Identifying groups of people with similar sociobehavioural characteristics in Malawi to inform HIV interventions: a Latent Class Analysis

Aziza Merzouki\textsuperscript{a,*,$}, Amanda Styles\textsuperscript{b,$}, Janne Estill\textsuperscript{a,c}, Zofia Baranczuk\textsuperscript{a,d,e}, Karen Petrie\textsuperscript{b}, Olivia Keiser\textsuperscript{a}

$ equal contribution

\textsuperscript{a}Institute of Global Health, University of Geneva, Geneva, Switzerland

\textsuperscript{b}University of Dundee, Dundee, United Kingdom

\textsuperscript{c}Institute of Mathematical Statistics and Actuarial Science, University of Bern, Bern, Switzerland

\textsuperscript{d}Department of Psychology, University of Zurich, Zurich, Switzerland

\textsuperscript{e}Institute of Mathematics, University of Zurich, Zurich, Switzerland

*Correspondence to:

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

Word count: Abstract 295 words; main text 3496 words; 2 tables; 3 figures; 1 appendix
Abstract

Background: Within many sub-Saharan African countries including Malawi, HIV prevalence varies widely between regions. This variability may be related to the distribution of population groups with specific sociobehavioural characteristics that influence the transmission of HIV and the uptake of prevention. In this study, we intended to identify groups of people in Malawi with similar risk profiles.

Methods: We used data from the Demographic and Health Survey in Malawi from 2016, and stratified the analysis by sex. We considered demographic, socio-behavioural and HIV-related variables. Using Latent Class Analysis (LCA), we identified clusters of people sharing common sociobehavioural characteristics. The optimal number of clusters was selected based on the Bayesian information criterion. We compared the proportions of individuals belonging to the different clusters across the three regions and 28 districts of Malawi.

Results: We found nine clusters of women and six clusters of men. Most women in the clusters with highest risk of being HIV infected were living in female-headed households and were formerly married or in a union. Among men, older men had the highest risk of being HIV infected, followed by young (20-25) single men. Generally, low HIV testing uptake correlated with lower risk of having HIV. However, rural adolescent girls had a low probability of being tested (48.7%) despite a relatively high HIV prevalence. Urban districts and Southern region had a higher percentage of high-prevalence and less tested clusters of individuals than other areas.

Conclusions: LCA is an efficient method to find clusters of people sharing common HIV risk profiles, identify particularly vulnerable population groups, and plan targeted interventions focusing on these groups. Tailored support, prevention and HIV testing programmes should focus particularly on female household heads, adolescent girls living in rural areas, older married men, and young men who have never been married.

Funding: The project was funded by the Swiss National Science Foundation (grant no 163878).

Keywords: Latent Class Analysis, Clustering, HIV prevalence, HIV testing, risk groups, sociobehavioural characteristics, spatial distribution.
Introduction

Diagnosis of HIV infected people is a critical step to improve access to treatment and control the HIV epidemic. In Eastern and Southern Africa, where HIV prevalence is the highest in the world [1], progress has been made to scale up HIV testing. The proportion of people living with HIV who know their status increased from 77% in 2015 to 85% in 2018[1]. In order to keep up with the progress and identify the remaining HIV infected people, testing programs should better target those who are most at risk of having HIV.

Within many sub-Saharan African (SSA) countries, HIV prevalence varies widely between regions. This heterogeneity may result from the spatial distribution of population groups with specific sociobehavioural characteristics that promote or impede the transmission of HIV and the uptake of preventive measures. Improving our understanding of the characteristics of different groups in the population, their risk of being HIV infected and their uptake of HIV testing is essential to design more targeted and tailored prevention and care programs. Clustering analysis, which is an unsupervised machine learning technique, allows us to identify similarity patterns and find hidden sub-groups of people with potentially different risk levels and drivers of having or acquiring HIV[2]. In a previous study, we investigated differences between SSA countries and identified sociobehavioural profiles that coincided with a high HIV prevalence and incidence at the country level[3]. However, this past study did not consider sub-national differences. Moreover, previous studies on high-risk groups in SSA countries mainly focused on women and excluded men[4–6]. Most of these studies also did not analyse the spatial distribution of the identified groups across the country. A recent study mapped high-risk men and women across Malawi to study whether their distribution could explain the spatial variability of HIV prevalence in the country, but the definition of high-risk behaviour was based only on the total number of lifetime sexual partners of individuals[7].

In the present study, we aimed to identify and characterize vulnerable groups who are most at risk of being HIV positive in Malawi. We used Latent Class Analysis (LCA) to identify clusters of men and women with different sociobehavioural characteristics, based on the latest Demographic and Health Survey (DHS)[8] of Malawi. We also compared the most common clusters across the three regions and 28 districts of Malawi.
Methods

Data

We used data from the latest available DHS (2015-2016) for Malawi. The DHS programme collects nationally representative data and covers a wide range of indicators for population and health.

We considered datasets (Table S1) that contained the survey responses of men (aged 15-54 years) and women (15-49), and their HIV status (positive, negative, unknown). Two researchers (AM and OK) preselected the following variables because they covered topics that could be related to the HIV epidemic: place of residence (rural; urban), age (≤19; 20 to 25; 26 to 35; >35 years), literacy (can read a complete sentence: yes; no), access to media at least once a week (yes; no), currently employed (yes; no), marital status (formerly married or lived in a union; currently married or living in a union; never married nor lived in a union), sex of household head (female; male), age at first sexual intercourse (<16; 16 to 18; 19 to 21; >21), condom use (wife justified to ask husband to use a condom when he has a sexually transmitted infection (STI); not justified), wife beating (at least one reason justifies wife beating; no reason justifies it), comprehensive correct knowledge about AIDS (yes; no), and prior HIV testing (ever been tested; never tested). We also considered the district and region of residence of respondents for mapping.

Analysis

We conducted separate analyses for women and men. Since our study focused on sexual transmission of HIV, we excluded persons who never had sexual intercourse. We used multiple imputation by chained equations (MICE)[9,10] to complete missing data (Supplementary Material). Some implausible or inconsistent values were set to missing before imputation.

We identified clusters of people sharing common sociobehavioural characteristics using LCA[11], a multivariate technique that groups individuals with similar characteristics in latent (i.e. not directly observable) classes (clusters). We used maximum likelihood estimation to calculate the probability of an individual belonging to a cluster. We selected the number of clusters that minimized the Bayesian information criterion (BIC)[12]. We performed the final analysis based on a subset of variables identified as the most important for clustering (Supplementary Material). We selected these variables for men and women separately based on the overall variance of their categories’ probabilities across the clusters, similarly to the method by Dean and Raftery[13]. Reducing the number of variables allows us to reduce the number of clusters, thus facilitating interpretability.

We analyzed the HIV prevalence and the uptake of HIV testing in each LCA cluster. We defined HIV prevalence as the proportion of positive tests among the positive and negative tests results. We also analyzed the distribution of the clusters across regions and districts, and assessed whether it could help to explain the wide spatial variability of HIV prevalence.

We conducted the analysis using the open source R language, version 3.5.1. We used the implementation of LCA provided in the poLCA package, version 1.4.1[14,15]. The code is available on GitLab (https://gitlab.com/AzizaM/dhs_malawi_lca).
Ethics approval

The DHS data are freely available on request. Thus, no ethical approval was needed.
Results

The survey we used included 24,562 women and 7,478 men. After imputing incomplete data and removing respondents who never had sexual intercourse, we performed the analysis on 21,564 women and 6,379 men. The HIV status was known for 6,750 women (5,889 HIV negative; 861 HIV positive) and 5,642 men (5,168 HIV negative; 474 HIV positive).

Women’s clusters

We found an optimal number of nine clusters for women based on seven variables: age, literacy, marital status, location of residence, sex of household head, having ever been tested for HIV, and employment (Table 1). Two clusters had a high HIV prevalence (Clusters 2 and 8). Cluster 8 had an HIV prevalence of 37.2%, and was predominantly urban (100.0%) and literate (87.8%). Cluster 2 had a prevalence of 21.2%, and was predominantly rural (96.6%). In comparison, other clusters had a median HIV prevalence of 10.0% (IQR 7.9%-12.3%). Clusters 2 and 8 both had high proportions of female-headed households (100.0% and 83.2%, respectively) and most women had lived in a union before (76.2% and 59.1%, respectively).

HIV testing uptake was relatively low in Clusters 1 and 9 (48.7% in Cluster 1; 47.1% in Cluster 9) compared to other clusters where the median HIV testing uptake was 96.8% (IQR 94.1%-99.1%). Cluster 9 had a low HIV prevalence (2.1%) and consisted mainly of older (46.5% over 35 years) women who lived in rural areas (97.3%) and had a low level of literacy (29.4%). These women generally lived in a union (95.4%) and their household was male-headed (88.7%). In Cluster 1, the HIV prevalence was higher (7.7%). This cluster included literate (73.3%), adolescent (77.8% under 20 years) girls who lived in rural (84.6%) areas and had never been in a union (94.2%). HIV testing uptake was approximately two-fold higher (96.8%) in Cluster 4 than Cluster 1 (48.7%), although their HIV prevalence were similar (8.0% in Cluster 4 vs. 7.7% in Cluster 1). Cluster 4 was characterized by young (83.1% aged between 20 and 35), rural (100.0%) and married (95.3%) women.

Figure 1 and Table S2 present the distributions of the clusters of women in the Northern, Central and Southern regions of Malawi. The Southern region included a higher percentage (7.2%) of rural, literate, adolescent girls, who had never been in a union and who had a low uptake of HIV testing (Cluster 1); and a lower percentage (10.8%) of well tested married women aged over 20 (Cluster 6) than the Central (Cluster 1, 6.0%; Cluster 6, 16.7%) and Northern (Cluster 1, 3.3%; Cluster 6, 15.1%) regions. The Southern region had also a higher percentage (17.0%) of rural divorced/widowed women with a female household head and a high HIV prevalence (Cluster 2) than the other regions (13.1% in the Central and 9.6% in the Northern regions). In contrast, the Northern region had a higher percentage (32.5%) of young rural married women (Cluster 4) with a high HIV testing uptake than the Central (21.6%) and Southern (23.2%) regions.

The distribution of clusters of women varied also across the 28 districts of Malawi (Table S3). Figure 2 presents the proportion of the nine clusters in five selected districts that cover all three regions of the country: Chitipa (Northern), Salima (Central), Lilongwe (Central), Blantyre (Southern) and Chikwawa (Southern region). We selected two neighbouring districts from the same region (Salima and Lilongwe in the Central region; Chikwawa and Blantyre in the Southern region) because they had remarkably different HIV prevalence (5.2% in Salima vs
9.2% in Lilongwe; 9.1% in Chikwawa vs 24.1% in Blantyre). Table S4 and Figure S1 present the HIV prevalence per district for women and men. In both Central and Southern regions, the districts with higher HIV prevalence (Lilongwe and Blantyre) contained the two largest cities of Malawi, and therefore were characterized by a lower level of rural clusters (Clusters 1, 2, 3 and 4), and a higher percentage of more urban clusters (Clusters 5, 6, 7 and 8). In the Southern region, Cluster 9, which is 97.3% rural and has the lowest HIV prevalence, was substantially more common in Chikwawa than in Blantyre (7.2% vs 1.0%), unlike in the Central region where its proportion was similar in Salima and Lilongwe (2.4%; 3.1%). Similarly to districts with lower HIV prevalence in the Central and Southern regions of Malawi, in Chitipa (Northern region) a relatively high percentage of people belonged to rural clusters, such as Cluster 3 (33.9%), Cluster 4 (29.4%) and Cluster 9 (5.6%), whereas urban clusters, such as Cluster 5 (1.7%), Cluster 7 (2.2%) and Cluster 8 (1.7%), were much less common. But Chitipa had also a lower percentage of Cluster 1 (1.7%; less tested and mainly rural with relatively high HIV prevalence, vs. 10.0% in Salima and 7.0% in Chikwawa) and a lower percentage of Cluster 2 (7.8%; rural with high HIV prevalence and female-headed households, vs. 16.9% in Salima and 12.2% in Chikwawa). Chitipa also had a higher percentage of women from Cluster 6 (16.1%; well tested, literate, employed women, who were living in union, with a male household head, vs. 9.6% in Salima and 5.9% in Chikwawa).

Men’s clusters

The optimal number of clusters for men was six based on 11 variables: age, age at first sex, marital status, location of residence, having ever been tested for HIV, media access, employment, literacy, sex of household head, wife beating justification, and willingness to use condom (Table 2). Compared to women, men had a lower HIV prevalence (8.4% vs 12.7%); and a lower uptake of HIV testing. The percentage of men who had never been tested for HIV was three times higher than for women (22.5% vs 7.6%). Two clusters of men, Clusters 3 and 5, had a particularly low testing uptake (52.5% and 35.0%; vs. a range of 80.5% to 92.2% in other clusters). Although literacy was substantially different between these clusters (97.7% in Cluster 3; 29.7% in Cluster 5), men in both clusters had many similar characteristics: they were mainly adolescents (77.7% under 20 years in Cluster 3; 70.3% in Cluster 5), had never lived in a union (96.0% in Cluster 3; 98.2% in Cluster 5), and lived in rural settings (80.3% in Cluster 3; 97.2% in Cluster 5). These clusters also had the lowest HIV prevalence (0.8% and 0.5%).

The clusters of men with the highest HIV prevalence were Clusters 1 and 4 (10.0% in Cluster 1; 13.0% in Cluster 4). These clusters represented 31.2% and 31.7% of all men, respectively, and had a relatively high HIV testing uptake (80.5% in Cluster 1; 86.6% in Cluster 4). They both had a high proportion of older men (48.3% aged over 35 in Cluster 1; 59.0% in Cluster 4), who were mainly living in a union (92.0% in Cluster 1; 97.6% in Cluster 4). The household heads were also mostly men (93.6% in Cluster 1; 97.6% in Cluster 4). In contrast, Cluster 4 was more urban than Cluster 2 (25.4% vs 3.2%), and access to media in it was better (74.4% vs 33.0%). The cluster with the third highest HIV prevalence was Cluster 2 (6.4%). This cluster had a high proportion of young men (72.7% between 20 and 25 years old), who were literate (87.1%) and had never lived in a union (83.1%). In comparison to Cluster 2, Cluster 6, which included mainly men who were literate (97.8%), aged between 20 and 35 (96.1%), had regular access to media (77.2%), and were living in a union (100.0%) and in rural settings (79.3%), had a lower HIV prevalence (3.9%) and a higher HIV testing uptake (92.2% in Cluster 6 vs. 86.3% in Cluster 2).
Figure S1 and Table S5 present the percentage of each cluster of men in the Northern, Central and Southern regions of Malawi. The distribution of these clusters did not vary much between the three regions. It varied more between districts (Table S6). Figure 3 presents the proportion of the six clusters in the districts of Chitipa (Northern), Salima (Central), Lilongwe (Central), Blantyre (Southern) and Chikwawa (Southern region). In both Central and Southern regions, the districts with higher HIV prevalence (Lilongwe; Blantyre) were characterized by a lower proportion of older men (over 35) living in rural areas, who were in a union and did not have regular access to media (22.8% of Cluster 1 in Lilongwe vs 36.5% in Salima; 15.0% in Blantyre vs 30.7% in Chikwawa). Lilongwe and Blantyre also had a lower percentage of rural, illiterate adolescents with a low HIV prevalence (4.9% of Cluster 5 in Lilongwe; 3.3% in Blantyre) than Salima and Chikwawa (14.5% of Cluster 5 in Salima; 12.2% in Chikwawa). Higher HIV prevalence coincided also with a higher percentage of young, literate men, who had never lived in a union (18.5% of Cluster 2 in Lilongwe vs 9.4% in Salima; 23.0% in Blantyre vs 7.8% in Chikwawa), and a higher proportion of older (over 35) men, who were living in a union and had regular access to media (36.3% of Cluster 4 in Lilongwe vs 21.0% in Salima; 42.0% in Blantyre vs 35.1% in Chikwawa). Both districts of the Southern region (Blantyre and Chikwawa) were characterized by a lower proportion of literate, married men, who were aged 20-35 and lived in rural settings (4.7% of Cluster 6 in Blantyre; and 4.9% in Chikwawa vs 10.2%), compared to Chitipa (Northern region). In contrast, Chitipa was characterized by a lower percentage of rural, illiterate, adolescent men (5.1% of Cluster 6) than the rural districts of Southern and Central regions (12.2% in Chikwawa; 14.5% in Salima).
Discussion

Using LCA, we categorized Malawian women into nine clusters and men into six clusters with common sociobehavioural characteristics. Among these clusters, we identified profiles of women and men who were most at risk of being HIV infected and less tested for HIV. Female household heads bore the heaviest burden of HIV infection, and rural adolescent girls had a significantly lower HIV testing uptake despite a relatively high HIV prevalence. Among men, older ones were the most affected by HIV, followed by young men who had never been married. Mapping the identified clusters across the regions and districts of Malawi showed that the clusters most vulnerable to HIV (higher prevalence and lower testing) were more frequent in the Southern region and the urban districts of the country.

The clusters of women with the highest HIV prevalence were characterized by an older age (over 25 years), living in a female-headed household and having lived formerly in a union. This finding is consistent with earlier studies that showed that women who were no longer in a union were most-at-risk of being HIV infected in Malawi[4]. The most common group of female household heads in Malawi are widows[16], and it is well possible that their husband died of AIDS. Indeed, the proportion of widows and orphans increased over time in the most HIV-affected countries[17]. Female household heads were also shown to live more frequently in poverty, requiring support from dedicated programmes[16,18].

The uptake of HIV testing in the identified clusters was generally consistent with the risk of being infected by HIV, with more persons being tested for HIV when HIV prevalence was high. Adolescents (<20 years) were tested much less frequently than adults[19–21]. The low HIV testing uptake among adolescents is concerning, because the proportion of young people (15-24) among all AIDS-related deaths in Malawi increased from 3.1% to 11.5% between 2004 and 2018[1]. In particular, rural adolescent girls, who have never lived in a union, were tested about half as much as older (20-35 years), rural, literate women, who were living in a union, and had a similar relatively high HIV prevalence. Yet, adolescent girls are a vulnerable cluster that often lack knowledge about HIV[19], tend to perceive themselves as being at low risk of HIV[22], and are prime targets of gender-based violence[23]. These observations confirm the need to accelerate tailored HIV prevention programming among adolescents in general, and specifically among rural adolescent girls[24].

Men tended to have a lower HIV prevalence than their female counterparts, confirming the gender disparity in HIV infection in SSA[25,26]. Men were also less often tested for HIV, which could be influenced by harmful masculinity stereotypes and cultural norms that inhibit men from seeking care[27,28]. It may also result from the focus of HIV prevention in recent years on women and girls, and HIV testing being routinely performed in antenatal care. UNAIDS’ report released in 2017[29] highlighted the low utilisation of health services by men and qualified it as a blind spot in the response to HIV. Men have therefore a higher risk of HIV transmission to their partners, and eventually a higher mortality rate due to AIDS-related illnesses, affecting directly their families.

HIV prevalence was associated with the age and partnership status of men across the clusters. Older men (>35 years) had the highest HIV prevalence regardless of their place of residence, literacy, or access to media. Compared to 2016, HIV incidence in Malawi was nearly three
times higher in year 2000[1], which can explain the higher risk of HIV acquisition among men who were young adults at that time. The current lower willingness to condom utilisation among rural, less educated, older and married men[30,31] may also contribute to higher risk of recent HIV acquisition. Among young men, living in a union and in a rural setting was associated with a reduced risk of HIV. This finding is in agreement with a recent study performed in South Africa, where single never married men had a high risk of being HIV infected, which may be associated with a higher frequency of less stable relations and multiple sexual partners[32].

Clusters of women and men with high prevalence of HIV and low HIV testing coverage were more frequent in urban areas of Malawi. Palk and Blower[7] showed that urban areas of Malawi also had a higher density of individuals with a high number of lifetime sexual partners. Women and men in the mostly rural Northern region were generally well tested. Our finding is in agreement with a recent study, which also reported a regional variance of HIV testing uptake in Malawi, with a better testing uptake in the North, although the latter study focused on men only[13].

Our work has some limitations. The analysis was based on a limited number of factors that were selected by the researchers based on prior knowledge from the available literature. In order to avoid the subjective selection of the variables in the future, we are working in parallel on an automatic method to select the most important variables to predict individual HIV status. Our preliminary results showed that most of the variables that we selected in this study were highly relevant to predict HIV. Our analysis allowed us to identify clusters of men and women with associated probabilities of being HIV infected and tested for HIV, but the cross-sectional nature of our analysis does not allow us to draw causal relations. There are other methods, such as Bayesian network analyses, that attempt to determine the possible causal pathways between various risk factors and HIV[33]. The data we used are also potentially biased because they are self-reported (ex. more educated people may be less keen to admit gender-based violence or non-willingness to use condoms).

In conclusion, using LCA to identify and map vulnerable clusters of women and men can provide key information to policy makers, enabling them to strategically design tailored support and prevention programs that target people who are most in need. Our results clearly demonstrated that rural adolescent girls, young single men, and married men aged over 35 years have a particularly high risk of being infected by HIV. Efforts should be made to increase the coverage of HIV testing in particular among these sub-populations.
Acknowledgement

We acknowledge the support of the Swiss National Science Foundation (SNF professorship grant n° 163878 to O Keiser) who funded this study. We thank Isotta Triulzi for helpful discussions and comments.

Conflict of interest

We declare no competing interests.
Table 1 - Clusters of women based on the seven most important variables: age, literacy, marital status, location of residence, sex of household head, having ever been tested for HIV and employment. Data from 21,564 women were analysed.
| Cluster | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---------|---|---|---|---|---|---|---|---|---|
| **Age** |   |   |   |   |   |   |   |   |   |
| 19 and under | 1044 (77.8%) | 85 (2.7%) | 121 (2.1%) | 687 (13.5%) | 77 (7.8%) | 53 (1.8%) | 229 (24.2%) | 0 (0.0%) | 154 (19.4%) |
| 20 to 25 | 267 (19.9%) | 477 (15.2%) | 1103 (19.4%) | 2157 (42.4%) | 410 (41.6%) | 620 (20.9%) | 592 (62.4%) | 0 (0.0%) | 151 (19.0%) |
| 26 to 35 | 31 (2.3%) | 1119 (35.8%) | 2019 (35.5%) | 2075 (40.7%) | 403 (40.9%) | 1544 (52.0%) | 127 (13.4%) | 314 (52.0%) | 121 (15.2%) |
| Over 35 | 0 (0.0%) | 1447 (46.3%) | 2452 (43.1%) | 174 (3.4%) | 97 (9.8%) | 752 (25.3%) | 0 (0.0%) | 290 (48.0%) | 371 (46.5%) |
| **Literacy** |   |   |   |   |   |   |   |   |   |
| Good Reader | 986 (73.4%) | 1452 (46.4%) | 1708 (30.0%) | 3900 (76.6%) | 799 (81.0%) | 2970 (100.0%) | 936 (98.7%) | 530 (87.8%) | 235 (29.4%) |
| Poor Reader | 357 (26.6%) | 1675 (53.6%) | 3987 (70.0%) | 1193 (24.3%) | 188 (19.0%) | 0 (0.0%) | 12 (1.3%) | 74 (12.2%) | 562 (70.6%) |
| **Current marital status** |   |   |   |   |   |   |   |   |   |
| Formerly married or lived in a union | 41 (3.0%) | 2384 (76.2%) | 60 (1.1%) | 237 (4.7%) | 0 (0.0%) | 111 (3.7%) | 184 (19.5%) | 357 (59.1%) | 37 (4.6%) |
| Married or living in a union | 37 (2.8%) | 678 (21.7%) | 5635 (98.9%) | 4856 (95.3%) | 987 (100.0%) | 2859 (96.3%) | 0 (0.0%) | 191 (31.6%) | 760 (95.4%) |
| Never married nor lived in a union | 1265 (94.2%) | 65 (2.1%) | 0 (0.0%) | 0 (0.0%) | 0 (0.0%) | 764 (80.5%) | 56 (9.3%) | 0 (0.0%) |   |
| **Location** |   |   |   |   |   |   |   |   |   |
| Rural | 1136 (84.6%) | 3021 (96.6%) | 5238 (92.0%) | 5093 (100.0%) | 0 (0.0%) | 1716 (57.8%) | 391 (41.3%) | 0 (0.0%) | 775 (97.3%) |
| Urban | 207 (15.4%) | 106 (3.4%) | 457 (8.0%) | 0 (0.0%) | 987 (100.0%) | 1254 (42.2%) | 557 (58.7%) | 604 (100.0%) | 22 (2.7%) |
| **Household head** |   |   |   |   |   |   |   |   |   |
| Female | 524 (39.1%) | 3127 (100.0%) | 250 (4.4%) | 1045 (20.5%) | 98 (9.9%) | 228 (7.7%) | 456 (48.1%) | 503 (83.2%) | 90 (11.3%) |
| Male | 819 (60.6%) | 0 (0.0%) | 5445 (95.6%) | 4048 (79.5%) | 889 (90.1%) | 2742 (92.3%) | 492 (51.9%) | 101 (16.8%) | 707 (88.7%) |
| **Ever tested for HIV** |   |   |   |   |   |   |   |   |   |
| Never Tested | 689 (51.3%) | 208 (6.7%) | 0 (0.0%) | 165 (3.2%) | 0 (0.0%) | 58 (2.0%) | 57 (6.1%) | 35 (5.8%) | 422 (52.9%) |
| Tested | 654 (48.7%) | 2919 (93.3%) | 5695 (100.0%) | 4928 (96.8%) | 987 (100.0%) | 2912 (98.1%) | 891 (93.9%) | 569 (94.2%) | 375 (47.1%) |
| **Employment status** |   |   |   |   |   |   |   |   |   |

1 Can read a whole sentence
| HIV status     | Negative | Positive | Unknown |
|---------------|----------|----------|---------|
|               | 407 (30.3%) | 34 (2.5%) | 902 (67.2%) |
|               | 759 (24.3%) | 204 (6.5%) | 2164 (69.2%) |
|               | 1526 (26.8%) | 228 (4.0%) | 3941 (69.2%) |
|               | 1529 (30.0%) | 133 (2.6%) | 3431 (67.4%) |
|               | 260 (26.3%) | 45 (4.5%) | 682 (69.1%) |
|               | 780 (26.3%) | 103 (3.5%) | 2087 (70.3%) |
|               | 262 (27.6%) | 29 (3.1%) | 657 (69.3%) |
|               | 135 (22.4%) | 80 (13.3%) | 389 (64.4%) |
|               | 231 (29.0%) | 5 (0.6%) | 561 (70.4%) |

| HIV prevalence | 7.7% | 21.2% | 13.0% | 8.0% | 14.8% | 11.7% | 10.0% | 37.2% | 2.1% |

2 Cannot read at all, or can read part of a sentence only
3 We defined the prevalence as the proportion of positive tests among the conclusive (positive and negative) tests results
Table 2 - Clusters of men based on the 11 most important variables: age, age at first sex, marital status, location of residence, having ever been tested for HIV, media access, employment, literacy, sex of household head, wife beating justification and willingness to use condom. Data from 6,379 men were analyzed.

| Cluster | 1         | 2         | 3         | 4         | 5         | 6         |
|---------|-----------|-----------|-----------|-----------|-----------|-----------|
| **Age** |           |           |           |           |           |           |
| 19 and under | 19 (1.0%) | 0 (0.0%)  | 570 (77.7%) | 0 (0.0%)  | 314 (70.3%) | 12 (2.7%) |
| 20 to 25 | 321 (16.1%) | 535 (72.7%) | 164 (22.3%) | 3 (0.1%)  | 125 (28.0%) | 215 (47.8%) |
| 26 to 35 | 689 (34.7%) | 167 (22.7%) | 0 (0.0%)  | 827 (40.9%) | 3 (0.7%)  | 217 (48.3%) |
| Over 35  | 960 (48.3%) | 34 (4.6%)  | 0 (0.0%)  | 1193 (59.0%) | 5 (1.1%)  | 6 (1.2%)  |
| **Age at first Sex** |           |           |           |           |           |           |
| Under 16 | 713 (35.9%) | 158 (21.4%) | 479 (65.3%) | 355 (17.5%) | 294 (65.8%) | 58 (12.8%) |
| 16 to 18 | 393 (19.8%) | 143 (19.4%) | 194 (26.4%) | 296 (14.7%) | 106 (23.8%) | 122 (27.1%) |
| 19 to 21 | 608 (30.6%) | 339 (46.0%) | 61 (8.3%)  | 692 (34.2%) | 47 (10.4%) | 225 (50.1%) |
| Over 21  | 275 (13.8%) | 97 (13.2%)  | 0 (0.0%)  | 680 (33.6%) | 0 (0.0%)  | 45 (10.0%) |
| **Current marital status** |           |           |           |           |           |           |
| Formerly married or lived in a union | 132 (6.7%) | 73 (9.9%)  | 0 (0.0%)  | 49 (2.4%)  | 2 (0.4%)  | 0 (0.0%)  |
| Married or living in a union | 1828 (92.0%) | 52 (7.1%)  | 30 (4.0%)  | 1974 (97.6%) | 7 (1.5%)  | 450 (100.0%) |
| Never married nor lived in a union | 27 (1.4%) | 612 (83.1%) | 704 (96.0%) | 1 (0.0%)  | 439 (98.2%) | 0 (0.0%) |
| **Location** |           |           |           |           |           |           |
| Rural | 1925 (96.8%) | 385 (52.3%) | 589 (80.3%) | 1509 (74.6%) | 435 (97.2%) | 357 (79.3%) |
| Urban | 63 (3.2%) | 352 (47.7%) | 145 (19.7%) | 514 (25.4%) | 12 (2.8%) | 93 (20.7%) |
| **Ever tested for HIV** |           |           |           |           |           |           |
| Never Tested | 387 (19.5%) | 101 (13.7%) | 348 (47.5%) | 272 (13.4%) | 290 (65.0%) | 35 (7.8%) |
| Tested | 1601 (80.5%) | 636 (86.3%) | 386 (52.5%) | 1751 (86.6%) | 157 (35.0%) | 415 (92.2%) |
| **Has media access** |           |           |           |           |           |           |
| No | 1333 (67.05%) | 201 (27.28%) | 306 (41.74%) | 518 (25.63%) | 283 (63.26%) | 103 (22.85%) |
| Yes          | 655 (33.0%) | 536 (72.7%) | 428 (58.3%) | 1505 (74.4%) | 164 (36.4%) | 347 (77.2%) |
|--------------|-------------|-------------|-------------|--------------|-------------|-------------|

**Employment status**

| Not currently employed | 185 (9.3%) | 143 (19.3%) | 303 (41.2%) | 30 (1.5%) | 128 (28.7%) | 0 (0.0%) |
|------------------------|------------|-------------|-------------|----------|-------------|--------|
| Currently employed     | 1803 (90.7%) | 594 (80.7%) | 431 (58.8%) | 1993 (98.5%) | 319 (71.3%) | 450 (100.0%) |

**Literacy**

| Good Reader⁴ | 863 (43.4%) | 642 (87.1%) | 717 (97.7%) | 1777 (87.9%) | 133 (29.7%) | 440 (97.8%) |
|--------------|-------------|-------------|-------------|-------------|-------------|-------------|
| Poor Reader⁵ | 1125 (56.6%) | 95 (12.9%) | 17 (2.3%) | 246 (12.2%) | 314 (70.3%) | 10 (2.2%) |

**Household head**

| Female | 127 (6.4%) | 212 (28.8%) | 270 (36.7%) | 48 (2.4%) | 120 (26.8%) | 32 (7.1%) |
|--------|------------|-------------|-------------|----------|-------------|--------|
| Male   | 1861 (93.6%) | 525 (71.2%) | 464 (63.3%) | 1975 (97.6%) | 327 (73.2%) | 418 (92.9%) |

**Beating wife justified**

| No | 1753 (88.2%) | 645 (87.6%) | 578 (78.8%) | 1972 (97.5%) | 308 (68.9%) | 345 (76.7%) |
|----|--------------|-------------|-------------|-------------|-------------|-------------|
| Yes | 235 (11.8%) | 92 (12.4%) | 156 (21.3%) | 51 (2.5%) | 139 (31.1%) | 105 (23.4%) |

**Wife justified asking husband to use condom if he has STI⁶**

| No | 365 (18.4%) | 45 (6.2%) | 51 (6.9%) | 38 (1.9%) | 107 (24.0%) | 45 (10.1%) |
|----|-------------|-----------|----------|---------|-----------|--------|
| Yes | 1623 (81.6%) | 692 (93.9%) | 683 (93.1%) | 1985 (98.1%) | 340 (76.0%) | 405 (89.9%) |

**Cluster Total**

| 1988 (31.2%) | 737 (11.6%) | 734 (11.5%) | 2023 (31.7%) | 447 (7.0%) | 450 (7.1%) |

**HIV status**

| Negative | 1573 (79.1%) | 613 (83.2%) | 652 (88.8%) | 1568 (77.5%) | 396 (88.6%) | 366 (81.3%) |
|----------|-------------|-------------|-------------|--------------|-------------|-------------|
| Positive | 175 (8.8%) | 42 (5.7%) | 5 (0.7%) | 235 (11.6%) | 2 (0.5%) | 15 (3.3%) |
| Unknown  | 240 (12.1%) | 82 (11.1%) | 77 (10.5%) | 220 (10.9%) | 49 (11.0%) | 69 (15.3%) |

**HIV prevalence⁷**

| 10.0% | 6.4% | 0.8% | 13.0% | 0.5% | 3.9% |

---

⁴ Can read a whole sentence
⁵ Cannot read at all, or can read part of a sentence only
⁶ Sexually Transmitted Infection
⁷ We defined the prevalence as the proportion of positive tests among the conclusive (positive and negative) tests results.
Figure 1 – Distribution of the nine identified clusters of women in the three regions of Malawi: Northern, Central, and Southern.

Figure 2 – Proportion of the nine female clusters in five selected districts of Malawi: (a) Chitipa, (b) Salima, (c) Lilongwe, (d) Blantyre, (e) Chikwawa.
**Figure 3** - Proportion of the six male clusters in five selected districts of Malawi: (a) Chitipa, (b) Salima, (c) Lilongwe, (d) Blantyre, (e) Chikwawa.
References

1. AIDSinfo | UNAIDS. Available at: http://aidsinfo.unaids.org/. Accessed 27 September 2019.

2. Engl E, Smittenaar P, Sgaier SK. Identifying population segments for effective intervention design and targeting using unsupervised machine learning: an end-to-end guide. Gates Open Res 2019; Available at: https://dqo52087pnd5x.cloudfront.net/manuscripts/14221/75e1d884-e2d5-4f97-ab59-d738a298d247_13029_. _sema_sgaier_v2.pdf?doi=10.12688/gatesopenres.13029.2&numberOfBrowsableCollections =2&numberOfBrowsableInstitutionalCollections=0&numberOfBrowsableGateways=4. Accessed 5 November 2019.

3. Merzouki A, Estill J, Tal K, Keiser O. Sociobehavioural characteristics and HIV incidence in 29 sub-Saharan African countries: Unsupervised machine learning analysis. Epidemiology, 2019. Available at: http://biorxiv.org/lookup/doi/10.1101/620450. Accessed 27 September 2019.

4. Emina JBO, Madise N, Kuepie M, Zulu EM, Ye Y. Identifying HIV most-at-risk groups in Malawi for targeted interventions. A classification tree model. BMJ Open 2013; 3:e002459.

5. Recent levels and trends in HIV incidence rates among adolescent girls and young women in ten high-prevalence African countries: a systematic review and meta-analysis | Elsevier Enhanced Reader. Available at: https://reader.elsevier.com/reader/sd/pii/S2214109X19304103?token=9D2B309BABF99EA0 C1D5A0D05A13639DB80F890A2F33A7F86E2474D4B91937D40F7887A53820CB0EA99979 6E4559E7. Accessed 9 December 2019.

6. Quinn C, Kadengye DT, Johnson CC, Baggaley R, Dalal S. Who are the missing men? Characterising men who never tested for HIV from population-based surveys in six sub-Saharan African countries. J Int AIDS Soc 2019; 22:e25398.

7. Palk L. Geographic variation in sexual behavior can explain geospatial heterogeneity in the severity of the HIV epidemic in Malawi. BMC Med 2018; :9.

8. The DHS Program - Quality information to plan, monitor and improve population, health, and nutrition programs. Available at: https://dhsprogram.com/. Accessed 27 September 2019.

9. Buuren S van, Groothuis-Oudshoorn K, Robitzsch A, et al. mice: Multivariate Imputation by Chained Equations. 2019. Available at: https://CRAN.R-project.org/package=mice. Accessed 19 August 2019.

10. Buuren S van, Groothuis-Oudshoorn K. mice : Multivariate Imputation by Chained Equations in R. J Stat Softw 2011; 45. Available at: http://www.jstatsoft.org/v45/i03/. Accessed 19 August 2019.

11. McCutcheon AL. Latent Class Analysis. SAGE, 1987.

12. Schwartz G. Estimating the Dimension of a Model. Ann Stat 1978; 6:461–464.

13. Dean N, Raftery AE. Latent class analysis variable selection. Ann Inst Stat Math 2010; 62:11–35.

14. Linzer DA, polLCA: Polytomous Variable Latent Class Analysis Version 1.4. :31.

15. Linzer DA, Lewis JB. polLCA : An R Package for Polytomous Variable Latent Class Analysis. J Stat Softw 2011; 42. Available at: http://www.jstatsoft.org/v42/i10/. Accessed 6 December 2019.

16. Mathiassen A, Eliasi R, Mahowe M, et al. Gender Assessment for Malawi. :47.

17. Heuveline P. Impact of the HIV epidemic on population and household structure: the
dynamics and evidence to date. 2014; :14.
18. Doctor HV. Layers of Socioeconomic Vulnerability in Malawi in the Context of the Millennium Development Goals. ISRN Econ 2013; 2013:1–7.
19. Mkandawire P. Assessing factors associated with HIV testing among adolescents in Malawi. Glob Public Health 2017; 12:927–940.
20. Mandiwa C, Namondwe B. Uptake and correlates of HIV testing among men in Malawi: evidence from a national population–based household survey. BMC Health Serv Res 2019; 19:203.
21. Olakunde BO, Adeyinka DA, Olawepo JO, Pharr JR. HIV testing among men in Nigeria: a comparative analysis between young people and adults. AIDS Care 2019; :1–8.
22. Price JT, Rosenberg NE, Vansia D, et al. Predictors of HIV, HIV Risk Perception, and HIV Worry among Adolescent Girls and Young Women in Lilongwe, Malawi: JAIDS J Acquir Immune Defic Syndr 2017; :1.
23. Women and HIV — A spotlight on adolescent girls and young women. :20.
24. UNAIDS HIV_prevention_among_adolescent_girls_and_young_women.pdf. Available at: https://www.unaids.org/sites/default/files/media_asset/UNAIDS_HIV_prevention_among_adolescent_girls_and_young_women.pdf. Accessed 16 September 2019.
25. Understanding the gender disparity in HIV infection across countries in sub-Saharan Africa: evidence from the Demographic and Health Surveys. 2011; :18.
26. Sia D, Onadja Y, Hajizadeh M, Heymann SJ, Brewer TF, Nandi A. What explains gender inequalities in HIV/AIDS prevalence in sub-Saharan Africa? Evidence from the demographic and health surveys. BMC Public Health 2016; 16. Available at: http://bmcpublichealth.biomedcentral.com/articles/10.1186/s12889-016-3783-5. Accessed 2 January 2019.
27. Skovdal M, Campbell C, Madanhire C, Mupambireyi Z, Nyamukapa C, Gregson S. Masculinity as a barrier to men’s use of HIV services in Zimbabwe. Glob Health 2011; 7:13.
28. Wyrod R. Masculinity and the persistence of AIDS stigma. Cult Health Sex 2011; 13:443–456.
29. UNAIDS. BLIND SPOT - Reaching out to men and boys - Addressing a blind spot in the response to HIV. 2017. Available at: https://www.unaids.org/sites/default/files/media_asset/blind_spot_en.pdf. Accessed 16 October 2019.
30. Chimbiri AM. The condom is an ‘intruder’ in marriage: Evidence from rural Malawi. Social Science & Medicine 2007; 64:1102–1115.
31. Anglewicz P, Clark S. The effect of marriage and HIV risks on condom use acceptability in rural Malawi. Soc Sci Med 2013; 97:29–40.
32. Tlou B. The influence of marital status on HIV infection in an HIV hyperendemic area of rural South Africa, 2000–2017. Afr J AIDS Res 2019; 18:65–71.
33. Baranczuk Z, Estill J, Blough S, et al. Socio-behavioural characteristics and HIV: findings from a graphical modelling analysis of 29 sub-Saharan African countries. J Int AIDS Soc 2019; 22:e25437.