PHIA Data Use Manual

Reference Guide for Using Data from the Population-based HIV Impact Assessments

This project is supported by the U.S. President's Emergency Plan for AIDS Relief (PEPFAR) through CDC under the terms of cooperative agreement #U2GGH001226.
The contents of this document do not necessarily represent the official position of the funding agencies.

PHIA Collaborating Institutions

ICAP at Columbia University
The United States Centers for Disease Control and Prevention (CDC)
Westat

Donor Support

This project has been supported by the President's Emergency Plan for AIDS Relief (PEPFAR) through the Centers for Disease Control and Prevention (CDC) under the terms of cooperative agreement #U2GGH001226. The findings and conclusions are those of the authors and do not necessarily represent the official position of the funding agencies.

Suggested Citation

Population-based HIV Impact Assessment (PHIA) Data Use Manual. New York, NY. April 2021.

Access this Manual Online

The PHIA Project: http://phia.icap.columbia.edu

Contact Information

ICAP at Columbia University
722 West 168th Street
New York, NY 10032
Website: icap.columbia.edu
Email: icap-communications@columbia.edu

Table of Contents

Abbreviations

AAPOR	American Association for Public Opinion Research
ARV	Antiretroviral
ART	Antiretroviral Therapy
BIO	Biomarker
CAMPHIA	Cameroon Population-Based HIV Impact Assessment
CD4	CD4+ T-Cell
CDC	Centers for Disease Control and Prevention
CHAID	Chi-square Automatic Interaction Detection
CSV	Comma-Separated Values
CASI	Computer-Assisted Self-Interview
CI	Confidence Interval
CONSORT	Consolidated Standard of Reporting Trials
CIPHIA	Cote d'Ivoire Population-Based HIV Impact Assessment
DF	Degrees of Freedom
DHS	Demographic and Health Surveys
EA	Enumeration Area
EPHIA	Ethiopia Population-Based HIV Impact Assessment
HAPHIA	Haiti Population-Based HIV Impact Assessment
HH	Household
HIV	Human Immunodeficiency Virus
ID	Identification
IND	Individual
JK	Jackknife
KENPHIA	Kenya Population-Based HIV Impact Assessment
LASSO	Least Absolute Shrinkage and Selection Operator
LEPHIA	Lesotho Population-Based HIV Impact Assessment
LAg-EIA	Limiting-Antigen Avidity Enzyme Immunoassay
MPHIA	Malawi Population-Based HIV Impact Assessment
MDRI	Mean Duration of Recent Infection
NAMPHIA	Namibia Population-Based HIV Impact Assessment
ODn	Normalized Optical Density
ODK	Open Data Kit
OVC	Orphans and Vulnerable Children
PHIA	Population-based HIV Impact Assessment
PSU	Primary Sampling Unit
PCA	Principal Components Analysis
PFR	Proportion of False Recents
RPHIA	Rwanda Population-Based HIV Impact Assessment
SRS	Simple Random Sampling
SACEMA	South African Centre for Epidemiological Modelling and Analysis
SHIMS2	Swaziland HIV Incidence Measurement Survey 2
THIS	Tanzania HIV Impact Survey
TB	Tuberculosis
PEPFAR	U.S. President's Emergency Plan for AIDS Relief
VM	Violence Module
VL	Viral Load
ZAMPHIA	Zambia Population-Based HIV Impact Assessment
ZIMPHIA	Zimbabwe Population-Based HIV Impact Assessment

1. Background

1.1. Purpose of this Manual

The purpose of the ***PHIA Data Use Manual*** (hereafter, "Manual") is to serve as a reference document to guide researchers in using Population-based HIV Impact Assessment (PHIA) data. Information contained herein applies generally to all PHIA surveys in all countries. The Manual is designed to provide information on: (i) survey design, sampling and measures; (ii) dataset structure and key variables; (iii) the interactive online ***Data Explorer***; (iv) practical guidance on how to download and use PHIA datasets; and (v) a list of other resources and documentation.

Some survey design specifications vary by PHIA survey, which are described in separate documentation. Survey-specific ***Data Use Manual Supplements*** are provided for each PHIA survey, which describe each survey's target population, sampling approach, data collection dates, optional survey measures and biomarkers collected, variations in eligibility criteria, and dataset and variable specifications. ***Sampling and Weighting Technical Reports*** provide further in-depth details on technical aspects of survey design and weighting not covered in these documents. References are provided in this Manual when users should consult these resources.

Detailed descriptions of survey data collection procedures, establishing participation by the household head, procedures for individual consent, maintaining confidentiality during data collection and testing procedures, procedures for returning/obtaining test results, and referral for or direct linkage to services are included in the final country report for each PHIA survey.

1.2. Overview of the Population-based HIV Impact Assessments

The PHIA Project, implemented by ICAP at Columbia University in partnership with the Ministries of Health and the US Centers for Disease Control and Prevention (CDC), is designed to measure the reach and impact of HIV programs in PEPFAR-support countries through national household surveys. The goals of the PHIA surveys are to provide population-level assessment of the burden of HIV disease at the national and sub-national level, and to document the achievement of HIV programs in participating countries. PHIA surveys are being conducted in 14 countries from 2015 – 2019.

Country (survey acronym)	Data collection dates
Zimbabwe (ZIMPHIA)	Oct 2015 – Aug 2016
Malawi (MPHIA)	Nov 2015 – Aug 2016
Zambia (ZAMPHIA)	Mar 2016 – Aug 2016
Uganda (UPHIA)	Aug 2016 – Mar 2017
Eswatini (SHIMS2)[1]	Aug 2016 – Mar 2017
Lesotho (LEPHIA)	Nov 2016 – May 2017
Tanzania (THIS)[2]	Nov 2016 – Jun 2017
Namibia (NAMPHIA)	Jun 2017 – Nov 2017
Cameroon (CAMPHIA)	Jun 2017 – Jan 2018
Cote d'Ivoire (CIPHIA)	Aug 2017 – Mar 2018
Ethiopia (EPHIA)	Oct 2017 – Apr 2018
Kenya (KENPHIA)	May 2018 – Mar 2019
Rwanda (RPHIA)	Oct 2018 – Mar 2019
Haiti (HAPHIA)	Jul 2019 – Nov 2020

Survey acronyms generally represent abbreviated country name and Population-based HIV Impact Assessment (PHIA), with exceptions noted below.

[1] Swaziland HIV Incidence Measurement Survey (SHIMS2).

[2] Tanzania HIV Impact Survey (THIS)

1.3. Survey sampling and measures

The PHIA Project consists of a set of cross-sectional household-based surveys designed to assess key HIV-related health indicators. PHIA uses a stratified multistage survey sampling design, with strata defined by sub-national geographic divisions used in each country's latest census (e.g., region/province). Within each stratum, census enumeration areas (EAs) are randomly selected with probability proportional to population size (1st stage), followed by a random sample of households within selected EAs (2nd stage). Consenting households are administered a household interview completed by the head of household or designee. For most countries, within selected households, a subset of households is selected to administer procedures targeting children under 15 (3rd stage). In each selected household, individuals are assessed for eligibility to complete individual interviews where they are administered a structured questionnaire, and individuals who complete individual interviews are offered biomarker testing (see *Section 3.4. Survey Weights* and each PHIA survey's ***Sampling and Weighting Technical Report*** for details).

As described elsewhere[1], biomarker testing involves collection of whole blood samples from participants either by venous blood draw, finger prick or heel stick, depending on age. Blood samples are used for home-based tests including HIV rapid testing and, depending on the survey, other point of care tests, such as syphilis, hepatitis B surface antigen and CD4. Additional biomarker tests, including viral load and early infant diagnosis testing are conducted on blood plasma, or dried blood spots at satellite or central laboratories, and results are returned to participants or at a nearby health facility.

PHIA surveys are designed to collect standardized data across countries. In some surveys, additional optional measures are collected—e.g., questionnaire modules on HIV knowledge, physical and sexual violence, biomarker tests for syphilis and hepatitis B and C, and weight measurement for children under 5 years of age. The list of measures and eligibility criteria may differ by PHIA survey (see each survey's ***Data Use Manual Supplement***).

1.3.2. Survey questionnaires

In participating households, a household questionnaire is administered to the household head. Then, individual questionnaires are administered to eligible and consenting individuals in the household. Adults (typically 15+ years) complete an adult questionnaire, and in some countries, adolescents (typically 10-14 years) complete an adolescent questionnaire. Adults also provide data on their children (typically 0-14 years) as part of the "children" module of the adult questionnaire. Emancipated minors may complete the household questionnaire as the household head but only complete the adult questionnaire if they meet age eligibility criteria. Modules potentially included in each questionnaire and their general eligibility criteria are listed in the table below. The age ranges for adults, adolescents and children, as well as content and order of each module may differ between PHIA surveys. Users can refer to each survey's ***Data Use Manual Supplement***, ***Survey Questionnaires*** and ***Codebooks***.

Questionnaire module	Eligibility criteria
Household questionnaire	Sample of households within selected EAs
Household roster	
Support for orphans and vulnerable children (OVC)	
Household spouses/live-in partners	
Deaths	
Household characteristics	
Economic support	
Individual questionnaire – adults (15+ years)	All rostered[1] and consenting adults
Respondent background	
Marriage	
Reproductive history	All women
Children	Parents or guardians of children or adolescents (age 0-14) in the household provide education, health and HIV-related data about each of their children
Male circumcision	All men
Sexual activity	
HIV knowledge	See each survey's ***Data Use Manual Supplement***
HIV testing history	
HIV status, care and treatment	All self-reporting HIV-positive adults
Tuberculosis and other health issues	
Alcohol use	
Injection drug use	
Gender norms	
Violence	A sub-sample of individuals, meeting varied sex and age criteria[2] in specific PHIA surveys
Computer-Assisted Self-Interview (CASI)	A sub-sample of all adults
Migration	
Network scale-up	A sub-sample of all adults
Individual questionnaire – adolescents (10-14 years[3])	All rostered[1] and assenting adolescents from selected households in specific surveys[2]
Sociodemographic characteristics	
HIV knowledge	All adolescents
HIV prevention interventions	
Sexual behavior	
Violence	A sub-sample of adolescents, meeting varied sex and age criteria[2] in specific PHIA surveys
HIV risk perceptions	
Social norms, intention to abstain, self-efficacy and assertiveness	
HIV testing	
Alcohol and drugs	
HIV stigma	

[1] Household members are eligible to be rostered if they were reported by the household head to have slept in the household the night before the interview.
[2] Eligibility criteria may vary by PHIA survey. See each survey's ***Data Use Manual Supplement*** for details.

1.3.3. Biomarker testing

Biomarker testing is offered to all adults and adolescents who completed an individual interview and consented or assented to provide blood samples, and children whose parents consented, depending on the PHIA survey. Biomarker tests potentially included in each PHIA survey and general eligibility criteria are listed in the table below, including tests performed in the field or at central laboratories. For details, users can refer to each survey's ***Data Use Manual Supplement***.

Biomarker	Eligibility criteria
HIV serostatus[1]	All participants
Recency of HIV infection[2]	All HIV+ individuals
CD4+ cell count	All HIV+ and a sub-sample of 2-5% of HIV- individuals[3] in specific PHIA surveys
HIV RNA viral load	All HIV+ individuals
Antiretroviral (ARV) drug presence	All HIV+ individuals
ARV drug resistance	All HIV+ individuals with recent infection
Syphilis antibody	Sub-sample of individuals meeting varied age and HIV status criteria[3] in specific PHIA surveys
Hepatitis B antigen	Sub-sample of individuals meeting varied age and HIV status criteria[3] in specific PHIA surveys
Hepatitis C antigen	Sub-sample of individuals meeting varied age and HIV status criteria[3] in specific PHIA surveys
Weight	Children 0-5 years, in selected PHIA surveys, varied sampling rates by survey[3]

[1] HIV serostatus is determined via pre-specified HIV testing algorithms that generally include a combination of home-based rapid HIV testing and confirmatory laboratory-based testing. The algorithm and specific tests used may vary by PHIA survey. See each survey's ***Data Use Manual Supplement*** for details.
[2] Recency of HIV infection is determined via a combination of Limiting Antigen Enzyme (LAg-Avidity) Immunoassay, viral load and ARV results. See *Section 3.6.3. New HIV infections and annual HIV incidence*.
[3] Eligibility criteria may vary by PHIA survey. See each survey's ***Data Use Manual Supplement*** for details.

1.4. Other documentation and resources

In addition to this Manual, researchers can find other documentation available for download on the PHIA Project website <https://phia-data.icap.columbia.edu/>.

General documentation applicable to all PHIA surveys includes:

- ***PHIA Tabulation Plan***: A list of formatted model tables used in published PHIA reports and descriptions of the datasets and variables used in these tables.

Additionally, each PHIA survey has supporting survey-specific documentation including:

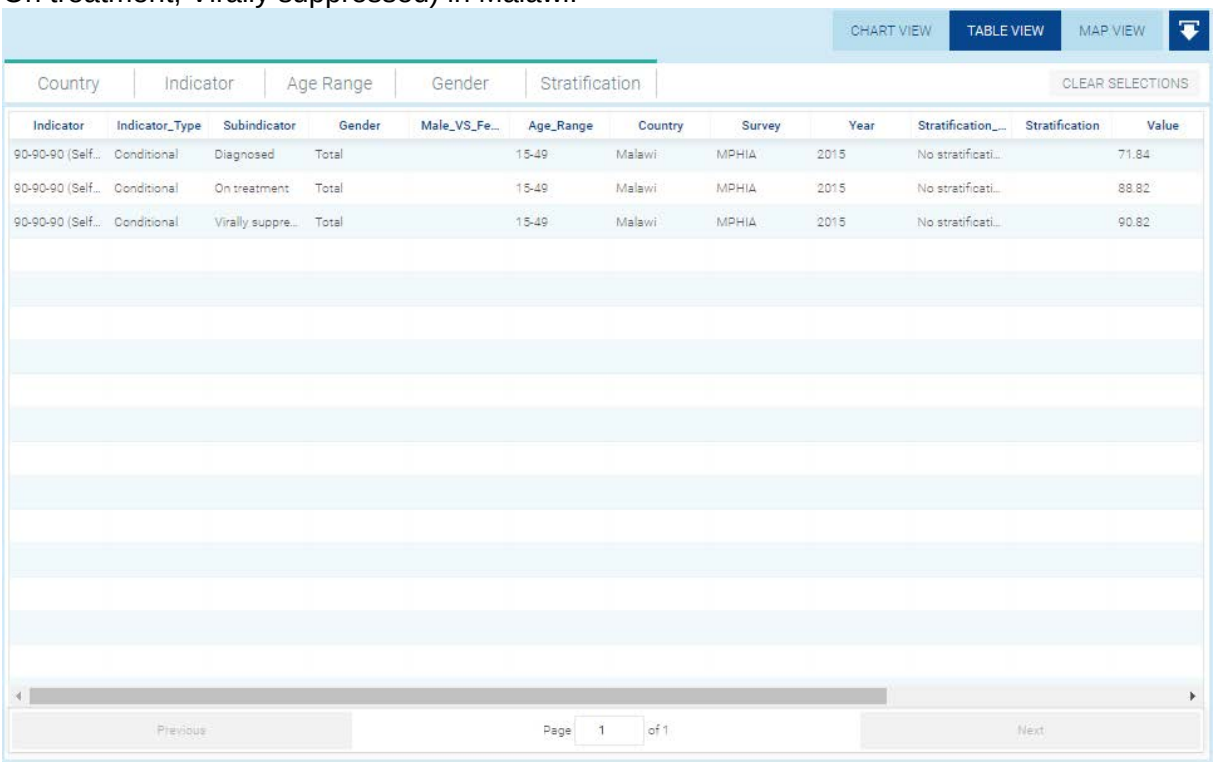
- ***Data Use Manual Supplement***: A supplement to this manual is provided for each PHIA survey, detailing survey specifications (e.g., variations in target population, sampling approach, data collection dates, and optional survey measures and biomarkers collected, etc.).
- ***Geospatial Data Use Manual****:* Description of how EA-level geospatial data are derived and masked, how to request access, and example code to link geospatial variables with other datasets.
- ***Survey Questionnaires***: Three questionnaires are provided per PHIA survey, one each for the household, adult and adolescent questionnaires. These questionnaires illustrate the questionnaire's structure, including the order that the questions were asked, each question's wording, variable names and labels, value coding and labels, and skip patterns.
- ***Codebooks***: Codebooks are provided for each dataset, indicating all variables contained within. These codebooks document each variable's name, category (i.e., the questionnaire module or source data of the variable), label (i.e., question wording or other label), type (e.g., integer, select one, select multiple, free text, and date/time) and coding values and labels.
- ***Variable Frequencies***: Variable frequencies are provided, which contain frequencies of all categorical variables in each dataset.
- ***CONSORT Diagrams***: CONSORT (CONsolidated Standard of Reporting Trials) style diagrams define key analytic variables that combine multiple source variables. A list of CONSORT Diagrams is provided for each PHIA survey.
- ***Sampling and Weighting Technical Reports***: Technical details of each PHIA survey's sampling and weighting procedures are provided in detail.
- ***PHIA Publications***: Preliminary PHIA data has been published and provided on the website, including summary sheets and final reports as they become available.

2. Data Explorer

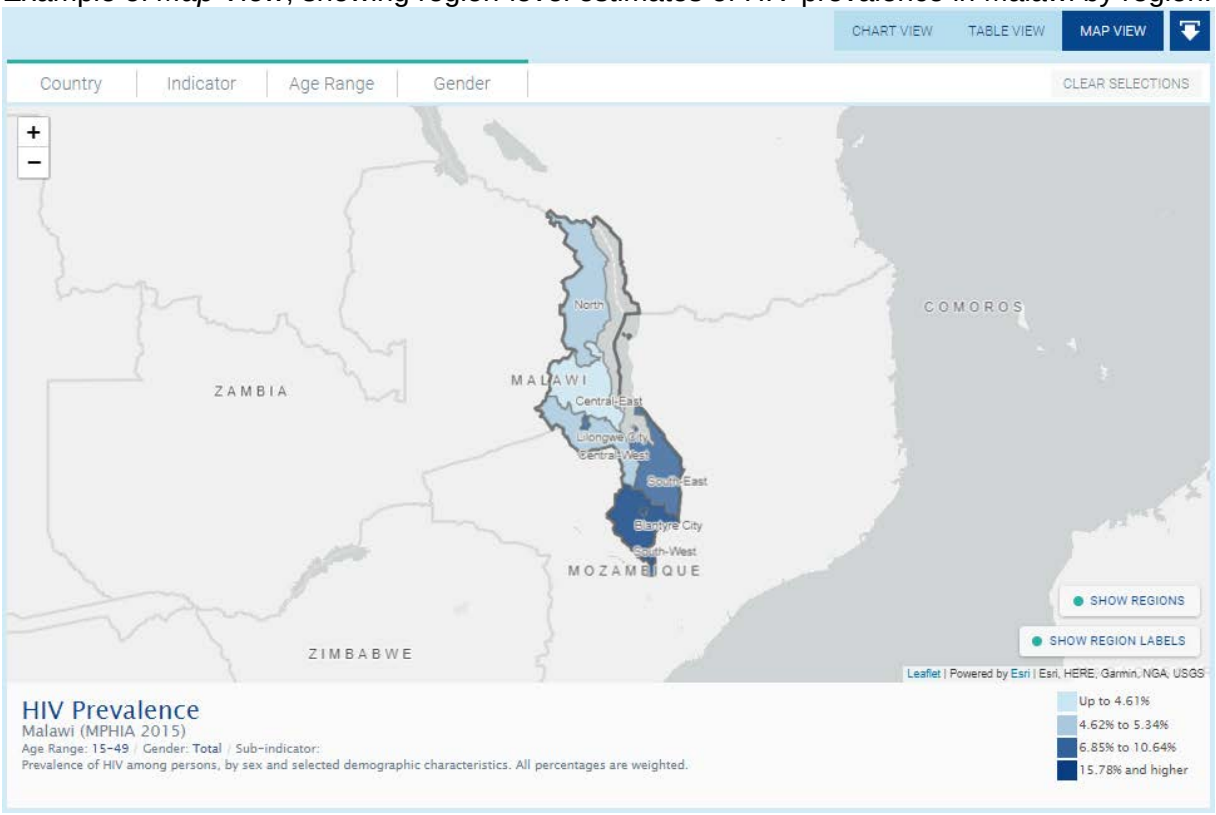
Researchers who are interested in exploring PHIA data can begin by using the online ***Data Explorer*** interface. This interactive tool enables exploration of PHIA data, including survey response rates and key indicators, such as prior HIV testing, receipt of antenatal care, male circumcision, and biomarkers such as HIV prevalence and incidence, ARV use, and viral suppression. The ***Data Explorer*** supports stratification by demographic groups and comparisons across countries. Users can create customized visualizations in chart, table, and map view forms and download them for external use (see example screenshots below). The ***Data Explorer*** can be accessed on the PHIA Project website at <phia-data.icap.columbia.edu>.

Example of *Chart View*, showing national-level estimates of 90-90-90 indicators (Diagnosed, On treatment, Virally suppressed) in Malawi. Multi-country comparisons are also possible.

Example of *Table View*, showing national-level estimates of 90-90-90 indicators (Diagnosed, On treatment, Virally suppressed) in Malawi.



Example of *Map View*, showing region-level estimates of HIV prevalence in Malawi by region.



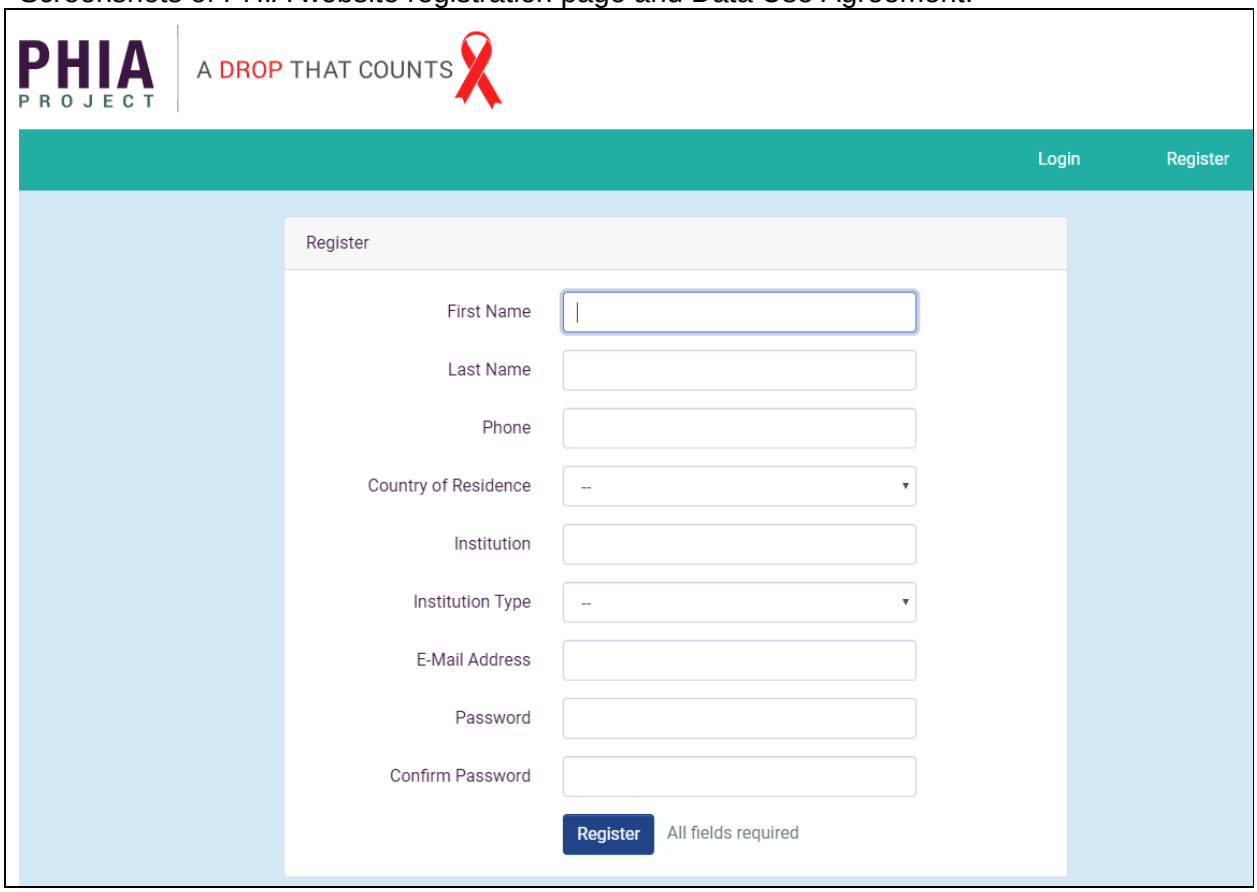
3. Guide to Analysis

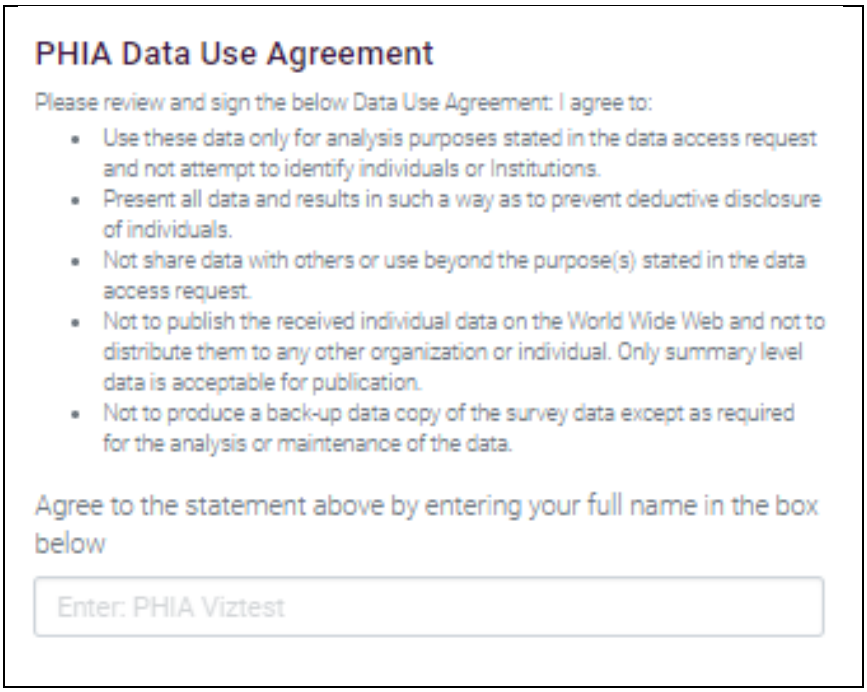
For users who are interested in taking a closer look at the PHIA data, this section provides practical information for accessing and managing respondent-level datasets and conducting analyses. Detailed background information and instructions are provided, and example code for SAS (SAS Institute Inc., Cary, NC, USA), Stata (StataCorp LLC., College Station, TX, USA), and R (R Foundation for Statistical Computing, Vienna, Austria) statistical software programs are provided in *Section 4. Example Code*.

3.1. How to access PHIA data

In order to access PHIA datasets, users will need to create an account and password and submit a request for dataset access on the *Registration* page of the PHIA website <https://phia-data.icap.columbia.edu/visualization-testing>. Users are required to specify the purpose of the analysis and the datasets needed and complete a Data Use Agreement when submitting a request for access. Upon approval of the request for access, registered users can then proceed to the *Data downloads* page to find links to each dataset. Download links will be available for 30 days from the date of approval.

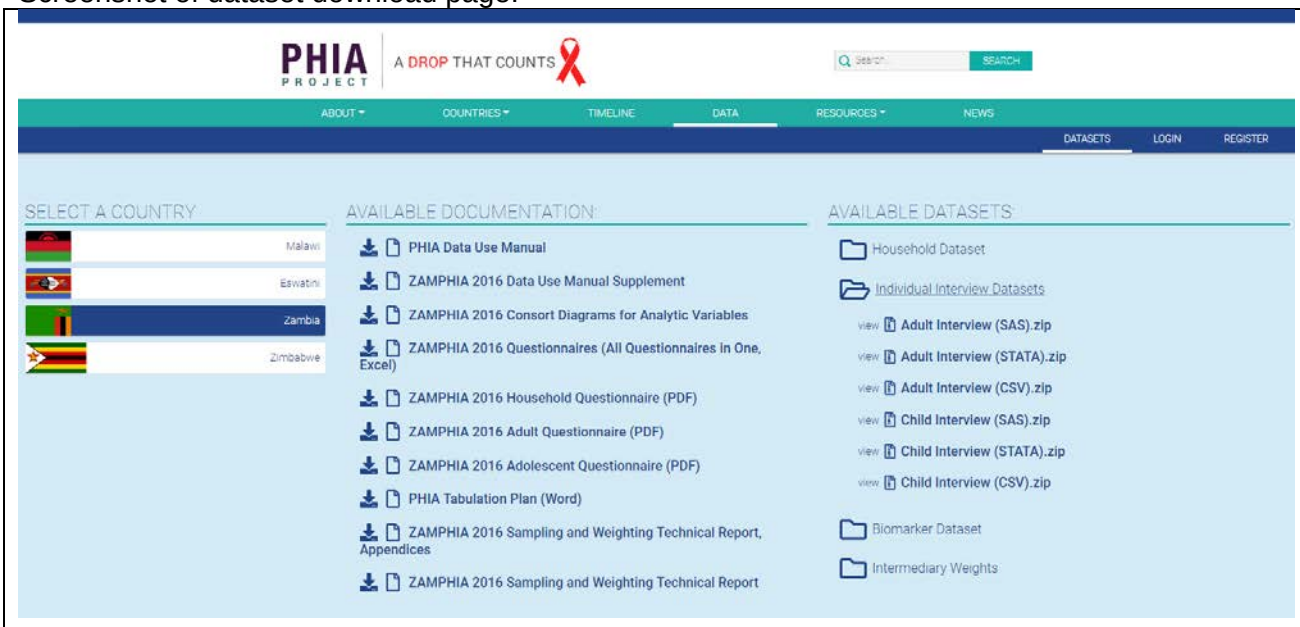
Screenshots of PHIA website registration page and Data Use Agreement.





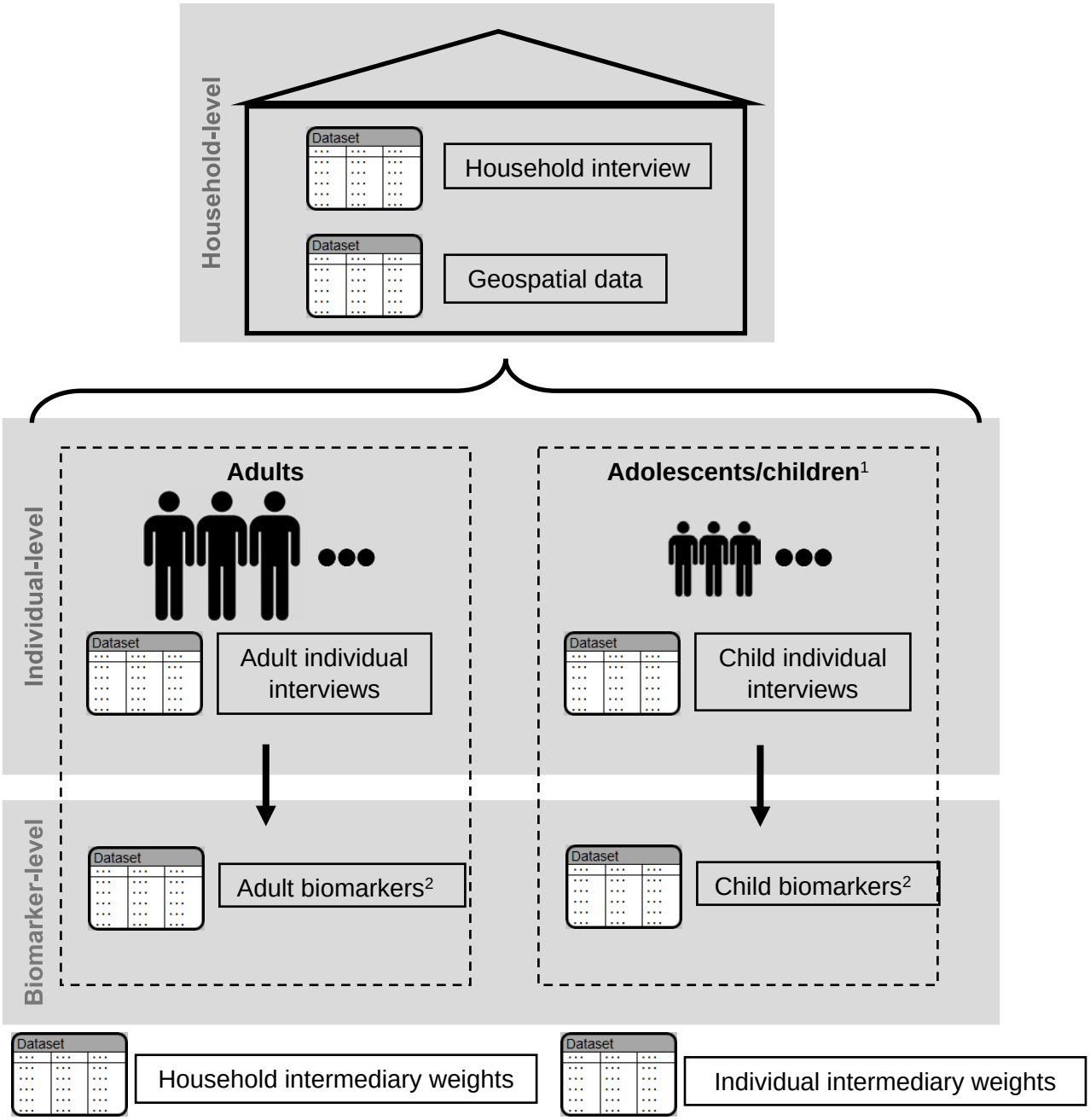
Data files are available in 3 formats: SAS (.sas7bdat), Stata (.dta), and comma-separated values (.csv). CSV files are provided to enable analysis outside of SAS and Stata software programs, but users should note that CSV file formats do not include useful metadata, such as variable labels. Users should also note that variable values may differ across PHIA surveys and should refer to each PHIA survey's supporting documentation (***Survey Questionnaires*** and ***Codebooks***) to compare across surveys.

Screenshot of dataset download page.



3.2. Structure of PHIA datasets

PHIA datasets are organized in a hierarchical structure, with each household record being associated with one or more individual records, and each individual record potentially having an associated biomarker record. Separate datasets are provided for records at different levels of analysis, and individual and biomarker tests are separated for adults and children.



[1]See each PHIA survey's **Data Use Manual Supplement** for survey-specific definitions of adolescents and children.
[2]Drug resistance data will be available at a later date.

3.2.1. Structure of datasets

Generally, dataset variables are ordered beginning with country name, followed by ID variables for the household or individual record, all questionnaire and/or biomarker variables, and finally all survey weights (base weights and jackknife replicate weights). Further information on how these survey weights were constructed and practical steps for using them in analysis are described in detail below (see *Section 3.4. Survey weights*).

- Households are the highest-level observation. Each sampled household is represented as a single record (row) on the *Household* dataset (e.g., *zamphia2016hh*), including sampled households that were ultimately determined to be ineligible or were non-responding households. Each household is identified with a 14-digit unique code `householdid`, beginning with a two-letter country code (e.g., ZM for Zambia), and a randomly-generated 12-digit numerical code. Households that participated in the household survey are indicated as eligible respondents by the variable `hhstatus = 1`.

country	householdid	hhstatus	<var1>	<var2>	…
Zambia	ZM000000000001	1	…	…	…
Zambia	ZM000000000002	2	…	…	…
…	…	…	…	…	…

- Individuals are the next level observation. There are three categories of individuals, adults 15+, adolescents 10-14 and children 0-9 (age definitions differ by PHIA survey). The adult interview records are contained within the adult interview data set (e.g., *zamphia2016adultind*), while both the adolescent and child interview records are contained within the child interview data set (e.g., *zamphia2016childind*). Note that adults provide data for children 0-14. On the child interview data set, records for children 0-9 contain data provided by their parent/guardian and not the children themselves, while records for adolescents 10-14 contain data provided by their parent/guardian and that they provided themselves as part of the adolescent questionnaire. Additionally, some individual-level data is collected during the household interview. Within each responding household, all individuals are represented as a single record (row) on the *Interview* datasets, including sampled individuals who were determined to be ineligible or were non-respondent. Each individual is identified with a 16-digit unique code `personid`, which begins with the `householdid` of their household followed by their "line number" on the household roster. Line numbers are unique identifiers for individual rostered household members, beginning with "01" for the household head and incrementing for each household member reported by the household head during the household interview. Individuals who participated in the individual interview are indicated as eligible respondents by the variable `indstatus = 1`.

country	householdid	personid	indstatus	<var1>	<var2>	…
Zambia	ZM000000000001	ZM00000000000101	1	…	…	…
Zambia	ZM000000000001	ZM00000000000102	2	…	…	…
Zambia	ZM000000000002	ZM00000000000201	1	…	…	…
Zambia	ZM000000000002	ZM00000000000202	2	…	…	…
Zambia	ZM000000000002	ZM00000000000203	1	…	…	…
…	…	…	…	…	…	…

- Biomarkers are the next level observation, and are available for all individuals who completed an individual interview and consented to undergo a blood draw and HIV and other point of care testing. Individuals who are eligible for biomarker testing are represented as a single record (row) on the *Biomarker* datasets (individuals may be ineligible based on age criteria; see each PHIA survey's ***Data Use Manual Supplement***). ID variables are identical to those in the *Interview* datasets. There is one data set for adults 15+ years of age and another for adolescents and children 0-14 years of age. Individuals who participated in biomarker testing and had valid laboratory test results are indicated by the variable `bt_status = 1`.

country	householdid	personid	bt_status	<var1>	<var2>	…
Zambia	ZM000000000001	ZM00000000000101	1	…	…	…
Zambia	ZM000000000001	ZM00000000000102	2	…	…	…
Zambia	ZM000000000002	ZM00000000000201	1	…	…	…
Zambia	ZM000000000002	ZM00000000000202	2	…	…	…
Zambia	ZM000000000002	ZM00000000000203	1	…	…	…
…	…	…	…	…	…	…

- Enumeration areas (EAs) are the unit of analysis available for geospatial data. Geospatial datasets include the latitude and longitude coordinates for the centroid of each EA. These centroid coordinates are masked using a random offset to protect participant privacy and confidentiality. See ***Geospatial Data Use Manual*** for further details. EAs are identified by a unique ID variable centroidid, which is available on all datasets to enable linking of the centroid coordinates.

country	centroidid	longitude	latitude
Zambia	ZM000001	…	…
Zambia	ZM000002	…	…
Zambia	ZM000003	…	…
Zambia	ZM000004	…	…
Zambia	ZM000005	…	…

3.2.2. Variable labels and formats

All information on PHIA dataset variables including variable labels, survey question wordings, value codes and value labels is provided in supporting documentation (users are advised to consult each PHIA survey's ***Questionnaires***, ***Codebook***, and ***Variable Frequencies***). For user convenience, a program to create a SAS format library is also provided for download with each SAS data set, containing value labels for all variables where applicable. Users can run this program to create format libraries in SAS, after which they can stored and used to format SAS output. Formats follow a naming convention where they are as the source variable followed by 'f'. As an example, the variable `hivstatusfinal` has values `1`, `2` and `99`, with format named `hivstatusfinalf`, containing `(1) HIV Positive`, `(2) Negative`, and `(99) Missing` (see Section 4. Example Code). Stata and R users should note that value labels are not provided in the .dta and .csv versions of the datasets; options to import SAS format libraries into those programs are user-generated and cannot be guaranteed to work.

3.2.3. Protecting participant confidentiality

To protect participant confidentiality, all participant IDs described above are scrambled to ensure that participants cannot be identified from the data. Specifically, `householdid` variables are randomly generated and cannot be linked to the participant except via a secure link file maintained by designated data managers at Ministries of Health and ICAP at Columbia University. All other identifying information such as participant names, addresses, phone numbers, as well as identifying information provided in free-text fields have been excluded from PHIA datasets. See section 3.3.3 for additional information about data confidentiality processes.

Geospatial data below the highest sub-national geographical area designation (e.g., region/province) are available upon request. To protect participant confidentiality, this data is limited to the latitude and longitude coordinates of the centroid of each EA, and has been masked by adding a random offset. Further information on masking procedures are provided in accompanying ***Geospatial Data Use Manual***.

3.3. Data management and cleaning

3.3.1. Missing data and other exceptions

PHIA surveys are administered using open data kit (ODK)-based software on electronic tablets, which enables forced responses. As a result, missing data for survey variables are minimal, except where participants explicitly responded “don’t know” (generally coded as “-8”, with some exceptions where “don’t know” is a valid response), refused to answer (“-9”), or responses were out of range (“-7”, e.g., when a woman who has been pregnant says she has never had sex), or where a question does not apply (“.”, e.g., number of prior pregnancies does not apply for men). Variables from check-all-that-apply questions are coded as character variables, with “don’t know” and refusal responses coded as “Y” and “Z”, respectively. Missing data for analytic variables (see *Section 3.6. Analytic variables*) are coded as “99” without distinguishing the reason for missingness (don’t know, refused or not applicable). For biomarker data, missing values (“.”) indicate that participants were not tested for the biomarker.

Users should take care when conducting analyses to check for and determine appropriate treatment for missing responses. Consult each PHIA survey’s ***Survey Questionnaires***, ***Codebook***, and ***Variable Frequencies*** for further information on each variable.

Users are also strongly advised to take caution and refer to each PHIA survey’s supporting documentation when conducting analyses using variables pertaining to sexual partnerships. Data on sexual partnerships are captured in PHIA surveys as part of the Sexual Activity module of the Individual Interview. It should be noted that in some countries (e.g., Lesotho, Namibia, Swaziland, Uganda, and Zambia), the total number of unique sexual partners reported in the last 12 months (`part12mo`) may exceed the total number of reported lifetime sexual partners (`partlifetm`). Consult each PHIA survey’s ***Survey Questionnaires*** and ***Codebook*** for specific information on each variable.

3.3.2. Age and date variables

Several age and date variables are provided in the PHIA datasets. Age variables include, but are not limited to, ages of participants and their children at the time of their interviews, and self-reported age at first sex and of recent sexual partners. Ages of all household members are first reported by the household head, which are subsequently confirmed by the individual if they participated in the individual interview. In cases of discrepancies, the participant’s confirmed age takes precedence. Household member ages that are above a maximum age threshold (typically 80 years) are “top-coded” to the maximum age to preserve confidentiality.

Other age data may be cleaned, though this varies on a variable-by-variable basis. Mother’s age, for example, is cleaned to missing (“.”) when mother and child ages are less than 10 years apart since this is considered to be biologically implausible (see *Section 3.5.3 Mother-to-child linking*). Conversely, ages of recent sexual partners are not cleaned; for example, if a participant reports that a recent sexual partner was >100 years old, this is considered unlikely but not implausible. PHIA data users are strongly advised to check distributions of variables to ensure appropriateness to the intended analytical purpose.

Date variables include prior pregnancies, HIV diagnosis/testing, and ART use, and birth year (birth day and month data are redacted to protect participant privacy) among others. Tablet-generated dates, such as start date for survey procedures at a given household, are retained as one date variable. However, dates based on responses to survey questions generally allow

participants to provide partial dates. For example, participants who remember the year but not month/day of a given event are allowed to provide the year only and month and day will be left as missing. Date variables are retained as answered in the dataset, with a few exceptions:

- Non-calendar dates. Non-real calendar dates are removed, such that if a non-real calendar date is reported, all date variables are blanked. For example, 31 February 2015 is an impossible day/month combination, thus in this instance, all date variables including the year are deleted.
- Future dates. Dates for events that are supposed to be in the past cannot be indicated as taking place in the future. For data collected in the household interview, HHQSTS is used to determine the current date. For data collected in the individual interview, surveystdt is used to determine the current date. surveystdt uses the interview start timestamp if available, and signature timestamps (for consents) to fill in missing values. All future dates are blanked.
- Impossible events. Events cannot happen in a child's life before the child is born. Similarly, pregnancy-related dates cannot exist for men, nor can HIV testing, HIV care and ARV dates that occur before 1985. Dates corresponding to impossible events are blanked.

3.3.3. Data confidentiality processes

The protection of participant privacy and confidentiality was maintained at each phase of PHIA data collection and processing. To ensure the protection of participant privacy and confidentiality, PHIA data processing encompasses various methods to reduce the risk of disclosure in the public use data. The mitigation of potential risk disclosure occurs at the household-level and individual-level and addresses both direct and indirect identifiers in the public use data.

In general, the following methods were used to minimize any privacy or confidentiality concerns in the PHIA data:

- Redaction: removal of specific variables or removal of elements within the data variable (e.g. day from date).
- Top-coding: process of re-coding continuous values above an upper bound to the value of the upper bound.
- Bottom-coding: process of re-coding continuous values below a lower bound to the value of the lower bound.
- Small case count: defined as the lowest category containing at least 25 cases or 1 percent of households or individuals reporting the category; may be managed through top-coding, bottom-coding, or redaction.
- Standardization of number/units variables: there are some questions for which the response is stored in two variables, the first which provides the units (days, weeks, months, years) and another with the number of the selected units. Prior to applying the methods above, the two variables are combined into a single measure with a consistent unit of days, weeks, months or years. For example, in some countries, the duration for which a child is breastfed is indicated by the variable ch_childbrstfddurw on the child's record and the variable childbrstfddurw1 on the mother's record. In both instances the "w" in the variable name indicates that the duration is given in weeks, which the user can confirm in the corresponding codebooks.

The following risk mitigation methods are applied across all PHIA public-use datasets:

- Removal of all direct identifiers (e.g. names, addresses, phone numbers)
- Household and participant IDs were randomly reassigned, as indicated in section 3.2.3.
- Days have been redacted from all date variables. Month and year were retained.
- All age variables have been top-coded to 80.
- In certain circumstances, age variables were bottom-coded. See each PHIA survey's ***Data Use Manual Supplement*** for specific details.
- For categorical variables, categories with counts of less than 25 were collapsed into "other", if "other" is an option. Response types "Don't know" and "Refused" were not collapsed into "other" because these response options are not identifying. Special circumstances may exist. See each PHIA survey's ***Data Use Manual Supplement*** for additional details, including variables with this method applied.
- For dichotomous variables (i.e. variables with yes/no response options), the variables may have been redacted from the data if there were no risk remediation measure possible. See each PHIA survey's ***Data Use Manual Supplement*** for additional details, including variables with this method applied.
- For continuous variables, top-coding or bottom-coding may have been used. See each PHIA's Supplement for additional details, including variables with this method applied.

Recodes and redactions may introduce some data limitations. Some variables were redacted altogether and collapsed categories lost some detail. It may not be possible to reproduce all standard analytic variables from the variables available on the public-use datasets.

For more information about redactions to specific variables, see each PHIA survey's ***Data Use Manual Supplement***.

3.4. Survey weights

Survey weights are provided for each survey assessment. Weights are based on sampling probabilities and adjusted for non-response and post-stratification to national population projections from the survey year based on age and sex. In addition, jackknife (JK) replicate weights are provided for JK variance estimation. This section describes the method used for constructing these weights and practical guidance on how to use these weights for analyses. Example code for Stata, SAS and R statistical programs are provided in *Section 4. Example Code*. For additional technical details, refer to each PHIA survey's ***Sampling and Weighting Technical Report***.

3.4.1. Weighting approach

Base weights. PHIA surveys use a stratified, multistage probability sample design. At the first stage, Enumeration Areas (EAs) are selected with probability proportional to size within strata, which usually consist of the first administrative division after country, such as region or province. Within each selected EA, lists of households are constructed for the second stage of sampling, which are drawn from updated household listing data collected by the PHIA team. Households are selected using an equal probability method. The actual number of households selected per cluster varied by PHIA survey, with an average of 30 and range of 15 to 60 (refer to each PHIA survey's ***Sampling and Weighting Technical Report***).

The overall household probability of selection is calculated as the product of the EA probability of selection and the household probability selection for each case, as follows:

$$p_{hij} = p_{hi} * p_{j|hi} ,$$

where:
p_{hij} = the household probability of selection for household *j* in EA *i* in EA stratum *h*
p_{hi} = the probability of selection for EA *i* in EA stratum *h* (adjusted for any substitution of EAs, if necessary)
$p_{j|hi}$ = the conditional probability of selection for household *j* in EA *i* in EA stratum *h*

Household base weights are calculated as the inverse of the overall household probability of selection, as follows:

$$w_{hij} = \frac{1}{p_{hij}}$$

Individual base weights are calculated based on the household base weights after adjustment for non-response (see below), multiplied by probability of selection based on age criteria that differ by PHIA survey. Usually, adults in the household were selected with 100% probability and children 0-14 years old were selected with some fixed probability in each household (see each PHIA survey's ***Data Use Manual Supplement*** for eligibility criteria):

$$w_{hijk} = K_k NR_{hi} W_{hij}$$

where:
W_{hij} = the household base weight for household *j* in EA *i* in EA stratum *h*
NR_{hi} = the non-response adjustment for households in EA *i* in EA stratum *h*
K_k = the inverse probability of selection for person *k*

Blood test base weights are calculated based on the individual base weights after adjustment for non-response (see below). Since all individuals who are eligible for the interview are selected for blood testing, no further probability sampling is taken into account to create the blood test base weights.

$$w_{hijk}^{bt} = NR_{hi}W_{hijk}$$

where:
W_{hijk} = the individual base weight for individual *k* in household *j* in EA *i* in EA stratum *h*
NR_{hi} = the non-response adjustment for individuals in EA *i* in EA stratum *h*

Nonresponse adjustments. Some nonresponse is anticipated for each of the three study components – the household questionnaire, the individual-level questionnaire, and a blood draw. Response status is nested, such that individual-level responses are only obtained within households that participate, and blood sample responses are only obtained within individual-level responses. Under these conditions, household-level data are available for individual-level nonresponse adjustments, and individual-level interview data are available for blood sample nonresponse adjustments.

Nonresponse weight adjustment follows the cell-weighting approach.[2] Nonresponse weight adjustment cells for households are EAs or groups of EAs. Nonresponse weight adjustment cells for individuals and blood samples are determined through the use of a CHAID (Chi-square Automatic Interaction Detection) tree classification scheme, which identifies predictors of response.[3] Response propensities are calculated within each cell defined by these response predictors, which are then used to adjust for non-response. The table below lists examples of potential independent (predictor) variables that are used to define the nonresponse adjustment cells, which are initially selected using Least Absolute Shrinkage and Selection Operator (LASSO) regression.[4] Ultimately, any and all variables available on each data source can be selected to define non-response adjustment cells. Further details on each survey's weighting variable specifications can be found in each PHIA survey's ***Sampling and Weighting Technical Report***.

Examples of sources and variables used in nonresponse adjustment

Component	Source	Potential independent variables
Household	EA sampling frame	Region, district, urban/rural, HIV prevalence rate (high/low, etc.)
Individual interview	EA sampling frame and household interview	Roster information about the individual such as age and sex of individual; roster information about the household, such as household size, recent deaths, sick parents, presence of parent/guardian, and assets (ownership of electronic equipment, various animals, water source, power source, etc.)
Blood test	EA sampling frame and household and individual interviews	Any and all individual characteristics such as age, sex, education, employment, and other demographics; HIV status, knowledge, HIV testing and care history; TB status and care history, circumcision status.

Nonresponse adjustment of the household weights uses EAs as nonresponse adjustment cells. For household *j* in EA *i*, the nonresponse adjusted (final) household weight is computed as:

$$w_{hij}^{HH} = w_{hij} * \frac{\sum_{sampled}^{cell\ HH} w_{hij}}{\sum_{responding}^{cell\ HH} w_{hij}}$$

If an EA has such a low response rate that the adjustment would result in excessively high adjusted weights, that EA is grouped with a similar EA with more respondents to form a single non-response adjustment cell.

Nonresponse adjustment of the individual-level weight begins with the household-level nonresponse adjusted weight. The initial individual-level weight for sampled individual *k* in household *j* in EA (PSU) *i* and stratum (region) *h* is computed as:

$$w_{hijk}^{indiv} = w_{hij}^{HH} / p_{hijk} ,$$

where w_{hij}^{HH} is the nonresponse adjusted household weight, and p_{hijk} is the conditional probability of selecting individual *k* in household *j* in EA *i* and stratum *h*. Note that generally $p_{hijk} = p_{hij}$ for adults, while $p_{hijk} < p_{hij}$ for children 0-14 years of age, since children are subsampled from households to participate in child interviews. For example, in surveys where half of eligible children are selected, $p_{hijk} = p_{hij}/2$ for children. Initial individual-level weights are then adjusted for nonresponse using data from the household interview as potential predictors of nonresponse, yielding the nonresponse-adjusted individual-level weight, $w_{hijk}^{indiv,nr}$.

Nonresponse adjustment of the blood sample weights begins with individual-level nonresponse-adjusted weights. These weights are adjusted for nonresponse using responses to the individual interview as possible predictors of nonresponse using a new, separate LASSO and CHAID-based model.

Poststratification (undercoverage) adjustments. PHIA surveys aim to provide estimates of number of persons affected by HIV in addition to proportions affected in various sub-groups. Thus, each set of nonresponse-adjusted weights are further adjusted for undercoverage to a set of population projections for the country. Undercoverage adjustments are made in similar fashion to nonresponse adjustments, by creating cells within which weights are adjusted for undercoverage. Undercoverage adjustment cells are defined by sex and age group distribution at the national level, with each cell having a known population total taken from the national census or population projections. The nonresponse-adjusted weights for all cases in each cell are summed, and multiplied by an adjustment factor, which is equal to the known total divided by the sum of weights for that cell. These weighting procedures are extended to the construction of survey weights for optional survey modules and biomarker tests.

Summary of sequence of weight adjustments

Weight	Input weight	Adjustment factor	Output weight
Household (final)	w_{hij}	$\frac{\sum_{sampled}^{cell\ HH} w_{hij}}{\sum_{responding}^{cell\ HH} w_{hij}}$	w_{hij}^{HH}

Individual-level, nonresponse	w_{hijk}^{Pers}	$\frac{\sum_{sampled}^{cell\ person-level} w_{hijk}^{Pers}}{\sum_{responding}^{cell\ person-level} w_{hijk}^{Pers}}$	$w_{hijk}^{Indiv,nr}$
Individual-level, undercoverage (final)	$w_{hijk}^{Indiv,nr}$	$\frac{Known\ total,cell}{\sum_{responding}^{cell\ person-level} w_{hijk}^{Indiv,nr}}$	w_{hijk}^{Indiv}
Blood sample, nonresponse	$w_{hijk}^{Indiv,nr}$	$\frac{\sum_{sampled}^{cell\ blood\ sample} w_{hijk}^{Indiv,nr}}{\sum_{responding}^{cell\ blood\ sample} w_{hijk}^{Indiv,nr}}$	$w_{hijk}^{BT,nr}$
Blood sample, undercoverage (final)	$w_{hijk}^{BT,nr}$	$\frac{Known\ total,cell}{\sum_{responding}^{cell\ blood\ sample} w_{hijk}^{BT,nr}}$	w_{hijk}^{BT}

3.4.2. Selected subsamples for optional modules

In some PHIA surveys, optional assessments are administered to selected subsamples of respondents. These include the HIV knowledge questionnaire module, violence questionnaire module, and Hepatitis B biomarker testing. Survey weights are provided in order to conduct weighted analyses for each subsample's target population of interest. The target population and eligibility criteria for each module are described briefly here. The final survey weights for each optional module are constructed using the final individual weights multiplied by each person's probability of selection for the module, and are adjusted for non-response and poststratification. Users should consult each PHIA survey's ***Sampling and Weighting Technical Report*** for survey-specific details on how these weights are constructed.

HIV knowledge module. In PHIA surveys that administered the HIV knowledge module, varied sampling strategies were applied to select adults at random. An additional sample of randomly selected adolescents per household meeting survey-specific age criteria may also be selected. Eligibility is thus determined by age, gender and eligibility for the individual interview. Response status is determined by whether the participant provided a valid response to a pre-selected question in the HIV knowledge module (usually `mosquito` "Can a person get HIV from mosquito bites?"). See each PHIA survey's ***Sampling and Weighting Technical Report*** for details on eligibility criteria and response status determination.

Child module. Special consideration should be given to analyses using child interview data, since only a subset of all sampled households was flagged for child data collection in each survey. These child-sampled households were randomly selected at a fixed probability per household (usually 50%, refer to each PHIA survey's ***Data Use Manual Supplement*** and ***Sampling and Weighting Technical Report*** for details). In child-sampled households, children are eligible for blood testing and may be asked additional questions. Interview and blood test weights included in child datasets (`intwt` and `btwt`, respectively) only apply to sampled children. On the other hand, while unsampled children (`indstatus=8`) are given 0 weight for individual interviews and blood tests (`intwt=0` and `btwt=0`), some data may have been provided for them by their parent or guardian. In general, this includes variables from the roster, such as age and gender, as well as questions from the adult questionnaire's children module that have been attached to the child records. These variables typically have the prefix "`ch_`" on the child dataset and can be identified by filtering the variable category in the child interview dataset codebook to "Adult questionnaire - Module 3A: children" (note that the module number may vary by PHIA survey). Analyses that intend to use data available for all children, sampled

and unsampled, should use child module weights (`chmodfwt`). See *3.4.3 Survey weight variables in PHIA datasets*.

Violence module. In PHIA surveys that administered the violence module (VM), a sub-sample of adult females in each household meeting survey-specific age criteria is randomly selected. An additional sample of randomly selected adolescents per household meeting survey-specific age criteria may also be selected. Eligibility is thus determined by age, gender and eligibility for the individual interview. Response status is determined by whether the participant provided valid responses to a minimum set of questions in the VM. See each PHIA survey's ***Sampling and Weighting Technical Report*** for details on eligibility criteria and response status determination.

Hepatitis B/C testing. In PHIA surveys that administered Hepatitis B (Hep B) and/or C (Hep C) biomarker testing, either all participants meeting survey-specific age criteria are selected, or a subsample of individuals meeting survey-specific age and HIV status criteria are randomly selected. In PHIA surveys where a full sample is drawn, no Hep B/C testing survey weights are needed. In PHIA surveys where a random subsample is drawn, eligibility is thus determined by age, HIV status and eligibility for the individual interview. Response status is determined by whether the results of the Hep B/C test is determinate (individuals with either reactive or non-reactive Hep B/C test results are considered non-respondents). See each PHIA survey's ***Sampling and Weighting Technical Report*** for details on eligibility criteria and response status determination.

Child nutrition. Weight and height measurements are collected on samples of children 0-5 years, depending on the PHIA survey. Typically, all children meeting this age criterion who test HIV positive and a random sample of 2% or 5% of those who test HIV negative (depending on the country) are sampled. Child nutrition sampling weights are provided to account for the subsampling of HIV negative cases. Eligibility is thus determined by having valid HIV test results. See each PHIA survey's ***Sampling and Weighting Technical Report*** for details on eligibility criteria and response status determination.

Computer-Assisted Self-Interview (CASI). In some PHIA surveys, a subset of potentially sensitive questions is re-administered via CASI. Questions are selected based on potential for social desirability (response) bias, such as HIV risk behaviors. The CASI module is administered to a random sample of adult males and females who participate in the individual interview. See each PHIA survey's ***Sampling and Weighting Technical Report*** for details on eligibility criteria and response status determination.

3.4.3. Survey weight variables in PHIA datasets

Eligibility and response indicators. PHIA datasets includes sets of survey weight variables to enable weighted analyses for each survey assessment. For each type of analytic weight, indicator variables identify each observation's eligibility and response status and thus whether weights are available. For example, `hhstatus`, `indstatus`, `bt_status` denote whether the household or individual is determined to be an eligible respondent, eligible non-respondent, ineligible or of unknown eligibility. For optional modules, including HIV knowledge/attitudes and violence modules, module-specific indicator variables identify sampling, eligibility and response status for the given module and thus whether weights are available. For the HIV knowledge module: `hivk_eligible` indicates eligibility and `hivk_status` indicates response status, and for the violence module: `vm_status` indicates eligibility and `vmflag` indicates response status. Refer to each PHIA survey's ***Data Use Manual Supplement*** for details on survey-

specific eligibility criteria, and ***Technical Report*** for details on how these eligibility and response indicators were derived.

Survey weight variables. The final nonresponse- and poststratification-adjusted sample weights are provided in each dataset and labeled accordingly (refer to the tables below for variable names of survey weights for each questionnaire module or biomarker test). Availability of survey estimation procedures varies by statistical software (see *Section 4. Example Code*). Users will need to use the appropriate weights for the specific analysis of interest, which is generally determined by the target population of inference.

- Household weights can be used for analyses conducted at the household level, for example, distribution of households by urban/rural residence. Household weights can also be used for analyses at the individual level which only depend on the household questionnaire, for example the percentage of orphaned children in the population, using the household roster variables momalive and dadalive. Household weights can be interpreted as the number of households that the participating household represents in the population, accounting for sampling and non-response at the EA and household levels.
- Interview weights can be used for analyses conducted at the individual level for data collected for all eligible interview participants. For example, self-reported HIV testing (i.e., ever received an HIV test prior to the survey) should be estimated using interview weights since all interview respondents received HIV testing questions. In this scenario, interview weights can be interpreted as the number of individuals that the respondent represents in the population who could have participated in the interview, accounting for sampling and non-response at the EA, household and individual levels.
- Blood weights can be used for analyses conducted only among blood test participants. For example, HIV prevalence should be estimated using blood test weights even if the analysis includes predictors at the household or individual level, since not all interview respondents participated in blood tests. In this scenario, each participant's blood weight can be interpreted as the number of individuals that the participant represents in the population who could have participated in blood testing, accounting for selection and non-response of EA, household, individual and blood testing. In addition, if the outcome of interest comes from the interview (e.g., HIV testing history), but the analysis is restricted to those who have blood test results, blood test weights should be used.
- Researchers who are interested in analyzing data for all children (sampled and unsampled) should not use interview weights for such analyses (see *3.4.2 Selected subsamples for optional modules* and each PHIA survey's ***Data Use Manual Supplement*** for details). An indicator variable `chmod_status` denotes whether the child is eligible for analysis using child module data, and child module weights (`chmodfwt`) are provided for all children with `chmod_status` = 1. However, care should be taken when conducting analyses that combine child module data with those collected only on sampled children (e.g., blood test outcomes), since those data will be missing for unsampled children. Child module weights can be interpreted as the number of children that the child represents in the population, accounting for non-response of EA, household and the responding adult.
- Violence weights can be used for analyses conducted among sub-samples of individuals who were sampled for and participated in the violence module (VM). For example,

prevalence of sexual violence in the past 12 months prior to the survey should be estimated using violence weights, since only a sub-sample of eligible interview participants per household were selected to receive the VM (eligibility criteria vary by PHIA survey). Violence weights can be interpreted as the number of individuals that the VM respondent represents in the population, accounting for selection and non-response of EA, household, individual, and VM.

- Data on sexual partners and marital relationships is collected, and couples may be a unit of analysis of interest to users (see *Section 3.5.5 Sexual / marital partner linking*). As is the case with other household-based surveys such as the Demographic and Health Surveys (DHS), we "did not identify eligible couples in the household listing, only eligible individuals. Therefore, the number of couples eligible to participate in the survey is unknown, and it is not possible to calculate a true couples' weight."[5] The man's individual sample weight is considered to be a reasonable proxy weight for the couples, on the basis that response rates tend to be lower among men. To maintain comparability, PHIA recommends using the man's individual interview or blood weight for couples, as appropriate for the analysis of interest.

Lastly, users interested in accessing the intermediary weights used for sample selection at each stage and for non-response and post-stratification adjustment will find these variables in each PHIA survey's *Intermediary Weights* datasets.

3.4.4. Jackknife variance estimation

Though multiple existing variance estimation methods can appropriately account for PHIA's complex survey design, Jackknife (JK) repeated replication is a preferred method and is used for all ***PHIA Publications***. JK variance estimation involves the creation of replicate weights, where one primary sampling unit (PSU) is omitted from the analysis in turn. This approach results in a number of replicate weights equal to the number of PSUs. In each set of replicate weights, observations within the omitted PSU have their survey weight set to 0, while other PSUs in the same stratum have weights adjusted upwards to make up for the missing PSU. PSUs in different strata retain the original survey weight. To estimate a quantity of interest, such as a proportion, each set of replicate weights is used to separately compute the proportion. The mean and variance of the resulting distribution of estimated proportions gives the final estimated mean and variance of the proportion. PHIA surveys use a special case in which each stratum contains exactly two PSUs, a variation of the JK method known as JK2. It has been shown that JK2 analyses can be simplified by generating one set of replicate weights per stratum, omitting a randomly selected PSU from each cluster in turn. This method yields asymptotically equivalent variance estimates with half the number of replicates required, and is thus more statistically efficient.[6] Variance estimates produced by the Jackknife method reflect nonresponse and poststratification adjustments since the replicate weights are based on the original final survey weights. Users should note two technical considerations when using JK2 variance estimation methods in their analyses (see Example 1 in *Section 4. Example Code*).

- JK coefficients: Since the JK replicate weights provided in the PHIA datasets follow the JK2 approach, JK coefficients must be set to 1, overriding the default option in most statistical packages. For further information, see Valliant.[7]
- Degrees of freedom (df): A widely accepted rule-of-thumb for calculating df for survey estimates from stratified cluster surveys is to use the number of clusters minus the number of strata.[8] This method is typically the default setting in most statistical software

packages and is appropriate for national-level estimates. For stratified analyses, since the number of clusters is often large, the default df may yield overly precise variance estimates. An alternative approach is recommended for stratified analysis, where the user should override the default df calculated by software and set df=25.[6]

Survey weight variables for JK variance estimation follow a consistent naming convention:

Module	Variable name	
	Base weight	JK replicate weights
Household	hhwt0	hhwt001-hhwt…
Individual interview	intwt0	intwt001-intwt…
Blood test	btwt0	btwt001-btwt…
Child module	chmodfwt0	chmodfwt001-chmodfwt…
(optional) HIV knowledge module	hivkpswt0	hivkpswt001-hivkpswt…
(optional) Violence module	vmpstw0	vmpstw001-vmpstw…
(optional) Hepatitis B testing	hepbw0	hepbw001-hepbw…
(optional) Hepatitis C testing	hepcw0	hepcw001-hepcw…
(optional) Hepatitis B/C testing	hepbcwt0	hepbcwt001-hepbcwt…
(optional) Child nutritional status	cwh_wt0	cwh_wt001-cwh_wt…
(optional) CASI module	casipswt0	casipswt001-casipswt…

Note: Refer to each PHIA survey's ***Data Use Manual Supplement*** for details on the number of JK replicates per survey.

3.4.5. Taylor series variance estimation

Alternatively, users can apply Taylor series linearization methods to estimate variances. This method requires specifying survey weights, strata and sampling units. For each set of survey weights, datasets include identifier variables for variance estimation stratum (`varstrat`) and primary sampling unit/cluster within variance estimation stratum (`varunit`). Users will need to specify the variance stratum and unit variables and base weights appropriate for the analysis of interest (see Example 1 in *Section 4. Example Code*).

Module	Variable name		
	Base weight	Sampling stratum	Sampling unit
Household	hhwt0	varstrat	varunit
Individual interview	intwt0	varstrat	varunit
Blood test	btwt0	varstrat	varunit
Child module	chmodfwt0	varstrat	varunit
(optional) HIV knowledge module	hivkpswt0	hivk_varstrat	hivk_varunit
(optional) Violence module	vmpstw0	vmpstw_varstrat	vmpstw_varunit
(optional) Hepatitis B testing	hepbw0	hepb_varstrat	hepb_varunit
(optional) Hepatitis C testing	hepcw0	hepc_varstrat	hepc_varunit
(optional) Hepatitis B/C testing	hepbcwt0	hepbc_varstrat	hepbc_varunit
(optional) Child nutritional status	cwh_wt0	varstrat	varunit
(optional) CASI module	casipswt0	varstrat	varunit

3.4.6. Calculating response rates

Response rates are reported in ***PHIA Publications*** tables. In order to calculate household and individual response rates, the following procedure is used.

Household response rates. Sampled households were visited by field workers who determined household eligibility status, primarily based on the type of building and occupancy status. Household response status also depends on sufficient information being collected during the household interview. The variable `hhstatus` categorizes each household into one of four eligibility and response status categories:

`hhstatus`	Indicator of household eligibility and response status	1 - Eligible Responding Household 2 - Eligible Nonresponding Household 3 - Ineligible (Vacant Household, Not a Dwelling, Dwelling Destroyed) 4 - Unknown Eligibility Status

To calculate household response rates, PHIA uses the following procedure. Let R be the number of responding households, NR the number of non-responding households, IE the number of ineligible households, and UE the number of households whose eligibility could not be determined. The estimated proportion of sampled households which are eligible is defined as PE = (R+NR)/(R+NR+IE). In other words, PE is the eligibility rate among households with known eligibility. Then, unweighted household response rates are calculated following AAPOR's Response Rate 4 (see page 62 of Standard Definitions (pdf)[9]):

$$\text{Response rate} = 100 \times \frac{R}{R + NR + PE \times UE},$$

To obtain weighted household response rates, households are weighted using the household base weight `hhbwt0`.

Individual response rates. Individual response rates are based on individual eligibility and response status. The variables `indstatus` and bt_status categorize each individual for interview and blood draw eligibility and response status.

indstatus	Indicator of individual eligibility and response status	1 - Eligible Respondent 2 - Eligible Non-Respondent 4 - Unknown Eligibility Status 6 - Collected in another tablet 7 - Rostered in Error 8 - Not Sampled 9 - Non-de facto participants
bt_status	Did lab blood test have definite result?	1 - Lab blood test has a definite result 2 - Lab blood test does not have a definite result 9 - Lab blood test has a definite result, non-de facto participant

Unweighted interview response rates are calculated by dividing the number of eligible respondents (`indstatus = 1`) divided by the total number of eligible respondents (`indstatus = 1 or 2`). To obtain weighted interview response rates, individuals are weighted using the interview base weight `indiv_bwt0`.

Unweighted blood draw response rates are calculated by dividing the number of individuals with definite lab blood test results (`bt_status = 1`) by the total number of interview respondents (`indstatus = 1`). To obtain weighted blood draw response rates, individuals are weighted using the trimmed, non-response adjusted individual weight not including post-stratification adjustments (`trmpnr1w0`). See each PHIA survey's ***Sampling and Weighting Technical Report*** for details on these additional survey weights.

3.4.7. Pooled multi-country analysis

Background. Users may be interested in conducting analyses using data from multiple countries in a pooled analysis. The following describes how to create a single file for multi-country analyses. Let G be the number of countries and $\hat{y}_g$ be the estimate from a country g, for example, the total number of persons who tested HIV+ in country g. The multi-country (or pooled) estimate $\hat{y}$ is computed as

$$\hat{y} = \frac{\sum_{g=1}^{G} \hat{N}_g \hat{y}_g}{\sum_{g=1}^{G} \hat{N}_g} = \sum_{g=1}^{G} \hat{W}_g \hat{y}_g$$

where $\hat{N}_g$ is the estimate of the population of country g and $\hat{N} = \sum_{g=1}^{G} \hat{N}_g$ is the estimate of total population of the g countries. This estimate is a linear combination of the individual country estimates $\hat{y}_g$

$$\hat{y} = \sum_{g=1}^{G} C_g \, \hat{y}_g,$$

where the coefficient $C_g = \hat{W}_g$ is the estimate of proportion of the population of country g among all countries computed as $\hat{W}_g = \frac{\hat{N}_g}{\sum_{g-1}^{G} \hat{N}_g}$. In the combined estimator $\hat{y}$, the estimates from countries with large population sizes have more influence on the estimate than smaller countries.

Since the sample from each country was drawn independently, the variance of the combined estimator for the combined population or domains controlled by the poststratification adjustment is the sum of the variance of the country estimates multiplied by $\hat{W}_g^2$. For subdomains not controlled by the poststratification adjustment, the variance is more complex because $\hat{W}_g$ is a random variable. Computing the variance separately by country and then combining them in an appropriate way is cumbersome. This process can be simplified using replication by concatenating (i.e., stacking) the files of each country in one single file. The combined variance can be computed in the same way as a single country analysis with an increased number of replicates (see below). Computing the variances using replication yields valid estimates of variance accounting for the additional variation when the factors $\hat{W}_g$ are estimated for some sub-populations.

Example. An illustrative example is described below with the aim of conducting a combined analysis of the Zambia, Zimbabwe and Malawi PHIA datasets. The method involves

concatenating the individual country files and combining the replicate weights variable across countries. The table below shows the number of replicate weights for each country. The combined file will contain one full sample weight and 751 new replicate weights.

Country	Number of Replicates
Zambia	253
Zimbabwe	248
Malawi	250
Total	751

In all of these countries, the variable for the full sample weight is `btwt0` and the replicate weights are denoted `btwt`*j* for the *j*-th replicate weight (for example, `btwt2` is the second replicate weight). The assignment of the 751 replicate weights in the combined file is illustrated in the next table.

Country	Full sample weight	New replicate weights		
		1-253	254-501	502-751
Zambia	`btwt0`	`btwt001-btwt253`	`btwt0`	`btwt0`
Zimbabwe	`btwt0`	`btwt0`	`btwt001-btwt248`	`btwt0`
Malawi	`btwt0`	`btwt0`	`btwt0`	`btwt001-btwt250`

The creation/assignment of the new replicate weights is as follows:

1. Create a new file to contain the combined replicate weights by appending the three countries. The number of records in this new file should be the sum of the records in the files of the three countries.
2. Retain the values of the full sample weight `btwt0` for each record in each country and create 751 new replicate weights.
3. For the records in Zambia, retain the values of the first 253 new replicate weights as the values of the replicate weights `btwt001-btwt253` from Zambia and set all the subsequent replicate weights (254-751) to `btwt0` for Zambia.
4. For the records in Zimbabwe, set the values of the first 253 replicate weights and the last 250 replicate weights (i.e., 502-751) to `btwt0` for Zimbabwe, and use the replicate weights `btwt001-btwt248` from Zimbabwe for the intervening 248 new replicate weights, 254-501.
5. For the records in Malawi, replace the values of the first 501 replicate weights by `btwt0` for Malawi, and use the replicate weights `btwt001-btwt250` from Malawi as the last 250 new replicate weights, 502-751.

With this combined country data file, analysis with JK variance estimation can proceed as usual. It is important to note that the increased number of replicate weights will change the value of the degrees of freedom (df) used by default in most statistical software packages. This needs to be properly adjusted depending on the analysis (see each PHIA survey's ***Sampling and Weighting Technical Report*** for additional detail). Additionally, although the new file contains more replicate weights, the replication method remains the same (JK2) and the default JK

coefficient must be overridden by the analyst and set to 1. Although the objective of the combined file is the production of multi-country estimates, the same file can be used to compute estimates and their variances of differences among countries. This can be done using software that estimates contrasts. However, this type of analysis requires appropriate changes in the degrees of freedom depending on the domains being analyzed (see Example 5 in *Section 4. Example Code*).

Researchers conducting multi-country analyses are strongly advised to consult each PHIA survey's ***Data Use Manual Supplement*** and documentation (e.g., ***Questionnaires*** and ***Codebooks***) to ensure that differences between PHIA surveys are understood prior to pooling data. Pooled analyses containing any PHIA survey with normalized weights (i.e. Kenya, Rwanda, or Haiti) will require that the researcher make an adjustment to the weights so that they sum to an estimated population total. Without this adjustment, pooled estimates will under-represent the countries with normalized weights.

There may also be differences in question wording and response options across surveys. For example, education categories may differ substantially between countries, even though the variable names are the same. These differences are likely to affect interpretation of results in pooled multi-country analyses.

In addition, researchers must use caution when using data from optional assessments that may have been administered to different samples across countries, since survey weights may not be provided depending on each country's sampling method. For example, suppose a researcher intends to conduct an analysis using Hep B testing data from two countries, *A* and *B*, and Hep B testing was administered to a targeted sub-sample in country *A* but to all blood test respondents in country *B*. Hep B weights would be provided for country *A* but not *B*; combining country *A* and *B*'s datasets as described above would produce missing survey weights for country *B*, dropping its observations from weighted analyses. The researcher would need to populate the Hep B weights for country *B* with the appropriate weights – in this case, blood weights would accurately reflect country *B*'s sampling of all blood test participants for Hep B testing.

3.5. Linkages

PHIA data users may be interested in combining information from different datasets to conduct analyses. For convenience, several commonly used variables are provided on multiple PHIA datasets, including key household characteristics, individual demographics and HIV status. In addition, general instructions are provided on how to merge across all PHIA datasets. Finally, we provide linking variables that facilitate analysis on mother-child pairs and sexual partnerships within the household. Descriptions of these special linkages are provided below.

3.5.1. Household, individual demographics, and HIV status

For convenience, commonly used variables have been copied across datasets, including:

1. From *EA (cluster)* to *household, interview*, and *biomarker* datasets
    a. EA ID (`centroidid`)
2. From *household* to *interview* and *biomarker* datasets
    a. Household ID (`householdid`)
    b. Country name (`country`)
    c. Sub-national geographic area (e.g., `province`, `region`)
    d. Urban/rural indicator (`urban`)
    e. Wealth index (`wealthscorecont`) and quintiles (`wealthquintile`)
3. From *interview* to *biomarker* datasets
    a. Person ID (`personid`) and mother's ID (`momid`)
    b. Survey start date (`surveystdt`)
    c. Age (`age`, `agem`, and derived variables) and gender (`gender`)
    d. Self-reported HIV status (`hivselfreport`)
4. From *biomarker* to *interview* datasets
    a. Blood test response status (`bt_status`)
    b. HIV status (`hivstatusfinal`)

3.5.2. General procedures for linking datasets

In order to link records across datasets, the EA (cluster), household and individual identifiers (`centroidid`, `householdid` and `personid`) should be used along with "merge" procedures available in standard statistical software. General procedures are detailed below (see *Section 4. Example Code* for examples in SAS, Stata and R).

- To link household information to individual interview or biomarker datasets, merge variables from the *household dataset* onto the *interview* or *biomarker* dataset using the household ID (`householdid`) as the merging variable. This is a one-to-many merge (see Example 3).
- To link biomarker information to interview datasets when you wish to keep all records in the interview dataset, use the individual ID (`personid`) as the merging variable. This is a one-to-one merge. Since biomarkers are only available on a sub-sample of individuals, note that while all individuals will have interview data, not all individuals will have biomarker data in the output dataset (see Example 2).
- To link individual interview information to biomarker datasets when you wish to keep only records in the biomarker dataset, use the individual ID (`personid`) as the merging variable. This is a one-to-one merge. Since biomarkers are only available on a sub-sample of individuals, those who only have interview data should be dropped in the output dataset (see Example 3).
- To link EA geospatial data to household, interview and biomarker datasets, merge variables using the Centroid ID (`centroidid`) as the merging variable. This is a one-to-one merge (see ***Geospatial Data Use Manual***).

3.5.3. Mother-to-child linking

Rationale. PHIA surveys capture data regarding mother-child relationships, which may be of interest for analyses on children, including mother-to-child HIV transmission, among other topics. These data include information about children provided by their mother (e.g., breastfeeding, HIV testing, care and treatment history, vital status), as well as information about the mother (e.g., mother's age, HIV status, HIV testing and care and treatment history of the mother).

Identification of mother-child pairs. During the household interview, the household head identifies all relationships between a child (0-14 years of age) and their mother or female guardian. During the adult interview, mothers corroborate the relationship by providing the "line number(s)" of their children on the household roster.

In the reproductive module of the adult interview, women report if they had delivered a child in the last 3 years prior to the survey. If so, they provide pregnancy, childbirth and postpartum data on the last pregnancy and children born as a result. This data includes HIV testing, care and ARV use during pregnancy and children's vital status, breastfeeding history, and HIV status, testing, care and ARV use. Thus, data to identify mother-child pairs is strongest for children born within the last 3 years of the survey, and reproductive module data is the primary source of information used to identify mother-child pairs for children under 3.

For other children (4-14 years of age and children 0-3 years of age who were not from the last pregnancy), reproductive module data is not available. The individual identified as the mother by the household head is considered to be the mother as long as she is both a rostered member and alive at the time of the interview. This linking approach yields a small number of mothers linked to children with discordant HIV status (mother is HIV-negative and child is HIV-positive). These discordant cases are likely due to misreporting of the biological maternal relationship by the mother, household head or both or other mode of HIV transmission to the child.

Finally, some identified children are not eligible to be rostered because they did not sleep in the household the night before or were deceased at the time of the interview. These children are not assigned a `personid` and are not represented in child interview or biomarker datasets. Data provided by the mother on these children can only be found on the mother's record.

Some data checks are performed to confirm the validity of mother-child relationships. Specifically, certain mother-child age combinations, such as when the mother's age is less than the child's age or less than 10 years older than the child, are considered impossible. In these cases, the identity of the mother is considered to be unknown.

Mother-to-child linking variables. The identity of the mother is captured on the child's record by the identifier variables `momid` and `momfemname`, which contain the `personid` and "line number" of the mother within the rostered household members, respectively. For children whose mothers are unknown, these ID variables are coded as `momfemname` = "-7" and `momid` = <blank>. To facilitate mother-child analyses, we have used these linking variables to copy data about the mother to the child's record and vice versa. (refer to each PHIA survey's ***Codebook***):

- Data provided by the mother about her childbirth and postpartum information about their last born child born within 3 years of the survey are copied onto the child's record and are denoted with the "`ch_`" prefix in the *child interview* dataset. Note that not all children are sampled for the survey; therefore, some children will have `indstatus` = 8 and no

corresponding biomarker data or biomarker weights. The questions are also retained in the adult record in the *adult interview* dataset.

- Data provided by parents/guardians about all children 0-14 years old under their care have similarly been copied onto the child's record and are denoted with the "`ch_kid`" prefix in the *child interview* dataset. Note that not all children were sampled for the survey; therefore, some children will have `indstatus` = 8 and have no corresponding biomarker data or biomarker weights. The questions are also retained in the adult record in the *adult interview* dataset.
- Selected variables about the biological mother have been copied onto the child's record and are denoted with the "`mom`" prefix.
- Selected variables about the biological child, including child's HIV status and viral load, have been copied onto the mom's record and are denoted with.

Users who would like to merge additional variables on linked mother-child pairs can follow these steps (see example 4 in *Section 4. Example Code*):

1. Create a copy of the Adult IND dataset, renaming `personid` as `momid` and any other desired variables with a new name (e.g., rename `education` as `momeducation`). This will serve as the source of mother data to merge onto the child record.
2. Merge the new "mother dataset" onto the Child (IND or BIO) dataset, by the `momid` variable. Ensure that no childless mothers are retained in the dataset by restricting output to child records. This is a one-to-many merge since mothers may have multiple children in the household.

3.5.5. Sexual / marital partner linking

Sexual and marital partnership data are collected as part of the Household Interview (Roster Information), and the Individual Interview (Marriage and Sexual Activity modules). In order to facilitate analyses of partners, three types of partner linkage variables have been provided. Note that survey weights are not provided for analyses with couples as the unit of analysis since sampling procedures did not identify couples during household listing. For couples analyses, we suggest the use of the men's individual interview or blood weight (see *Section 3.4.3. Survey weight variables in PHIA datasets*).

HusID. The variable `husid` contains the `personid` of the husband reported by each female participant in the marriage module. If the husband is not a rostered household member, `husid` is blank. There is no analogous `wifeid` variable in the PHIA data sets. Husband-wife pair and polygamous relationships are identified only from `husid`.

PartID1-3. Three variables `partid1 partid2 partid3` contain the `personid` of up to 3 most recent sexual partners within the household as reported by the participant in the sexual activity module in the adult interview.

Lastpartner. The variable `lastpartner` contains the `partid` (1, 2 or 3) of the most recent sexual partner, if it is ascertainable from the data. Variables that contribute to `lastpartner` may differ by PHIA survey (refer to CONSORT diagram in each survey's ***Data Use Manual Supplement*** for details).

PartnerClusterID. Researchers may be interested in analyzing groups of 3+ individuals in the same household who are linked by sexual partnerships. Additionally, because HIV is a sexually transmitted disease, groupings of persons in a household who have had either direct sexual contact or indirect exposure via a mutual sexual partner or spouse are a potential unit of interest for study. These "partnership clusters" are relevant where an individual has multiple wives and/or sexual partners, thus pairs of individuals who are not themselves sexually partnered are connected indirectly through the common partner. The variable `partnerclusterid` captures these complex partnerships by assigning a unique ID to all individuals who are linked directly or indirectly by some marital or sexual relationship to any other individual in the same household. For partnership clusters formed by combinations of marital and sexual partnerships, linking individual records in a dataset is complex: the chains of partnership may require multiple links or joins (see case 7 below). The inclusive definition of a partnership cluster and the addition of the unique number to the dataset enables analysts to easily examine both these complex linked groups and simple partnerships without having to do their own complex joining and sorting. This definition also avoids assigning persons to more than one cluster, which would require multiple partnership grouping variables. Note that only relationships within the same sampled households are included. Any relationships reported outside the household are not identified.

Construction. The variable `partnerclusterid` uniquely identifies each partnership cluster across the whole dataset. Partnership clusters are defined using the following rules:

1. All wives linked to their husbands using the `husid` variable are a part of the same cluster.
2. Any persons reported as sexual partners by a given person will be a part of that person's cluster.

3. A person can only be in one cluster: if a person is linked to two or more other people then all of them, and anyone linked to them as a sexual or marital partner, will be combined into a single larger cluster.
4. Self-reported information will be assumed to be correct, even if only one side of the partnership reports the partnership.

The table below lists expected types of partnership cluster structures. In all of these examples, other persons, such as children, grandparents, or other unrelated adults may be present in the household, but only the members related by spouse/sexual partner links are shown.

Case 1 is a household with two pairs of partners: a husband and wife who are recorded as spouses and who reported each other as their only recent sexual partners, and an unmarried couple who also recorded each other as their only recent sexual partners. In this case the married couple are assigned a cluster number of 1 and the second couple is assigned to cluster 2.

Case 2 shows another relatively simple situation. Each partner has reported another sexual partner who is outside the household. This example demonstrates the utility in distinguishing between null/none responses and 'individual outside household' responses to the sexual partner questions. The presence of the other partners does not change the cluster numbering. Note that the husband/wife does not need to be the primary or most recent partner and could be identified under `partid2` or `partid3`.

Case 3 shows a husband with multiple wives. All of the husband's wives are linked to the husband, and to each other, through the partnership cluster number.

Case 4 is similar to case 3, but there is an additional woman in the household who is linked to person 401 by sexual partnership reports. All three are linked in one partnership cluster.

Case 5 illustrates inconsistent reports of partnership within a household. Person 503 has reported a sexual partnership with person 501, but there was no reciprocal report by person 501. In this method, self-reports are treated as correct regardless of whether the relationship is reciprocated, so these people will be linked. In this example, person 501 is also married to and a reciprocal sexual partner with person 502. As a result, person 502 is linked together with person 503 in the same partnership cluster.

Case 6 demonstrates the (relatively rare) case of households with more complex connections. Here there are two married couples, but these are also connected by an additional non-marital sexual partnership. All four persons are part of the same partnership cluster. Note that in this case person 602 is linked to 603 and 604 by the partnership cluster number, but that neither of these numbers occur on her record at all, and 603 does not occur on either her or her husband's record. That is, persons 602 and 604 are indirectly linked through 603.

	personID	gender	husID	partid1	partid2	partid3	partnerclusterid
Case 1. Two simple couples in same household							
	101	M	.	102	.	.	1
	102	F	101	101	.	.	1
	103	M	.	104	.	.	2
	104	F	.	103	.	.	2
Case 2. Husband and wife with other partners outside household							
	201	M	.	202	*(not in hh)*	.	3
	202	F	201	201	*(not in hh)*	.	3
Case 3. Husband and two wives							
	301	M	.	302	303	.	4
	302	F	301	301	.	.	4
	303	F	301	301	.	.	4
Case 4. Husband and wife with another partner in the household							
	401	M	.	403	402	.	5
	402	F	401	401	.	.	5
	403	F	.	401	.	.	5
Case 5. Inconsistently reported partnership							
	501	M	.	502	.	.	6
	502	F	501	501	.	.	6
	503	F	.	501	.	.	6
Case 6. Complex/chained partnership							
	601	M	.	602	604	.	7
	602	F	601	601	.	.	7
	603	M	.	604	.	.	7
	604	F	603	603	601	.	7

3.6. Analytic variables

This section includes background and technical information on analytic variables and statistics and provided in datasets.

3.6.1. CONSORT diagrams for analytic variables

Recoded ("analytic") variables have been provided in each of the datasets. As a guide for PHIA data users, CONSORT (CONsolidated Standard Of Reporting Trials)-style diagrams are provided for each variable in each PHIA survey's ***CONSORT Diagrams*** document. These diagrams illustrate which source variables are incorporated into each analytic variable and how participants are categorized based on data from these variables. PHIA data users should consult CONSORT diagrams to ensure proper interpretation of analyses when using the analytic variables provided.

3.6.2. Wealth index

Rationale. Wealth index methods using survey data on household assets, materials and durable goods have become an established measure of socioeconomic status since their adoption by the Demographic and Health Surveys Program (DHS).[10,11] These wealth measures are widely considered to be a more accurate construct than income to quantify socioeconomic status in resource-limited settings and are easily discernible via survey questionnaire. DHS has provided commonly accepted guidelines for wealth index construction.[11] Wealth index variables (continuous scores and quintiles) have been constructed for analysis using an easily reproducible method across PHIA surveys. Household dwelling characteristic and asset variables used to construct wealth indices vary by PHIA survey and are noted in each survey's ***Data Use Manual Supplement***. In PHIA datasets, two wealth index variables have been provided: a continuous score (`wealthscorecont`) and categorical wealth quintile (`wealthquintile`).

Method. To construct wealth quintiles via DHS methods, the following steps are used:

1. **Recode asset variables.** Household data include categorical variables about household characteristics, such as construction materials for walls, floors and roof of the household dwelling, source of water, availability of electricity and type of sanitation facilities used, and binary variables indicating ownership of durable goods such as beds, vehicles, and livestock, etc. The specific assets and question wording vary across PHIA surveys (see each PHIA survey's ***Data Use Manual Supplement*** and ***Survey Questionnaire***). Categorical variables are recoded as binary indicator variables (e.g., one variable was created for each floor type and a household receives 1 for the variable indicating their floor type and 0s for all others). Binary variables are coded as 1 (yes) or 0 (no). Generally, missing data are treated as the absence of that asset, and households that do not have data on any assets are not assigned wealth index scores or wealth quintiles.
2. **Select the asset variables for inclusion.** Asset variables are analyzed using PCA, which is a statistical technique that transforms a number of (correlated) variables into uncorrelated components that capture variability (information) in decreasing order; thus, PCA is a useful dimension reduction technique. DHS recommends using the first component of the model as a summary indicator for wealth (the wealth index). Since assets may vary in relevance in urban and rural settings, PCAs are run separately for urban and rural households, and then for all households combined. Decisions to include or exclude asset variables from either setting may be made on the basis of contextual

knowledge; for parsimony, all asset variables that have any variability are included in each analysis.

3. **Run PCA and combine results.** Three PCAs are run: a "common" model across all households, and models restricted to "urban" and "rural" households. As per convention, the first factor from each model is extracted to obtain three separate wealth indices. The common model wealth index is regressed separately on the urban or rural wealth index for households in those areas, and this regression model is then used to convert each household's (rural or urban) wealth index into a final "composite" wealth index (`wealthscorecont`).
4. **Generate wealth quintiles.** Households are classified into quintiles (`wealthquintile`) using the composite wealth index. In order to account for the complex survey design, the weighted cumulative distribution of the wealth index is used to identify weighted quintile cut-points. Weights represent the normalized household sampling weight (hhwt0 divided by the mean hhwt0 across all households).

Caveats and other considerations. Wealth indices and quintiles derived using this methodology are intended to represent relative measures of wealth as compared to other households in the same country. It is important to note that the underlying Principal Components model simply finds the factors that best capture the variation in the data, and does not guarantee a straightforward interpretation. On average, households in higher wealth quintiles should be more wealthy, but there is considerable uncertainty due to limitations of the available asset data and modeling procedure. Wealth is a complex concept that cannot be captured fully in the model, thus wealth indices should be treated as approximate estimates rather than precise measures.

The value of the wealth index should not be thought of as directly proportional to household wealth, or as being measured along a standard baseline that can be compared between different countries or sub-populations. Relative measures should not be applied to subsets of the population; doing so implicitly assumes that the relative distribution of wealth is similar between the total and subsetted population.

For simplicity and to facilitate replication, variables were not selected differently in urban and rural models on the basis of contextual or subjective knowledge. However, this may not be valid if assets are differentially related to wealth across contexts. Sensitivity analyses excluding variables considered to be context-specific (e.g. livestock) or which scored the most differently in the rural and urban models have typically shown that wealth indices are not sensitive to model specification. Alternative socioeconomic indicators are available and the merits of these alternatives are the subject of ongoing debate.[12,13]

3.6.3. New HIV infections and annual HIV incidence

Rationale. This section summarizes the methods used by PHIA to estimate HIV incidence and the number of new HIV cases occurring per year.

PHIA uses blood test results to determine whether HIV positive people were infected with HIV during a certain time period prior to testing. A specialized estimator is used to convert the number of people infected during this recent infection window into a standardized annual incidence rate. The population at risk is calculated as the weighted number of HIV-negative people, using the survey HIV test results. These two figures are multiplied to obtain an estimate of the number of people newly infected with HIV per year. This section explains the blood tests used, the parameters used for estimation, and other details of the methods used to identify recent HIV infections and calculate annual HIV incidence from PHIA survey data, accounting for the complex survey design.

Definitions. The following definitions are used throughout this memo to explain the properties of the recent infection testing and are also crucial for the estimation of incidence.

- MDRI: Mean Duration of Recent Infection (ω) – the amount of time on average between a person being first seropositive with HIV and the recency test no longer registering them as recently infected. The technical definition in Kassanjee et al. (2012) is "the average time spent both alive and 'recently' infected, within a time T postinfection".[14] The term 'recently' is in quotation marks because it refers to recency derived from the test, rather than true recency; in other words, this acknowledges possible false recent results.
- Cutoff time (T) – a time period that is set with regard to the recency test being used. Ideally T is set such that it is as short as possible while ensuring that as few people as possible test positive for recency after time T post infection.
- PFR: Proportion of false recents (ε) – given a cutoff time T, this is defined as "the probability that a randomly chosen person infected for more than time T will be classified as 'recently' infected by the recency test."

Identification of recently infected people. The PHIA surveys determine whether an HIV-positive person was recently infected through two blood test results: normalized optical density (`lagodnfinal`) from the Limiting-Antigen (LAg) Avidity Enzyme Immunoassay (LAg-avidity EIA), and HIV viral load (`resultvlc`).

The LAg-Avidity blood test was developed as a point-in-time blood test for recent HIV infection.[15,16] It uses a specially created protein which binds with a wide range of HIV antibodies. It takes some time after a new HIV infection for a person's immune system to produce antibodies with high avidity (strength of binding) to these HIV proteins. The test measures the avidity of antibodies to the test protein by allowing antibodies in the blood to bind to a small amount of the protein, then washing away weakly-bound antibodies. The remaining strongly-bound antibodies are chemically stained, and the optical density (ability to absorb light) is measured versus a control specimen. The final measurement is a normalized optical density or ODn.

A person whose measured LAg-Avidity EIA ODn ≤1.5 is classified as recently infected. The measured ODn value from the assay increases over time as an HIV infection progresses, and measurements of specimens with known times of infection show that measured ODn reaches a

value of 1.5 after 130 days since HIV infection on average for subtype C. This characteristic time is called the Mean Duration of Recent Infection (MDRI). PHIA incidence analysis has used MDRI = 130 days (95% CI 118-142 days), as determined by Duong et al.[17], in all surveyed countries to date except Uganda. Uganda has a relatively high proportion of people infected with HIV subtype D, which has a higher MDRI value, so an MDRI value of 153 days (95% CI 127-178 days) was used for incidence estimation there. This value is a weighted average combining various MDRI studies with the measured subtype distribution in the PHIA survey in Uganda. See Kassanjee et al.[18] and Longosz et al.[19] for more details on the effect of HIV subtype on the LAg-EIA MDRI assay.

Kassanjee et al.[20] reported that viral load, a measure of the concentration of HIV virus copies in the blood sample, can be used to help reduce the number of false recents particularly among long-term ART users. For PHIA recency determination, people with a measured ODn ≤ 1.5 must also have a viral load measured at ≥ 1000 copies/mL to be classified as recently infected.

Use of ARV test results. Some people with long term infections and who test positive for antiretroviral drugs (ARVs) can appear to be recently infected using the LAg and VL criteria described above. This can be a result of inadequate adherence to treatment or the development of drug resistance, and has also been observed in adolescents who started ARV during infancy.[21]

Based on these findings, PHIA used an alternative recent infection algorithm that incorporates ARV blood test results. This alternative algorithm reclassifies people who are recently infected according to the LAg and VL criteria as long-term cases if they test positive for ARVs. Although most people reclassified in this way are expected to truly have long-term infections, some false non-recents could be introduced where people have started treatment soon after infection. With more countries starting to implement ‘test and start’ treatment strategies the assumption that recently infected people will not be on treatment may become less reliable over time.

In the PHIA final reports, measures of recent infection using both algorithms are provided. Because of the false recent cases discovered through ARV testing, PHIA recommends the algorithm using the LAg+VL+ARV criteria as our most accurate measure of recent infection.

Proportion of false recents. False recent results can inflate incidence estimates, but it is infeasible to directly estimate the proportion of false recents (PFR) specific to each survey. The approach taken by PHIA to address this challenge is to use the available data from survey participants to minimize the number of false recents at the person level and to set the PFR equal to zero in the estimation stage.

Incidence estimation

The PHIA Project uses the following approach to estimate annual HIV incidence. Because the MDRI is less than one year, adjustments are required to estimate the annual incidence and the number of new infections per year from the raw number of recent infections in the survey data. Kassanjee et al. derived an estimator for instantaneous incidence which can be expressed as

$$I_r = \frac{R - \varepsilon Q}{\left(1 - \frac{\varepsilon T}{\omega}\right)\left(\frac{\omega}{T}\right) N'}$$

where *R* is the number of recent cases, ε is the proportion of false recent cases, *Q* is the number of HIV positive people tested, ω is the MDRI, and *T* is a cutoff time for the assay set at 365 days.[14] *N'* is the adjusted number of HIV negative people in the sample, taking into account the possibility that in practice not all HIV positive people are tested for recency:

$$N' = N\frac{Q}{P}$$

In this equation *N* and *P* are the numbers of negative and positive people in the sample. If all HIV-positive participants were tested for recency, *N'* = *N*. In situations where recency results are unavailable for a certain proportion of HIV-positive participants, the count of HIV-negative people in the sample is scaled down by the same proportion.

As explained in the previous section, we set the proportion of false recent cases $\varepsilon = 0$ for PHIA incidence estimation. This means the equation for instantaneous incidence becomes

$$I_r = \frac{R}{N'} * \frac{T}{\omega}$$

This simplified estimator effectively rates up the number of recent cases as a proportion of the population at risk by a factor of 365/130, or approximately 2.81, to calculate the instantaneous incidence rate. The annual incidence rate is calculated from the instantaneous incidence using

$$I_a = 1 - \exp(-I_r).$$

To obtain our final incidence estimate, we first calculate the number of people in the sample, the number of HIV positive and negative people, and the number of recent cases. The PHIA survey data also allows the number of people at risk (that is, HIV-negative people) to be estimated. This figure is multiplied by the annual incidence estimate to obtain an estimate of the number of new HIV cases per year.

Detailed steps for incidence estimation

SAS and Stata programs have been developed by PHIA and provided to users to calculate both point estimates and confidence intervals for incidence from PHIA survey data (see *Section 4.4 HIV incidence calculators*). To calculate annual incidence, three basic steps are necessary:

1. Use the final blood test weights, HIV status and recency status variables to estimate *R*, *N'*, *Q* and ω (via PROC SURVEYFREQ in SAS).
2. Use the equations above to compute instantaneous incidence (I_r) and then annual incidence rate (I_a).
3. Multiply the annual incidence by the estimated population at risk, that is, the total HIV negative population.

Steps 1-3 are carried out for each age group and gender sub-population required. Confidence intervals are calculated using the formulae in the appendices of Kassanjee et al. (2012).[14] The standard deviation of the MDRI is an important parameter for calculating these confidence intervals. The values of the key parameters used in the estimation are:

Parameter	Value
Cutoff time (*T*)	365 days
MDRI (ω)	130 days
95% CI for ω	118-142 days
Proportion of false recents (ε)	0%

Note that as mentioned earlier, for Uganda the cutoff time and proportion of false recents are the same but a higher MDRI is used: 153 days, with a 95% CI of 127-178 days.

SAS and Stata code used to perform the incidence calculations is included in *Section 4.4 HIV incidence calculators*. The full code for producing PHIA incidence tables is available on request.

Incidence variables included in the PHIA survey datasets

Provided in PHIA datasets are two separate recent infection indicator variables to facilitate the choice of recency algorithm. The `recentlagvlarv` variable has a value of 1 for people determined to be recent infections using the LAg ODn, viral load, and ARV test results, a 2 for long-term infections, and missing where there is insufficient data or the person is not HIV positive. The `recentlagvl` variable has the same set of values, but uses the LAg ODn and viral load results alone.

Accounting for the PHIA survey design in incidence estimation

Use of weights. The estimated incidence (I_r above) depends on the parameters *Q, R*, and *N* measured in the survey population. Using our standard survey blood test weights, we can estimate these population counts at a national level. However, if we use these weighted figures directly in the estimator as if they are counts from a simple random sample we will tend to underestimate the level of error in the estimation. In order to account for unequal final blood test weights while preserving the overall sample size, we normalize the blood test weights by taking the sum of weights divided by the population size in each cell of our incidence table (in this example, cells are strata defined by age group and gender). The *Q, R*, and *N* values that we use in the estimator equation reflect the proportions of positive, recent, and negative status people estimated using our full sample design and calibrated weights, normalized so that they sum to the actual number of people tested in each cell.

Design effect adjustment. In addition to using normalized weights, we also use the survey design effect to adjust the variance in cases where the sample design underperforms the simple random sample assumed by Kassanjee et al.[14] Because of the complexity of the estimator and the assumptions required to derive it, we have not attempted to rigorously calculate the effects of the sampling design on the variance. Instead, we take a conservative approach which aims to avoid giving overly optimistic estimates of precision.

For each gender/age group cell, we estimate the design effect for the proportion of people with recent infections using jackknife replicate weights. When the estimated design effect is greater than one we multiply the variance estimate by the design effect and use this adjusted variance to compute the final confidence interval for the incidence rate estimate. When the design effect is less than one, we set it equal to 1.0 for this variance estimation step.

Estimation of the annual number of new infections. The number of new infections per year is equal to the annual incidence rate multiplied by the at-risk population, i.e. the number of HIV

negative people in the country or sub-population of interest. The most straightforward way to estimate this population is a weighted total of the HIV negative people in the survey population. PHIA blood test weights are adjusted for non-response and post-stratified, so this weighted total will be calibrated appropriately to external population estimates from the national census or official projections. The methods used to calculate and adjust PHIA survey weights are described in detail in each PHIA survey's ***Sampling and Weighting Technical Report***.

Confidence interval calculation for zero cells. The PHIA estimation method is based on assumptions of simple random sampling (SRS) and normally-distributed errors, correcting for the non-SRS sample design using the estimated design effect and weight normalization; however, the assumption of normality is retained. In most PHIA surveys, the total number of participants classified as recently infected was around 30-40, and at least one recently infected person was identified in each required age by gender incidence estimation cell. However, in some PHIA surveys, one or more of the age group by gender estimation cells had no recently infected persons, resulting in a degenerate confidence limit when variance is based on normal approximation. Accordingly, PHIA calculates upper confidence limits based on the Clopper-Pearson binomial confidence interval in these 0 cells.

The Clopper-Pearson upper confidence limit for a proportion when zero successes were observed in n trials is given by

$$1 - \left(\frac{\alpha}{2}\right)^{\frac{1}{n}},$$

where α is the confidence level, set to 0.95.[22] This interval is based on the binomial distribution and is generally conservative. The choice of a conservative estimator is justified given that variance cannot easily be incorporated in the Mean Duration of Recent Infection (MDRI) without using simulations or other computationally intense methods. See Newcombe[23] and Brown et al.[24] for alternative binomial confidence intervals and comparisons.

To apply the Clopper-Pearson equation, the PHIA estimation method uses the weighted number of HIV-negative people in the estimation cell, normalized to the total sample size in the cell (*N'*) as the sample size *n*. This accounts for unequal weights of the sampled people in the calculation. The upper confidence limit (UCL) for the number of recent infections is calculated as

$$r_{\mathrm{UCL}} = N' \times \left[1 - \left(\frac{\alpha}{2}\right)^{\frac{1}{N'}}\right].$$

Finally, this upper limit for the number of recent cases in the cells is substituted into the standard incidence estimation equations

$$I_r = \frac{r}{N'} \times \frac{T}{\omega}$$

and

$$I_a = 1 - \exp\left(-I_r\right),$$

where ω is the MDRI and *T* is the cutoff time (365 days), as used in the Kassanjee incidence estimator. The resulting annual incidence I_a becomes the UCL for incidence in the estimation cell in question.

Application to number of new infections. The number of new infections per year is another estimate of interest reported by PHIA. It is derived directly from the incidence and generally presented only by age group, resulting in fewer occurrences of zero cells; nonetheless, some zero cells have occurred. PHIA uses the UCL for the annual incidence derived above, multiplied by the total weighted HIV-negative population, to derive an upper limit for the number of new infections. To incorporate the variance in the HIV negative population, this upper limit is multiplied by the relative standard error of this estimated population (see *Section 4.4 HIV incidence calculators*).

Other methods and calculators. The South African Centre for Epidemiological Modelling and Analysis (SACEMA) manages the *inctools* R package which is a comprehensive set of tools designed for incidence estimation using tests for recent infection like the LAg assay used in PHIA. *Inctools* also contains other tools, such as methods for estimating the MDRI from study data and for designing incidence surveys. The incidence estimation in *inctools* is based on the same estimator used in the PHIA SAS code for incidence estimation.[14]

Prior to the development of this R package, both SACEMA and CDC developed Excel-based calculators to carry out incidence estimation and to help in the design of incidence surveys (the SACEMA version is available here).

4. Example Code

The following pages contain example SAS, Stata, and R code illustrating how to load in and merge datasets, declare survey designs, and conduct simple survey-weighted analyses using jackknife (JK) replicate weights using Zambia PHIA datasets. Users should take care to change filepaths and filenames in each example to the appropriate names, and tailor analyses to the question of interest. Users should also note where to specify the appropriate number of JK replicate weights for each PHIA survey (see each survey's ***Data Use Manual Supplement***). The examples illustrate six scenarios:

1. Estimate adult HIV prevalence and 90-90-90 indicators (HIV care cascade steps: awareness, ARV use, and viral suppression). Point estimates and 95% confidence intervals are generated for each indicator, conditional on the prior cascade step. Variances are estimated using Jackknife (JK2) and Taylor series methods.
2. Estimate adult HIV care by Hepatitis B status. This analysis requires merging the Hepatitis B variable (`hepb`) from the Adult BIO to Adult IND datasets.
3. Estimate HIV prevalence among adults and children, stratified by wealth quintile and 5-year age groups. This analysis requires appending the Adult and Child BIO datasets, then merging wealth quintile variable (`wealthquintile`) from the HH dataset and 5-year age group variable (`agegroup5population`) from the Adult and Child IND datasets onto it.
4. Estimate HIV prevalence among children by mother's education status. This analysis requires linking mothers' IND data to children using the mother ID variable (`momid`) and merging the mother's education (`education`, renamed as `momeducation`) to the Child BIO dataset.
5. Estimate HIV prevalence in a multi-country analysis using data from Zambia, Zimbabwe and Malawi.
6. Analyses of data from optional modules: Estimate proportion of HIV knowledge module respondents answering all five knowledge questions correctly, and estimate proportion of violence module respondents reporting experiencing any physical or sexual violence from a partner in the past 12 months.

4.1. SAS code examples

SAS code for examples 1-6 are shown below. Note that in examples where datasets are merged, the `(in=a)` option and `if a;` statements specify that the output dataset only retains matched records or unmatched records that were in the master dataset. This ensures that when, for example, merging adult variables onto the biomarker dataset, SAS does not output rows of individuals with no biomarker data.

```
*************************************************************************;
***** Example 1. Estimate 90-90-90 indicators among adults          *****;
*************************************************************************;

*** Load in Adult BIO dataset;
LIBNAME filepath "C:\Desktop\";
DATA adultbio;
      SET filepath.zamphia2016adultbio;
RUN;

*** Conduct analyses using survey weights, JK2;
*   Note REPWEIGHT statement specifies 253 jackknife weights;
*   Check that no. JK weights matches country specifications;
* HIV prevalence;
PROC SURVEYMEANS DATA = adultbio VARMETHOD = jackknife df mean clm nobs;
      WHERE hivstatusfinal ne 99;
      WEIGHT btwt0;
      REPWEIGHT btwt001-btwt253 / JKCOEFS = 1 DF = 25;
      CLASS hivstatusfinal;
      VAR hivstatusfinal;
RUN;
* HIV awareness, conditional on HIV positive status;
PROC SURVEYMEANS DATA = adultbio VARMETHOD = jackknife df mean clm nobs;
      WHERE tri90 = 1 AND hivstatusfinal = 1 AND aware ne 99;
      WEIGHT btwt0;
      REPWEIGHT btwt001-btwt253 / JKCOEFS = 1 DF = 25;
      CLASS aware;
      VAR aware;
RUN;
* ARV use, conditional on awareness;
PROC SURVEYMEANS DATA = adultbio VARMETHOD = jackknife df mean clm nobs;
      WHERE tri90 = 1 AND aware = 1 AND art ne 99;
      WEIGHT btwt0;
      REPWEIGHT btwt001-btwt253 / JKCOEFS = 1 DF = 25;
      CLASS art;
      VAR art;
RUN;
* Viral suppression, conditional on ARV use;
PROC SURVEYMEANS DATA = adultbio VARMETHOD = jackknife df mean clm nobs;
      WHERE tri90 = 1 AND aware = 1 AND art = 1 AND vls ne 99;
      WEIGHT btwt0;
      REPWEIGHT btwt001-btwt253 / JKCOEFS = 1 DF = 25;
      CLASS vls;
      VAR vls;
RUN;


*** Conduct analyses using survey weights, Taylor series linearization;
```

```
*   Note strata and cluster specifications;
* HIV prevalence;
PROC SURVEYMEANS DATA = adultbio VARMETHOD = taylor mean clm nobs;
      WHERE hivstatusfinal ne 99;
      WEIGHT btwt0;
      STRATA varstrat;
      CLUSTER varunit;
      CLASS hivstatusfinal;
      VAR hivstatusfinal;
RUN;
* HIV awareness, conditional on HIV positive status;
PROC SURVEYMEANS DATA = adultbio VARMETHOD = taylor mean clm nobs;
      WHERE tri90 = 1 AND hivstatusfinal = 1 AND aware ne 99;
      WEIGHT btwt0;
      STRATA varstrat;
      CLUSTER varunit;
      CLASS aware;
      VAR aware;
RUN;
* ARV use, conditional on awareness;
PROC SURVEYMEANS DATA = adultbio VARMETHOD = taylor mean clm nobs;
      WHERE tri90 = 1 AND aware = 1 AND art ne 99;
      WEIGHT btwt0;
      STRATA varstrat;
      CLUSTER varunit;
      CLASS art;
      VAR art;
RUN;
* Viral suppression, conditional on ARV use;
PROC SURVEYMEANS DATA = adultbio VARMETHOD = taylor mean clm nobs;
      WHERE tri90 = 1 AND aware = 1 AND art = 1 AND vls ne 99;
      WEIGHT btwt0;
      STRATA varstrat;
      CLUSTER varunit;
      CLASS vls;
      VAR vls;
RUN;
```

```
*************************************************************************;
***** Example 2. Estimate adult HIV care by Hep B status            *****;
*****    Note: Hep B testing was not performed in all PHIA surveys  *****;
*****          hepb variable used for demonstration purposes only   *****;
*************************************************************************;

*** Load in Adult IND dataset;
LIBNAME filepath "C:\Desktop\";
DATA adultind;
      SET   filepath.zamphia2016adultind;
RUN;
DATA adultbio;
      SET filepath.zamphia2016adultbio;
RUN;

* Sort before merge;
PROC SORT DATA = adultbio;
      BY personid;
RUN;
PROC SORT DATA = adultind;
      BY personid;
RUN;

* Merge Hep B variable from Adult BIO dataset;
DATA ind;
      MERGE adultind
            adultbio (keep=personid hepb);
      BY personid;
RUN;

*** Conduct analyses using survey weights, JK2;
*   Note REPWEIGHT statement specifies 253 jackknife weights;
*   Check that no. JK weights matches country specifications;
* HIV care by Hep B status;
PROC SURVEYMEANS DATA = ind VARMETHOD = jackknife df mean clm nobs;
      DOMAIN hepb;
      WHERE hivcare IN (1,2);
      WEIGHT intwt0;
      REPWEIGHT intwt001-intwt253 / JKCOEFS = 1 DF = 25;
      CLASS hivcare;
      VAR hivcare;
RUN;
```

```
*************************************************************************;
***** Example 3. Estimate HIV prevalence among adults & children    *****;
*****                  by 5-year age groups and wealth quintiles     *****;
*************************************************************************;

*** Load in Adult and Child BIO datasets;
LIBNAME filepath "C:\Desktop\";
DATA bio;
      SET   filepath.zamphia2016adultbio
            filepath.zamphia2016childbio;
RUN;
DATA ind;
      SET   filepath.zamphia2016adultind
            filepath.zamphia2016childind;
RUN;
DATA hh;
      SET   filepath.zamphia2016hh;
RUN;

* Sort before merge;
PROC SORT DATA = bio;
      BY householdid;
RUN;
PROC SORT DATA = hh;
      BY householdid;
RUN;


* Merge wealth quintile variable from HH dataset;
DATA bio;
      MERGE bio
            hh (keep=householdid wealthquintile);
      BY householdid;
RUN;

* Sort before merge;
PROC SORT DATA = bio;
      BY personid;
RUN;
PROC SORT DATA = ind;
      BY personid;
RUN;

* Merge agegroup5population variable from Adult and Child IND datasets;
DATA bio;
      MERGE bio (in=a)
            ind (keep=personid agegroup5population);
      BY personid;
      IF a;
RUN;

*** Conduct analyses using survey weights, JK2;
*   Note REPWEIGHT statement specifies 253 jackknife weights;
*   Check that no. JK weights matches country specifications;
* HIV prevalence by wealth quintile;
PROC SURVEYMEANS DATA = bio VARMETHOD = jackknife df mean clm nobs;
      DOMAIN wealthquintile;
```

```
      WHERE hivstatusfinal ne 99 AND wealthquintile ne 99;
      WEIGHT btwt0;
      REPWEIGHT btwt001-btwt253 / JKCOEFS = 1 DF = 25;
      CLASS hivstatusfinal;
      VAR hivstatusfinal;
RUN;
* HIV prevalence by 5-year age groups;
PROC SURVEYMEANS DATA = bio VARMETHOD = jackknife df mean clm nobs;
      DOMAIN agegroup5population;
      WHERE hivstatusfinal ne 99 AND agegroup5population ne 99;
      WEIGHT btwt0;
      REPWEIGHT btwt001-btwt253 / JKCOEFS = 1 DF = 25;
      CLASS hivstatusfinal;
      VAR hivstatusfinal;
RUN;
```

```
*************************************************************************;
***** Example 4. Estimate child HIV prevalence by mother's education *****;
*************************************************************************;

*** Load in Child BIO dataset;
LIBNAME filepath "C:\Desktop\";
DATA childbio;
      SET filepath.zamphia2016childbio;
RUN;

* Prepare mother dataset for merging using Adult IND dataset;
DATA momind;
      SET filepath.zamphia2016adultind (keep=personid education);
      RENAME personid = momid;
      RENAME education = momeducation;
RUN;

* Sort before merge;
PROC SORT DATA = childbio;
      BY momid;
RUN;

PROC SORT DATA = momind;
      BY momid;
RUN;

* Merge mother's education variable from Adult IND dataset;
DATA childbio;
      MERGE childbio (in=a)
            momind;
      BY momid;
      IF a;
RUN;

*** Conduct analyses using survey weights, JK2;
*   Note REPWEIGHT statement specifies 253 jackknife weights;
*   Check that no. JK weights matches country specifications;
* Child's HIV status by mother's education;
PROC SURVEYMEANS DATA = childbio VARMETHOD = jackknife df mean clm nobs;
      DOMAIN momeducation;
      WHERE hivstatusfinal ne 99 AND momeducation ne 99;
      WEIGHT btwt0;
      REPWEIGHT btwt001-btwt253 / JKCOEFS = 1 DF = 25;
      CLASS hivstatusfinal;
      VAR hivstatusfinal;
RUN;
```

```
*************************************************************************;
***** Example 5. Estimate HIV prevalence in multi-country analysis   *****;
*************************************************************************;

*** Load in Adult BIO datasets for 3 countries;
LIBNAME filepath "C:\Desktop\";
DATA zamadultbio;
      SET filepath.zamphia2016adultbio;
RUN;
DATA zimadultbio;
      SET filepath.zimphia2015adultbio;
RUN;
DATA maladultbio;
      SET filepath.mphia2015adultbio;
RUN;

* Combine countries;
DATA combined;
      SET   zamadultbio
            zimadultbio
            maladultbio;
      KEEP country hivstatusfinal btwt:;
RUN;

* Generate replicate weights for each country;
PROC SORT data=combined out=combined;
      BY country;
RUN;

DATA combined;
      SET combined;

      * Declare arrays for each set of weights;
      ARRAY zam btwt001-btwt253;
      ARRAY zim btwt001-btwt248;
      ARRAY mal btwt001-btwt250;

      ARRAY newzam b2wt0001-b2wt0253;
      ARRAY newzim b2wt0254-b2wt0501;
      ARRAY newmal b2wt0502-b2wt0751;

      ARRAY newall b2wt0001-b2wt0751;

      * Set weights to original JK replicate weights within each country;
      IF country="Zambia"     THEN DO OVER zam; newzam=zam; END;
      IF country="Zimbabwe"   THEN DO OVER zim; newzim=zim; END;
      IF country="Malawi"     THEN DO OVER mal; newmal=mal; END;
      * Set remaining weights to base weight;
      DO OVER newall; IF newall=. THEN newall=btwt0; END;

      LABEL country = "Country";
RUN;

*** Conduct analyses using survey weights, JK2;
*   Note REPWEIGHT statement specifies combined 751 jackknife weights;
* HIV prevalence in 3 countries;
PROC SURVEYMEANS DATA = combined VARMETHOD = jackknife df mean clm nobs;
```

```
      WHERE hivstatusfinal ne 99;
      WEIGHT btwt0;
      REPWEIGHT b2wt0001-b2wt0751 / JKCOEFS = 1 DF = 25;
      CLASS hivstatusfinal;
      VAR hivstatusfinal;
RUN;
```

```
*************************************************************************;
***** Example 6. Analyses of data from optional modules             *****;
*****     (HIV knowledge and violence)                              *****;
*************************************************************************;

*** Load in Adult BIO dataset;
LIBNAME filepath "C:\Desktop\";
DATA adultind;
      SET filepath.zamphia2016adultind;
      *** Code new outcomes;
      * All knowledge questions correct;
      IF hivk_status = 1 AND onepartnr = 1 AND mosquito = 1 AND condoms = 1
AND sharefood = 1 AND healthyinf = 1 THEN knowledge = 1;
      ELSE IF hivk_status = 1 THEN knowledge = 0;
      * Sexual or physical violence by partner in last 12 months;
      IF vmflag = 1 AND (sexualviolencepart12mo = 1 OR
physicalviolencepart12mo = 1) THEN violencepart12mo = 1;
      ELSE IF vmflag = 1 THEN violencepart12mo = 0;
RUN;

*** Conduct analyses using survey weights, JK2;
*   Note REPWEIGHT statement specifies 253 jackknife weights;
*   Check that no. JK weights matches country specifications;
* Proportion of HIV knowledge module respondents answering all 5 questions
correctly;
PROC SURVEYMEANS DATA = adultind VARMETHOD = jackknife df mean clm nobs;
      WHERE hivk_status = 1;
      WEIGHT hivkpswt0;
      REPWEIGHT hivkpswt1-hivkpswt253 / JKCOEFS = 1 DF = 25;
      CLASS knowledge;
      VAR knowledge;
RUN;

*** Conduct analyses using survey weights, JK2;
*   Note REPWEIGHT statement specifies 253 jackknife weights;
*   Check that no. JK weights matches country specifications;
* Proportion of violence module respondents experiencing either sexual or
physical violence from a partner in the past 12 months;
PROC SURVEYMEANS DATA = adultind VARMETHOD = jackknife df mean clm nobs;
      WHERE vmflag = 1;
      WEIGHT vmpstw0;
      REPWEIGHT vmpstw1-vmpstw253 / JKCOEFS = 1 DF = 25;
      CLASS violencepart12mo;
      VAR violencepart12mo;
RUN;
```

4.2. Stata code examples

Stata code for examples 1-6 are shown below. Note that Stata survey design and jackknife estimation procedures are declared prior to analysis using the `svyset` command. Note that Stata does not provide detailed error messages on procedures when run using the `svy` prefix for weighted analyses. Users are advised to test analyses without the `svy` specification for errors before running survey estimation. For example, yes/no variables are coded 1=yes and 2=no, but logistic regression using the `logit` command requires the outcome variable to be coded as 1=yes and 0=no; if not, Stata will produce unspecified errors if using survey estimation procedures (e.g., `svy: logistic hivstatusfinal if hivstatusfinal!=99`).

In the merge commands, the `keep(match master)` option specifies that the output dataset only retains matched records or unmatched records that were in the master dataset (the dataset in memory). This ensures that when, for example, merging adult variables onto the biomarker dataset, Stata does not output rows of individuals with no biomarker data. Also note that merge commands must explicitly specify if the matching function is `1:1` (one-to-one) or `m:1` (many-to-one).

```
*********************************************************************
***** Example 1. Estimate 90-90-90 indicators among adults          *****
*********************************************************************

*** Load in Adult BIO dataset
use "C:\Desktop\zamphia2016adultbio.dta", clear

*** Declare survey design, JK2
*     Note jkrweight option specifies 253 JK weights;
*     Check that no. JK weights matches country specifications;
svyset [pw=btwt0], jkrweight(btwt001-btwt253, multiplier(1)) vce(jackknife) dof(25)

*** Conduct analyses using survey weights, JK2
* HIV prevalence
svy: tab hivstatusfinal if hivstatusfinal!=99, se ci obs format(%8.3g)
* HIV awareness, conditional on HIV positive status
svy: tab aware if tri90==1 & hivstatusfinal==1 & aware!=99, se ci obs format(%8.3g)
* ARV use, conditional on awareness
svy: tab art if tri90==1 & aware==1 & art!=99, se ci obs format(%8.3g)
* Viral suppression, conditional on ARV use
svy: tab vls if tri90==1 & aware==1 & art==1 & vls!=99, se ci obs format(%8.3g)

*** Declare survey design, Taylor series linearization
*     Note jkrweight option specifies 253 JK weights;
*     Check that no. JK weights matches country specifications;
svyset varunit [pw=btwt0], strata(varstrat) vce(linearized) singleunit(scaled)

*** Conduct analyses using survey weights, Taylor series linearization
* HIV prevalence
svy: tab hivstatusfinal if hivstatusfinal!=99, se ci obs format(%8.3g)
* HIV awareness, conditional on HIV positive status
svy: tab aware if tri90==1 & hivstatusfinal==1 & aware!=99, se ci obs format(%8.3g)
* ARV use, conditional on awareness
svy: tab art if tri90==1 & aware==1 & art!=99, se ci obs format(%8.3g)
* Viral suppression, conditional on ARV use
svy: tab vls if tri90==1 & aware==1 & art==1 & vls!=99, se ci obs format(%8.3g)
```

```
*************************************************************************
***** Example 2. Estimate adult HIV care by Hep B status            *****
*****    Note: Hep B testing was not performed in all PHIA surveys  *****
*****          hepb variable used for demonstration purposes only   *****
*************************************************************************

*** Load in Adult IND dataset
use "C:\Desktop\zamphia2016adultind.dta", clear

*** Merge Hep B variable from Adult BIO dataset
merge m:1 personid using "C:\Desktop\zamphia2016adultbio.dta", keep(match master)
    keepusing(hepb)

*** Declare survey design
*     Note jkrweight option specifies 253 JK weights;
*     Check that no. JK weights matches country specifications;
svyset [pw=intwt0], jkrweight(intwt001-intwt253, multiplier(1)) vce(jackknife)
    dof(25)

*** Conduct analyses using survey weights
*     Note for bivariate analyses, outcome is specified first
*     "col" option obtains column proportions

* HIV care by Hep B status
svy: tab hivcare hepb if inlist(hivcare,1,2) & inlist(hepb, 1,2), se ci col obs
    format(%8.3g)
```

```
*************************************************************************
***** Example 3. Estimate HIV prevalence among adults & children    *****
*****                  by 5-year age groups and wealth quintiles     *****
*************************************************************************

*** Prepare combined adult and child IND dataset for merging
use "C:\Desktop\zamphia2016adultind.dta", clear
append using "C:\Desktop\zamphia2016childind.dta"
save "C:\Desktop\zamphia2016ind.dta", replace

*** Prepare combined adult and child BIO dataset for merging
use "C:\Desktop\zamphia2016adultbio.dta", clear
append using "C:\Desktop\zamphia2016childbio.dta"

*** Merge wealth quintile variable from HH dataset
merge m:1 householdid using "C:\Desktop\zamphia2016hh.dta", keep(match master)
   keepusing(wealthquintile) nogen

*** Merge agegroup5population variable from Adult and Child IND datasets
merge m:1 personid using "C:\Desktop\zamphia2016ind.dta", keep(match master)
   keepusing(agegroup5population) nogen

*** Declare survey design
*     Note jkrweight option specifies 253 JK weights;
*     Check that no. JK weights matches country specifications;
svyset [pw=btwt0], jkrweight(btwt001-btwt253, multiplier(1)) vce(jackknife) dof(25)

*** Conduct analyses using survey weights
*     Note for bivariate analyses, outcome is specified first
*     "col" option obtains column proportions

* HIV prevalence by wealth quintile
svy: tab hivstatusfinal wealthquintile if hivstatusfinal!=99 & wealthquintile!=99,
   se ci col obs format(%8.3g)
* HIV prevalence by 5-year age groups
svy: tab hivstatusfinal agegroup5population if hivstatusfinal!=99 &
agegroup5population!=99, se ci col obs format(%8.3g)
```

```
*******************************************************************************
***** Example 4. Estimate child HIV prevalence by mother's education          *****
*******************************************************************************

*** Prepare mother dataset for merging using Adult IND dataset
use "C:\Desktop\zamphia2016adultind.dta", clear
rename personid momid
rename education momeducation
label var momeducation "Mother's education"
save "C:\Desktop\zamphia2016momind.dta", replace

*** Load in Child BIO dataset
use "C:\Desktop\zamphia2016childbio.dta", clear

*** Merge mother's education variable from Adult IND dataset
merge m:1 momid using "C:\Desktop\zamphia2016momind.dta", keep(match master)
    keepusing(momeducation)

*** Declare survey design
*     Note for bivariate analyses, outcome is specified first
*     "col" option obtains column proportions
svyset [pw=btwt0], jkrweight(btwt001-btwt253, multiplier(1)) vce(jackknife) dof(25)

*** Conduct analyses using survey weights
*     Note for bivariate analyses, outcome is specified first
*     "col" option obtains column proportions

* Child's HIV status by mother's education
svy: tab hivstatusfinal momeducation if hivstatusfinal!=99 & momeducation !=99, se
    ci col obs format(%8.3g)
```

```
*******************************************************************************
***** Example 5. Estimate HIV prevalence in multi-country analysis            *****
*******************************************************************************

*** Combine countries
use "C:\Desktop\zamphia2016adultbio.dta", clear
append using "C:\Desktop\zimphia2015adultbio.dta"
append using "C:\Desktop\mphia2015adultbio.dta"

*** Generate replicate weights for each country
* Specify number replicate weights per country
local zamnwts = 253
local zimnwts = 248
local malnwts = 250
local totwts = `zamnwts'+`zimnwts'+`malnwts'
* Set weights to original JK replicate weights within each country
forvalues n = 1/`totwts' {
       local newsuffix = string(`n',"%04.0f")
       gen b2wt`newsuffix' = .
}
* Set weights to original JK replicate weights within each country
forvalues n = 1/`zamnwts' {
       local oldsuffix = string(`n',"%03.0f")
       local newsuffix = string(`n',"%04.0f")
       replace b2wt`newsuffix' = btwt`oldsuffix' if country=="Zambia"
}
forvalues n = 1/`zimnwts' {
       local oldsuffix = string(`n',"%03.0f")
       local newsuffix = string(`n'+`zimnwts',"%04.0f")
       replace b2wt`newsuffix' = btwt`oldsuffix' if country=="Zimbabwe"
}
forvalues n = 1/`malnwts' {
       local oldsuffix = string(`n',"%03.0f")
       local newsuffix = string(`n'+`zamnwts'+`zimnwts',"%04.0f")
       replace b2wt`newsuffix' = btwt`oldsuffix' if country=="Malawi"
}
* Set remaining weights to base weight
forvalues n = 1/`totwts'{
       local newsuffix = string(`n',"%04.0f")
       replace b2wt`newsuffix' = btwt0 if b2wt`newsuffix'==.
}

*** Declare survey design
*     Note for bivariate analyses, outcome is specified first
*     "col" option obtains column proportions
local newtotwts = string(`totwts',"%04.0f")
svyset [pw=btwt0], jkrweight(b2wt0001-b2wt`newtotwts', multiplier(1)) vce(jackknife)
   dof(25)

*** Conduct analyses using survey weights
* HIV prevalence in 3 countries
svy: tab hivstatusfinal if hivstatusfinal!=99, se ci col obs format(%8.3g)
```

```
*******************************************************************************
***** Example 6. Analyses of data from optional modules                   *****
***** (HIV knowledge and violence)                                        *****
*******************************************************************************

*** Load in Adult IND dataset
use "C:\Desktop\zamphia2016adultind.dta", clear

*** Code new outcomes
* All knowledge questions correct
gen knowledge = 1 if hivk_status==1 & (onepartnr==1 & mosquito==1 & condoms==1 &
    sharefood==1 & healthyinf==1)
replace knowledge = 0 if hivk_status==1 & (onepartnr!=1 | mosquito!=1 | condoms!=1 |
    sharefood!=1 | healthyinf! =1)

* Sexual or physical violence by partner in last 12 months
gen violencepart12mo = 1 if vmflag==1 & (sexualviolencepart12mo==1 |
    physicalviolencepart12mo==1)
replace violencepart12mo = 0 if vmflag==1 & (sexualviolencepart12mo!=1 &
    physicalviolencepart12mo!=1)

*** Declare survey design, JK2
*     Note jkrweight option specifies 253 JK weights;
*     Check that no. JK weights matches country specifications;
svyset [pw=hivkpswt0], jkrweight(hivkpswt1-hivkpswt253, multiplier(1))
    vce(jackknife) dof(25)

*** Conduct analyses using survey weights, JK2
* Proportion of HIV knowledge module respondents answering all 5 questions correctly
svy: tab knowledge, se ci obs format(%8.3g)

*** Declare survey design, JK2
*     Note jkrweight option specifies 253 JK weights;
*     Check that no. JK weights matches country specifications;
svyset [pw=vmpstw0], jkrweight(vmpstw1-vmpstw253, multiplier(1)) vce(jackknife)
    dof(25)

*** Conduct analyses using survey weights, JK2
* Proportion of violence module respondents experiencing either sexual or physical
    violence from a partner in the past 12 months
svy: tab violencepart12mo, se ci obs format(%8.3g)
```

4.3. R code examples

R code for examples 1-6 are shown below (using .csv files). Note that R survey design and jackknife estimation procedures are declared prior to analysis by generating a `survey` object. Columns from the original dataset are separated into analytic variables (`variables`), base weights (`weights`), and JK replicate weight (`repweights`) components. In Taylor Series variance estimation, strata (`strata`) and PSUs (`ids`) are declared.

```
################################################################
### Example 1. Estimate 90-90-90 indicators among adults ###
################################################################

### Install and load survey analysis and dplyr packages
memory.limit(size=56000)
#install.packages("survey")
library(survey)
#install.packages("dplyr")
library(dplyr)
#install.packages("foreign")
library(foreign)

### Load in Adult BIO dataset
adultbio <- read.csv('C:/Desktop/zamphia2016adultbio.csv')

### Ensure that weight variables are converted to numeric type, and drop
### records with no blood weight
wtname <- 'btwt'
adultbio <- adultbio %>%
  filter(bt_status == 1) %>%
  mutate(across(contains(wtname), as.double))

### Create survey object, JK2
# Recode analytic outcomes to 0/1
vars <- c('hivstatusfinal', 'aware', 'art', 'vls', 'tri90')
svydata <- adultbio
for(i in 1:length(vars)){
  varname <- vars[i]
  if(is.factor(svydata[,varname])) {
    svydata[,varname] <- as.numeric(levels(svydata[,varname]))[svydata[,varname]]
    # coerces factor variables to numeric before recoding
  }
  svydata[,varname][(svydata[,varname]==2)] <- 0
}

# Extract analytic variables, base weights and JK replicate weights
#   and create survey object,
#   specifying variance estimation method "JK" and JK coefficients=1
svydata1 <- svrepdesign(
  variables=svydata[,vars],
  weights=svydata[,paste0(wtname,0)],
  repweights=svydata[,grep(wtname,names(svydata))][,-1],
  type="JKn", scale=1, rscales=1)

### Conduct analyses using survey weights
# HIV prevalence
res <- svyciprop(~I(hivstatusfinal==1),
                 design=svydata1,
                 method="mean",level=0.95,df=25)
(c(res[[1]],attr(res,"ci"))) # display results
table(svydata1$variables$hivstatusfinal)
```

```
# HIV awareness
svydata1_aware <- subset(svydata1, tri90 == 1 & hivstatusfinal != 99 & aware != 99)
res <- svyciprop(formula=~I(aware==1),
                 design= svydata1_aware,
                 method="mean",level=0.95,df=25)
(c(res[[1]],attr(res,"ci"))) # display results
table(svydata1_aware$variables$aware)

# ARV use
svydata1_art <- subset(svydata1, tri90 == 1 & aware == 1 & art != 99)
res <- svyciprop(formula=~I(art==1),
                 design= svydata1_art,
                 method="mean",level=0.95,df=25)
(c(res[[1]],attr(res,"ci"))) # display results
table(svydata1_art$variables$art)

# Viral suppression
svydata1_vls <- subset(svydata1, tri90 == 1 & aware == 1 & art == 1 & vls != 99)
res <- svyciprop(formula=~I(vls==1),
                 design=svydata1_vls,
                 method="mean",level=0.95,df=25)
(c(res[[1]],attr(res,"ci"))) # display results
table(svydata1_vls$variables$vls)

### Create survey object, Taylor series linearization
# Recode analytic outcomes to 0/1
vars <- c('hivstatusfinal', 'aware', 'art', 'vls', 'tri90')
wtname <- 'btwt'
strataname <- 'varstrat'
clustername <- 'varunit'
svydata <- adultbio
for(i in 1:length(vars)){
  varname <- vars[i]
  if(is.factor(svydata[,varname])) {
    svydata[,varname] <- as.numeric(levels(svydata[,varname]))[svydata[,varname]]
    # coerces factor variables to numeric before recoding
  }
  svydata[,varname][(svydata[,varname]==2)] <- 0
}

# Extract analytic variables, base weights and variance unit/strata variables
#   and create survey object
svydata1 <- svydesign(
  variables=svydata[,vars],
  weights=svydata[,paste0(wtname,0)],
  strata=svydata[,strataname],
  ids=~svydata[,clustername],
  nest=TRUE)

### Conduct analyses using survey weights
# HIV prevalence
res <- svyciprop(~I(hivstatusfinal==1),
                 design=svydata1,
                 method="mean",level=0.95,df=25)
(c(res[[1]],attr(res,"ci"))) # display results
table(svydata1$variables$hivstatusfinal)

# HIV awareness
sbydata1_aware <- subset(svydata1, tri90 == 1 & hivstatusfinal != 99 & aware != 99)
res <- svyciprop(formula=~I(aware==1),
                 design=svydata1_aware,
                 method="mean",level=0.95,df=25)
(c(res[[1]],attr(res,"ci"))) # display results
```

```r
table(svydata1_aware$variables$aware)

# ARV use
svydata1_art <- subset(svydata1, tri90 == 1 & aware == 1 & art != 99)
res <- svyciprop(formula=~I(art==1),
                 design=svydata1_art,
                 method="mean",level=0.95,df=25)
(c(res[[1]],attr(res,"ci"))) # display results
table(svydata1_art$variables$art)

# Viral suppression
svydata1_vls <- subset(svydata1, tri90 == 1 & aware == 1 & art == 1 & vls != 99)
res <- svyciprop(formula=~I(vls==1),
                 design=svydata1_vls,
                 method="mean",level=0.95,df=25)
(c(res[[1]],attr(res,"ci"))) # display results
table(svydata1_vls$variables$vls)
```

```r
############################################################
### Example 2. Estimate adult HIV care by Hep B status ###
###    Note: Hep B testing was not performed in all    ###
###       PHIA surveys. hepb used for demonstration    ###
###       purposes only                                ###
############################################################

### Load in Adult IND dataset
adultind <- read.csv('C:/Desktop/zamphia2016adultind.csv')
### Merge Hep B variable from Adult BIO dataset
adultbio <- read.csv('C:/Desktop/zamphia2016adultbio.csv')
adultbio <- adultbio[,c('personid','hepb')]
# Merge personid and hepb, all.x=TRUE option retains unmatched rows from master
adult <- merge(adultind, adultbio, by='personid', all.x = TRUE)

### Ensure that weight variables are converted to numeric type, and drop
### records with no interview weight
wtname <- 'intwt'
adult <- adult %>%
  filter(indstatus == 1) %>%
  mutate(across(contains(wtname), as.double))

### Create survey object
# Recode analytic variables to 0/1
vars <- c('hivcare', 'hepb')
svydata <- adult
for(i in 1:length(vars)){
  varname <- vars[i]
  if(is.factor(svydata[,varname])) {
    svydata[,varname] <- as.numeric(levels(svydata[,varname]))[svydata[,varname]]
    # coerces factor variables to numeric before recoding
  }
  else {
    svydata[,varname] <- as.numeric(svydata[,varname])
  }
  svydata[,varname][(svydata[,varname]==2)] <- 0
}

# Extract analytic variables, base weights and JK replicate weights
#   and create survey object,
#   specifying variance estimation method "JK" and JK coefficients=1
svydata1 <- svrepdesign(
  variables=svydata[,vars],
  weights=svydata[,paste0(wtname,0)],
  repweights=svydata[,grep(wtname,names(svydata))][,-1],
  type="JKn", scale=1, rscales=1)

### Conduct analyses using survey weights
# HIV care by Hep B status
svydata1_hepb <- subset(svydata1, hivcare %in% c(0, 1))
svyby(formula=~I(hivcare==1),by=~hepb,
      design=svydata1_hepb,
      FUN=svyciprop,vartype="ci",method="beta",df=25)
table(svydata1_hepb$variables$hepb,
      svydata1_hepb$variables$hivcare)
```

```r
#########################################################################
### Example 3. Estimate HIV prevalence among adults & children ###
###               by 5-year age groups and wealth quintiles    ###
#########################################################################

### Load in and merge Adult and Child BIO datasets
adultbio <- read.csv('C:/Desktop/zamphia2016adultbio.csv')
childbio <- read.csv('C:/Desktop/zamphia2016childbio.csv')
bio <- bind_rows(adultbio,childbio)

### Merge wealth quintile variable from HH dataset
hh <- read.csv('C:/Desktop/zamphia2016hh.csv')
# Select columns
hh2 <- select(hh, householdid, wealthquintile)
bio2 <- select(bio, householdid, personid, hivstatusfinal, bt_status,
starts_with("btwt"))
# Merge
bio3 <- merge(bio2,hh2,by='householdid')

### Merge agegroup5population variable from Adult and Child IND datasets
adultind <- read.csv('C:/Desktop/zamphia2016adultind.csv')
childind <- read.csv('C:/Desktop/zamphia2016childind.csv')


# Select columns
indvars <- c("personid", "agegroup5population", "country")
ind <- bind_rows(select(adultind, all_of(indvars)),
                 select(childind, all_of(indvars)))
data <- merge(bio3, ind, by='personid', all.x=TRUE)

### Create survey object
# Recode analytic variables to 0/1
vars <- c('hivstatusfinal','wealthquintile','agegroup5population')
wtname <- 'btwt'
svydata <- data
for(i in 1:length(vars)){
  varname <- vars[i]
  if(is.factor(svydata[,varname])) {
    svydata[,varname] <- as.numeric(levels(svydata[,varname]))[svydata[,varname]]
    # coerces factor variables to numeric before recoding
  }
  svydata[,varname][(svydata[,varname]==2)] <- 0
}

### Ensure that weight variables are converted to numeric type, and drop
### records with no blood weight
svydata <- svydata %>%
  filter(bt_status == 1) %>%
  mutate(across(contains(wtname), as.double))

# Extract analytic variables, base weights and JK replicate weights
#   and create survey object,
#   specifying variance estimation method "JK" and JK coefficients=1
svydata1 <- svrepdesign(
  variables=svydata[,vars],
  weights=svydata[,paste0(wtname,0)],
  repweights=svydata[,grep(wtname,names(svydata))][,-1],
  type="JKn", scale=1, rscales=1)

### Conduct analyses using survey weights
# HIV prevalence by wealth quintile
```

```r
#svydata1_hivbywealth <- svydata1[!(svydata1$variables$wealthquintile %in%
c(99,NA))]
svydata1_hivbywealth <- subset(svydata1, !(wealthquintile %in% c(99, NA)))
svyby(formula=~I(hivstatusfinal==1),by=~wealthquintile,
      design=svydata1_hivbywealth, FUN=svyciprop,vartype="ci",method="beta",df=25)
table(svydata1_hivbywealth$variables$wealthquintile,
      svydata1_hivbywealth$variables$hivstatusfinal)
# HIV prevalence by 5-year age groups
svyby(formula=~I(hivstatusfinal==1),by=~agegroup5population,
      design=svydata1_hivbywealth, FUN=svyciprop,vartype="ci",method="beta",df=25)
table(svydata1_hivbywealth$variables$agegroup5population,
      svydata1_hivbywealth$variables$hivstatusfinal)
```

```
#############################################################################
### Example 4. Estimate child HIV prevalence by mother's education ###
#############################################################################

### Prepare mother dataset for merging using Adult IND dataset
adultind <- read.csv('C:/Desktop/zamphia2016adultind.csv')
momind <- adultind[,c('personid','education')]
# Rename mom's variables
names(momind)[names(momind)=='personid'] <- 'momid'
names(momind)[names(momind)=='education'] <- 'momeducation'
### Load in Child BIO dataset
childbio <- read.csv('C:/Desktop/zamphia2016childbio.csv')

### Merge mother's education variable from Adult IND dataset
childbio2 <- merge(childbio,momind,by='momid')

### Create survey object
# Recode analytic variables to 0/1
vars <- c('hivstatusfinal','momid','momeducation')
wtname <- 'btwt'
svydata <- childbio2
for(i in 1:length(vars)){
  varname <- vars[i]
  if(is.factor(svydata[,varname])) {
    svydata[,varname] <- as.numeric(levels(svydata[,varname]))[svydata[,varname]]
    # coerces factor variables to numeric before recoding
  }
  svydata[,varname][(svydata[,varname]==2)] <- 0
}

### Ensure that weight variables are converted to numeric type, and drop
### records with no blood weight
svydata <- svydata %>%
  filter(bt_status == 1) %>%
  mutate(across(contains(wtname), as.double))

# Extract analytic variables, base weights and JK replicate weights
#   and create survey object,
#   specifying variance estimation method "JK" and JK coefficients=1
svydata1 <- svrepdesign(
  variables=svydata[,vars],
  weights=svydata[,paste0(wtname,0)],
  repweights=svydata[,grep(wtname,names(svydata))][,-1],
  type="JKn", scale=1, rscales=1)

### Conduct analyses using survey weights
# Child's HIV status by mother's education
svydata1_childhiv <- subset(svydata1, !(momeducation %in% c(99, NA)))
svyby(formula=~I(hivstatusfinal==1),by=~momeducation,
      design=svydata1_childhiv,
      FUN=svyciprop,vartype="ci",method="beta",df=25)
table(svydata1_childhiv$variables$momeducation,
      svydata1_childhiv$variables$hivstatusfinal)
```

```
#########################################################################
### Example 5. Estimate HIV prevalence in multi-country analysis   ###
#########################################################################

### Combine countries
zambio <- read.csv('C:/Desktop/zamphia2016adultbio.csv')
zimbio <- read.csv('C:/Desktop/zimphia2015adultbio.csv')
malbio <- read.csv('C:/Desktop/mphia2015adultbio.csv')

zambio2 <-
cbind(zambio[,c('country','hivstatusfinal')],zambio[,grep('btwt',names(zambio))])
zimbio2 <-
cbind(zimbio[,c('country','hivstatusfinal')],zimbio[,grep('btwt',names(zimbio))])
malbio2 <-
cbind(malbio[,c('country','hivstatusfinal')],malbio[,grep('btwt',names(malbio))])
combined.0 <- bind_rows(zambio2,zimbio2,malbio2)

### Generate replicate weights for each country
# Manually enter these parameters into matrices
#   Country name, number replicate weights per country (retain same order)
countries <- c('Zambia', 'Zimbabwe', 'Malawi')
numwts <- c(253, 248, 250)

# Create new data.frame with new combined weights “b2wt” that are set equal to
#   original JK replicate weights if replicate belongs to that country
#   full sample weight if replicate belongs to a different country
newwts <- data.frame(matrix(data=NA,nrow=nrow(combined.0),ncol=sum(numwts)))
names(newwts) <- sprintf('b2wt%04d',1:sum(numwts)) # name newwts with leading zeroes
combined <- cbind(combined.0,newwts)

w <- 0 # This is a cumulative counter for the number of total weights that have been
generated
# Loop over each country
for (c in 1:length(countries)) {

  countryname <- countries[c]

  # Loop over number of weights in country
  for (i in 1:numwts[c]) {

    w <- w+1 # Increment cumulative counter by one

    # Replace b2wt JK replicate weight with original JK replicate weight if in
country
    combined[combined$country== countryname,sprintf('b2wt%04d',w)] <-
      combined[combined$country== countryname,sprintf('btwt%03d',i)]
    # Replace b2wt JK replicate weight with full sample weight btwt0 if not in
country
    combined[combined$country!= countryname,sprintf('b2wt%04d',w)] <-
      combined[combined$country!= countryname,'btwt0']
  }
}

### Ensure that weight variables are converted to numeric type, and drop
### records with no blood weight
wtname <- 'b2wt'
svydata <- combined %>%
  mutate(across(contains(c(wtname, "btwt0")), as.double)) %>%
  filter(!is.na(btwt0))


### Create survey object, JK2
# Recode analytic outcomes to 0/1
```

```
vars <- c('country','hivstatusfinal')
for(i in 1:length(vars)){
  varname <- vars[i]
  if(is.factor(svydata[,varname])) {
    svydata[,varname] <- as.numeric(levels(svydata[,varname]))[svydata[,varname]]
    # coerces factor variables to numeric before recoding
  }
  svydata[,varname][(svydata[,varname]==2)] <- 0
}

# Extract analytic variables, base weights and JK replicate weights
#   and create survey object,
#   specifying variance estimation method "JK" and JK coefficients=1
svydata1 <- svrepdesign(
  variables=svydata[,vars],
  weights=svydata[,'btwt0'],
  repweights=svydata[,grep(wtname,names(svydata))][,-1],
  type="JKn", scale=1, rscales=1)

### Conduct analyses using survey weights
# HIV prevalence
res <- svyciprop(~I(hivstatusfinal==1),
                 design=svydata1,
                 method="mean",level=0.95,df=25)
(c(res[[1]],attr(res,"ci"))) # display results
table(svydata1$variables$hivstatusfinal)
```

```
##############################################################################
### Example 6. Analyses of data from optional modules                      ###
###          (HIV knowledge and violence)                                  ###
##############################################################################

### Load in Adult IND dataset
adultind <- read.csv('C:/Desktop/zamphia2016adultind.csv')

### Code new outcomes
# All knowledge questions correct
svydata <- select(adultind, -contains(c("intwt", "vmpswt"))) %>%
  mutate(across(all_of(c("hivk_status", "onepartnr", "mosquito", "condoms",
"sharefood", "healthyinf")), as.numeric))
svydata$knowledge[svydata$hivk_status==1 & svydata$onepartnr==1 &
svydata$mosquito==1 & svydata$condoms==1 & svydata$sharefood==1 &
svydata$healthyinf==1] <- 1
svydata$knowledge[svydata$hivk_status==1 & !(svydata$onepartnr==1 &
svydata$mosquito==1 & svydata$condoms==1 & svydata$sharefood==1 &
svydata$healthyinf==1)] <- 0

### Ensure that weight variables are converted to numeric type and
### drop those with missing weights
wtname <- 'hivkpswt'
svydata <- svydata %>%
  mutate(across(contains(wtname), as.double)) %>%
  filter(!is.na(hivkpswt0))

# Extract analytic variables, base weights and JK replicate weights
#   and create survey object,
#   specifying variance estimation method "JK" and JK coefficients=1
vars <- c('knowledge','hivk_status')
svydata1 <- svrepdesign(
  variables=svydata[,vars],
  weights=svydata[,paste0(wtname,0)],
  repweights=svydata[,grep(wtname,names(svydata))][,-1],
  type="JKn", scale=1, rscales=1)

### Conduct analyses using survey weights
# Proportion of HIV knowledge module respondents answering all 5 questions correctly
svydata1_hivknowledge <- subset(svydata1, hivk_status == 1)
res <- svyciprop(~I(knowledge==1),
                 design=svydata1,
                 method="mean",level=0.95,df=25)
(c(res[[1]],attr(res,"ci"))) # display results
table(svydata1_hivknowledge$variables$knowledge)

### Code new outcomes
# Sexual or physical violence by partner in last 12 months
svydata <- adultind %>%
  mutate(across(all_of(c("vmflag", "sexualviolencepart12mo",
"physicalviolencepart12mo")), as.numeric))

svydata$violencepart12mo[svydata$vmflag==1 & (svydata$sexualviolencepart12mo==1 |
svydata$physicalviolencepart12mo==1)] <- 1
svydata$violencepart12mo[svydata$vmflag==1 & !(svydata$sexualviolencepart12mo==1 |
svydata$physicalviolencepart12mo==1)] <- 0

### Ensure that weight variables are converted to numeric type and
### drop those with missing weights
wtname <- 'vmpstw'
svydata <- svydata %>%
  mutate(across(contains(wtname), as.double)) %>%
  filter(!is.na(vmpstw0))
```

```
# Extract analytic variables, base weights and JK replicate weights
#   and create survey object,
#   specifying variance estimation method "JK" and JK coefficients=1
vars <- c('violencepart12mo','vmflag')
svydata1 <- svrepdesign(
  variables=svydata[,vars],
  weights=svydata[,paste0(wtname,0)],
  repweights=svydata[,grep(wtname,names(svydata))][,-1],
  type="JKn", scale=1, rscales=1)

### Conduct analyses using survey weights
# Proportion of violence module respondents experiencing either sexual or physical
violence from a partner in the past 12 months
svydata1_viol <- subset(svydata1, vmflag == 1)
res <- svyciprop(~I(violencepart12mo==1),
                 design=svydata1_viol,
                 method="mean",level=0.95,df=25)
(c(res[[1]],attr(res,"ci"))) # display results
table(svydata1_viol$variables$violencepart12mo)
```

4.4. HIV incidence calculators

4.4.1. SAS macro for HIV incidence estimation

```
/*
***  This macro calculates the weighted counts to be used as inputs to the CDC
Incidence Calculator and the design effect to be used.
It also directly calculates the annual incidence, which can be checked against the
incidence calculator result, and confidence
       intervals for this incidence. Finally, it calculates the estimated number of
new infections in the previous year.

      The required input dataset variables are:
            hivstatusfinal - hiv status (positive or negative) from blood testing
            recentlagvlarv - Recent infection indicator derived from LAg Avidity
testing and our recent infection algorithm
            btwt0, btwt001-btwtNNN - blood test weights and replicate weights

            gender and age variables for table breakdowns

      Input Parameters:
      diffvar - variable for which differentials will be calculated separately
(i.e., gender = male, female)
       filter_var - variable preset to 1 for observations to be included in
calculation (i.e., use age15_49 to select all ages 15-49)
      nrep - Number of replicates to be used in proc surveyfreq with
varmethod=jackknife to get the design effect
            omega - the MDRI (Mean Duration of Recent Infection)
            sig_omega2 - the standard deviation (sigma**2) of omega, derived as
=(12/(NORMSINV(0.975)))**2 using CDC figures
***;
*/
*Uganda;
*%macro incidence_calc(diffvar, filter_var, nrep, omega = 153, sig_omega2 =
162.6926, pfr = 0, sig_pfr2 = 0);

*All countries except for Uganda;

%macro incidence_diff(diffvar, filter_var, nrep, incidence_var = recentlagvlarv,
omega = 130, sig_omega2 = 37.48575911, pfr = 0, sig_pfr2 = 0);


data temp;
      set all_biomarker;
      where &filter_var. = 1;
run;

Proc sort data=temp; by &diffvar.;
run;

proc summary data = temp;
      by &diffvar.;
      var btwt0;
      output out = totals sum = sumbtwt0;
run;

Proc sort data=totals;
      by &diffvar.;
run;
```

```
** Data step creates the normalized weights for observations and the variables for
getting aggregate counts ***;
data temp2;
       merge temp totals;
       by &diffvar.;
       norm_btwt = btwt0*(_freq_ / sumbtwt0); **normalized weight such that sum of
the weights = number of observation;
       if hivstatusfinal = 1 then do;
              if &incidence_var. = 1 then when_infected = 1;
              else if &incidence_var. = 2 then when_infected = 2;
              else when_infected = 3; ** Positive but not classified by Lag test;
       end;
       else when_infected = 4; ** HIV negative;

       PplusN = 1;
       n = (when_infected = 4);
       p = (when_infected in (1,2,3));
       q = (when_infected in (1,2));
       r = (when_infected = 1);

       PplusN_wtd = PplusN * norm_btwt;
       n_wtd = n * norm_btwt;
       p_wtd = p * norm_btwt;
       q_wtd = q * norm_btwt;
       r_wtd = r * norm_btwt;

       PplusN_pop = PplusN * btwt0;
       p_pop = p * btwt0;
       q_pop = q * btwt0;
       r_pop = r * btwt0;

       Recent_over_all = 100 * (&incidence_var. = 1);
run;

* Counts for table 2.1.X;
proc summary data=temp2;
       by &diffvar.;
       var PplusN n_wtd p_wtd q_wtd r_wtd;
       output out=counts_aux sum=;
run;

data counts_aux;
       set counts_aux;
       rename n_wtd = n
              p_wtd = p
                 q_wtd = q
                 r_wtd = r;
run;

*** Summary gets aggregate weighted and unweighted counts ***;
proc summary data=temp2;
       by &diffvar.;
       var PplusN p q r PplusN_wtd p_wtd q_wtd r_wtd PplusN_pop p_pop q_pop r_pop;
       output out=counts sum=;
run;

data calc_incidence;
       set counts;
       neg_wtd = (PplusN_wtd - p_wtd) * (q_wtd / p_wtd);
       incid_instant = (r_wtd - &pfr. * q_wtd) / ((1 - &pfr. / &omega.) * (&omega. /
365) * neg_wtd);
       ** reduces to (r_wtd/neg_wtd)*(365/130) if false recent rate = 0;
       incid_annual = 100 *( 1 - exp(-incid_instant));
```

```
       ** expressed as annual percent;

       * Estimate number of new cases based on incidence and population at risk;
       neg_pop = PplusN_pop - p_pop;
       NewCases = neg_pop * incid_annual / 100;
run;

* Compute design effect for recent infection using blood test weights;
* Note: the &nrep parameter gives the number of replicate weights in the biomarker
dataset;
proc surveyfreq data=temp2 VARMETHOD=JACKKNIFE nosummary;
       by &diffvar.;
       ods output oneway=sfreq_results; ** OneWay is a table produced by Surveyfreq
with some results in a one way table;
       weight BTWT0;
       repweights BTWT001-BTWT&nrep.; **Note: defaults used for df and jkcoefs
parameters;
       table Recent_over_all / DEFF;
run;

proc sort data=sfreq_results;
       by &diffvar.;
run;

data sfreq_results2;
       set sfreq_results;
       by &diffvar.;
       if first.&diffvar.; *select first line of results for each subpopulation;
run;

*** add the design effect for the proportion recent/all to the counts needed for
Incidence Calculator ***;
data combine;
       merge calc_incidence sfreq_results2;
       by &diffvar.;
run;

* This formula comes from Kassanjee et al, Epidemiology, Vol 23, No 5, September
2012;
* A, B, and C are the 3 main terms in equation (e7) in the online appendix;
data variance;
       set combine;
       A = (1/q_wtd)*( (1/neg_wtd) + (r_wtd * ( p_wtd - r_wtd)) / (r_wtd - &pfr. *
p_wtd * (&omega. - &pfr. * 365))**2);
       B = &sig_omega2. * 1 / (&omega. - &pfr. * 365 )**2;
       C = &sig_pfr2. * (
              ( (p_wtd - r_wtd) * &omega. - r_wtd * (365 - &omega.) ) /
              ( (r_wtd - &pfr. * p_wtd) * (&omega. - &pfr. * 365) )
              )**2;

       * Confidence interval not including the design effect;
       * Note this formula is for instantaneous incidence;
       c_sq = (A + B + C);
       UCL_i = incid_instant * (1 + sqrt(c_sq) * probit(0.975));
       LCL_i = incid_instant * (1 - sqrt(c_sq) * probit(0.975));

       * Clopper-Pearson CI for cases where R = 0;
       if r_wtd = 0 then do;
              R_CPUCL = neg_wtd * (1 - (0.05/2)**(1/neg_wtd));
              UCL_i = (r_CPUCL - &pfr. * q_wtd) / ((1 - &pfr. / &omega.) * (&omega. /
365) * neg_wtd);
              LCL_i = 0.0;
       end;
```

```
        **;

        * Convert upper and lower limits of the interval into annual incidence
limits;
        UCL_a = 100 *( 1 - exp(-UCL_i));
        LCL_a = 100 *( 1 - exp(-LCL_i));

        * Confidence interval adjusting for the design effect;
        * Design effect < 1 is treated as equal to 1;
        if (DesignEffect > 1) then
                adj_c_sq = DesignEffect * (A + B + C);
        else if (DesignEffect <= 1) then
                adj_c_sq = 1.0 * (A + B + C);

        adj_UCL_i = incid_instant * (1 + sqrt(adj_c_sq) * probit(0.975));
        adj_LCL_i = incid_instant * (1 - sqrt(adj_c_sq) * probit(0.975));

        * Use Clopper-Pearson CI for cases where R = 0;
        if r_wtd = 0 then do;
                R_CPUCL_adj = neg_wtd * (1 - (0.05/2)**( max(1.0, DesignEffect) /
neg_wtd));
                adj_UCL_i = (r_CPUCL_adj - &pfr. * q_wtd) / ((1 - &pfr. / &omega.) *
(&omega. / 365) * neg_wtd);
                                        *(R_CPUCL_adj / neg_wtd) * (365 / &omega.);
                adj_LCL_i = 0.0;
        end;

        * Convert upper and lower limits of the interval into annual incidence;
        adj_UCL_a = 100 *( 1 - exp(-adj_UCL_i));
        adj_LCL_a = 100 *( 1 - exp(-adj_LCL_i));
run;

proc print data=variance;
        var &diffvar. PplusN p q r PplusN_wtd p_wtd q_wtd r_wtd DesignEffect neg_wtd
incid_instant incid_annual adj_LCL_a adj_UCL_a NewCases;
run;

* Create tables in appropriate format with the desired output variables;
data tab_output_incidence;
        length Row $ 8 gend $ 8;
        set variance;
        Row = translate( substr("&filter_var.", index("&filter_var.","e") + 1), '-',
'_');

        * User should customize to variable categories chosen;
        if gender = 1 then gend = 'Male';
        else if gender = 2 then gend = 'Female';
        else gend = 'Total';
run;

* Output weighted counts (normalized to the sample size);
data tab_output_x (keep = Row gend n p q r);
        length Row $ 8 gend $ 8;
        set Counts_aux;
        Row = translate( substr("&filter_var.", index("&filter_var.","e") + 1), '-',
'_');

        if gender = 1 then gend = 'Male';
        else if gender = 2 then gend = 'Female';
        else gend = 'Total';
run;
```

```
** Create columns for number of new cases;
proc surveyfreq data = all_biomarker method = jackknife;
       where &filter_var. = 1 and bt_status = 1;
       weight BTWT0;
       repweights BTWT001-BTWT&nrep. / jkcoefs=1 df = 25;
       tables &diffvar. * hivstatusfinal / row CL CLWT;
    ods output crosstabs = AdultPrev_&filter_var._out;
run;

data AdultPrev_&filter_var.;
       length gend $10;
       set AdultPrev_&filter_var._out;
       where hivstatusfinal = 1;

       if gender = 1 then gend = 'Male';
       else if gender = 2 then gend = 'Female';
       else gend = 'Total';

       if gend ne 'Total' then    do;
              Percent = RowPercent;
              StdErr = RowStdErr;
              LowerCL = RowLowerCL;
              UpperCL = RowUpperCL;
       end;

       drop Table F_: _SkipLine RowPercent RowStdErr RowLowerCL RowUpperCL
hivstatusfinal;
       rename wgtfreq = PLHIV
                 lowerclwgtfreq = PLHIVLCL
                 upperclwgtfreq = PLHIVUCL
                 StdDev = PLHIVStdErr
                 Percent = Prevalence
                 LowerCL = PrevalenceLCL
                 UpperCL = PrevalenceUCL
                 StdErr = PrevalenceStdErr;
run;

proc sql;
       create table PLHIV_new as
       select i.*, p.PLHIV, p.PLHIVStdErr, p.PLHIVLCL, p.PLHIVUCL,
                     p.Prevalence, p.PrevalenceStdErr,
                     p.PrevalenceLCL, p.PrevalenceUCL
       from tab_output_incidence i
       left join adultprev_&filter_var. p
       on i.gend = p.gend;
       quit;
run;

data tab_output_newcases;
       set PLHIV_new;

       neg_pop = PLHIV*100.0/Prevalence - PLHIV;

       if incid_annual = 0 then do;
              NewCasesLCL = 0.0;
              NewCasesUCL = neg_pop * (adj_UCL_a/100) * (1 + (Prevalence -
PrevalenceLCL) / Prevalence);

              NewCasesRelErr = .;
       end;
       else do;
              NewCasesRelErr = SQRT( ((Prevalence - PrevalenceLCL) / Prevalence)**2
```

```
                                          + ((Incid_annual - adj_LCL_a) /
Incid_annual)**2
                                            );

              NewCasesLCL = MAX(0.0, NewCases * (1 - NewCasesRelErr));
              NewCasesUCL = NewCases * (1 + NewCasesRelErr);

              NewCasesStdErr = (NewCases - NewCasesLCL) / 2.0635;
       end;
run;

* Filter and rearrange output variables in final tables;

data tab_output_incidence;
       retain Row gend Designeffect incid_annual LCL_a UCL_a adj_LCL_a adj_UCL_a;

       set tab_output_incidence;

       if LCL_a < 0 then LCL_a = 0;
       if adj_LCL_a < 0 then adj_LCL_a = 0;

       keep Row gend Designeffect incid_annual LCL_a UCL_a adj_LCL_a adj_UCL_a;
run;

data tab_output_newcases;
       retain Row gend PLHIV PLHIVStdErr PLHIVLCL PLHIVUCL
              NewCases NewCasesStdErr NewCasesLCL NewCasesUCL;

       set tab_output_newcases;
       keep Row gend PLHIV PLHIVStdErr PLHIVLCL PLHIVUCL
              NewCases NewCasesStdErr NewCasesLCL NewCasesUCL;
run;

** End new cases calculation;
%mend incidence_diff;

** Load in dataset (user must customize for dataset location);
libname read "P:\PHIA Project\Data Analysis and
Dissemination\SAC\Dissemination\First Report\Ethiopia" access=readonly;

proc datasets;
       delete tab_output:;
run;

data all_biomarker;
       set read.zamphia2016adultbio;
       where hivstatusfinal in (1,2) AND bt_status = 1;

       if age >= 15 and age <= 64 then
              age15_64 = 1;
       else
              age15_64 = 2;
       if age >= 15 and age <= 49 then
              age15_49 = 1;
       else
              age15_49 = 2;
       all = 1; ** define by variable for getting totals when using the macro;
run;

%incidence_diff(diffvar = gender, filter_var = age15_64, nrep = 253) ;
%incidence_diff(diffvar = gender, filter_var = age15_49, nrep = 253) ;
```

4.4.2. Stata program for HIV incidence estimation

```
// Specify filepath and dataset name
global filepath "C:\Desktop"
global dataset "zamphia2016adultbio"

// Import dataset
use "$filepath\\$dataset.dta", clear


*************************
*** Specify parameters ***
*************************

// Data-related parameters
	// Filter variables: for example, if estimate of interest is among subsetted
population (e.g., aged 15 to 49)
	keep if age15_49 == 1
	// diffvar - stratification variable by which incidence will be calculated
separately (e.g., gender)-- one at a time
	global diffvar "gender"
	// Recency variable
	global recent "recentlagvlarv"
	// Number of jackknife weights
	global nrepw "253"

// Incidence estimator parameters
	// Mean duration of recent infection, MDRI (omega)
	global omega = 130
	// Variance(MDRI)
	global sig_omega2 = 37.48575911
	// Proportion of false recents
	global pfr = 0
	// Variance(PFR)
	global sig_pfr2 = 0


***************************
*** Incidence calculator ***
***************************

// Subset to participants with valid data on HIV status
keep if inlist(hivstatusfinal,1,2)
keep if bt_status == 1

//Sort data by stratification variable
sort $diffvar

//Calculate sum of blood weights by diffvar
by $diffvar: egen sumbtwt0 = sum(btwt0)
	format sumbtwt0 %15.0f
egen _freq_ = count(1), by ($diffvar)

//Normalized weights such that sum of the weights = number of observations
gen norm_btwt = btwt0 * (_freq_ / sumbtwt0)

//Create indicator for HIV/recency status (when_infected)
gen when_infected = 1 if hivstatusfinal == 1 & $recent == 1
replace when_infected = 2 if hivstatusfinal == 1 & $recent == 2
replace when_infected = 3 if hivstatusfinal == 1 & !inlist($recent,1,2)
replace when_infected = 4 if hivstatusfinal == 2

gen pplusn = 1
gen n_ = 1 if when_infected == 4
	replace n_ = 0 if when_infected != 4
gen p_ = when_infected
	recode p_ (1/3 = 1) (4=0)
gen q_ = when_infected
	recode q_ (1/2 = 1) (3=0) (4=0)
gen r_ = 1 if when_infected == 1
	replace r_ = 0 if when_infected != 1
```

```
gen pplusn_wtd = pplusn * norm_btwt
gen n_wtd = n_ * norm_btwt
gen p_wtd = p_ * norm_btwt
gen q_wtd = q_ * norm_btwt
gen r_wtd = r_ * norm_btwt

gen pplusn_pop = pplusn * btwt0
gen p_pop = p_ * btwt0
gen q_pop = q_ * btwt0
gen r_pop = r_ * btwt0

gen recent_over_all = 100 if $recent == 1
      replace recent_over_all = 0 if $recent != 1

// Estimate design effects and save for later use when calculating variance
      //Declare survey design
      svyset [pw=btwt0], jkrweight(btwt001-btwt$nrepw, multiplier(1))
vce(jackknife) //Note this calculator uses default df, instead of dof(25) used in
final report calculations

      //For loop to calculate design effect over categories of stratification
variable
      levelsof $diffvar, local(categories)
            foreach ctg in `categories' {
                  svy jackknife: tab recent_over_all if $diffvar == `ctg', deff
                  matrix deff_`ctg' = e(Deff)
                  }



// Obtain aggregate counts
collapse (sum)      pplusn_sum = pplusn ///
                        p_sum = p_ ///
                        q_sum = q_ ///
                        r_sum = r_ ///
                        pplusn_wtd_sum = pplusn_wtd ///
                        p_wtd_sum = p_wtd ///
                        q_wtd_sum = q_wtd ///
                        r_wtd_sum = r_wtd ///
                        pplusn_pop_sum = pplusn_pop ///
                        p_pop_sum = p_pop ///
                        q_pop_sum = q_pop ///
                        r_pop_sum = r_pop ///
                        , by($diffvar)
format *_pop_sum %15.2f

//Calculate incidence
gen neg_wtd = (pplusn_wtd_sum - p_wtd_sum) * (q_wtd_sum / p_wtd_sum) // this weights
up P&Q to make up for indeterminate results on recency test
gen incid_instant = (r_wtd_sum - $pfr * q_wtd_sum) / ((1 - $pfr / $omega) * ($omega
/ 365) * neg_wtd) // reduces to (r_wtd_sum/neg_wtd)*(365/$omega.) if false recent
rate (pfr) = 0
gen incid_annual = 100 *( 1 - exp(-incid_instant)) // expressed as annual percent

// Estimate number of new cases based on incidence and population at risk
gen neg_pop= pplusn_pop_sum - p_pop_sum
gen new_cases_pop = neg_pop*incid_annual/100
format neg_pop new_cases_pop %15.2f


//Variance
gen A = (1/q_wtd_sum)*((1/neg_wtd) + (r_wtd_sum * (p_wtd_sum - r_wtd_sum))/
(r_wtd_sum - $pfr * p_wtd_sum * ($omega - $pfr * 365))^2)
gen B = $sig_omega2 * 1/ ($omega - $pfr * 365)^2
gen C = $sig_pfr2 * ( ( ( p_wtd_sum - r_wtd_sum) * $omega - r_wtd_sum * (365 -
$omega))/ ((r_wtd_sum - $pfr * p_wtd_sum) * ($omega - $pfr *365))) ^ 2

//Formula for instantaneous incidence
gen c_sq = (A + B + C)
gen probit_test = invnormal(0.975)
```

```
gen lcl_i = incid_instant * (1 - sqrt(c_sq) * probit_test)
gen ucl_i = incid_instant * (1 + sqrt(c_sq) * probit_test)


//Convert upper and lower limits of the interval into annual incidence
gen lcl_a = 100 * (1 - exp(-lcl_i))
gen ucl_a = 100 * (1- exp(-ucl_i))


// design effect
// uses for loop to read in estimated design effects over levels of the
stratification variable
gen designeffect = .
levelsof $diffvar, local(categories)
foreach ctg in `categories' {
       replace designeffect = deff_`ctg'[1, 1] if _n ==`ctg'
       }

//Second confidence interval adjusts for the design effect
// When design effect < 1, use 1
gen adj_c_sq = max(1.0, designeffect) * (A + B + C)
gen probit_calc = invnormal(0.975)
gen adj_lcl_i = incid_instant * (1 - sqrt(adj_c_sq) * probit_calc)
gen adj_ucl_i = incid_instant * (1 + sqrt(adj_c_sq) * probit_calc)

//Clopper-Pearson CI for cases where R = 0
replace adj_lcl_i = 0 if r_wtd == 0
gen r_cpucl_adj = neg_wtd * (1- (0.05/2)^( max(1.0, designeffect) / pplusn_wtd_sum))
                          if r_wtd_sum == 0
replace adj_ucl_i = (r_cpucl_adj - $pfr * q_wtd_sum) / ((1 - $pfr/ $omega) *
($omega/365)*neg_wtd)      if r_wtd == 0

//Convert upper and lower limits of the interval into annual incidence
gen adj_lcl_a = 100 *(1 - exp(-adj_lcl_i))
gen adj_ucl_a = 100 *(1 - exp(-adj_ucl_i))
```

5. References

1. Saito S, Duong YT, Metz M, et al. Returning HIV-1 viral load results to participant-selected health facilities in national Population-based HIV Impact Assessment (PHIA) household surveys in three sub-Saharan African Countries, 2015 to 2016. *J Int AIDS Soc.* 2017;20(Suppl 7. doi):10.1002/jia1002.25004.
2. Kalton G, Flores Cervantes I. Weighting methods. *Journal of Official Statistics.* 2003;19:81-97.
3. Kass GV. An exploratory technique for investigating large quantities of categorical data. *Appl Statist.* 1980;29(2):119-127.
4. Hastie T, Tibshirani R, Friedman J. *The Elements of Statistical Learning.* New York, NY: Springer-Verlag New York; 2009.
5. Fishel JD, Bradley SEK, Young PW, Mbofana F, Botão C. *HIV among couples in Mozambique: HIV status, knowledge of status, and factors associated with HIV serodiscordance. Further analysis of the 2009 Inquérito Nacional de Prevalência, Riscos Comportamentais e Informação sobre o HIV e SIDA em Moçambique 2009. .* Calverton, Maryland, USA: ICF International;2011.
6. Westat. *WesVar® User's Guide.* 2007.
7. Valliant R, Dever JA, Kreuter F. *Practical tools for designing and weighting survey samples.* Vol 51. New York, NY: Springer-Verlag; 2013.
8. Burns AM, Morris RJ, Liu J, Byron MZ. Estimating degrees of freedom for data from complex surveys. *2003 Joint Statistical Meetings - Section on Survey Research Methods.* 2003.
9. AAPOR. *The American Association for Public Opinion Research. Standard Definitions: Final Dispositions of Case Codes and Outcome Rates for Surveys.* 2015.
10. Filmer D, Pritchett LH. Estimating wealth effects without expenditure data - or tears: An application to educational enrollments in states of India. *Demography.* 2001;38(1):115-132.
11. Rutstein SO, Kiersten J. *The DHS Wealth Index.* Calverton, Maryland: ORC Macro;2004.
12. Howe LD, Hargreaves JR, Huttly SR. Issues in the construction of wealth indices for the measurement of socio-economic position in low-income countries. *Emerg Themes Epidemiol.* 2008;5:3.(doi):10.1186/1742-7622-1185-1183.
13. Kaiser BN, Hruschka D, Hadley C. Measuring material wealth in low-income settings: A conceptual and how-to guide. *Am J Hum Biol.* 2017;29(4).(doi):10.1002/ajhb.22987. Epub 22017 Feb 22925.
14. Kassanjee R, McWalter TA, Barnighausen T, Welte A. A new general biomarker-based incidence estimator. *Epidemiology.* 2012;23(5):721-728. doi: 710.1097/EDE.1090b1013e3182576c3182507.
15. Duong YT, Qiu M, De AK, et al. Detection of recent HIV-1 infection using a new limiting-antigen avidity assay: potential for HIV-1 incidence estimates and avidity maturation studies. *PLoS One.* 2012;7(3):e33328. doi: 33310.31371/journal.pone.0033328. Epub 0032012 Mar 0033327.
16. Wei X, Liu X, Dobbs T, et al. Development of two avidity-based assays to detect recent HIV type 1 seroconversion using a multisubtype gp41 recombinant protein. *AIDS Res Hum Retroviruses.* 2010;26(1):61-71. doi: 10.1089/aid.2009.0133.
17. Duong YT, Kassanjee R, Welte A, et al. Recalibration of the limiting antigen avidity EIA to determine mean duration of recent infection in divergent HIV-1 subtypes. *PLoS One.* 2015;10(2):e0114947. doi: 0114910.0111371/journal.pone.0114947. eCollection 0112015.

18. Kassanjee R, Pilcher CD, Keating SM, et al. Independent assessment of candidate HIV incidence assays on specimens in the CEPHIA repository. *AIDS.* 2014;28(16):2439-2449. doi: 2410.1097/QAD.0000000000000429.
19. Longosz AF, Morrison CS, Chen PL, et al. Comparison of antibody responses to HIV infection in Ugandan women infected with HIV subtypes A and D. *AIDS Res Hum Retroviruses.* 2015;31(4):421-427. doi: 410.1089/AID.2014.0081. Epub 2014 Nov 1019.
20. Kassanjee R, Pilcher CD, Busch MP, et al. Viral load criteria and threshold optimization to improve HIV incidence assay characteristics. *AIDS.* 2016;30(15):2361-2371. doi: 2310.1097/QAD.0000000000001209.
21. Parekh B. Case for including ARV in recent infection testing algorithm (RITA). Paper presented at: WHO Incidence Meeting2018.
22. Clopper CJ, Pearson ES. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika.* 1934;26(4):404-413.
23. Newcombe RG. Two-sided confidence intervals for the single proportion: comparison of seven methods. *Statist Med.* 1998;17:857-872.
24. Brown LD, Cai TT, DasGupta A. Interval estimation for a binomial proportion. *Statist Sci.* 2001;16(2):101-133.