# The Demographic, Social, and Economic Correlates of HIV Infection Status in Sub-Saharan Africa

# Supplementary Information

Chirag J. Patel, Kajal Claypool, Eric Chow, Jake Chung, Don Mai, Jessie Chen, Eran Bendavid

# Supplementary Methods

DHS Data sources

The DHS datasets were downloaded as zipped Stata files (available at <https://dhsprogram.com>). Up to six files were obtained per country-survey if they were available: household member recode (PR), individual recode (IR), men’s recode (MR), household recode (HR), HIV test results recode (AR), and geographic data (GE).

Appending the women and men’s recode

While appending the individual and men’s recode datasets (IR and MR), we also systematically identified common variables to map between MR and IR. The following prefix changes were made to variables in MR in this sequence: removing “\_”, “mv” to “v”, removing “s”, “mcase” to “case”, “m” to “v”, “sm” to “mm” or “s” if no “mm” candidates exist, “dm” to “d”, and “sm6” to “s8” and “mv304a” to “v304a\_”.

Additionally, IR variables that had at least a 70% syntactic similarity to MR variables (computed using Levenshtein distance) were mapped using Amazon Mechanical Turk (MTurk) Master Workers. These workers were trained briefly and presented with an MR variable and a set of potential IR candidates. They selected the IR candidate that they thought was the best match, or indicated that the MR variable was an individual category of the question header label, or that none of the IR candidates were appropriate. We validated the MTurk responses by excluding any workers who only selected the same positional answer (or semantically similar answers), or whose responses were not submitted through the MTurk interface. A majority vote rule (70%) was used to select the MR-IR variable mapping, and any candidate mappings that were indeterminate remained unmapped.

Harmonizing the Household, Male, and Female surveys

Each DHS contains several parts: a census of all household members, in-depth interviews with reproductive-age (15-49 years old) women, men (15-59 years old), a ledger of household information, and HIV test results for each tested individual. We harmonized the surveys to match individuals in the household with available in-depth data, household data, and HIV test results. Variables that represented the same concept were identified by similarities in their variable name or the variable description, and merged together. Amazon Mechanical Turk (mTurk) was used to match men’s and women’s variables when the relationship between the variable names was intermediate.[16] A harmonized dataset was created for each country-survey, and each of these merged surveys was then analyzed.

We analyzed the harmonized surveys from all sub-Saharan African countries and survey years that performed the DHS HIV testing protocol. All surveys that reported an HIV prevalence ≥ 0.01 were included in our analysis (see Table S1).

Resolving value-label conflicts

When mapping categorical variables together in IR and MR, we resolved conflicting value labels by: 1) obtaining a complete list of all possible values of categorical variables, 2) merging all labels to the values if value-labels are available in the native Stata data dictionary, 3) appending both complete lists of value-labels from IR and MR, 4) removing duplicates from this new value-label dictionary, 5) finding conflicting value-labels (where the same value in a variable has two different labels, e.g. 1 = “Yes” in IR but 1 = “Definitely” in MR) and re-levelling one of the conflicting value-labels. If a conflict is found, then the value without a label receives priority (and the labelled conflict is relevelled); otherwise the IR value keeps its level while the MR value gets relevelled to the next larger level number in the set of levels in that variable. All value-labels are prefixed with their original value in square brackets in the label (e.g. the relevelled value of 2 might have the label “[1] Yes” because the value of 1 had two conflicting labels).

Merging recode datasets

The household member recode (PR) is then left-joined with the appended IR/MR file (with variables mapped and levels relevelled) in a 1:1 merge using (1) the cluster ID (v001), (2) household ID (v002), (3) respondent line number (v003), and occasionally, (4) structure ID (sstruct, sconces, svivi, sqnumber, or snumber) if it is needed to uniquely identify a respondent. Not all PR records are merged to a record in the appended IR/MR file, but are retained in the dataset with empty cells for the IR/MR variables. The HR dataset is left-joined to the PR/IR/MR dataset in a 1:m merge using (1) the cluster ID, and (2) household ID. The AR dataset (which contains the HIV test result of a subset of respondents) is left-joined to the PR/IR/MR/HR dataset in a 1:1 merge and left as missing if not available. In the AR file, v001 = hivclust, v002 = hivnumb and v003 = hivline. The GE dataset is left-joined in a 1:m merge using cluster ID only.

Handling missing recode datasets

The IR dataset is the minimum file needed for merging. If the IR dataset is not present, the datasets are not merged. Alerts are given if <75% of the IR/MR records are merged to PR, or if >25% of the IR/MR records are not mergeable to PR. Similarly, alerts are given if <75% of HR records are merged to PR, or if >25% of HR records are not mergeable to PR. Alerts are given if <50% of HIV test result data are merged to PR. The final merged dataset of PR, IR/MR, HR, and AR was saved as a Stata 15 file with the embedded data-dictionary and named after the country-survey code and given the “flattened” suffix.

Processing of duplicate and categorical variables

We removed variables that exhibited no variation or were “duplicates”. We then computed pairwise Pearson correlations for all variables in a given survey across all 50 surveys. We identified groups of variables whose pairwise Pearson correlation was 0.9 or greater. For each of these groups of variables that were highly correlated with one another, we selected one to represent the group that had the largest average sample size for the variable across all surveys. These decision rules eliminated redundant variables, preserved most meaningfully continuous variables, and discretized most non-ordinal variables. All continuous variables were scaled and centered.

Number of surveys available per country

Five countries had 3 surveys available for analysis (Lesotho, Malawi, Rwanda, Tanzania, Zimbabwe), 11 had 2 surveys (Burkina Faso, Democratic Republic of Congo, Cote d'Ivoire, Ethiopia, Ghana, Guinea, Kenya, Liberia, Mali, Sierra Leone, Zambia). The 13 remaining countries had 1 survey (Angola, Burundi, Cameroon, Chad, Congo, Gabon, Mozambique, Namibia, Sao Tome and Principe, Senegal, eSwatini, Togo, and Uganda).

Pan Sub-Saharan sex-specific systematic meta-analyses across all years and countries

For each variable, we combined associations across all of the surveys (year and country combination) for males and females with a random effects meta-analysis procedure. Specifically, given an association between a variable (e.g., marital status, call it ) and HIV+ in a survey for a sex (e.g., Zambia, 2013-2014, females denoted by ; Zimbabwe, 2005, females, denoted by ), we estimated the overall association for each of the variables and measures of their heterogeneity, including the I2 for males and females using a DerSimonian-Laird random effects meta-analysis model, arriving at an overall estimate of association (e.g., for our example,).[17] Out of the 7,251 and 6,288 variables unique available for females and males, 2,830 and 2,307 variables for females and males were measured in greater than one survey and thus available for a meta-analysis (Supplementary Figure 1AB). As examples, 4421 and 3921 variables were only available in 1 survey for males and females respectively. 79 and 57 variables were available across all 50 surveys for females and males respectively. For the 4421 and 3921 variables that appeared only in one survey (for females and males), we retained the survey-specific estimate. We used the *metafor* package in R to compute the meta-analytic association estimates.[18]

Sex-specific meta-analyses within countries

For each variable, we combined associations of all of the surveys within a country (e.g., Zambia, 2007 and Zambia 2013-2014) to estimate a country specific association for each sex. Given an association between a variable (i.e, marital status, call it ) and HIV+ in a survey for a sex (i.e, Zambia, 2013-2014, females denoted by in the equation; Zambia, 2007, females, denoted by ), we estimated the overall association for each of the variables for males and females using a DerSimonian-Laird random effects meta-analysis model, arriving at an overall estimate of association ().

Comparison of associations assayed across multiple countries

To facilitate comparison of associations that are measured in multiple countries, we binned by the number of countries in which the variables appear, including 1 country, 2-10 countries, 11-19 countries, and 20-29 countries (Figure 1, Supplementary Figure 2A-C, Table 3). Variables that were identified and assessed across a larger number of countries exhibited similar Nagelkerke R2 distributions (Figure 1A, Supplementary Figure S2A); however, their odds ratios were attenuated (Figure 1B, Supplementary Figure S2B).

Prediction of HIV+ across Sub-Saharan Africa

We calculated the predicted probability of HIV as a function of the variables that were identified in common in 29 African countries. First, we identified the latest survey with HIV status for each country and then built a logistic regression model for the top 10 identified factors (had p-value less than 1x10-6 and R2 greater than 0.001) present in all 29 countries. We removed DHS participants who did not have all 10 variables available. Using this model, we estimated the area under the curve and the predicted probability for each individual.

The Area Under the Curve (AUC) estimates were calculated by varying the threshold at which a person was considered “HIV+” based on their predicted probability, estimating the sensitivity and specificity based on these thresholds, and calculating the area under the receiver operator curve (AUC). Precision-recall curves are preferred when the proportion of cases are much lower than non-cases by focusing on positive cases and are estimated by computing the area under the precision (positive-predictive value) versus recall (sensitivity) curve (PRAUC). For random predictors, the AUC will be equal to approximately 0.5 and the PRAUC will be equal to the prevalence of HIV+. Therefore, a predictor that achieves an AUC greater than 0.5 or a PRAUC greater than the prevalence is considered to be better than random. A perfect AUC or PRAUC is 1.

Finally, we assess the concentration of HIV risk. We do this by calculating the proportion of the population that carries the predicted probability of HIV from our model. This allows us to calculate a “gini” coefficient of HIV risk by estimating the cumulative HIV risk distribution that is accounted for by the portion of the sample population using the *ineq* package in R.[19] For example, the gini coefficient allows us to see if 20% of HIV risk is accounted for by 80% of the population as ranked by predicted risk.[20]

Hardware and software environment

Data preparation was performed using Stata 15 MP on a MacBook Pro with a 2.7 GHz Intel Core i7 processor and 16GB of RAM. In Stata 15, setting the maxvar to 18000 (or higher) is necessary given the high number of variables. The codebase for data preparation is available at <https://github.com/ekhco/dhs_hiv>, <https://github.com/donasaur/dhs-processing>, and <https://github.com/donasaur/hiv-aws>

The source code for intra-survey associations are available at <https://bitbucket.org/kajalc/dhs_hiv/src/master/>, and the source code for meta-analysis across surveys are available here : <https://github.com/chiragjp/dhs_hiv_meta>. A website of the findings is here: <https://www.chiragjpgroup.org/dhs_hiv_meta/>

**Supplementary Figures:**

**Figure S1. Number of Variables assessed as a function of number of surveys and countries.**

**Figure S2. Empirical CDF of (A) Nagelkerke R2, (B) exp(absolute value(beta)) or OR, and (C) Heterogeneity (I2).** Red line depicts CDF for those not identified, blue line identified (e.g., pvalue < 1e-6 and R2 > 0.001).

**Supplementary Tables:**

**Table S1. Distributions of odds ratios, Nagelkerke R2, and I2 (heterogeneity) estimates for variables appearing in 1 country and at least 2 countries.**

**Online Supplementary Tables:**

To enhance reuse and reproducibility, we have placed all of our summary statistics, including the meta-analytic odds ratio, R2, and I2, their standard errors, and p-values in tables located here: https://github.com/chiragjp/dhs\_hiv\_meta/blob/master/meta\_data/Online%20Supplementary%20Tables.xlsx?raw=true

The table worksheets include the following:

**Online Table S1. Distributions of odds ratios, Nagelkerke R2, and I2 (heterogeneity) estimates for variables appearing in 1 country and at least 2 countries.**

**Online Table S2. Summary statistics (meta-analytic OR, Nagelkerke R2, and I2) for females across all surveys**

**Online Table S3. Summary statistics (meta-analytic OR, Nagelkerke R2, and I2) for males across all surveys**

**Online Table S4. Country-specific summary statistics for males.**

**Online Table S5. Country-specific summary statistics for females.**

**Supplementary Figures**

Chart

Description automatically generated

**Figure S1. Number of Variables assessed as a function of number of surveys and countries.**

**Graphical user interface, diagram, application

Description automatically generated**

**Figure S2. Empirical CDF of (A) Nagelkerke R2, (B) exp(absolute value(beta)) or OR, and (C) Heterogeneity (I2).** Red line depicts CDF for those not identified, blue line identified (e.g., pvalue < 1e-6 and R2 > 0.001).

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **gender/number** | **Num Associations** | **25th OR** | **Median OR** | **75th OR** | **25th R2** | **Median R2** | **75th R2** | **25th I2** | **Median I2** | **75th I2** |
| **1 Survey** |  |  |  |  |  |  |  |  |  |  |
| f | 4543 | 1.2 | 1.62 | 3.69 | 8.56 × 10−5 | 3.58 × 10−4 | 1.17 × 10−3 | . | . | . |
| m | 4089 | 1.24 | 1.78 | 12.67 | 8.39 × 10−5 | 3.29 × 10−4 | 1.12 × 10−3 | . | . | . |
| **>1 Surveys** |  |  |  |  |  |  |  |  |  |  |
| f | 2708 | 1.18 | 1.73 | 73.83 | 3.33 × 10−4 | 7.12 × 10−4 | 1.36 × 10−3 | 17.56 | 75.77 | 99.22 |
| m | 2199 | 1.24 | 2.6 | 263.09 | 3.01 × 10−4 | 6.49 × 10−4 | 1.30 × 10−3 | 17.77 | 81.94 | 99.32 |

**Table S1. Distributions of odds ratios, Nagelkerke R2, and I2 (heterogeneity) estimates for variables appearing in 1 country and at least 2 countries.**