and this is reflected in the exceedingly diverse range of cultural and sociobehavioral characteristics. Now 35 years since the World Health Organization (WHO) has launched the Special Programme on AIDS, the combination above is the primary reason why renewed and sustained action is needed to get closer to "ending the AIDS epidemic as a public health threat" (Lee et al., 2016).

Previous studies have generally looked at HIV risk factors with rather narrow geographical scopes, limited to single countries (Brockerhoff and Biddlecom, 1999), single regions (Crampin et al., 2003), or specific key populations (Wirtz et al., 2013). Other studies have managed to broaden the range, but have generally looked at a narrow set of risk factors (Hajizadeh et al., 2014), or specific HIV-related attributes. These studies have most commonly used standard descriptive statistics, linear or logistic

regression. Of particular interest is the study published by Merzouki et al. (2021) which assesses sociobehavioral heterogeneity across 29 SSA countries in relation to HIV incidence, but does so from a cross-sectional perspective.

In the pursuit of epidemic control and reduction of the burden the disease causes, numerous programmes and interventions have been implemented. It has become increasingly complex to gauge the impact these various efforts have, all while becoming ever more important to do so in order to prioritize efforts (Galvani et al., 2018). Most of the studies above use HIV incidence as a standard metric to gauge the various risk factors and outcomes, but this may fall short in a context of changing epidemiology and transition from pandemic to low level endemic epidemics (Jones et al., 2014).

In this study we gathered historical sociobehavioral data aggregated at the national-level for 29 SSA countries on which we used unsupervised machine learning techniques to identify major sociobehavioral trends and assess heterogeneity across SSA, both spatially and across time. In parallel we gauge the progression of the various HIV epidemics using the effective contact rate and compare such a metric against the more commonly used HIV incidence.

1

Methods

1.1 Data

The DHS has been gathering nationally representative data at regular intervals since its inception in 1984 (USAID, 2019). The data span social, behavioral, geographic, economic, and health related issues, including HIV-related data and sexual behavior, especially since the emergence of HIV/AIDS as a public health crisis (Curtis and Sutherland, 2004). This has naturally made the DHS the primary source of data for this study.

Data was collected with the intent of including all indicators possibly influencing the risk of acquiring HIV for individuals aged 15-49 years old, or possibly affecting the course of HIV epidemics on local, regional, and national levels. We used the StatCompiler tool (USAID, 2020) to collect the surveys of 29 SSA countries between 2000 and 2018, that were available up to September 2020. While we aimed to go as far back in time as possible to include the emergence of HIV, earlier surveys lacked reliability in the specific indicators of interest, resulting in a total of 80 surveys collected, with each of the 29 countries having between 1 and 6 surveys over that time period; see Table 2.1.

Similarly as Merzouki et al. (2021), we retained only those variables that varied significantly across regions and did not strongly correlate with other variables. The data were represented as percentages, with mean number of sexual partners in lifetime scaled using min-max normalization. The resulting 46 attributes create a 46 dimensional image of each survey collected, in turn each representing a snapshot of a country at a specific point in time.

While most of these attributes were collected using the DHS StatCompiler tool (USAID, 2020), we also sourced data from the Institute of Health Metrics and Evaluation for data pertaining to male circumcision (IHME, 2020), the World Bank for urban/rural divide data (World Bank, 2020b), Gini wealth coefficient (World Bank, 2020c), ART coverage (World Bank, 2020a), and from The Correlates of War Project for religious data (Zeev Maoz and Errol A. Henderson, 2013). Attributes with missing variables for which alternative sources were unavailable were imputed using additional topic-related variables available from the DHS using multiple iterative chained equations – while the detailed analysis of the imputation technique lies outside the scope of this study, all details can be found in Appendix A. Finally, estimates for national HIV incidence and prevalence between the years 1990 and 2019 were obtained from the latest (2020) UNAIDS' AIDSinfo (UNAIDS, 2020, 2018).

1.2 Ethics approval

This study uses publicly accessible data aggregated through StatCompiler, the World Bank, the Institute for Health Metrics and Evaluation, the Correlates of War Project, and UNAIDS' AIDSinfo. No ethics approval was needed from our side.

1.3 Analysis

1.3.1 Sociobehavioral characteristics

We reduced the dimensionality of each survey using Principal Component Analysis (PCA) (Pearson, 1901). PCA allows the transformation of the 46 original dimensions of each survey into a smaller subset of uncorrelated indices called principle components (PCs) along which the variation of data is maximized and information loss is minimized (Jolliffe and Cadima, 2016). As the PCs consist of a linear combination of the initial 46 dimensions, they can be interpreted in terms of the original demographic, socioeconomic, and behavioral characteristics, and can thus be used to construe sociobehavioral heterogeneity or homogeneity across SSA, both spatially and over time. To capture the longitudinal trends in our dataset, and in an effort to facilitate comparison with the results obtained by Merzouki et al. (2021), we obtained the rotation matrix using only the most recent surveys i.e. only those collected since 2015. This rotation matrix was then used to project all other surveys onto the dimensionally reduced PCA space.

While the first 2 PCs, which explain the most variance, are used to represent the axes on a 2D-space to provide a visual perspective of similarity between the surveys, we used a consensus clustering method to provide a more robust means of identifying groups of similar surveys in terms of sociobehavioral characteristics. We predefined 3 as the number of clusters to obtain as was obtained by Merzouki et al. (2021). The consensus clustering method consisted of three hierarchical clustering and three k-means clustering (Lloyd, 1982) partitions, each run on a different number of PCs: the first

with only the 2 main PCs, the second with the n PCs that accounted for 95% of the overall variance, and the third with the entire 46 dimensional data set. We identified longitudinal trends and determined the key sociobehavioral characteristics driving change by visualizing both the trajectories of countries and of the cluster centroids on the 2 dimensional PCA space over time. Finally, we compared the HIV incidence of countries between these clusters by visualizing the HIV incidence using box plots.

1.3.2 Course of HIV epidemics and effective contact rate β

We calculated for each country their effective contact rate β from the estimates of the basic HIV indicators of incidence and prevalence - details of the derivation of this indicator can be found in Appendix B. The effective contact rate β is a measure of incidence given the prevalence (i.e. the number of new infections given the local context of the HIV epidemic) and is effectively both a proxy for the sociobehavioral characteristics of the population while also directly gauging the transmission dynamics of HIV within that population, doing so independently of the local context of the HIV epidemic. We compared the progression of the effective contact rates β to that of HIV incidence across SSA and across the clusters obtained previously and visualized them again with box plots. We compared them over two time frames, first over the 2000-2018 time frame as to have a direct comparison with the DHS data set, and once again over the 1990-2019 time frame so as to have a broader view.

Lastly, we identified clusters of countries based on the evolution of their effective contact rates by first scaling the time-series such that they had zero mean and unit variance (following the assumption that their amplitudes are not as informative as their relative shapes) and used k-means clustering

(Lloyd, 1982) with a dynamic time warping barycenter averaging algorithm (Petitjean et al., 2011). We measured the quality of the clustering results using a silhouette score (Rousseeuw, 1987) and determined the optimal number of clusters - to differentiate with the sociobehavioral clusters above, we called these clusters β clusters. We then compared this new set of β clusters with the previously obtained sociobehavioral clusters and compared their respective HIV incidence and effective contact rates.

1.4 Tools and software

We used the open source Python language (Van Rossum and Drake Jr, 1995), Pandas (McKinney, 2010), NumPy (Harris et al., 2020), SciPy (Virtanen et al., 2020), TSLearn (Tavenard et al., 2020), Scikit-learn (Pedregosa et al., 2011b) for all data aggregating, data wrangling, and data analysis, and used Plotly (Inc., 2015) for data visualizations. All code, data, and other material are available on RenkuLab (`https://renkulab.io/gitlab/jeffrey.post/ssa_hiv_ml`).

2

Results

2.1 Data

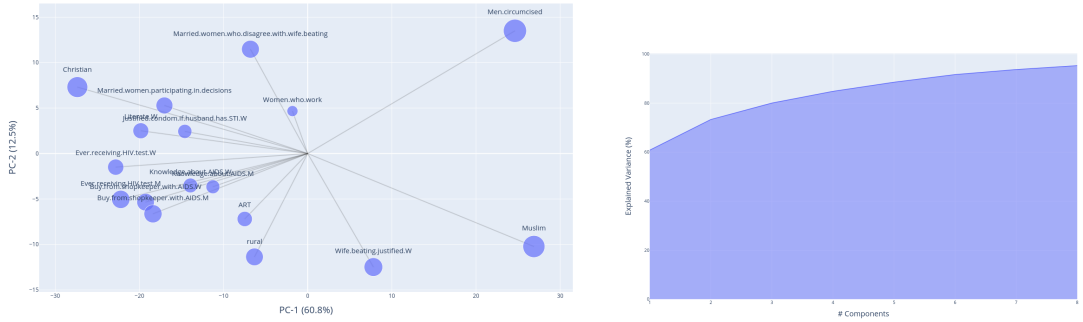
We analyzed 80 surveys from 29 SSA countries, each having between 1 and 6 surveys between the years 2000 and 2018, see Table 2.1. HIV incidence ranged from a minimum of 0.09 new cases per 1000 population in Niger in 2018 to 50.04 new cases per 1000 population in Zimbabwe in 1991. HIV prevalence ranged from less than 0.1% in Gambia and Senegal in 1990 and Gambia in 1991, to 25.4% in Zimbabwe in 1996. HIV mortality ranged from 0.01 per 1000 population in Benin in 1990 to 12.83 per 1000 population in Zimbabwe in 2003. Likewise, sociobehavioral characteristics varied substantially across SSA countries and surveys.

Country	**# of surveys**	**Years of surveys**
Angola	1	2015
Benin	4	2001
		2006
		2011
		2017
Burkina Faso	2	2003

		2010
Burundi	2	2010
		2016
Cameroon	3	2004
		2011
		2018
Chad	2	2004
		2014
Congo	2	2005
		2011
Congo Democratic Republic	2	2007
		2013
Cote d'Ivoire	1	2011
Ethiopia	4	2000
		2005
		2011
		2016
Gabon	2	2000
		2012
Gambia	1	2013
Ghana	3	2003
		2008
		2014
Kenya	3	2003
		2008
		2014
Lesotho	3	2004
		2009
		2014
Liberia	2	2007
		2013
Malawi	4	2000
		2004
		2010
		2015
Mali	4	2001
		2006
		2012
		2018
Mozambique	2	2003

		2011
Namibia	3	2000
		2006
		2013
Niger	2	2006
		2012
Nigeria	4	2003
		2008
		2013
		2018
Rwanda	4	2000
		2005
		2010
		2014
Senegal	6	2005
		2010
		2014
		2015
		2016
		2017
Sierra Leone	2	2008
		2013
Togo	1	2013
Uganda	4	2000
		2006
		2011
		2016
Zambia	4	2001
		2007
		2013
		2018
Zimbabwe	3	2005
		2010
		2015

Table 2.1: Number of surveys per country



(a) (b)

Figure 2.1: Visualization of contributions of sociobehavioral indicators. (a) Projection of the original sociobehavioral indicators onto the reduced 2 dimensional space given by PC-1 and PC-2. The size of the dots represent their contribution (in %). (b) Cumulative contribution of the PCs to total variance in the data set - the first 8 PCs account for over 95% of the variance.

2.2 Analysis

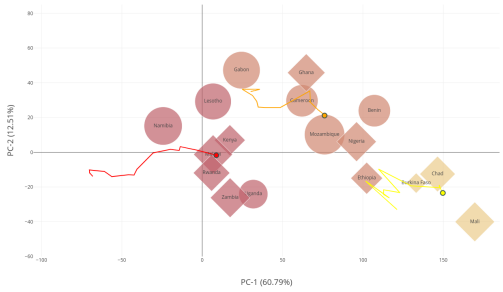
2.2.1 Sociobehavioral characteristics and Principal Component Analysis

Using PCA we found that 95% of the variance found in the original 46 dimension data set can be explained by only the first 8 PCs, with the first, PC-1, explaining 60.7% and the second, PC-2, explaining 12.5% of the total variance across those surveys done in SSA after 2015 (Fig. 2.1b). Of the 46 original sociobehavioral indicators, male circumcision (10.7%) contributed the most to these 2 main PCs, followed by religion (9.8% for Muslim and 8.6% for Christian), acceptance of domestic violence (6.9% for women who think wife beating can be justified and 6.3% for married women who disagree with wife beating), HIV testing (6.6% for men and 5.2% for women), an accepting attitude towards people living with HIV/AIDS (PLWHA) (6.5% for men and 6.2% for women), rurality (6.2%), women participating in decisions (5.7%), literacy (5.0% for women and 4.1% for men), and ART coverage (4.6%) (Fig. 2.1a).

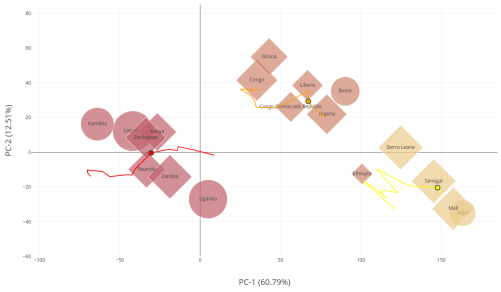
Figure 2.1a shows how the rotation matrix projects the original sociobehavioral characteristics onto the 2 dimensional space of PC-1 and PC-2. Similarly, Figure 2.2a shows how this transformation places the SSA surveys since 2015 onto the same 2D-space. We discern an upside down V-shape, and distinguish on the lower left hand side the first cluster of countries, identified in red, of mostly Southern and Eastern Africa (Burundi, Kenya, Lesotho, Malawi, Rwanda, Uganda, Zambia, and Zimbabwe). These countries are characterized by more accepting attitudes towards PLWHA and better knowledge about HIV, higher literacy rates, higher ART coverage and HIV testing, but notably lower rates of male circumcision. On the lower right hand side, identified in yellow, a second cluster of countries of the Sahel region (Chad, Mali, and Senegal) characterized by a high percentage of Muslim populations, lower mean number of sexual partners, and fewer women participating in decisions. In between and higher on the PC-2 axis, and identified in orange, is the third cluster of countries of mostly Western and Central Africa (Angola, Benin, Cameroon, Ethiopia, Ghana, and Nigeria). These countries are characterized by a less rural population, stronger women empowerment (women who work and disagree with domestic violence), and high rates of male circumcision.

Projecting earlier surveys onto the same 2D-space gives a longitudinal perspective and insight into the evolution of sociobehavioral characteristics in SSA (Figure 2.2) since the turn of the millennium. We notice the upside down V-shape is mostly preserved through time, and although there are some edge cases like Ethiopia (between Sahel and western/central countries) and Gabon (between western/central and eastern/southern countries), the geographical and sociobehavioral similarities remain strong and the clusters easily distinguishable. Since 2000, while no clear movement can be seen on the PC-2 axis, we observe a clear trend of the three clusters to shift towards the negative PC-1 direction, although to a varying degree across the clusters.

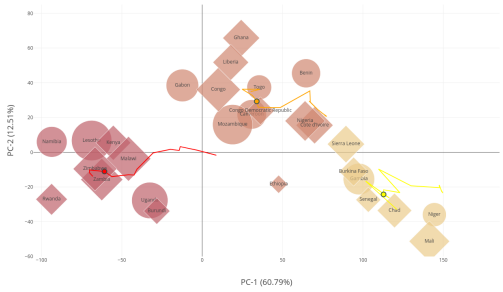
This shift corresponds in all three clusters to an increase in HIV testing, particularly for women (increase from 15% to 85% for the first cluster, 5% to 42% for the second cluster, and 7% to 46% for the third cluster), an increase in ART coverage (increasing to 60%, 42% , and 54% coverage in the



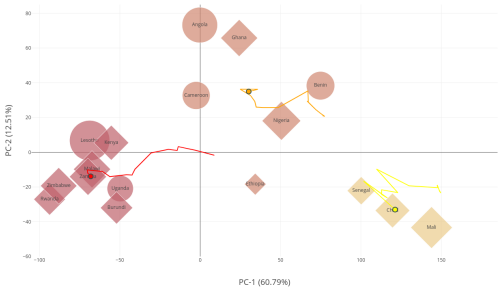
(a) Surveys [2000-2004]



(b) Surveys [2005-2009]



(c) Surveys [2010-2014]



(d) Surveys [2015-2019]

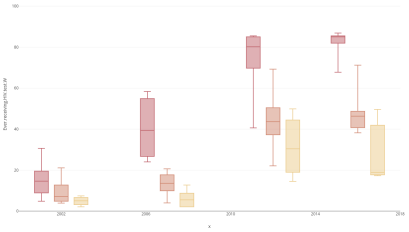
Figure 2.2: Visualization of progression of countries on 2 dimensional space given by PC1- and PC-2. Size represents β - Color represents SB cluster - Shape represents β cluster - Lines represent trajectory of cluster centroids over 2000-2018 with current position marked with a circle. (a) Surveys gathered between the years 2000 and 2004. (b) Surveys gathered between the years 2005 and 2009. (c) Surveys gathered between the years 2010 and 2014. (d) Surveys gathered since 2015.

first, second, and third clusters respectively since its introduction in the early 2000s), and an increase in acceptance of PLWHA (51% to 80% for the first cluster, 26% to 45% for the second cluster, and 27% to 50% for the third cluster). While this leftwards shift is slowed in the second cluster (countries of Sahel region) by a decrease in working women (70% to 50%), the trend is accentuated in the other two clusters by increases in knowledge about HIV (28% to 51% for the first cluster and 18% to 30% for the third cluster), increases in women empowerment indicators (women participating in decision making and disagreeing with domestic violence), and an increase in working men for the first cluster in red (Figure 2.3). From 2000 to 2018 we found the first cluster to move the most, a Cartesian distance of almost 80 units, the second cluster drifts the least with only a change of 30 units, and the third cluster moves almost 50 units. We found the sociobehavioral heterogeneity across clusters tended to increase over time as measured by the 2D Cartesian distance between them: from 2000 to 2018, the first and second clusters moved 142 to 190 units away, the first and third clusters moved from 70 to 110 units away, and the second and third clusters moved from 90 to 110 units away.

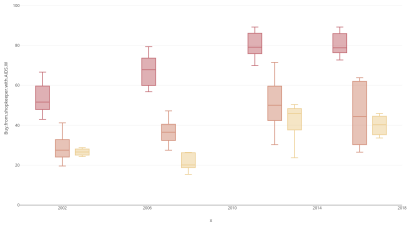
While overall levels of HIV incidence have tended to decrease between 2000 and 2019 we found that countries of the same clusters had mostly similar HIV incidence and that HIV incidence differed across clusters. Countries of the first cluster in red tended to have a very high HIV incidence, countries of the second cluster identified in yellow had low HIV incidence, and the third cluster in orange tended to land somewhere in between with rather low levels of HIV incidence. We found that this is not something that developed over the time period 2000-2019, and found no association between changes in sociobehavioral characteristics since 2000 and the decreases in HIV incidence over the same period (Figure 2.4a).

2.2.2 Course of HIV epidemics and effective contact rate β

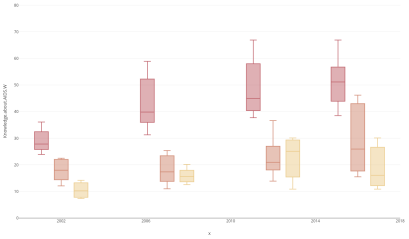
Figure 2.4 also show the progression of the effective contact rates across SSA since 1990. While we discern a sharp decrease from 1990 to 2000, with substantial variation across SSA in 1990 to a point



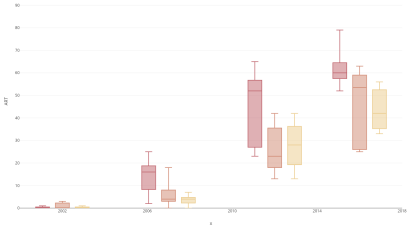
(a) Women who ever received an HIV test



(b) Accepting attitudes towards PLWHA (women)



(c) Knowledge about AIDS (women)

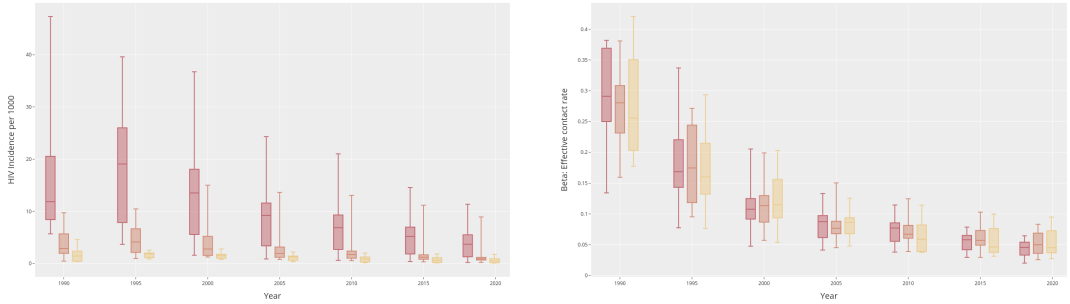


(d) ART coverage

Figure 2.3: Visualization of progression of HIV-related attributes between 2000 and 2018. Color represents SB cluster. (a) Progression of women who ever received an HIV test. (b) Progression of attitudes towards PLWHA (women). (c) Progression of knowledge about AIDS (women). (d) Progression of ART coverage.

where low values are standard across SSA in 2019, we found that the three clusters had a remarkably similar progression of effective contact rates. We did notice a slightly higher average initial value of effective contact rate for countries of the first cluster compared to the other two clusters (0.30 compared to 0.27 and 0.26 for countries of Sahel region and countries of central/western SSA respectively). At the other end, in 2019, countries of the first cluster tend to have a slightly lower average value of β when compared to the other two clusters (0.044 on average compared to 0.052 and 0.054 for countries of western/central SSA and countries of Sahel region respectively).

Our silhouette score indicates countries of SSA formed 2 clusters based on their effective contact rate progression. The first cluster identified in green includes countries whose effective contact rates were already decreasing in 1990 and continued doing so until rapidly reaching low value steady-states around the year 2000 (from an average value of 0.266 in 1990 to 0.094 in 2000, and finally reach-



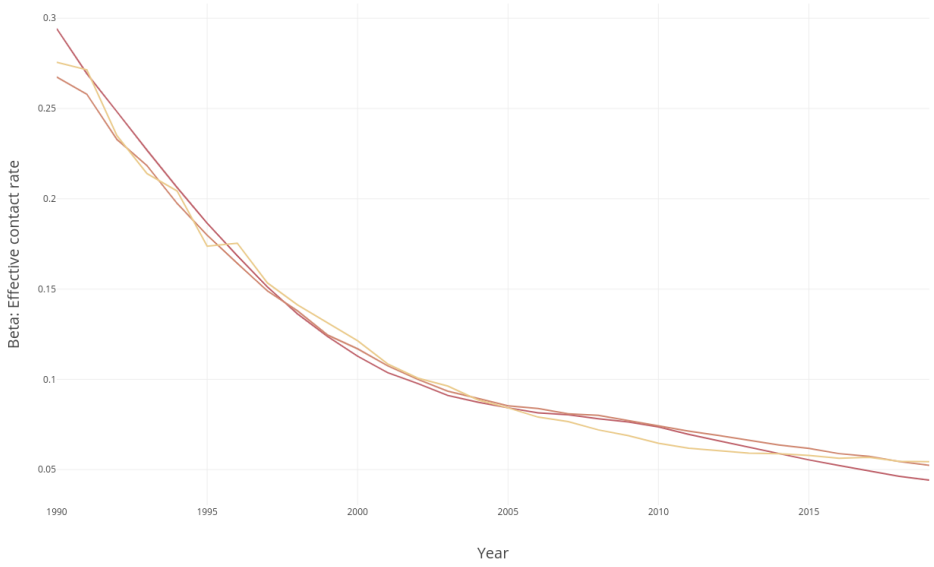
(a) HIV incidence progression per cluster

(b) Effective contact rate progression per cluster

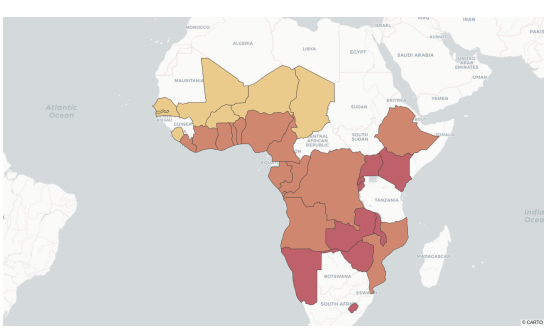
Figure 2.4: Visualization of progression of epidemics across SSA. Color represents SB cluster. (a) Progression of HIV incidence across the 3 clusters between the years 1990 and 2019. (b) Progression of effective contact rates β across the 3 clusters between the years 1990 and 2019.

ing 0.046 in 2019). The second group of countries identified in purple are countries whose effective contact rates tended to only start decreasing later around 1995, and only reaching the low steady-state values between 2005 and 2010 (an average of 0.296 in 1990, 0.153 in 2000, and finally 0.057 in 2019)(Figure 2.6). It should be noted that without first scaling the effective contact rate time-series to have zero mean and unit variance, the clustering reduces to clustering the effective contact rates of the year 1990 as all countries trend to a similar low value over time; a trivial result that does not add information.

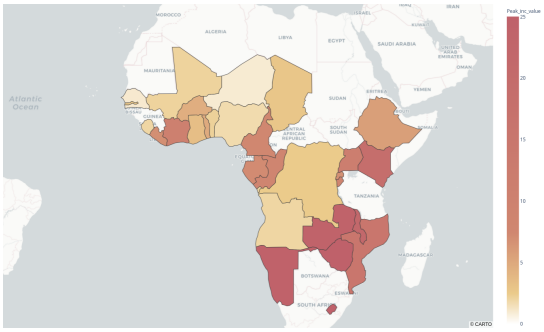
We found that while the 3 original clusters based on sociobehavioral characteristics associate well with how high the peak HIV incidence were, the 2 new clusters based on effective contact rate progression associate with the timing of the epidemics and when those peak HIV incidence levels were reached (Figures 2.5 & 2.6).



(a) Effective contact rates β per SB cluster between 1990 and 2019

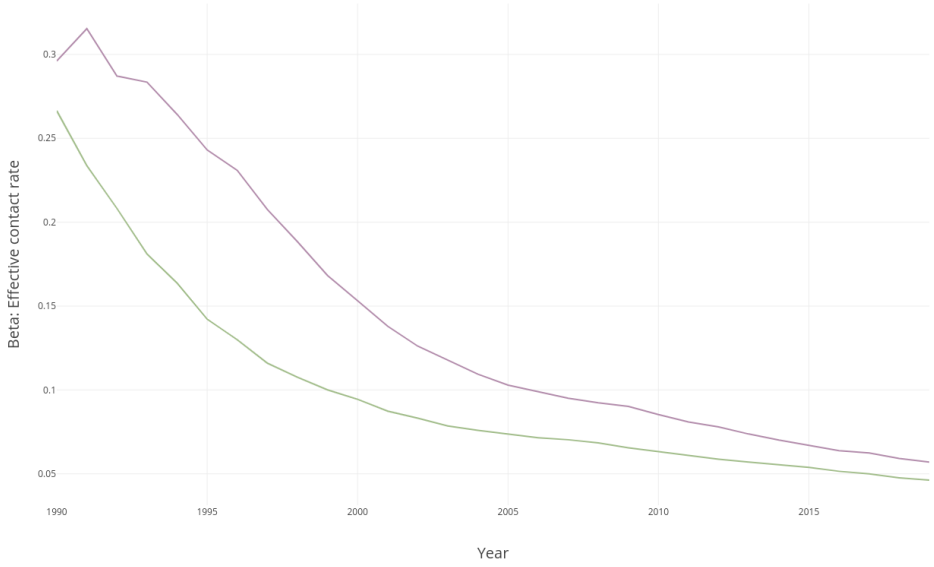


(b) Mapping the SB clusters

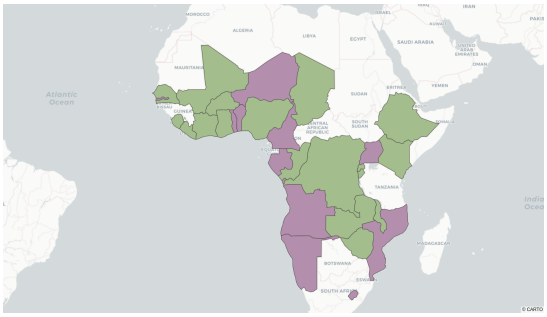


(c) Mapping of peak levels of HIV incidence across SSA

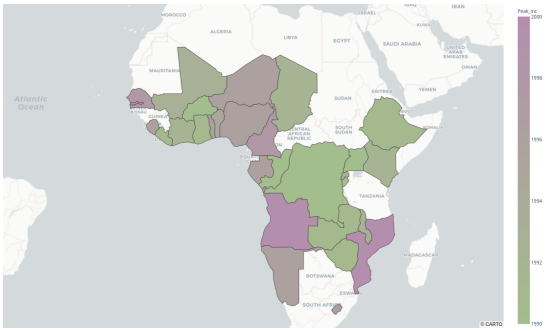
Figure 2.5: Visualization of SB clusters across SSA. (a) Progression of effective contact rates per SB cluster. (b) Map of SSA showing SB clusters. (c) Map of SSA showing peak levels of HIV incidence.



(a) Effective contact rates β per β cluster between 1990 and 2019



(b) Mapping the β clusters



(c) Mapping of peak levels of HIV incidence across SSA

Figure 2.6: Visualization of β clusters across SSA. (a) Progression of effective contact rates per β cluster. (b) Map of SSA showing β clusters. (c) Map of SSA showing year of peak levels of HIV incidence.

3

Discussion

Using PCA and consensus clustering we found three groups of countries with similar sociobehavioral characteristics, with heterogeneity across clusters being explained for the most part (73.2%) by religion, rates of male circumcision, women empowerment, rates of HIV testing, acceptance towards PLWHA, rurality, literacy, knowledge about HIV, and ART coverage. We found HIV incidence to be similar between countries of the same clusters but dissimilar across clusters, in line with the findings of Merzouki et al. (2021). From a longitudinal perspective, we found that while overall HIV incidence rates have significantly decreased across SSA over the 2000-2019 time period, the dissimilar levels persisted across the clusters while levels remained similar within clusters. On the other hand, the effective contact rates, while also decreasing significantly, were remarkably similar across clusters and over time, although they varied substantially within clusters. Over the same time period, heterogeneity on the sociobehavioral space has slightly increased across clusters but homogeneity within clusters remained.

The difference in levels of HIV incidence across the three clusters has been present since at least the early 1990s. Further, the effective contact rates have evolved in a similar manner across clusters over the same period. The different levels of HIV epidemics seen today across the clusters are therefore unlikely to be a direct result of the different evolutions of the clusters as seen on the sociobehavioral space since 2000, and are more likely to be the result of the different initial conditions in which the early years of the HIV epidemics took place. This is perfectly exemplified by Chad, Cote d'Ivoire, and

Malawi, who have an almost identical progression of their effective contact rates between 1990 and 2019, but the very different initial conditions in 1990 (HIV incidence of 2.53 per 1000 for Chad, 9.73 per 1000 for Cote d'Ivoire, and 19.51 per 1000 for Malawi) have lead to large differences in levels of their respective epidemics to this day (HIV incidence of 0.5 per 1000 for Chad, 0.8 per 1000 for Cote d'Ivoire, and 3.71 per 1000 for Malawi).

Small differences in effective contact rates in nascent epidemics can lead to very different levels of epidemics over time (Koopman et al., 1997) (Appendix C). It can thus be hypothesized that the starting effective contact rates of the three clusters were slightly different and the cause of the long-term differences over the course of the entire epidemic. There is some evidence of this in the effective contact rates of the early 1990s, where we noted that countries of the first cluster had a value of 0.30 compared to a slightly lower value of 0.27 and 0.26 for the other two clusters (countries of Sahel region and countries of central/western SSA respectively). Further work on reconstructing the effective contact rates of the 1980s across SSA would allow for better insight into this hypothesis.

Male circumcision has been shown to have a protective effect for seronegative men (Bailey et al., 2007, LEI et al., 2015, Sharma et al., 2018), which could in turn reduce the effective contact rates, all else being equal, of countries with high rates of male circumcision versus countries with low rates of male circumcision. There is some evidence this could be one of the drivers of the early differences in effective contact rates across SSA, where countries with high rates of male circumcision (defined as over 75%) had on average an effective contact rate value lower by 0.0189 when compared to countries with low rates of male circumcision between the years 1990 and 2000. It should be noted that although high rates of male circumcision might have had a protective effect and reduced effective contact rates in nascent HIV epidemics, it is unclear whether voluntary medical male circumcision (VMMC) would have as much of an effect in later stages of epidemics where the effective contact rates have already been drastically lowered (Koopman et al., 1997) - VMMC is a prevention priority of PEPFAR and has been an official recommendation of the WHO since 2007 (Reed et al., 2012).

Since 2005, SSA has also seen a large increase in coverage of ART with countries of the first cluster (eastern/southern SSA) reaching 81% coverage while countries of the other two clusters increased to 52% coverage on average since its introduction in the early 2000s. With high ART coverage being associated with declines in both risk of HIV acquisition and HIV mortality (Tanser et al., 2013), countries of the first cluster would be expected to see a quicker decrease in effective contact rates relative to countries of the other two clusters. Evidence of this can be seen in the latest estimates from UNAIDS (from 2019) showing that countries of the first cluster (eastern/southern SSA) now have the lowest effective contact rates of SSA, with an average value of 0.044 compared to values of 0.052 and 0.054 for countries of western/central SSA and countries of Sahel region respectively. An almost 20% lower value which, if sustained over time, is likely to lead to a more rapid control of the epidemic in those countries.

Our use of effective contact rates as a metric against which to gauge and compare progression of HIV epidemics across SSA is somewhat unusual. As public health programmes have been implemented, the importance of the criteria against which these are evaluated has increased (Galvani et al., 2018). Incidence-based criteria, which are easily understood, have typically been used, while incidence-to-mortality ratio, which follows the decline or increase of total number of people with HIV, or incidence-to-prevalence ratio, which tends to convey information about incidence reduction and survival extension, have increasingly been used (Ghys et al., 2018). While these metrics are critical in order to prioritize focus and optimize programme and intervention implementations, they should only be used with extreme caution to compare different HIV epidemics without first an extensive look at each local context. The effective contact rate has the advantage of being able to extract and convey information about the level of the HIV epidemics given the local context and lends itself more easily as a comparison tool. It however relies on national estimates of incidence and prevalence which may produce large errors (Nsanzimana et al., 2017).

PCA has allowed us to visualize the complex patterns of behavior across SSA and their evolution

over time and while we used more indicators than is typical (Kidman and Anglewicz, 2016, Lakew et al., 2015, Hajizadeh et al., 2014) to account for as many indicators that can influence the course of HIV epidemics, research has shown that using a subset of socioeconomic indicators can provide better results than a broader set (Homenauth et al., 2017). This may be the reason behind the apparent lack of association seen between movement on the reduced PCA space and reduction in effective contact rate. Further, the use of nationally aggregated data allows us to compare countries and clusters across SSA and are useful in the context of generalized epidemics, but are prone to ecological fallacy (Levin, 2006) and overlook the salience of so-called high-risk key populations (Barr et al., 2021), despite the latter being of critical importance in the appropriate allocation of resources for effective interventions, especially so in the context of epidemiological transition and HIV incidence decline across SSA (Garnett, 2021).

4

Conclusion

While sociobehavioral factors play a key role, especially early on in nascent epidemics, latent indicators seem to play an important role in transmission dynamics and reduction of effective contact rates. Our use of principal component analysis and clustering techniques allowed us to identify the complex ways in which sociobehavioral characteristics evolved across SSA as the HIV epidemics progressed. Our methods and findings can help guide further research, especially for targeted country-specific interventions in the pursuit of epidemic control. It should be noted that with the trend of HIV epidemics shifting from generalized epidemics to epidemics of smaller key populations, a people-centered approach is of the utmost importance to avoid risking stigmatisation.

A

Missing variables

A.1 Introduction

Missing values can occur for multiple reasons, from non-responders, to changes in questionnaires overtime, or simply database corruptions (Grung and Manne, 1998). Some techniques of analysis allow for missing variables but analysis of incomplete data sets can often result in misleading conclusions (Dray and Josse, 2015). Further, our decision to use PCA meant that missing data was not allowed as complete datasets are required for the algorithm to work. While there are some related techniques that do work with missing variables, the purpose of our study was also to compare with previous research that used PCA.

The goal of this appendix is to detail the steps taken to mitigate the impact of missing values found in the DHS surveys.

A.2 Methods

A.2.1 Indicators for which we found an alternate source

The raw DHS data set sourced from STATCompiler (USAID, 2020) contained 37 columns, each corresponding to one of our 46 chosen indicators, and 83 rows, each corresponding to a specific survey

(there were 83 total surveys for our 29 SSA countries between the years of 2000 and 2018). The 9 other indicators that were included in our study were sourced from alternative sources as they were either already known to be unreliable or were simply not included in the DHS surveys. These 9 indicators are male circumcision rates ("Men.circumcised" sourced from the Institute for Health Metrics and Evaluation (IHME, 2020)), ART coverage ("ART" sourced from World Bank World Bank (2020a)), data pertaining to rurality ("rural" sourced from World Bank (World Bank, 2020b)), gini wealth index ("Wealth.index.Gini" sourced from World Bank (World Bank, 2020c)), and data pertaining to religion (5 indicators overall: "Christian", "Muslim", "Folk.Religion", "Unaffiliated.Religion", and "Other.Religion" sourced from the Correlates of War Project (Zeev Maoz and Errol A. Henderson, 2013)) and had no missing values.

A.2.2 Identifying missing values from data set

The first step is of course to analyze the data set and get a sense of the impact missing values may have on our plans for analysis. The goal here is twofold, to see how many missing variables there are and where they are missing. As stated above, the raw data contained 46 columns, each corresponding to one of our 46 chosen indicators, and 83 rows, each corresponding to a specific survey. Each missing variable thus has a row and column position corresponding to the survey they belong to and which indicator they represent. It is of interest to list the missing values per country (Table A.1).

Country	Survey	# Missing Values	Percentage missing
Angola	2015	0	0.00
Benin	2001	8	17.39
	2006	3	6.52
	2011	0	0.00
	2017	0	0.00
Burkina Faso	2003	8	17.39
	2010	0	0.00
Burundi	2010	0	0.00

	2016	0	0.00
Cameroon	2004	4	8.70
	2011	0	0.00
	2018	0	0.00
Chad	2004	6	13.04
	2014	0	0.00
Congo	2005	6	13.04
	2011	0	0.00
Congo Democratic Republic	2007	3	6.52
	2013	0	0.00
Cote d'Ivoire	2011	0	0.00
Ethiopia	2000	13	28.26
	2005	3	6.52
	2011	0	0.00
	2016	0	0.00
Gabon	2000	20	43.48
	2012	0	0.00
Gambia	2013	0	0.00
Ghana	2003	4	8.70
	2008	1	2.17
	2014	0	0.00
Kenya	2003	2	4.35
	2008	3	6.52
	2014	0	0.00
Lesotho	2004	5	10.87
	2009	4	8.70
	2014	0	0.00
Liberia	2007	2	4.35
	2013	0	0.00
Malawi	2000	7	15.22
	2004	3	6.52
	2010	0	0.00
	2015	0	0.00
Mali	2001	9	19.57
	2006	3	6.52
	2012	0	0.00
	2018	0	0.00
Mozambique	2003	2	4.35
	2011	0	0.00
Namibia	2000	9	19.57

	2006	1	2.17
	2013	0	0.00
Niger	2006	3	6.52
	2012	0	0.00
Nigeria	2003	4	8.70
	2008	1	2.17
	2013	0	0.00
	2018	2	4.35
Rwanda	2000	6	13.04
	2005	3	6.52
	2007	27	58.70
	2010	0	0.00
	2014	0	0.00
Senegal	2005	5	10.87
	2010	0	0.00
	2012	19	41.30
	2014	0	0.00
	2015	0	0.00
	2016	0	0.00
	2017	0	0.00
	2018	8	17.39
Sierra Leone	2008	1	2.17
	2013	0	0.00
Togo	2013	0	0.00
Uganda	2000	6	13.04
	2006	3	6.52
	2011	0	0.00
	2016	0	0.00
Zambia	2001	7	15.22
	2007	1	2.17
	2013	0	0.00
	2018	0	0.00
Zimbabwe	2005	3	6.52
	2010	2	4.35
	2015	0	0.00

Table A.1: Missing values per country and survey

Removing surveys that don't add information

Table A.1 allows us to identify 2 surveys in particular:

- Senegal/2018 has 8 missing values while Senegal/2017 has none
- Senegal/2012 has 19 missing values while Senegal/2010 has none

Both of these surveys Senegal/2018 and Senegal/2012 can thus be entirely discarded as Senegal will have reliable data in close time proximity.

Table A.1 also allows us to identify the surveys Rwanda/2007 and Rwanda/2005 which have 27 and 3 missing values respectively. Combining them is unfortunately unhelpful as the 3 missing values in Rwanda/2005 are also missing in Rwanda/2007, which can thus be discarded entirely as well.

Identifying indicators for which imputation is necessary

We are now left with the same 46 columns but only 80 rows in which we can easily identify problematic (in the sense that they have many missing values) indicators by summing the number of missing values column-wise (Table A.2), or surveys by summing the number of missing values row-wise (Table A.1 again).

Indicator	**# Missing Values**	**Percentage missing**
Mean.number.of.sexual.partners.W.Normalized	22	26.51
Mean.number.of.sexual.partners.M.Normalized	21	25.30
Justified.condom.if.husband.has.STI.M	20	24.10
Ever.paid.for.sex	19	22.89
Justified.condom.if.husband.has.STI.W	15	18.07
Knowledge.about.AIDS.M	12	14.46
Buy.from.shopkeeper.with.AIDS.M	10	12.05
Buy.from.shopkeeper.with.AIDS.W	10	12.05
Wife.beating.justified.M	8	9.64
Married.women.participating.in.decisions	6	7.23
Ever.receiving.HIV.test.W	5	6.02

Married.women.who.disagree.with.wife.beating	5	6.02
Knowledge.about.AIDS.W	4	4.82
Wife.beating.justified.W	3	3.61
Ever.receiving.HIV.test.M	2	2.41
Number.of.co.wives.1	2	2.41
Unprotected.paid.sex	2	2.41
Number.of.co.wives.0	2	2.41
Number.of.co.wives.2	2	2.41
Literate.M	1	1.20
Access.to.media.W	1	1.20
Access.to.media.M	1	1.20
Number.of.wives.2	1	1.20
Number.of.wives.1	1	1.20
Literate.W	1	1.20

Table A.2: Missing values per indicator

A.2.3 Imputation strategy

It is important to note that no country has all missing values for any particular indicator, or in other words, every country has at least one non-missing value for each of the 46 indicators between 2000 and 2018 – this is critical as it substantially improves our imputation results.

We devised an imputation strategy for the missing values above as follows:

1. Indicators that have 3 or more missing values and that have related indicators in the DHS surveys are imputed individually, these indicators are:

    - Wife.beating.justified.[W/M]
    - Knowledge.about.AIDS.[W/M]
    - Buy.from.shopkeeper.with.AIDS.[W/M]
    - Justified.condom.if.husband.has.STI.[W/M]
    - Mean.number.of.sexual.partners.[W/M]
    - Married.women.participating.in.decisions
    - Married.women.who.disagree.with.wife.beating

- Ever.paid.for.sex

2. The remaining indicators that have only 1 or 2 missing values, or indicators that do not have related indicators in the DHS surveys are then imputed using the entire data set, these indicators are:

    - Literate.[W/M]
    - Access.to.media.[W/M]
    - Ever.receiving.HIV.test.[W/M]
    - Number.of.wives.1
    - Number.of.wives.2
    - Number.of.co.wives.0
    - Number.of.co.wives.1
    - Number.of.co.wives.2
    - Unprotected.paid.sex

A.2.4 Categorical imputation

For those indicators of the first group, we notice we can group them categorically as follows:

- Risk
- Knowledge about AIDS
- Spousal relationship
- Stigma - Accepting attitudes towards PLWHA

For these indicators the optimal imputation procedure is as follows:

1. Categorize the indicator
2. Find other indicators related to the category from STATCompiler (USAID, 2020)
3. Create a data set consisting of all original variables (including missing variables) with the addition of the extra category-related indicators
4. Run the imputation algorithm (see A.2.5 for details)
5. Use box plots to visualize the imputation results (see A.3)

Risk

Indicator to impute	**Extra indicators to be used for imputation**
Ever.paid.for.sex	Higher risk sex in the last year
Mean.number.of.sexual.partners	Condom use at last higher risk sex (with a non-marital, non-cohabiting partner)
	Higher-risk Sex (with multiple partners among all respondents)
	Condom use during higher-risk sex (with multiple partners)

Knowledge about AIDS

Indicator to impute	Extra indicators to be used for imputation
Knowledge.about.AIDS	Men/Women who have heard of HIV or AIDS
	Knowledge of HIV prevention methods
	Use of condoms (prompted)
	Only one partner (prompted)
	Composite of 2 components (prompted)
	No incorrect beliefs about AIDS
	Healthy-looking person can have the AIDS virus
	AIDS cannot be transmitted by mosquito bites
	AIDS cannot be transmitted by supernatural means
	Cannot become infected by sharing food with someone who has AIDS
	Composite of 3 components
	Comprehensive correct knowledge about AIDS
	Knowledge of MTCT
	Knowledge of MTCT risk reduction

Stigma

Indicator to impute	Extra indicators to be used for imputation
Buy.from.shopkeeper.with.AIDS	Willing to care for family member sick with AIDS
	Female HIV+ teacher but not sick should be allowed to teach
	Not secretive about family member's HIV status
	Composite of 4 components
	Adult support of education on condom use

Spousal relationship

Indicator to impute	Extra indicators to be used for imputation
Wife.beating.justified	Justified in refusing sex with her husband if he has sex with other women
Justified.condom.if.husband.has.STI	Justified in asking for condom if she knows that her husband has an sti
Married.women.participating.in.decisions	Justified in refusing sex if she knows husband has sexually transmitted disease
Married.women.who.disagree.with.wife.beating	Wife beating justified for at least one specific reason
	Husband jealousy [7 indicators]

Non-categorical missing variables

Following imputation of the variables above, the remaining missing variables were then imputed without additional indicators.

A.2.5 Imputation algorithm and tools

We used scikt-learn's (Pedregosa et al., 2011a) implementation of a multiple iterative chained equation (van Buuren and Groothuis-Oudshoorn, 2011) imputation technique (Buck, 1960) with an extra-trees regressor with 100 estimators, and bounded between 0 and 100 as we are working with percentages.

Earlier surveys had a larger amount of missing values than more recent surveys, meaning the missing variables were not missing completely at random (Pedersen et al., 2017) and so we took the extra step to display missing variables and imputation results in 5 year groups.

For each categorical imputation we ran the imputation algorithm with 3 different datasets:

1. Using only the original data set (46 columns and 80 rows)
2. Using only the indicators in question with the extra categorical indicators
3. Using all the original data set with the additional categorical indicators

This was only done as a comparison measure, only imputation done as 3. above were kept.

The results can be seen below and are shown in 2 ways:

- Per indicator (i.e. column-wise)
- Per country (i.e. row-wise)

A.3 Results

A.3.1 Knowledge about AIDS

The results of the imputation can be seen here: `https://renkulab.io/gitlab/jeffrey.post/ssa_hiv_ml/-/tree/master/markdown/writing/A_Knowledge`

A.3.2 Stigma - Accepting attitudes towards PLWHA

The results of the imputation can be seen here: `https://renkulab.io/gitlab/jeffrey.post/ssa_hiv_ml/-/tree/master/markdown/writing/A_Stigma`

A.3.3 Risk

The results of the imputation can be seen here: `https://renkulab.io/gitlab/jeffrey.post/ssa_hiv_ml/-/tree/master/markdown/writing/A_Risk`

A.3.4 Spousal relationship

The results of the imputation can be seen here: https://renkulab.io/gitlab/jeffrey.post/ssa_hiv_ml/-/tree/master/markdown/writing/A_Spousal

A.3.5 Non-categorical missing variables

The results of the imputation can be seen here: https://renkulab.io/gitlab/jeffrey.post/ssa_hiv_ml/-/tree/master/markdown/writing/A_Other

B

HIV indicators

B.1 Introduction

In the fight against HIV, numerous programmes and interventions have been implemented in the pursuit of epidemic control and reduction of the burden the disease causes. One the one hand it has become increasingly complex to gauge the impact these various efforts have, and on the other hand it has become ever more important in order to prioritise efforts (Galvani et al., 2018).

While "Ending the AIDS epidemic as a public health threat by 2030" (Lee et al., 2016) is reflected in the SDGs, it has not been formally defined in scientific terms (Ghys et al., 2018). While there are multiple metrics that are commonly used to gauge HIV epidemics, we use this appendix to propose a novel approach, one using the effective contact rate.

B.2 HIV epidemic metrics

The question of which metric to use to gauge HIV epidemics is not a new one (Ghys et al., 2018, Galvani et al., 2018). The most easily understood metrics are those based on HIV incidence. Absolute rates of HIV incidence give a clear indication of how many people are being newly infected with the virus. It is however highly dependent on the local context of the epidemic, as the higher number of infected people there are the higher the chance becomes of being infected (Galvani et al., 2018). Using

a percentage reduction of HIV incidence has the same advantages and drawbacks.

Incidence-to-mortality ratio can also be used as it is clear that if this ratio falls below 1, the total living number of people living with the virus will decrease. While the incidence-to-prevalence ratio conveys information about incidence reduction and survival extension. Both have increasingly been used as HIV epidemic transition metrics (Ghys et al., 2018).

B.3 Effective contact rate as epidemiological transition metric

While the indicators above are easily understood and widely used, neither really take into account the transmission dynamics of HIV. For the purpose of our study, which was not to be gauge the progression of a single HIV epidemic but to compare the progression of man epidemics across SSA, a simpler metric was needed.

B.3.1 Deriving the metric

In its simplest formulation, an individual becomes exposed after an exposure event, from which it can become infected with the virus given a certain probability.

Let r be the number of exposure events an individual has per year (this means number of drug injections, number of sexual partners, etc).

And let ρ be the chance for a susceptible individual to contract the disease after such an exposure event.

ρ is a generic factor which combines many into one, including:

- Male circumcision
- Use of condoms
- Other STDs which may increase chance of HIV infection
- Use of *PrEP*

- It can also include ART coverage (if not already included above in r)
- and so on..

We can combine the two terms into one to give:

$$\beta = \rho * r$$

On a population-level however, the number of susceptible individuals (S) that will become infected and infectious (I) over the course of a year also depends on the proportion of S itself in the population.

Note, any factor impacting susceptibility of an individual (*PrEP*, condom use, etc) is already factored into β.

The total number of new infections in a given year can thus be given by:

$$\text{New infections[n]} = \beta[n] * \frac{I[n]}{N[n]} * S[n]$$

where:

- $S[n] =$ number of susceptible in given year
- $I[n] =$ number of infectious in given year
- $N[n] = S[n] + I[n] =$ total population in given year
- $\beta[n] =$ effective contact rate

UNAIDS provides HIV incidence in "new cases per 1000 population" so:

$$HIV_{Incidence}[n] = \text{New infections[n]} * \frac{1000}{N[n]}$$

$$\iff HIV_{Incidence}[n] = \beta[n] * \frac{I[n]}{N[n]} * \frac{S[n]}{N[n]} * 1000$$

- $\frac{I[n]}{N[n]} = HIV_{Prevalence}$ which is the prevalence given by UNAIDS as a proportion (from 0 to 1)
- $\frac{S[n]}{N[n]} = 1 - \frac{I[n]}{N[n]} = 1 - HIV_{Prevalence}$ (same from 0 to 1 depending on above)

We find that:

$$\beta = \frac{HIV_{Incidence}}{HIV_{Prevalence} * (1 - HIV_{Prevalence})} * 10^{-3}$$

Estimates from UNAIDS of ART coverage, HIV incidence, and HIV prevalence are easily accessible and go back to 1990, with which we can easily calculate the β.

From the description given above, β can easily be interpreted as a proxy for all sociobehavioral characteristics.

B.3.2 Issues

The main issue is that the effective contact rate relies on both data pertaining to HIV incidence and HIV prevalence. While UNAIDS and partners have undertaken a major process for the development of these modelled estimates, the quality of the estimates depend on many underlying factors (Ghys et al., 2018), and large margins of error can be seen which can vastly alter the results of the effective contact rates over time.

B.4 Conclusion

When using HIV indicators it should be clearly defined whether they are being used to gauge effectiveness of programme implementation within a specific epidemic context, or rather used to perhaps

compare different epidemics (either epidemics across countries in different contexts, or simply different sub-populations within the same local epidemic).

While the use of the effective contact rate can be a good proxy for sociobehavioral characteristics in specific situations where some assumptions of the underlying population are met (15-49 year old population for example), and it can be a good metric to compare progression of epidemics across contexts, choosing it as a metric is non-trivial and requires a thorough analysis.

C

Nascent epidemic modeling

C.1 Introduction

The large variability in levels of HIV epidemics across SSA is the source of large amounts of research (Merzouki et al., 2021). While a lot of the research focuses on the epidemiology of HIV in relation to sociobehavioral or socioeconomic characteristics in order to target interventions and prevention measures to reach epidemic control, it is also of interest to research the initial stages of the nascent epidemics to understand how different initial conditions can give way to such differences in the long term.

As we don't have easily accessible data pertaining to HIV incidence and prevalence prior to 1990 for SSA, the goal of this appendix is to show how we used a stochastic agent-based HIV model to evaluate how much the initial conditions of nascent epidemics impact their long-term progression. We did so specifically in relation to the epidemiological data that we found in our main paper.

C.2 Methods

C.2.1 Agent-based modeling

Compartmental models are widely used to model infectious diseases and their transmission dynamics. Compartments describe the state of the individuals, and the transmission dynamics describe the ways in which the compartments interact with each other.

A thorough examination of these compartmental models is outside the scope of this paper but it is important here to differentiate between individual-level disease dynamics and population-level dynamics. The former describes the progression of individual entities between compartments based on individual attributes inherent to them, while population-level dynamics describes on a population-level the total number of individuals in each compartment over time.

An agent-based model means we are focusing on individual-level dynamics. In order to do this, we need to simulate an entire population made up of individuals, who each have unique attributes. We have seen in Appendix B which attributes these are.

C.2.2 Stochastic modeling

Standard compartmental models result in analytical solutions and standard differential equations, meaning they are deterministic in nature with exponentially distributed infectious periods (for SIR model). One advantage of using an agent-based model is that it is trivial to make it stochastic, but the main advantage is that we can use any probability distributions to model individual disease progression (in this case from acute HIV infection after primary infection, followed by clinical latency, and eventually progression to AIDS and death).

C.2.3 Building the model

We were interested in modeling the nascent epidemics of HIV across SSA. In our main paper we found initial values of the effective contact rates β to be between 0.25 and 0.45. It was hypothesized that slight differences in the early values of β can result in large long-term differences in levels of epidemics.

We create a dataframe of 1000000 rows and 3 columns. Each row represents an individual in the population. The first column contains information pertaining to the individual's age, the second column is the status (susceptible, infectious, removed), and the third column records the number of years since the last change of state occurred (to model individual progression of disease).

Some rules:

1. The initial dataframe is composed only of susceptible people and is homogeneously distributed between 15-49 years old
2. We initially infect 10 individuals at Year 0
3. Each individual has a chance of being infected based on the effective contact rate value
4. Each individual follows an individual disease progression
5. After an individual passes 49 years old, it is automatically replaced with a 15 year old susceptible individual

Furthermore, we see from the UNAIDS estimates that the effective contact rate decreases abruptly after a certain numbers of years, or once the prevalence reaches a certain level. It is assumed that prior to the epidemic, each country has an initial steady-state value of $\beta_{Initial}$ which then tends to a final endemic value of β_{Final}. We model this with the following equation:

$$\beta^* = \beta_{Final} + (\frac{\beta_{Initial}}{\exp((j + (\beta_{Initial} * 2.9) - (j + 15) + 1)/(\beta_{Initial} * 27)))}$$

Where j represents the number of years after Year 0 after which we want β to decrease (data from SSA tends to show a value close to 30, so this is what we use).

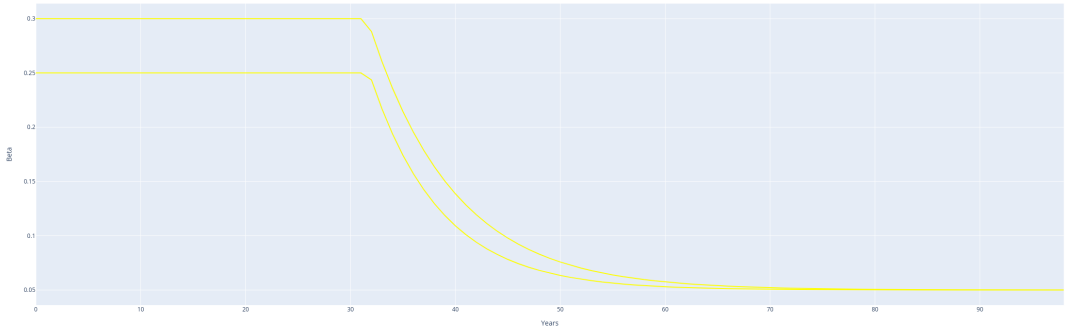


Figure C.1: Visualization of β for both sets of initial conditions.

We modeled two different initial conditions, and thus two sets of progression of β as seen in Figure C.1, and ran the model 10 times for each:

- $\beta_{Initial} = 0.30$ which is close to the average value of the first cluster
- $\beta_{Initial} = 0.25$ which is closer to the average value seen for the other 2 clusters

C.3 Results

A working model can be found here: `https://colab.research.google.com/github/jeffufpost/stochastic_SEIR_model/blob/master/2021_05_05_stochastic_HIV_model.ipynb`

Figure C.2 shows the results of the 20 simulations run.

We notice two different levels of long-term epidemic levels. All 10 of the simulations with $\beta_{Initial}$ of 0.30 resulted in high levels of HIV incidence (up to 50 per 1000 population) and HIV prevalence (up to 4%) for long periods of time. With a slightly smaller $\beta_{Initial}$ of 0.26, we see that the long term trends are largely reduced, with maximum HIV incidence of 7 per 1000 population and maximum HIV prevalence of 1%.

Another interesting result is that we see large variability of long term HIV prevalence and incidence within groups that have the same initial conditions, simply due to the stochasticity of the model,

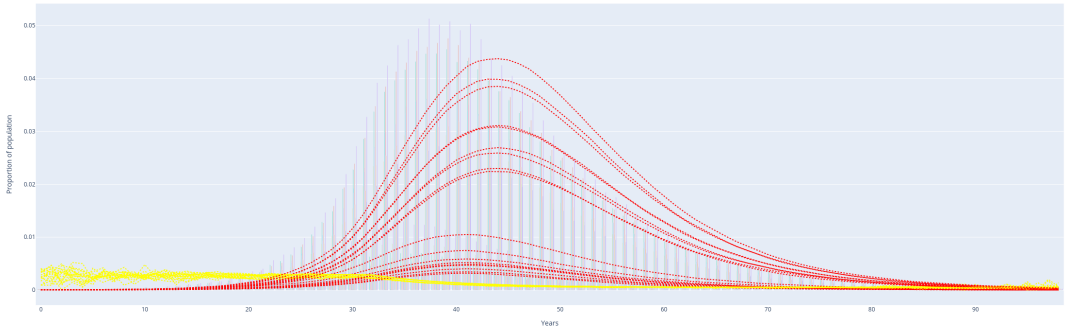


Figure C.2: Results of the simulation. Yellow lines are β - Red dotted lines are prevalence - Bars are incidence.

and the slight perturbations that the populations go through in the early years of the nascent epidemic.

C.4 Conclusion

Although this is a very simplistic simulation that relies on many assumptions, the results tend to be in line with the data and trends seen across SSA and the notion that sustained small differences in nascent epidemics can lead to high levels of long-term variability.

Africa being an exceedingly diverse continent populated by in-homogeneous sub-populations, it is easy to envisage how this could have impacted the continent at the start of the HIV epidemics.

References

R. C. Bailey, S. Moses, C. B. Parker, K. Agot, I. Maclean, J. N. Krieger, C. F. Williams, R. T. Campbell, and J. O. Ndinya-Achola. Male circumcision for hiv prevention in young men in kisumu, kenya: a randomised controlled trial. *The Lancet*, 369(9562):643–656, 2007. ISSN 0140-6736. doi: https://doi.org/10.1016/S0140-6736(07)60312-2. URL https://www.sciencedirect.com/science/article/pii/S0140673607603122.

D. Barr, G. P Garnett, K. H. Mayer, and M. Morrison. Key populations are the future of the african hiv/aids pandemic. *Journal of the International AIDS Society*, 24(S3):e25750, 2021. doi: https://doi.org/10.1002/jia2.25750. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/jia2.25750.

M. Brockerhoff and A. E. Biddlecom. Migration, sexual behavior and the risk of hiv in kenya. *International Migration Review*, 33(4):833–856, 1999. doi: 10.1177/019791839903300401. URL https://doi.org/10.1177/019791839903300401.

S. F. Buck. A method of estimation of missing values in multivariate data suitable for use with an electronic computer. *Journal of the Royal Statistical Society. Series B (Methodological)*, 22(2):302–306, 1960. ISSN 00359246. URL http://www.jstor.org/stable/2984099.

A. C. Crampin, J. R. Glynn, B. M. Ngwira, F. D. Mwaungulu, J. M. Pönnighaus, D. K. Warndorff, and P. E. Fine. Trends and measurement of hiv prevalence in northern malawi. *AIDS*, 17(12), 2003. ISSN 0269-9370. URL https://journals.lww.com/aidsonline/Fulltext/2003/08150/Trends_and_measurement_of_HIV_prevalence_in.11.aspx.

S. L. Curtis and E. G. Sutherland. Measuring sexual behaviour in the era of hiv/aids: the experience of demographic and health surveys and similar enquiries. *Sexually Transmitted Infections*, 80(suppl 2):ii22–ii27, 2004. ISSN 1368-4973. doi: 10.1136/sti.2004.011650. URL https://sti.bmj.com/content/80/suppl_2/ii22.

S. Dray and J. Josse. Principal component analysis with missing values: a comparative survey of methods. *Plant Ecology*, 216(5):657–667, May 2015. ISSN 1573-5052. doi: 10.1007/s11258-014-0406-z. URL https://doi.org/10.1007/s11258-014-0406-z.

A. P. Galvani, A. Pandey, M. C. Fitzpatrick, J. Medlock, and G. E. Gray. Defining control of hiv epidemics. *The Lancet HIV*, 5(11):e667–e670, 2018. ISSN 2352-3018. doi: https://doi.org/10.1016/S2352-3018(18)30178-4. URL https://www.sciencedirect.com/science/article/pii/S2352301818301784.

G. P. Garnett. Reductions in hiv incidence are likely to increase the importance of key population programmes for hiv control in sub-saharan africa. *Journal of the International AIDS Society*, 24(S3): e25727, 2021. doi: https://doi.org/10.1002/jia2.25727. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/jia2.25727.

P. D. Ghys, B. G. Williams, M. Over, T. B. Hallett, and P. Godfrey-Faussett. Epidemiological metrics and benchmarks for a transition in the hiv epidemic. *PLOS Medicine*, 15(10):1–10, 10 2018. doi: 10.1371/journal.pmed.1002678. URL https://doi.org/10.1371/journal.pmed.1002678.

B. Grung and R. Manne. Missing values in principal component analysis. *Chemometrics and Intelligent Laboratory Systems*, 42(1):125–139, 1998. ISSN 0169-7439. doi: https://doi.org/10.1016/S0169-7439(98)00031-8. URL https://www.sciencedirect.com/science/article/pii/S0169743998000318.

M. Hajizadeh, D. Sia, S. J. Heymann, and A. Nandi. Socioeconomic inequalities in hiv/aids prevalence in sub-saharan african countries: evidence from the demographic health surveys. *International Journal for Equity in Health*, 13(1):18, Feb 2014. ISSN 1475-9276. doi: 10.1186/1475-9276-13-18. URL https://doi.org/10.1186/1475-9276-13-18.

C. R. Harris, K. J. Millman, S. J. van der Walt, R. Gommers, P. Virtanen, D. Cournapeau, E. Wieser, J. Taylor, S. Berg, N. J. Smith, R. Kern, M. Picus, S. Hoyer, M. H. van Kerkwijk, M. Brett, A. Haldane, J. Fernández del Río, M. Wiebe, P. Peterson, P. Gérard-Marchant, K. Sheppard, T. Reddy, W. Weckesser, H. Abbasi, C. Gohlke, and T. E. Oliphant. Array programming with NumPy. *Nature*, 585:357–362, 2020. doi: 10.1038/s41586-020-2649-2.

E. Homenauth, D. Kajeguka, and M. A. Kulkarni. Principal component analysis of socioeconomic factors and their association with malaria and arbovirus risk in tanzania: a sensitivity analysis. *Journal of Epidemiology & Community Health*, 71(11):1046–1051, 2017. ISSN 0143-005X. doi: 10.1136/jech-2017-209119. URL https://jech.bmj.com/content/71/11/1046.

IHME. Sub-Saharan Africa Male Circumcision Geospatial Estimates 2000-2017, 2020. URL http://ghdx.healthdata.org/record/ihme-data/sub-saharan-africa-male-circumcision-geospatial-estimates-2000-2017. [Online; accessed 11. Apr. 2021].

P. T. Inc. Collaborative data science. 2015. URL https://plot.ly.

I. T. Jolliffe and J. Cadima. Principal component analysis: a review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374(2065):20150202, 2016. doi: 10.1098/rsta.2015.0202. URL https://royalsocietypublishing.org/doi/abs/10.1098/rsta.2015.0202.

A. Jones, I. Cremin, F. Abdullah, J. Idoko, P. Cherutich, N. Kilonzo, H. Rees, T. Hallett, K. O'Reilly, F. Koechlin, B. Schwartlander, B. de Zalduondo, S. Kim, J. Jay, J. Huh, P. Piot, and M. Dybul. Transformation of hiv from pandemic to low-endemic levels: a public health approach to combination prevention. *The Lancet*, 384(9939):272–279, 2014. ISSN 0140-6736. doi: https://doi.org/10.1016/S0140-6736(13)62230-8. URL https://www.sciencedirect.com/science/article/pii/S0140673613622308.

R. Kidman and P. Anglewicz. Are adolescent orphans more likely to be hiv-positive? a pooled data analyses across 19 countries in sub-saharan africa. *Journal of Epidemiology & Community Health*, 70(8):791–797, 2016. ISSN 0143-005X. doi: 10.1136/jech-2015-206744. URL https://jech.bmj.com/content/70/8/791.

J. S. Koopman, J. A. Jacquez, G. W. Welch, C. P. Simon, B. Foxman, S. M. Pollock, D. Barth-Jones, A. L. Adams#, and K. Lange. The role of early hiv infection in the spread of hiv through populations. *JAIDS Journal of Acquired Immune Deficiency Syndromes*, 14(3), 1997. ISSN 1525-4135. URL https://journals.lww.com/jaids/Fulltext/1997/03010/The_Role_of_Early_HIV_Infection_in_the_Spread_of.9.aspx.

Y. Lakew, S. Benedict, and D. Haile. Social determinants of hiv infection, hotspot areas and subpopulation groups in ethiopia: evidence from the national demographic and health survey in 2011. *BMJ Open*, 5(11), 2015. ISSN 2044-6055. doi: 10.1136/bmjopen-2015-008669. URL https://bmjopen.bmj.com/content/5/11/e008669.

B. X. Lee, F. Kjaerulf, S. Turner, L. Cohen, P. D. Donnelly, R. Muggah, R. Davis, A. Realini, B. Kieselbach, L. S. MacGregor, I. Waller, R. Gordon, M. Moloney-Kitts, G. Lee, and J. Gilligan. Transforming our world: Implementing the 2030 agenda through sustainable development

goal indicators. *Journal of Public Health Policy*, 37(1):13–31, Sep 2016. ISSN 1745-655X. doi: 10.1057/s41271-016-0002-7. URL https://doi.org/10.1057/s41271-016-0002-7.

J. h. LEI, L. r. LIU, Q. WEI, S. b. YAN, L. YANG, T. r. SONG, H. c. YUAN, X. LV, and P. HAN. Circumcision status and risk of hiv acquisition during heterosexual intercourse for both males and females: A meta-analysis. *PLOS ONE*, 10(5):1–9, 05 2015. doi: 10.1371/journal.pone.0125436. URL https://doi.org/10.1371/journal.pone.0125436.

K. A. Levin. Study design vi - ecological studies. *Evidence-Based Dentistry*, 7(4):108–108, Dec 2006. ISSN 1476-5446. doi: 10.1038/sj.ebd.6400454. URL https://doi.org/10.1038/sj.ebd.6400454.

S. Lloyd. Least squares quantization in pcm. *IEEE Transactions on Information Theory*, 28(2):129–137, 1982. doi: 10.1109/TIT.1982.1056489.

W. McKinney. Data structures for statistical computing in python. In S. van der Walt and J. Millman, editors, *Proceedings of the 9th Python in Science Conference*, pages 51 – 56, 2010.

A. Merzouki, J. Estill, E. Orel, K. Tal, and O. Keiser. Clusters of sub-Saharan African countries based on sociobehavioural characteristics and associated HIV incidence. *PeerJ*, 9:e10660, Jan 2021. ISSN 2167-8359. doi: 10.7717/peerj.10660.

S. Nsanzimana, E. Remera, S. Kanters, A. Mulindabigwi, A. B. Suthar, J. P. Uwizihiwe, M. Mwumvaneza, E. J. Mills, and H. C. Bucher. Household survey of hiv incidence in rwanda: a national observational cohort study. *The Lancet HIV*, 4(10):e457–e464, 2017. ISSN 2352-3018. doi: https://doi.org/10.1016/S2352-3018(17)30124-8. URL https://www.sciencedirect.com/science/article/pii/S2352301817301248.

K. Pearson. Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11):559–572, 1901. doi: 10.1080/14786440109462720.

A. B. Pedersen, E. M. Mikkelsen, D. Cronin-Fenton, N. R. Kristensen, T. M. Pham, L. Pedersen, and I. Petersen. Missing data and multiple imputation in clinical epidemiological research. *Clinical epidemiology*, 9:157–166, Mar 2017. ISSN 1179-1349. doi: 10.2147/CLEP.S129785. URL https://pubmed.ncbi.nlm.nih.gov/28352203. 28352203[pmid].

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011a.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and Édouard Duchesnay. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12(85):2825–2830, 2011b. URL http://jmlr.org/papers/v12/pedregosa11a.html.

F. Petitjean, A. Ketterlin, and P. Gançarski. A global averaging method for dynamic time warping, with applications to clustering. *Pattern Recogn.*, 44(3):678–693, Mar. 2011. ISSN 0031-3203. doi: 10.1016/j.patcog.2010.09.013. URL https://doi.org/10.1016/j.patcog.2010.09.013.

J. B. Reed, E. Njeuhmeli, A. G. Thomas, M. C. Bacon, R. Bailey, P. Cherutich, K. Curran, K. Dickson, T. Farley, C. Hankins, K. Hatzold, J. Justman, Z. Mwandi, L. Nkinsi, R. Ridzon, C. Ryan, and N. Bock. Voluntary medical male circumcision: An hiv prevention priority for pepfar. *JAIDS Journal of Acquired Immune Deficiency Syndromes*, 60, 2012. ISSN

1525-4135. URL https://journals.lww.com/jaids/Fulltext/2012/08153/Voluntary_Medical_Male_Circumcision__An_HIV.7.aspx.

P. J. Rousseeuw. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65, 1987. ISSN 0377-0427. doi: https://doi.org/10.1016/0377-0427(87)90125-7. URL https://www.sciencedirect.com/science/article/pii/0377042787901257.

S. C. Sharma, N. Raison, S. Khan, M. Shabbir, P. Dasgupta, and K. Ahmed. Male circumcision for the prevention of human immunodeficiency virus (hiv) acquisition: a meta-analysis. *BJU International*, 121(4):515–526, 2018. doi: https://doi.org/10.1111/bju.14102. URL https://bjui-journals.onlinelibrary.wiley.com/doi/abs/10.1111/bju.14102.

F. Tanser, T. Bärnighausen, E. Grapsa, J. Zaidi, and M.-L. Newell. High coverage of art associated with decline in risk of hiv acquisition in rural kwazulu-natal, south africa. *Science*, 339(6122): 966–971, 2013. ISSN 0036-8075. doi: 10.1126/science.1228160. URL https://science.sciencemag.org/content/339/6122/966.

R. Tavenard, J. Faouzi, G. Vandewiele, F. Divo, G. Androz, C. Holtz, M. Payne, R. Yurchak, M. Rußwurm, K. Kolar, and E. Woods. Tslearn, a machine learning toolkit for time series data. *Journal of Machine Learning Research*, 21(118):1–6, 2020. URL http://jmlr.org/papers/v21/20-091.html.

UNAIDS. Estimates methods, 2018. URL http://aidsinfo.unaids.org/documents/estimates_methods_2018.pdf. [Online; accessed 26. Oct. 2020].

UNAIDS. 2020. aidsinfo, 2020. URL http://aidsinfo.unaids.org/. [Online; accessed 26. Oct. 2020].

USAID. Quality information to plan, monitor, and improve population, health, and nutrition progrems.estimates methods. 2019. URL https://dhsprogram.com/.

USAID. Statcompiler, Aug 2020. URL https://www.statcompiler.com/en. [Online; accessed 26. Oct. 2020].

S. van Buuren and K. Groothuis-Oudshoorn. mice: Multivariate imputation by chained equations in r. *Journal of Statistical Software, Articles*, 45(3):1–67, 2011. ISSN 1548-7660. doi: 10.18637/jss.v045.i03. URL https://www.jstatsoft.org/v045/i03.

G. Van Rossum and F. L. Drake Jr. *Python tutorial*. Centrum voor Wiskunde en Informatica Amsterdam, The Netherlands, 1995.

P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. J. Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. J. Carey, İ. Polat, Y. Feng, E. W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa, P. van Mulbregt, and SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272, 2020. doi: 10.1038/s41592-019-0686-2.

A. L. Wirtz, V. Jumbe, G. Trapence, D. Kamba, E. Umar, S. Ketende, M. Berry, S. Strömdahl, C. Beyrer, and S. D. Baral. Hiv among men who have sex with men in malawi: elucidating hiv prevalence and correlates of infection to inform hiv prevention. *Journal of the International AIDS Society*, 16(4S3):18742, 2013. doi: https://doi.org/10.7448/IAS.16.4.18742. URL https://onlinelibrary.wiley.com/doi/abs/10.7448/IAS.16.4.18742.

World Bank. Antiretroviral therapy coverage (Technical report, The World Bank, Washington, DC, May 2020a. URL https://data.worldbank.org/indicator/SH.HIV.ARTC.ZS.

World Bank. Rural population (% of total population). Technical report, The World Bank, Washington, DC, May 2020b. URL https://data.worldbank.org/indicator/SP.RUR.TOTL.ZS.

World Bank. Gini index (world bank estimate). Technical report, The World Bank, Washington, DC, May 2020c. URL https://data.worldbank.org/indicator/SI.POV.GINI.

Zeev Maoz and Errol A. Henderson. The world religion dataset, 1945-2010: Logic, estimates, and trends. Technical report, The World Bank, 2013. URL https://correlatesofwar.org/data-sets/world-religion-data/world-religion-data-v1-1.