Sociobehavioural characteristics and HIV incidence in 29 sub-Saharan African countries: Unsupervised machine learning analysis

Aziza Merzouki[a,*], Janne Estill[a,b], Kali Tal[c], Olivia Keiser[a]

[a]Institute of Global Health, University of Geneva, Geneva, Switzerland

[b]Institute of Mathematical Statistics and Actuarial Science, University of Bern, Bern, Switzerland

[c]Institute of Primary Health Care (BIHAM), University of Bern, Bern, Switzerland

*Corresponding author:

Aziza Merzouki, PhD
Institute of Global Health, University of Geneva
Chemin des Mines 9, 1202 Geneva, Switzerland
Tel. +41 78 712 56 46
FatmaAziza.Merzouki@unige.ch

*Alternate corresponding author:

Olivia Keiser, PhD
Institute of Global Health, University of Geneva
Chemin des Mines 9, 1202 Geneva, Switzerland
Tel. +41 22 379 41 79
olivia.keiser@unige.ch

Word count: Abstract 247 words; main text 2944 words; 1 table; 4 figures; 1 supplementary material file

Abstract

Objective: HIV incidence varies widely between sub-Saharan African (SSA) countries. This variation coincides with a substantial sociobehavioural heterogeneity, which complicates the design of effective interventions. In this study, we investigated how socio-behavioural heterogeneity in sub-Saharan Africa could account for the variance of HIV incidence between countries.

Methods: We used unsupervised machine learning to analyse data from the Demographic and Health Surveys of 29 SSA countries completed after 2010. We preselected 48 demographic, socio-economic, behavioural and HIV-related attributes to describe each country. We used Principle Component Analysis to visualize sociobehavioural similarity between countries, and to identify the variables that accounted for most sociobehavioural variance in SSA. We used hierarchical clustering to identify groups of countries with similar sociobehavioural profiles, and we compared the distribution of HIV incidence and sociobehavioural variables within each cluster.

Findings: The most important characteristics, which explained 69% of sociobehavioural variance across SSA among the variables we assessed were: religion; male circumcision; number of sexual partners; literacy; uptake of HIV testing; women's empowerment; accepting attitude toward people living with HIV/AIDS; rurality; ART coverage; and, knowledge about AIDS. Our model revealed three groups of countries, each with characteristic sociobehavioural profiles. HIV incidence was mostly similar within each cluster and different between clusters (median(IQR); 0.5/1000(0.6/1000), 1.8/1000(1.3/1000) and 5.0/1000(4.2/1000)).

Conclusion: Our findings suggest that sociobehavioural factors play a key role in determining the course of the HIV epidemic, and that similar techniques can help to design and predict the effects of targeted country-specific interventions to impede HIV transmission.

Research in context

Knowledge before this study

We searched PubMed with the terms: “HIV”, “inequality”, “factors” and “sub-Saharan Africa” for articles published in English before February 28th, 2019. The reviewed literature was usually limited to a certain sub-population, sub-national region, or country; but some recent studies covered up to 31 sub-Saharan African countries. Based on a relatively small number of variable (5 to 13), and using descriptive statistics, regressions and concentration indices, previous works analysed the association between socio-economic inequalities, male circumcision, high-risk sexual behaviour, or HIV-related stigma, with HIV testing, uptake of treatment, ART adherence, or HIV prevalence.

Contribution of this study

To our knowledge, this is the first study where unsupervised machine learning techniques (Principle Component Analysis and hierarchical clustering) were used to analyse the sociobehavioural heterogeneity in sub-Saharan Africa (SSA) and how it associates with the variability of HIV incidence in the region. We identified three distinct sociobehavioural profiles, which were associated with different geographical regions and different levels of HIV incidence in SSA. Because the association between the variability of HIV incidence across SSA and its underlying sociobehavioural factors is still not well understood, we believe that our analysis that compares 29 SSA countries based on 48 sociobehavioural characteristics brings significant value to the field. Identifying and comparing sociobehavioural profiles of countries helps to design and predict the effect of tailored country-specific interventions to impede HIV transmission.

Introduction

The burden of HIV in sub-Saharan Africa (SSA) is the heaviest in the world; in 2017, 70% of HIV-infected people lived in this region [1]. HIV prevalence and incidence vary widely between SSA countries. The region is heterogeneous and sociobehavioural and cultural factors vary widely within and between countries, complicating the design of effective interventions. This heterogeneity ensures that no “one-size-fits-all" approach will stop the epidemic. This is why WHO [2] highlights the need to use data and numerical methods to tailor interventions for specific populations and countries based on quantitative evidence.

So far, studies of HIV risk factors or risk factors for the uptake of interventions against HIV have generally been limited to specific sub-populations [3-5], sub-national regions [6-9] or countries [10-17]. Recent studies included up to 31 SSA countries, but narrowly focused their inquiries to examine, for example, the association between socio-economic inequalities [18], high-risk sexual behaviour [19], or HIV-related stigma [17, 20] with HIV testing, treatment uptake, ART adherence, or HIV prevalence. Most used standard statistical methods like descriptive statistics [5, 13], linear or logistic regression [3, 4, 20, 21], or concentration indices [6, 10, 18], to assess health inequity and the impact of 5 to 13 variables on the HIV epidemic. But, these methods do not tell us how HIV risk factors vary across SSA and which characteristic patterns are actually associated with different rates of new HIV infections in the region. Comparing and characterising SSA countries would allow us to test the hypothesis that sociobehavioural heterogeneity might account for spatial variance of HIV epidemic, and inform effective country-specific interventions.

We thus used unsupervised machine learning techniques (Principle Component Analysis and hierarchical clustering) to identify the most important factors of 48 national attributes that might account for variability of HIV incidence across sub-Saharan Africa, and identified the sociobehavioural profiles that characterized different levels of HIV incidence, based on Demographic and Health Survey [22] data from 29 SSA countries.

Methods

Data

We used Demographic and Health Surveys (DHS) that contained data from 2010 or later. These DHS contained the most recent data that came from 29 SSA countries up to July 2018 (**Table S1**). DHS typically gathers nationally representative data on health (including HIV-related data) and population (including social, behavioural, geographic and economic data) every 5 years, and provides individual- and country-level data.

We pre-selected the following variables because they covered topics that could relate to HIV and were available for all selected countries: age (under 25 vs older); rurality (rural vs urban); religion (Christian, Muslim, Folk/Popular religions, unaffiliated, others); marital status (married or in union vs widowed/divorced/other), number of wives (1, ≥2) or co-wives (0, 1, ≥2); literacy (literate vs illiterate); media access (with access to newspaper, television and radio at least once a week vs without such access); employment (worked in the last 12 months and currently working vs others); wealth (Gini coefficient); age at first sexual intercourse (first sexual intercourse by age 15 vs older); general fertility (number of births to women of reproductive age in the last 3 years); contraception use (using any method of contraception vs not using any); condom use (belief that a woman is justified in asking condom use if she knows her husband has an STI vs belief that she is not justified); number of sexual partners in lifetime; unprotected higher risk sex (men who had sex with a non-marital, non-cohabiting partner in the last 12 months and did not use condom during last sexual intercourse vs not); paid sex (men who ever paid for sexual intercourse vs never paid for sex); unprotected paid sex (men who used condom during the last paid sexual intercourse in the last 12 months vs did not use condom); gender-based violence (wife beating justified for at least one specific reason vs not justified for any reason); married women participation to decision making (yes vs no); gender of household head (female vs male); comprehensive correct knowledge about AIDS (yes vs no); HIV testing (ever receiving an HIV test vs never tested); male circumcision (yes vs no); ART coverage (i.e. percentage of people on antiretroviral treatment among those living with HIV); and accepting attitudes toward people living with HIV/AIDS (would buy fresh vegetables from a shopkeeper with AIDS vs would not); see **Table 1** for a complete summary of the variables.

We represented each country using 48 dimensions. Each dimension corresponded to an attribute in **Table 1**, such as the percentage of women married or in union, the mean number of sexual partners in a lifetime for men, the percentage of Christian populations and the Gini coefficient in this country. Data were represented as percentages; the mean number of sexual partners in lifetime was normalised using min-max normalisation. Most of these country-level data were exported from the DHS with the StatCompiler tool, except for data on religion that we obtained from Pew-Templeton Global Religious Futures Project [23], and ART coverage that we obtained from UNAIDS' AIDSinfo [24]. We used the latest (2018) UNAIDS estimates of national HIV incidence for the year 2016 [24, 25].

Analysis

We used Principle Component Analysis (PCA) [26, 27] to reduce the data from 48 to two dimensions (2D) so we could visualize sociobehavioural similarity between SSA countries; countries closest to each other on the 2D space corresponded to similar countries in terms of demographic, socio-economic and behavioural characteristics. The principle components (PCs) consist of a linear combination of the initial 48 dimensions and can therefore be interpreted in terms of the original variables. The first two PCs, which explain the most variance, represent the axes of the 2D-space used for visualization.

We used hierarchical clustering to identify similar SSA countries in terms of sociobehavioural characteristics. Pairwise countries dissimilarity was calculated using the Euclidian distance (**Equation S1**). These distances were used by the hierarchical clustering algorithm to create a *dendrogram* with 29 terminal nodes representing the countries to be grouped. Cutting the dendrogram at a certain height produces clusters of similar countries. The number of clusters depends on the height at which the tree is cut. To measure the quality of the clustering results and to select the final number of clusters, we used the Silhouette Index (**Equation S4**).

Having clustered countries based on sociobehavioural variables, we then determined if countries with similar sociobehavioural patterns tend to have similar HIV incidence. We used *box plots* to visualize the distribution of the HIV incidence within each cluster of countries. To identify the sociobehavioural variables that characterize the resulting clusters, we visualized and compared the distribution of these variables within each cluster with *density plots*.

We used the open source R language, version 3.5.1 for our analysis. Code and country-level data are available on GitLab (https://gitlab.com/AzizaM/dhs_ssa_countries_clustering).

Results

The surveys we used in this analysis included 594’644 persons (183’310 men and 411’334 women), ranging from 9’552 in Lesotho to 56’307 in Nigeria. Adult HIV incidence ranged from 0.14/1000 in Niger to 19.7/1000 in Lesotho in 2016. HIV prevalence ranged from 0.4% in Niger to 23.9% in Lesotho (**Table S1**). Sociobehavioural characteristics varied widely between SSA countries (**Table 1**).

Visualizing the SSA countries: Geographical and sociobehavioural similarities

Using PCA, we found that the first principle component (PC) explained 49.5% and the second 19.5% of the total sociobehavioural variance across SSA among the 48 variables we considered (**Figure 1**). The original sociobehavioural variables that contributed most to these PCs were religion (12.6% for Muslim and 12.1% for Christian populations), male circumcision (9.4%), number of sexual partners (7.8% for men and 3.4% for women), literacy (6.1 % for women and 3.2% for men), HIV testing (5.5% for men and 5.4% for women), women’s participation in decision making (3.8%), an accepting attitude towards those living with HIV/AIDS (3.6% for women and 3.2% for men), rurality (3.0% for women and 2.7% for men), ART coverage (2.5%), and women’s knowledge about AIDS (2.5%) (**Figure 1, right panel** and **Figure S1**).

Projecting the 29 SSA countries in two dimensions produced a roughly V-shaped scatterplot (**Figure 1, left panel**). As the two dimensions combine the 48 original sociobehavioural variables, we explored the scatterplot given sociobehavioural trends over the 2D-space (**Figure 1, right panel**). At the end of the V-shape’s left branch, Eastern and Southern African countries (such as Namibia, Zimbabwe, Malawi, Zambia and Uganda) lied next to each other. In these countries, less men are circumcised, but the percentage of literate people who had accepting attitudes toward people living with HIV/AIDS (PLWHA) was higher and so was uptake of HIV testing. Knowledge about AIDS and ART coverage were also high. The end of the right branch, in the upper right quadrant, included countries from the Sahel region, like Senegal, Burkina Faso, Mali, Niger and Chad, where the percentage of Muslims is higher and people have fewer sexual partners. The lower tip of the V-shape included countries in West and Central Africa, like Liberia, Ghana, Côte d'Ivoire, Democratic Republic of the Congo, and Gabon, where people have more sexual partners, more men are circumcised, and the rural population is smaller.

Clustering the SSA countries and analysis of the associated HIV incidence

The hierarchical clustering of the 29 SSA countries built a dendrogram (**Figure 2, left panel)**. Cluster compactness and separation were optimal (maximum silhouette index = 0.3) when we cut the dendrogram at a height that separated countries into three groups (**Figure 2, right panel**).

The countries of the first cluster, in yellow, had the lowest HIV incidence (median of 0.5/1000 population) (**Figure 3**). This cluster included countries from the Sahel Region, where the population was mostly rural (median of 71.1% for men) and Muslim (median of 86.2%). On the one hand, many of the factors that characterized this cluster could account for low HIV incidence and prevalence in these countries. Countries were characterized by high proportions of circumcised men (median of 95.0%), high percentages of women who were married or lived in union (median of 70.6%), late sexual initiation for men (median of 1.9% of men who had their first sexual intercourse by the age of 15), low numbers of sexual partners (median of 3.5 partners for men), low percentages of unprotected higher-risk sex (median of 9.7% for men) and low percentages of men having ever paid for sex (median of 3.9%).

Polygyny [9, 28], an institutionalized form of sexual concurrency, was also frequent in this region (median of 22.3 %). On the other hand, this cluster was also characterized by frequent belief that wife beating is justified (median of 61.2% for women), and low levels of literacy (median of 29.0% for women). Participation of married women in decision making (median of 18.5%), contraceptive prevalence (median of 13.9%), and knowledge about AIDS (median of 23.7 % for women) was also low. These countries had low percentages of people ever tested for HIV (median of 19.2% for men; 36.6% for women), low ART coverage (median of 38.0%) and low levels of acceptance of PLWHA (Median of 47.4% for men); see **Figure 4**.

The countries of the second cluster, coloured in orange, included countries from West and Central Africa. These countries had a rather low HIV incidence (median of 1.8/1000 population), though Mozambique was a remarkable outlier, with a high HIV incidence (9.8/1000 population) (**Figure 3**). Like the first cluster, these countries had a high percentage of circumcised men (median of 97.0%, except in Mozambique where only 48.4% of men were circumcised). However, these countries were also characterized by the lowest proportions of rural populations (median of 49.0% for men), the highest numbers of sexual partners (median of 10.1 for men), early sexual initiation (median of 12.0 % of men who had their first sexual intercourse by the age of 15), and more frequent unprotected high-risk sex (median of 24.3% for men) and paid sexual intercourse (median of 9.5% for men). HIV testing uptake (median of 25.8% for men and 48.6% for women), knowledge about AIDS (median of 23.6% for women), and ART coverage (median of 31.0%) were all low.

The third cluster, in red, included Southern and East African countries. These countries had high HIV incidence (median of 5.0/1000 population), except two countries that had a lower HIV incidence: Rwanda (1.1/1000 population) and Burundi (0.5/1000) (**Figure 3**). Countries belonging to the third cluster were characterized by the lowest percentage of circumcised men (median of 27.9%). But they were also the ones with the highest uptake of HIV testing (median of 65.2% for men; 83.3% for women) and ART (median of 61.0%), and the highest percentage with knowledge about HIV (median of 54.6% for women) and accepting attitudes towards PLWHA (median of 84.4% for men). This cluster was also characterized by the highest percentage of literacy (median of 80.2% for women), high use of contraceptives (median of 42.6%), low percentages of unprotected high-risk sex (median of 9.8% for men) and higher percentages of married women participating in decision making (median of 67.7%) and women-headed households (median of 31.0%). Rwanda and Burundi had the lowest HIV incidence and were characterized by a lower number of sexual partners (Rwanda, 2.6; Burundi, 2.1) vs a median of 6.3 partners for men in the other countries of the third cluster. They also had larger per capita rural populations (Rwanda, 80.4%; Burundi, 89.4%) vs a median of 61.3% for women in the other countries of the same cluster.

Discussion

Using hierarchical clustering, we identified most important characteristics that explained 69% of the sociobehavioural variance among the variables we assessed in SSA. We discovered three groups of countries with similar sociobehavioural patterns, and HIV incidence was also similar within each cluster.

In the first cluster, PLWHA were not widely accepted, and the population had an overall low-level knowledge about HIV. Stigma may be more widespread in this region and explain the lower uptake of interventions among people who are HIV-positive. The relatively low number of people who are living with HIV lowers the general public's exposure to this group and may increase stigma [29]. Stigma can also result from cultural and religious beliefs that link HIV/AIDS with sexual transgressions, immorality and sin [30, 31].

We speculate that the apparent contradiction between the presence of many high-risk factors and low HIV incidence in most countries of the second cluster could be explained by the high proportion of circumcised men. In line with this theory, Mozambique, the only country in this cluster with very high HIV prevalence and incidence, had few circumcised men. Previous observational studies and trials have confirmed the protective effect of male circumcision [7, 8, 32, 33].

Countries of the third cluster, with the highest HIV incidence, were also the ones with the highest knowledge about AIDS [29], ART coverage, uptake of HIV testing, and with the most accepting attitudes toward PLWHA. They also had the lowest percentage of unprotected higher risk sex. These findings are consistent with earlier studies that found broad ART coverage may reduce social distancing towards PLWHA and HIV-related stigma in the general population [20, 34]. Reduced social distancing and stigma is associated with higher uptake of voluntary HIV counselling and testing [17, 35], and less sexual risk-taking among HIV positive people [21].

The high HIV incidence in Mozambique could be caused by any combination of the following factors: a high number of sexual partners; a low level of male circumcision; a low level of literacy and knowledge about AIDS. These, in turn, could be responsible for low uptake of HIV testing and ART. In contrast, many West and Central African countries with population characteristics like Mozambique, e.g., sexual practices, literacy, knowledge about AIDS, HIV testing and ART coverage, had much lower HIV prevalence and incidence, possibly because males were circumcised at twice the rate. It is also possible that despite a low uptake of male circumcision, the combination of lower numbers of sexual partners, higher per capita rural populations, more literacy, more accurate knowledge about AIDS, more HIV testing, and broader ART coverage could account for the lower HIV incidence in Rwanda and Burundi.

The cross-sectional nature of our data makes it impossible to determine precedence and causality between the sociobehavioural characteristics we measured and HIV prevalence and incidence. But the associations we identified can open lines of inquiry for researchers. Our study had the advantage of allowing us to compare countries and regions, but ecological studies that use aggregated data are prone to confounding and ecological fallacy [36]. Africa is an exceedingly diverse continent with many distinct sub-populations, so a study based on national population averages cannot explain HIV variation within countries. Therefore, we intend to repeat the study at a lower level of granularity, using regional- and individual-level data to capture differences within countries and learn more about sociobehavioural factors that affect the sub-populations that are most at risk.

Our work has some other limitations. We used model estimates for HIV incidence, which may diverge from reality [37]. And even though we included many more variables from the DHS and other sources than is common practice [3, 4, 10, 11, 18, 19], we still had to exclude many more, including other sexually transmitted diseases, alcohol consumption, ART adherence and drug resistance data. Some of the variables we wanted to include were not collected in the DHS or were missing from some countries.

Our use of unsupervised machine learning allowed us to identify the most important characteristics among the variables we assessed that explained 69% of the sociobehavioural variance in SSA countries. We captured complex patterns of sociobehavioural characteristics shared by countries with similar HIV incidence, suggesting that the combination of sociobehavioural factors play a key role in determining the course of the HIV epidemic, and that similar techniques can be used to design and predict the effect of targeted country-specific interventions to impede HIV transmission.

Acknowledgements

We thank Zofia Baranczuk for helpful discussions.

Funding

This work was supported by the Swiss National Science Foundation [grant n° 163878].

Conflict of interest

We declare no competing interests.

References

1. Fact sheet - World AIDS Day 2018. Available at: http://www.unaids.org/sites/default/files/media_asset/UNAIDS_FactSheet_en.pdf

2. Global Health Sector Strategy on HIV 2016-2021: Towards Ending AIDS. Available at: http://apps.who.int/iris/bitstream/handle/10665/246178/WHO-HIV-2016.05-eng.pdf?sequence=1.

3. Ashaba S, Cooper-Vince C, Maling S, Rukundo GZ, Akena D, Tsai AC. Internalized HIV stigma, bullying, major depressive disorder, and high-risk suicidality among HIV-positive adolescents in rural Uganda. Global Mental Health **2018**; 5.

4. Kidman R, Anglewicz P. Are adolescent orphans more likely to be HIV-positive? A pooled data analyses across 19 countries in sub-Saharan Africa. Journal of Epidemiology and Community Health **2016**; 70(8): 791-7.

5. Sangowawa AO, Owoaje ET. Experiences of discrimination among youth with HIV/AIDS in Ibadan, Nigeria. Journal of Public Health in Africa **2012**; 3(1): 10.

6. Pons-Duran C, González R, Quintó L, et al. Association between HIV infection and socio-economic status: evidence from a semirural area of southern Mozambique. Tropical Medicine & International Health **2016**; 21(12): 1513-21.

7. Bailey RC, Moses S, Parker CB, et al. Male circumcision for HIV prevention in young men in Kisumu, Kenya: a randomised controlled trial. **2007**; 369: 14.

8. Gray RH, Kigozi G, Serwadda D, et al. Male circumcision for HIV prevention in men in Rakai, Uganda: a randomised trial. **2007**; 369: 10.

9. Eaton JW, Takavarasha FR, Schumacher CM, et al. Trends in Concurrency, Polygyny, and Multiple Sex Partnerships During a Decade of Declining HIV Prevalence in Eastern Zimbabwe. The Journal of Infectious Diseases **2014**; 210(suppl_2): S562-S8.

10. Kim SW, Skordis-Worrall J, Haghparast-Bidgoli H, Pulkki-Brännström A-M. Socio-economic inequity in HIV testing in Malawi. Global Health Action **2016**; 9(1): 31730.

11. Lakew Y, Benedict S, Haile D. Social determinants of HIV infection, hotspot areas and subpopulation groups in Ethiopia: evidence from the National Demographic and Health Survey in 2011. BMJ Open **2015**; 5(11): e008669.

12. Antelman G, Kaaya S, Wei RL, et al. Depressive symptoms increase risk of HIV disease progression and mortality among women, in Tanzania. Jaids-J Acq Imm Def **2007**; 44(4): 470-7.

13. Smith Fawzi MC, Ng L, Kanyanganzi F, et al. Mental Health and Antiretroviral Adherence Among Youth Living With HIV in Rwanda. PEDIATRICS **2016**; 138(4): e20153235-e.

14. Tsai AC, Venkataramani AS. The causal effect of education on HIV stigma in Uganda: Evidence from a natural experiment. Social Science & Medicine **2015**; 142: 37-46.

15. McGillen JB, Stover J, Klein DJ, et al. The emerging health impact of voluntary medical male circumcision in Zimbabwe: An evaluation using three epidemiological models. PLOS ONE **2018**; 13(7): e0199453.

16. Gregson S, Gonese E, Hallett TB, et al. HIV decline in Zimbabwe due to reductions in risky sex? Evidence from a comprehensive epidemiological review. International Journal of Epidemiology **2010**; 39(5): 1311-23.

17. Kelly JD, Weiser SD, Tsai AC. Proximate Context of HIV Stigma and Its Association with HIV Testing in Sierra Leone: A Population-Based Study. AIDS and Behavior **2016**; 20(1): 65-70.

18. Hajizadeh M, Sia D, Heymann S, Nandi A. Socioeconomic inequalities in HIV/AIDS prevalence in sub-Saharan African countries: evidence from the Demographic Health Surveys. International Journal for Equity in Health **2014**; 13(1): 18.

19. Kenyon C, Buyze J, Schwartz IS. Strong association between higher-risk sex and HIV prevalence at the regional level: an ecological study of 27 sub-Saharan African countries. F1000Research **2018**; 7: 1879.

20. Chan B, Tsai A. Trends in HIV-Related Stigma in the General Population During the Era of Antiretroviral Treatment Expansion: An Analysis of 31 Sub-Saharan African Countries. Open Forum Infectious Diseases **2015**; 2(suppl_1).

21. Delavande A, Sampaio M, Sood N. HIV-related social intolerance and risky sexual behavior in a high HIV prevalence environment. Social Science & Medicine **2014**; 111: 84-93.

22. The DHS Program - Quality information to plan, moniotr, and improve population, health and nutrition programs. Available at: http://www.dhsprogram.com.

23. Religions in Africa | African Religions | PEW-GRF.

24. AIDSinfo | UNAIDS. Available at: http://aidsinfo.unaids.org/.

25. Estimates Methods 2018. Available at: http://aidsinfo.unaids.org/documents/estimates_methods_2018.pdf.

26. Hastie T, Tibshirani R, Friedman J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Second Edition ed: Springer.

27. James G, Witten D, Hastie T, Tibshirani R. An introduction to statistical learning: with applications in R. New York: Springer, **2013**.

28. Reniers G, Tfaily R. Polygyny, Partnership Concurrency, and HIV Transmission in Sub-Saharan Africa. Demography **2012**; 49(3): 1075-101.

29. Chan BT, Tsai AC. Personal contact with HIV-positive persons is associated with reduced HIV-related stigma: cross-sectional analysis of general population surveys from 26 countries in sub-Saharan Africa. Journal of the International AIDS Society **2017**; 20(1): 21395.

30. Campbell C, Foulis CA, Maimane S, Sibiya Z. “I Have an Evil Child at My House”: Stigma and HIV/AIDS Management in a South African Community. American Journal of Public Health **2005**; 95(5): 808-15.

31. Mbonu NC, van den Borne B, De Vries NK. Stigma of People with HIV/AIDS in Sub-Saharan Africa: A Literature Review. Journal of Tropical Medicine **2009**; 2009: 1-14.

32. Lei Jh, Liu Lr, Wei Q, et al. Circumcision Status and Risk of HIV Acquisition during Heterosexual Intercourse for Both Males and Females: A Meta-Analysis. PLOS ONE **2015**; 10(5): e0125436.

33. Sharma SC, Raison N, Khan S, Shabbir M, Dasgupta P, Ahmed K. Male circumcision for the prevention of human immunodeficiency virus (HIV) acquisition: a meta-analysis. BJU International **2018**; 121(4): 515-26.

34. Chan BT, Tsai AC, Siedner MJ. HIV Treatment Scale-Up and HIV-Related Stigma in Sub-Saharan Africa: A Longitudinal Cross-Country Analysis. American Journal of Public Health **2015**; 105(8): 1581-7.

35. Kalichman SC. HIV testing attitudes, AIDS stigma, and voluntary HIV counselling and testing in a black township in Cape Town, South Africa. Sexually Transmitted Infections **2003**; 79(6): 442-7.

36. Levin KA. Study Design VI - Ecological Studies. Evidence Based Dentistry **2006**; 7: 108.

37. Nsanzimana S, Remera E, Kanters S, et al. Household survey of HIV incidence in Rwanda: a national observational cohort study. The Lancet HIV **2017**; 4(10): e457-e64.

Table 1 - Socio-economic and behavioural variables included in the analysis. Median values across all 29 countries are shown with the minimum and maximum values. *ART = antiretroviral therapy.

Attribute	Topic	Variable	Stratification	Categories	Median (min - max)
1	Demographic	Age under 25	Men		37.6% (28.1%-44.1%)
2			Women		39.9% (34.4%-45.0%)
3		Rurality	Men		56.5% (12.9%-85.1%)
4			Women		59.7% (11.3%-89.4%)
5		Religion		Christian	74.9% (0.8%-97.8%)
6				Muslim	13.9% (0.0%-98.5%)
7				Folk/Popular	1.7% (0.0%-35.7%)
8				Unaffiliated	2.5% (0.0%-18.0%)
9				Others	0.2% (0.0%-2.7%)
10		Married or in union	Men		50.5% (28.8%-65.2%)
11			Women		63.5% (34.0%-88.5%)
12		Number of wives or co-wives	Men	1	87.5% (72.0%-97.5%)
13				≥2	12.5% (2.5%-28.0%)
14			Women	0	75.5% (57.6%-93.2%)
15				1	17.2% (1.9%-30.4%)
16				≥2	4.3% (0.4%-12.3%)
17		Female headed household			28.0% (9.3%-43.9%)
18		Literacy	Men		79.0% (37.6%-94.2%)
19			Women		58.1% (14.0%-97.0%)
20		Access to media at least once a week	Men		9.9% (1.7%-47.5%)

21			Women		5.6% (0.3%-21.3%)
22	Employment	Worked in the last 12 months and is currently working	Men		76.9% (55.9%-92.8%)
23			Women		61.8% (24.5%-77.8%)
24	Wealth	Gini coefficient[1]			30.0% (10.0%-50.0%)
25	Sexual behaviour	First sex by age 15	Men		8.0% (0.8%-25.4%)
26			Women		18.0% (2.6%-28.8%)
27		Fertility rate	Women		17.5% (11.8%-26.9%)
28		Use of contraception	Women		21.7% (5.4%-50.2%)
29		Woman is justified asking for condom if husband has a sexually transmitted infection (STI)	Men		88.2% (70.3%-98.5%)
30			Women		81.5% (14.3%-97.3%)
31		Mean number of sexual partners in lifetime	Men		6.3 (1.9-15.3)
32			Women		2.2 (1.2-5.1)
33		Unprotected higher risk sex	Men		15.7% (1.6%-43.2%)
34			Women		11.10% (0.3%-30.3%)
35		Ever paid for sexual intercourse	Men		7.7% (1.4%-35.0%)
36		Unprotected paid sexual intercourse	Men		0.8% (0.1%-8.1%)
37	Gender-based violence	Wife beating justified	Men		32.3% (12.5%-59.5%)
38			Women		45.7% (16.2%-76.3%)
39	Women empowerment	Married women participating in decision making			49.9% (9.1%-78.0%)
40		Married women who disagree with all reason justifying wife beating			47.7% (18.7%-80.9%)
41	HIV/AIDS	Correct knowledge about AIDS	Men		35.8% (17.4%-68.8%)

42			Women		27.8% (10.9%-66.9%)
43		Ever received an HIV test	Men		30.5% (7.8%-80.8%)
44			Women		49.6% (14.5%-85.5%)
45		Male circumcision			94.0% (14.3%-99.4%)
46		ART* coverage 2015			41.0% (18.0%-76.0%)
47	Accepting attitudes toward PLWHA	Would buy vegetables from shopkeeper with AIDS	Men		57.5% (32.4%-92.1%)
48			Women		53.1% (23.7%-89.2%)

[1] The Gini coefficient indicates the level of wealth concentration in a country.

Figure 1 - Visualization of the sociobehavioural similarity between SSA countries using PCA.

Left panel: Projection of the SSA countries on a 2D-space, based on their socio-economic and behavioural factors. The two dimensions (first two PCs), Dim1 and Dim2, explained 69% of the variance in the data. Countries are coloured based on their HIV incidence per 1000 population (15-49) in 2016.

Right panel: Correlation plot of the original variables with the first and second dimensions (Dim1, Dim2). The variable transparency represents its contribution (in %) to the two dimensions. Moving along a variable's vector leads toward a region of the 2D-space where the variable levels tend to be higher, e.g. upper right quadrant contains mainly Muslim countries, while upper left quadrant contains countries with higher levels of HIV testing and knowledge about AIDS.

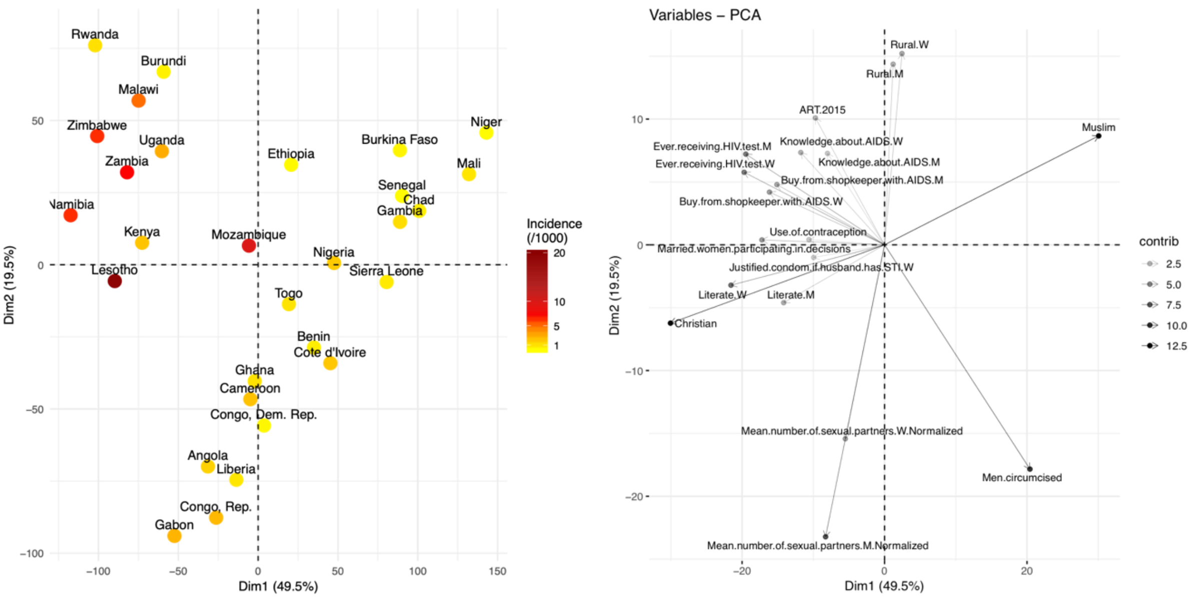


Figure 2 - Hierarchical clustering of 29 sub-Saharan African countries

Left panel: Dendrogram. Cutting the tree at the height of the red dashed line results in three clusters, highlighted in yellow, orange and red.

Right panel: Average Silhouette width for different numbers of clusters. The number of clusters (X axis), from 2 to 10, corresponds to different heights at which the dendrogram was cut. The maximum average Silhouette width was obtained for 3 clusters (red circle).

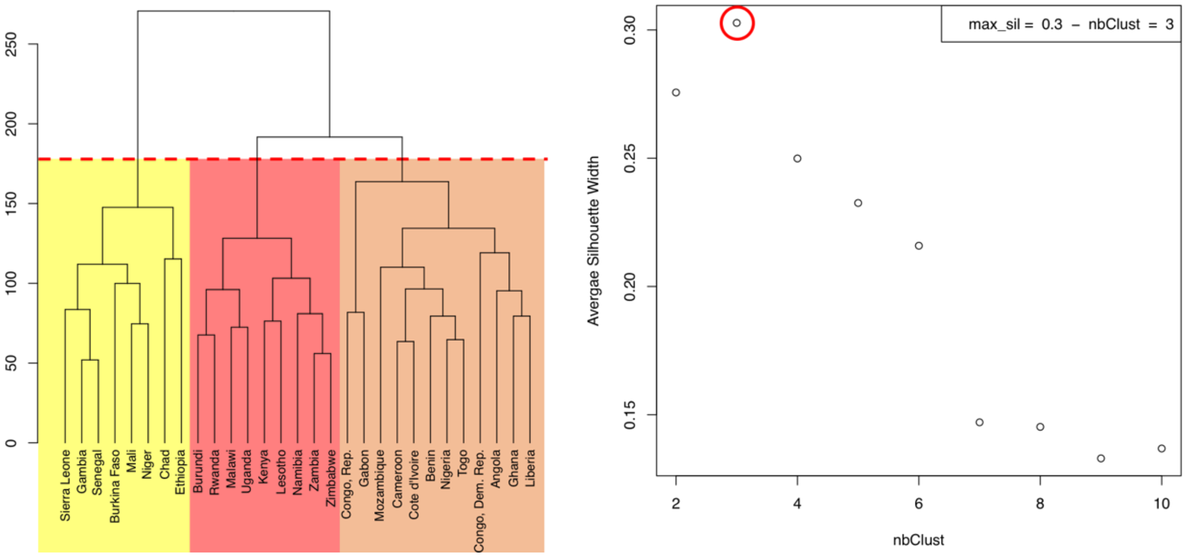


Figure 3 - Analysis of the resulting clusters.

Left panel: Map of clustered sub-Saharan countries. Countries are coloured based on the cluster to which they belong.

Right panel: Box plots of the HIV incidence distribution within each cluster.

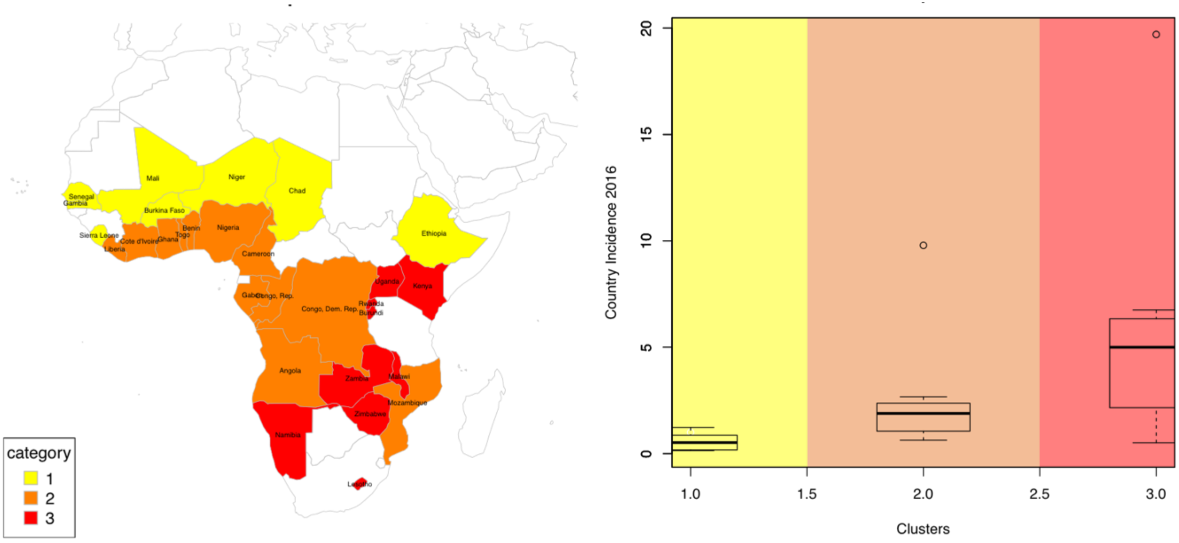
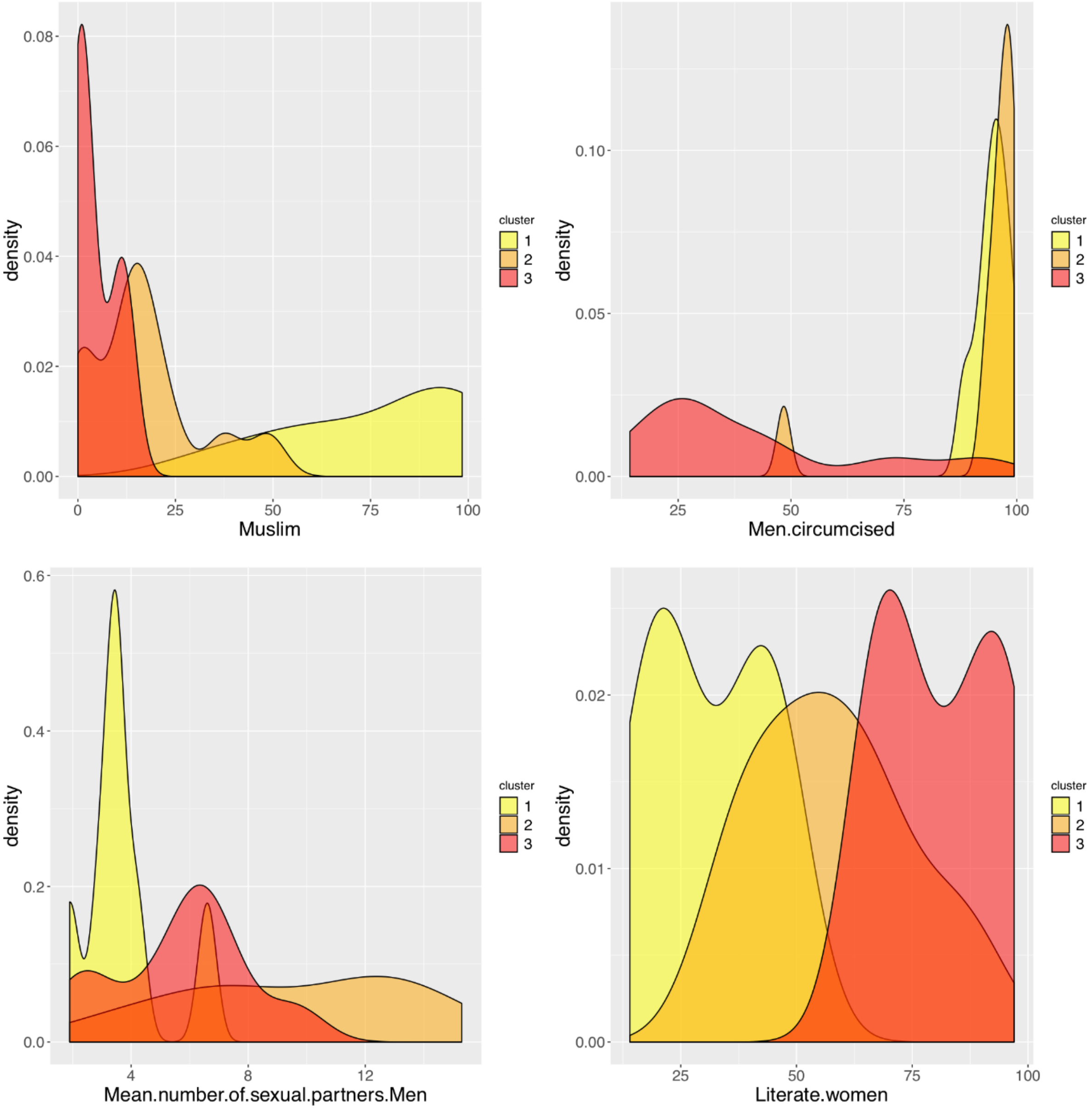
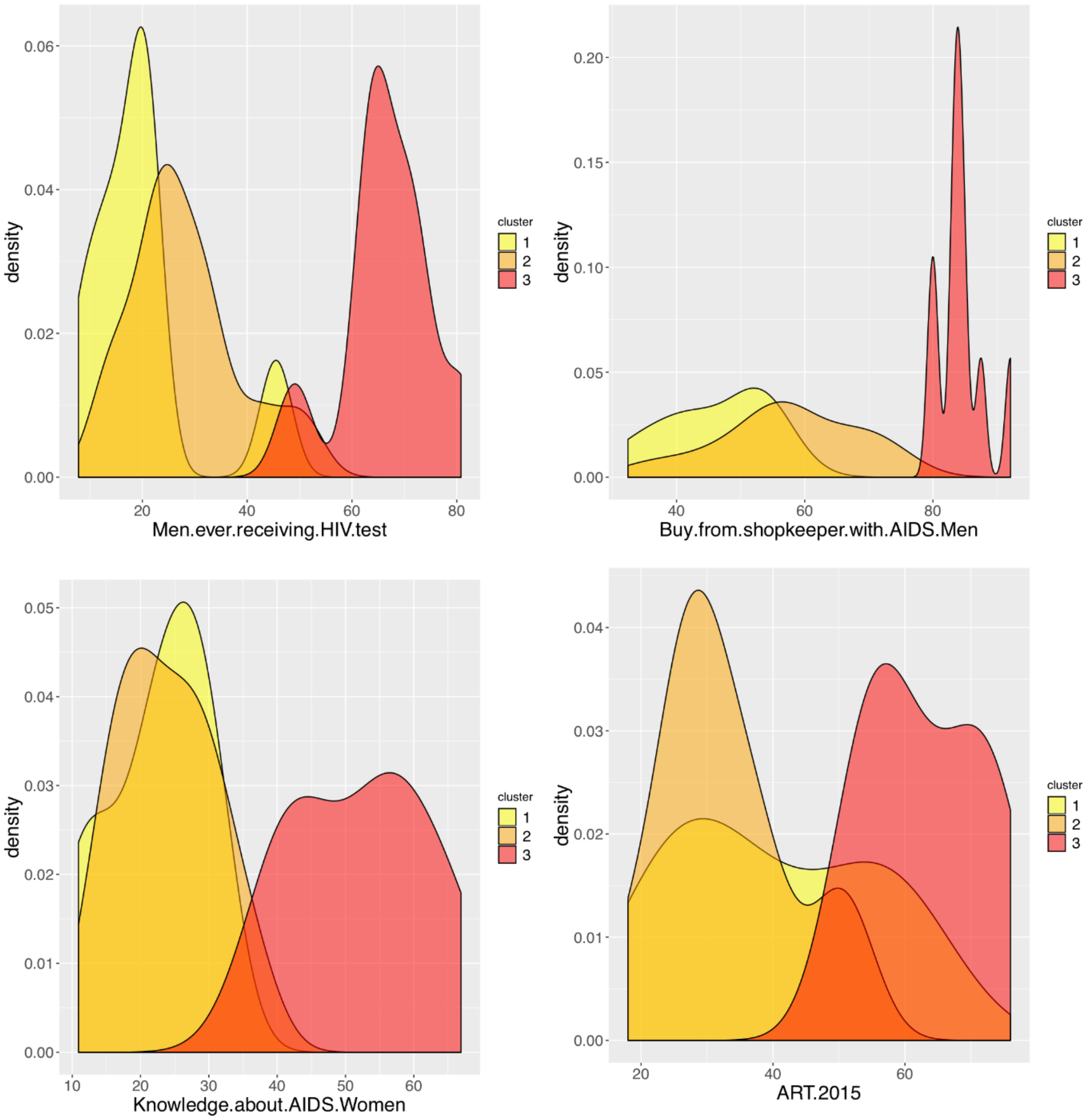


Figure 4 - Analysis of the resulting clusters in terms of their sociobehavioural characteristics. Density plots per cluster of (a) the percentage of Muslim population, (b) the percentage of circumcised men, (c) the mean number of sexual partners in a man's lifetime, (d) the percentage of literate women, (e) the percentage of men who have ever received an HIV test, (f) the percentage of men who say they would buy fresh vegetables from a vendor whom they knew was HIV+, (g) the percentage of women with a comprehensive knowledge about AIDS and (h) the ART coverage in 2015.




density
0.06
0.04
0.02
0.00
20
40
60
80
Men.ever.receiving.HIV.test
cluster
1
2
3
density
0.20
0.15
0.10
0.05
0.00
40
60
80
Buy.from.shopkeeper.with.AIDS.Men
cluster
1
2
3
density
0.05
0.04
0.03
0.02
0.01
0.00
10
20
30
40
50
60
Knowledge.about.AIDS.Women
cluster
1
2
3
density
0.04
0.03
0.02
0.01
0.00
20
40
60
ART.2015
cluster
1
2
3