

PHIA Data Use Manual

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

Surveys 2020-2023

PHIA Collaborating Institutions

ICAP at Columbia University

CIHEB at University of Maryland, Baltimore

The United States Centers for Disease Control and Prevention (CDC)

Westat

ICF

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 agreements #U2GGH002172 and #U2GGH002173. 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 Manual. Baltimore, MD and New York, NY. April 2023.

Access this Manual Online

The PHIA Project at ICAP – <http://phia-data.icap.columbia.edu>

RESPONSE at – <http://ciheb.org/PHIA/>

Contact Information

ICAP at Columbia University

722 West 168th Street

New York, NY 10032

Website: icap.columbia.edu

Email: icap-communications@columbia.edu

CIHEB at University of Maryland, Baltimore

725. W. Lombard Street

Baltimore, MD 21201

Website: ciheb.org

Email: ciheb@ihv.umaryland.edu

Table of Contents

Table of Contents	3
1. Introduction	6
1.1 What is the PHIA?	6
1.2 Purpose of the PHIA Data Manual.....	6
1.3 Purpose of PHIA Data Manual Supplement	6
1.4 Other documentation and resources	6
2. Overview of public release data contents	7
2.1 Datasets	7
2.2 Documentation	7
2.3 Statistical Programs	8
3. Files and Variables	9
3.1 Structure of PHIA 2.0 standard datasets	9
3.2 ID Variables	12
3.3 Eligibility and Response Variables.....	13
3.4 Original Variables	14
3.4.1. Single-response variables.....	14
3.4.2. Multi-response variables	14
3.4.3. Write-in variables.....	14
3.4.4. Continuous response variables.....	14
3.5 Analytic Variables	15
3.5.1 Flow diagrams for analytic variables	15
3.5.2 Wealth index	15
3.6 Missing Data.....	17
3.7 Survey Weights.....	17
3.7.1. Weighting Process	18
3.7.2. Using PHIA survey weight variables	22
3.8. Linkage Variables	24
3.8.1 Sexual partner linkages.....	24
3.8.2 Parent and Child Linkage.....	27
4. Data Management and Cleaning.....	28
4.1 Retaining stage	28

4.2 Reviewing stage.....	29
4.2.1 Checking questionnaires.....	29
4.2.2 Structural checks.....	29
4.2.3 Frequency checks.....	30
4.2.4 Outliers.....	30
4.3 Data Monitoring.....	30
4.4 Additional Data Management Issues.....	31
4.4.1 Write-in responses.....	31
4.4.2 Multi-response questions.....	31
4.4.3 Withdrawals and survey stopped.....	31
5. Data confidentiality processes.....	32
6. Statistical Guidance.....	34
6.1 Response rate calculations.....	34
6.1.1 Household response rates.....	34
6.1.2 Individual response rates.....	34
6.1.3 Overall response rates.....	35
6.2 Jackknife variance estimation.....	35
6.3 Taylor series variance estimation.....	36
6.4. Multi-country analysis.....	37
6.4.1 Example using the side-by-side method.....	38
6.4.2 Example using the concatenation method.....	39
6.5 New HIV infections and annual incidence.....	41
6.5.1 Definitions.....	42
6.5.2 Identification of recently infected people.....	42
6.5.3 Use of ARV test results.....	43
6.5.4 Proportion of false recents.....	43
6.5.5 Incidence estimation.....	43
6.5.6 Detailed steps for incidence estimation.....	45
6.5.7 Incidence variables included in the PHIA datasets.....	45
6.5.8 Accounting for the PHIA sample design in incidence estimation.....	45
6.5.9 Estimation of the annual number of new infections.....	46
6.5.10 Confidence interval calculation for zero cells.....	46
6.5.11 Application to number of new infections.....	48

7. References.....	49
8. Attachments.....	51
8.1 Tabulation plan for standard tables	51
8.2 Guide to getting started with PHIA data.....	51

1. Introduction

1.1 What is the PHIA?

The Population-based HIV Impact Assessment (PHIA) surveys were designed to measure the reach and impact of HIV programs in PEPFAR-support countries through national household surveys. The PHIA Project was implemented by ICAP at Columbia University and the University of Maryland (UMB) in partnership with the Ministries of Health and the US Centers for Disease Control and Prevention (CDC); with additional technical support provided to ICAP by Westat, and to UMB by ICF.

PHIA results have been published online in **Summary Sheets** and **Final Reports**, at phia.icap.columbia.edu and <http://ciheb.org>. In addition to these reports, de-identified person and household level data have been made publicly available to researchers to conduct their own analyses.

1.2 Purpose of the PHIA Data Manual

The **PHIA Data Manual** (hereafter, “**Manual**”) guides users in using PHIA data. The manual applies to all PHIAs conducted between 2020 and 2022 and describes survey data details such as the data structure, types of variables included on the files, and PHIA statistical guidance.

1.3 Purpose of PHIA Data Manual Supplement

In addition to this **Manual**, data users should refer to the PHIA-specific **Data Manual Supplement** (hereafter, “**Supplement**”) for each PHIA they plan to analyze. The PHIA-specific supplements describe survey elements that varied by PHIA or implementing partner. Each **Supplement** contains survey-specific information on the survey design, sample size, biomarker testing, and documentation such as questionnaires and codebooks.

1.4 Other documentation and resources

Users may also find it useful to refer to PHIA publications including the **Summary Sheet** and **Final Report** for each PHIA. Each PHIA’s **Final Report** contains detailed results from the PHIA along with information on 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.

2. Overview of public release data contents

2.1 Datasets

The PHIA data have been organized in four main datasets: household (hh), roster (roster), adult interview (adultind) and adult biomarker (adultbio). The datasets are available as SAS, Stata, and CSV files. The main data files available for each survey are:

- Household dataset (SAS)
- Household dataset (Stata)
- Household dataset (CSV)
- Roster dataset (SAS)
- Roster dataset (Stata)
- Roster dataset (CSV)
- Adult individual interview dataset (SAS)
- Adult individual interview dataset (Stata)
- Adult individual interview dataset (CSV)
- Adult biomarker dataset (SAS)
- Adult biomarker dataset (Stata)
- Adult biomarker dataset (CSV)

See section 3 “Files and Variables” of this **Manual** for more information about the structure of the four main datasets.

Additional datasets are available upon request: household intermediary weights, individual intermediary weights, and geospatial. These datasets are available as SAS, Stata, and CSV files.

2.2 Documentation

Additional PHIA documentation is included as attachments to this **Manual** and to the **Supplement** for each PHIA.

Attachments to this **Manual** include:

- **General PHIA tabulation plan:** Table shells for PHIA **Final Report** and **Summary Sheet** tabulations, accompanied by details on the datasets and variables used for each table including the analytic (outcome) variable, row and column stratifiers, subset criteria, and weights used in the table calculations.
- **Guide to getting started with the PHIA data:** Guide showing suggested approach for becoming familiar with the PHIA data and documentation for users new to PHIA data.

Attachments to the **Supplement** include:

- **Survey Questionnaires:** The survey-specific household, roster, and adult questionnaires for each PHIA. 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. The question number on the questionnaire is referenced in the variable label on the datasets and in the "variable label" of the codebook, where applicable.
- **Codebook with Frequencies:** Codebooks for each dataset. Codebooks document each variable's name, category (i.e., the questionnaire module or source data of the variable), full question text or variable description, variable label (i.e., a condensed label used on the datasets), type and width (e.g., numeric, text), coding values and labels, and the frequency and percent of records containing each value. Summary statistics have been provided in the coding values and labels for selected numeric variables, such as counts.
- **Analytic Variable Flow Diagrams:** Flow diagrams illustrating the logic used to create key analytic variables.
- **Testing Methodology Diagram:** Flow diagram illustrating household-based HIV testing algorithm.
- **Sampling and Weighting Technical Report:** Details of sampling and weighting procedures for each PHIA.
- **Survey-Specific Table Specifications (where applicable):** Table shells and technical specifications for report tabulations customized for each PHIA.

2.3 Statistical Programs

When users request datasets, they receive a data package that includes datasets in SAS, Stata, or CSV format. The package also includes statistical programs for getting started with the data, which includes topics such as reading in the data, merging files, and using the survey weights.

3. Files and Variables

The PHIA data have been organized into four main datasets: household (hh), household roster (roster), adult interview (adultind) and adult biomarker (adultbio) datasets. Any exceptions to this general structure (for example, in countries where data is collected for children) have been noted in the **Supplement** for each PHIA.

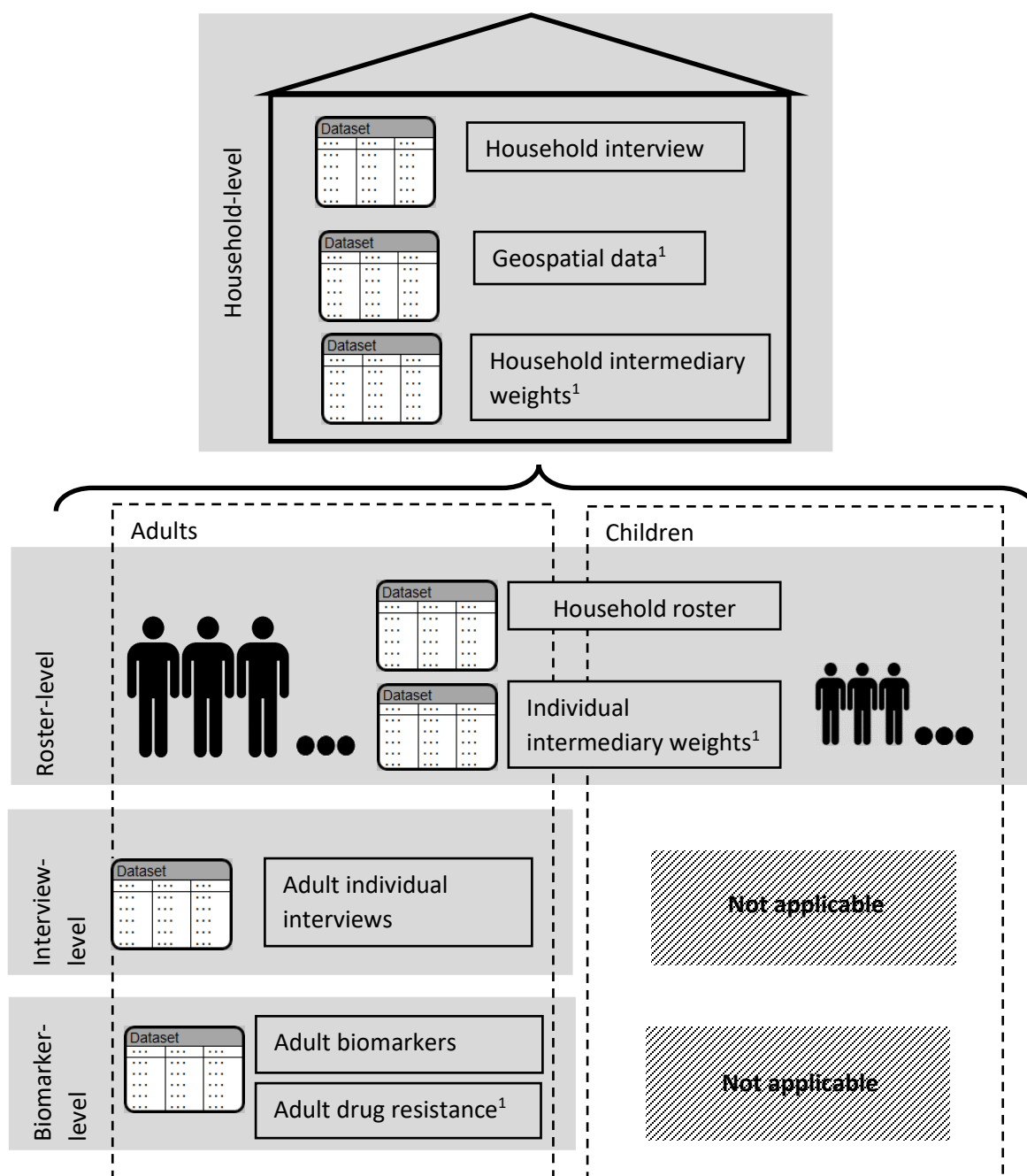
3.1 Structure of PHIA 2.0 standard datasets

PHIA datasets have been organized in a hierarchical structure, with each household record being associated with one or more records from the household roster, and individual and biomarker records provided for eligible and responding adults at each level (Figure 1).

The *Household* dataset contains records for all households that were selected to participate in the survey, regardless of eligibility and response status. Records for non-responding households were kept in the file to support calculating response rates. The *Roster* dataset contains records for all rostered individuals (children and adults), regardless of survey eligibility or response status. Records for non-responding and ineligible individuals were kept in the roster file to support calculating household characteristics.

All eligible responding adults have an individual interview record. Adults who consented and provided a blood sample have an associated biomarker record. Children were included on the roster file but did not have interview or biomarker data (consult the **Supplement** for exceptions).

Figure 1. Standard PHIA Dataset Structure



¹Refer to the **Supplement** for data availability and data request details.

Generally, dataset variables were ordered beginning with country name, followed by ID variables for the household or individual record, original questionnaire and/or biomarker variables, analytic variables and finally survey weights. The household, individual interview, and biomarker datasets contain the final weights that should be used for analyses of responding individuals on each file.

- Households were the highest-level observation. Each sampled household has been represented as a single record (row) on the household dataset (e.g., *xphia2020hh*), including sampled households that were ultimately determined to be ineligible (i.e. vacant, not a dwelling, or dwelling destroyed) or were non-responding households. Each household has been identified by a unique *householdid*. Households that participated in the household survey were indicated as eligible respondents by the variable *hhstatus* = 1.

Figure 2. Example Household records

country	year	householdid	hhstatus	<var1>	<var2>	...
Country	2020	CC0000000000001	1
Country	2020	CC0000000000002	1
Country	2020	CC0000000000003	2
...

- Roster records were the next level observation. Each individual rostered has been represented by a single record (row) on the roster dataset (e.g., *xphia2020roster*). All individuals who slept in the household the night before or who were usual residents were included on the household roster; this information was provided by the head of the household. Roster data contains individual-level roster data collected during the household interview for individuals of all ages, including those who were not eligible, who did not consent or who were not interviewed. Individuals who participated in the individual interview were indicated as eligible respondents by the variable *indstatus* = 1. Individuals who participated in biomarker testing and had valid laboratory test results were indicated by the variable *bt_status* = 1. Each person on the *Roster* dataset has been identified by a unique *personid*.

Figure 3. Example roster records

country	year	householdid	personid	age	indstatus	bt_status	...
Country	2020	CC0000000000001	CC000000000000101	24	1	1	...
Country	2020	CC0000000000001	CC000000000000102	26	2	99	...
Country	2020	CC0000000000002	CC000000000000201	40	1	2	...
Country	2020	CC0000000000002	CC000000000000202	35	1	1	...
Country	2020	CC0000000000002	CC000000000000203	12	8	99	...
Country	2020	CC0000000000002	CC000000000000204	5	8	99	...
...

- Individual Interviews were the next level observation. The adult interview records have been contained within the adult individual interview data (e.g. *xphia2020adultind*). Within each responding household, eligible responding individuals as identified by the variable *indstatus* = 1 have been represented as a single record (row) on the interview datasets. ID variables were identical to those in the *Roster* dataset.

Figure 4. Example adult interview records

country	year	householdid	personid	age	indstatus	<var1>	...
Country	2020	CC0000000000001	CC000000000000101	24	1
Country	2020	CC0000000000002	CC000000000000201	40	1
Country	2020	CC0000000000002	CC000000000000202	35	1
...

- Biomarkers were the lowest level observation. Individuals who were interviewed and consented to the biomarker testing as identified by the variable *bt_status* = 1 have been represented as a single record (row) on the *Biomarker* dataset (e.g., *xphia2020adultbio*). ID variables were identical to those in the *Roster* dataset.

Figure 5. Example adult biomarker records

country	year	householdid	personid	age	bt_status	<var1>	...
Country	2020	CC0000000000001	CC000000000000101	24	1
Country	2020	CC0000000000002	CC000000000000202	35	1
...

3.2 ID Variables

The ID variables on the PHIA datasets include country, PHIA year (*year*), *householdid*, *personid*, and *centroidid*. The *householdid* (and by extension *personid*) and *centroidid* were randomly assigned and do not include any embedded geographically identifying information.

- The variables *country* and *year* were the first two variables in each file. They contain the country and year in which the PHIA was conducted and can be used in pooled PHIA analyses to identify the PHIA.
- Each household has been identified with a 14-character unique code *householdid*, beginning with a two-letter country code (e.g., ZM for Zambia), and a randomly generated 12-digit numerical code. For countries where PHIA data has been collected in multiple rounds, one digit indicating the PHIA survey round follows the two-letter country code, and the remainder of the ID is a randomly generated 11-digit numerical code.
- Each individual participant has been 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.

- Each Enumeration Area (EA) was identified with an 8-character unique code *centroidid*, beginning with a two-letter country code followed by a randomly-generated 6-digit numerical code. Centroid IDs are used to link to geospatial data. For countries where PHIA data has been collected in multiple rounds, one digit indicating the PHIA survey round follows the two-letter country code, and the remainder of the ID is a randomly generated 5-digit numerical code.

The values of ID variables are unique across PHIA rounds, so that the IDs can be used in concatenated datasets with minimal manipulation by the user.

3.3 Eligibility and Response Variables

In PHIA datasets, the variables *hhstatus*, *indstatus*, and *bt_status* indicate eligibility and participation.

Figure 6. Eligibility and response indicators

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
indstatus	Indicator of individual eligibility and response status	1 - Eligible Respondent 2 - Eligible Non-Respondent 4 - Unknown Eligibility Status 5 - Roster Ineligible 8 - Not Sampled 9 - Non-defacto individuals
bt_status	Did lab blood test have definitive result?	1 - Lab blood test has a definite result 2 - Lab blood test does not a definite result 9 - Lab blood test has a definite result, non-defacto participants

These variables must be used in all analyses to ensure proper inclusion in or exclusion from analyses. They are used to examine response rates at the household, individual, and biomarker levels, and are fundamental to the calculation of survey weights used in weighted analyses. Detailed descriptions of how these variables are utilized in survey weighting and response rate calculations have been provided later in this manual.

Categories included in the status variables may vary across PHIA. Users should refer to the supplement and codebook for each PHIA for details.

3.4 Original Variables

Original variables were those that directly correspond to the survey questionnaire and were captured during the interview or lab testing process. They represent the variables from the field. This section describes how these variables were collected, edited, and their formats and processes for analytical use.

3.4.1. Single-response variables

Single response variables were questions in which a pre-defined set of categorical responses has been provided to respondents. These take the form of a list of potential responses in which one and only one response was permitted. Respondents can also answer with a “don’t know” or refuse to provide a response. In the PHIA data, -8 represents “don’t know” and -9 is a refusal to answer.

3.4.2. Multi-response variables

Multi-response variables were questions where the respondent answers openly and the interviewer selects responses mentioned from a pre-selected list. The interviewer may also ask for additional information to assure that the list of responses from the respondent is complete. Multi-response questions were recorded with an alpha character corresponding to each possible response in one field. Refer to section 4.4.2 for more information on how multi-response variables are coded.

3.4.3. Write-in variables

Write-in responses, also called open-ended responses or other/specify questions, were those where none of the pre-determined response categories capture the respondent’s answer and the interviewer had the option of typing in the respondent’s answer. In these situations, the interviewer selects “Other” and types in the response. This response was recorded as a character string. Write-in responses are not available in PHIA datasets to maintain participant confidentiality. Refer to section 4.4.1 for more information on how write-in responses are coded.

3.4.4. Continuous response variables

Continuous response variables are questions where the possible responses are continuous. Examples of questions with continuous response options include ages, dates of birth, and number of sexual partners. For continuous response questions, the interviewer records the response given by typing in the number. The response was recorded as a numeric field.

Some continuous response questions may include a maximum or minimum allowable response that was selected in place of entering a number. For example, any responses provided that exceed 100 were recorded as “100 and above” and took a value of 100 in the data. Other

questions may include nominal categories with specific meaning, for example “Never had sex” as a selection option for age at first sex. Responses of “don’t know” and “refused” when allowed, were also be included in the continuous response question as a nominal category. For all continuous response questions with nominal categories, the nominal categories were coded to a value that is outside the normal expected range. For example, “don’t know” takes a value of -8 and “refused” takes a value of -9 (see Section 4.6 on missing data for more information). When using continuous variables, refer to the codebook to ensure that any nominal values were being treated appropriately.

The original variables for the results of the HIV viral load test (*resultvlc*) and the normalized Optical Density (ODn) from the HIV-1 limiting antigen (LAG)-avidity assay testing (*lagodnfinal*) were continuous and corresponded to lab testing processes. The result of the viral load test had nominal values for viral load not detected (TND) and for viral load under the lower limit of quantification, dependent upon viral load test used.

3.5 Analytic Variables

Analytic variables refer to variables created during data processing after the conclusion of the survey. Analytic variables were derived by combining or collapsing information from original variables and are included in PHIA datasets to facilitate analyses. PHIA analytic variables are documented in data codebooks supplied with each country’s data.

3.5.1 Flow diagrams for analytic variables

Flow diagrams are available for selected analytic variables. Flow diagrams detail the exact process by which a variable was derived. Before utilizing an analytic variable in analysis, data users should review the corresponding flow diagram to ensure the variable is being interpreted correctly. Because the diagrams include frequency counts, they were constructed on a country-specific basis and are included in each PHIA’s **Supplement**.

3.5.2 Wealth index

Wealth index methods that utilize survey data on household assets, materials and durable goods have been an established measure of socioeconomic status since their adoption by the Demographic and Health Surveys Program (DHS). These wealth measures have been widely considered as superior to income in quantifying socioeconomic status in resource-limited settings and were easily discernible via the survey questionnaire.

PHIA wealth index variables have been constructed using the same method as DHS surveys. The household dwelling characteristic and asset variables that were used to construct wealth indices vary by PHIA and were noted in each PHIA’s **Supplement**. In PHIA datasets, two wealth index variables have been provided: a continuous score (*wealthscorecont*) and categorical wealth quintile (*wealthquintile*).

3.5.2.1 Wealth Index Construction

To construct wealth quintiles via DHS methods, the following steps were 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 such as ownership of durable goods such as beds, vehicles, and livestock. The specific assets and question wording vary across PHIA's **Supplement**. Multi-response categorical variables were 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 were coded as 1 "Yes" or 0 "No". Generally, missing data were treated as the absence of that asset, and households that did not have any asset data were not assigned wealth index scores or wealth quintiles.
2. Select the asset variables for inclusion. Asset variables were analyzed using Principal Component Analysis (PCA), a statistical technique that transforms a number of correlated variables into uncorrelated components that captured variability (information) in decreasing order. Thus, PCA has been a useful dimension reduction technique, and DHS recommends using the first component of the model as a summary indicator for wealth (the wealth index). Since asset relevance may vary between urban and rural settings, PCAs were run separately for urban and rural households, and then for all households combined. Decisions on whether to include or exclude certain asset variables from either setting may be made a priori. For parsimony, all asset variables that have any variability were included in each analysis.
3. Run PCA and combine results. Three PCAs were run: a "common" model across all households, and models restricted to "urban" and "rural" households. Per convention, the first factor from each model was extracted to obtain three separate wealth indices. The common model wealth index was regressed separately on the urban or rural wealth index for households in those areas, and this regression model was then used to convert each household's (rural or urban) wealth index into a final "composite" wealth index (wealthscorecont).
4. Generate wealth quintiles. Households were classified into quintiles (wealthquintile) using the composite wealth index. To account for the complex survey design, the weighted cumulative distribution of the wealth index was used to identify weighted quintile cut points.

3.5.2.2 Important considerations

Wealth indices and quintiles derived using this methodology were intended to represent measures of wealth relative to other households in the same country. It is not advisable to use wealth scores or quintiles from individual countries datasets with pooled data from multiple

countries. If wealth quintiles are required for the pooled data from multiple countries, the process above should be repeated on the pooled dataset.

It is important to note that the underlying PCA model simply identifies the factors that optimally capture the variation in the data and does not guarantee a straightforward interpretation. On average, households in higher wealth quintiles should be wealthier, 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 subsampled population.

For simplicity and to facilitate replication, variables were not selected differently in urban and rural models according to contextual or subjective knowledge. However, this approach may not be valid if assets were 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 were not sensitive to model specification. Alternative socioeconomic indicators were available, and the merits of these alternatives are the subject of ongoing debate.

3.6 Missing Data

PHIAs were administered using the Census and Survey Processing System (CSPro) package on electronic tablets, which permit forced responses. As a result, missing data for survey variables were 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” (coded as “-9”), or responses that were determined to be out of range (“-7”, e.g., when a woman who has been pregnant says she has never had sex), or where a non-applicable question was intentionally skipped (“.”, e.g., number of prior pregnancies does not apply for men and therefore that question is skipped for men altogether). A code of “-6” is used to indicate situations where a question was missing that the respondent should have answered.

Missing data for analytic variables (see Section 3.5 Analytic variables) were coded as “99” without distinguishing the reason for missingness (“don’t know”, “refused” or not applicable).

Users should take care when conducting analyses to check for and determine appropriate treatment for missing responses. Consult each PHIA’s **Supplement** for specific information on missing data.

3.7 Survey Weights

PHIA data were weighted based on sampling probabilities adjusted for non-response and post-stratified to national population projections from the survey year based on age and sex. This

section describes the method used for constructing PHIA weights and provides practical guidance on how to use these weights for analyses.

3.7.1. Weighting Process

The main steps of the weighting process include:

- Initial checks to confirm that the probabilities of selection associated with the sampled units are computed correctly.
- Creation of jackknife replicates to be used for variance estimation.
- Calculation of PSU base weights to reflect the overall PSU probabilities of selection.
- Calculation of household weights to reflect the probabilities of selecting households within PSUs, and to compensate for household nonresponse.
- Calculation of person-level interview weights to reflect the differential probabilities of selecting individuals within households, and to compensate for nonresponse to the interview.
- Poststratification of the person-level interview weights to calibrate the weighted counts of persons completing the interview so that they match external population counts.
- Calculation of person-level blood test weights to reflect the differential probabilities of selecting individuals within households, compensate for nonresponse to the blood test, and adjust for potential undercoverage through poststratification.

General weighting information is available in subsequent sections of this **Manual**. For additional technical details, refer to each PHIA's **Supplement** and to the **Sampling and Weighting Technical Report**.

3.7.1.1 Base weights

Each PHIA used a stratified, multistage probability sample design. At the first stage, EAs were 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 were constructed for the second stage of sampling, which were drawn from updated household listing data collected by the PHIA team. Households were selected using an equal probability method. The actual number of households selected per cluster varied by PHIA, with an average of 35 and range of 15 to 70 (refer to each PHIA's **Supplement**).

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

Equation 1. Household probability of selection

$$p_{hij} = p_{hi} \times p_{j/hi}$$

where

p_{hij} : the household probability of selection for household j in EA i in stratum h

p_{hi} : the probability of selection for EA i in 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 stratum h

Household base weights w_{hij}^{HH} were calculated as the inverse of the overall household probability of selection (also known as the design weight d_{hij}) adjusted for household-level nonresponse as follows:

Equation 2. Household non-response adjusted weights

$$w_{hij}^{HH} = R_{hij}^{HH} \frac{1}{p_{hij}} = R_{hij}^{HH} d_{hij}$$

where R_{hij}^{HH} denote the non-response adjustment factor adjusting for nonresponse among households selected for the survey. The non-response adjustment factor is based on the weighted number of households.

Usually, all adults in all completed household were eligible for the survey (see each PHIA's **Supplement** for eligibility criteria). Individual base weights w_{hijk}^{indv} were calculated based on the household base weights w_{hij}^{HH} after adjustment for non-response among individuals. Individual base weight w_{hijk}^{indv} for individual k in household j in EA i in stratum h was calculated as:

Equation 3. Individual base weights

$$w_{hijk}^{indv} = \frac{1}{p_{j/hi}^s} R_{hijk}^{indv} w_{hijk}^{HH}$$

where

$p_{j/hi}^s$: the probability of selection for household j in a subsample s . In case of individual base weights for adults, $p_{j/hi}^s = 1$

R_{hijk}^{indv} : the non-response adjustment factor applied on the household base weight for individual k in household j in EA i in stratum h , adjusting for nonresponse among individuals eligible for the survey.

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

base weights. Blood test base weight w_{hijk}^{bt} for individual k in household j in EA i in stratum h is calculated as:

Equation 4. Blood test base weights

$$w_{hijk}^{bt} = R_{hijk}^{bt} w_{hijk}^{indv}$$

where R_{hijk}^{bt} denote the non-response adjustment factor applied on the individual base weight for individual k in household j in EA i in stratum h , to adjust for nonresponse that happened in the blood testing.

3.7.1.2 Nonresponse adjustments

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

Nonresponse weight adjustment followed the cell-weighting approach (Kalton and Cervantes, 2003). Nonresponse weight adjustment cells for households were EAs or groups of EAs. Nonresponse weight adjustment cells for individuals and blood samples were determined through the use of a CHAID (Chi-square Automatic Interaction Detection) tree classification scheme, which identified predictors of response (Kass, 1980). Response propensities were calculated within each cell defined by these response predictors, which were then used to adjust for non-response. The table below lists examples of potential independent (predictor) variables that were used to define the nonresponse adjustment cells, which were initially selected using Least Absolute Shrinkage and Selection Operator (LASSO) regression (Hastie et al., 2009). 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's **Supplement**.

Figure 7. Examples of sources and variables used in nonresponse adjustment

Component	Source	Potential independent variables
Household	EA sampling frame	Region, district, urban/rural
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	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 any household j in EA i , the nonresponse adjustment factor was computed as:

Equation 5. Nonresponse adjustment factor for households

$$R_{hij}^{HH} = \sum_{j=1}^{n_c} d_{hij} / \sum_{j=1}^{n_c} R_{hij} d_{hij}$$

where R_{hij} denote the response status of household j in EA i in stratum h , where $R_{hij} = 1$ if the household j responded, and $R_{hij} = 0$ otherwise, and n_c denote the number of sampled households in adjustment cell c (EA i in case of adjustment for household nonresponse). If an EA had such a low response rate that the adjustment would result in excessively high adjusted weights, that EA was grouped with a similar EA with more respondents to form a single non-response adjustment cell.

Nonresponse adjustment of the individual-level weight began with the household base weight. The adjustment factor to adjust for individual-level nonresponse was calculated as:

Equation 6. Nonresponse adjustment factor for individuals

$$R_{hijk}^{indv} = \sum_{k=1}^{m_c} w_{hijk}^{HH} / \sum_{k=1}^{m_c} R_{hijk} w_{hijk}^{HH}$$

where R_{hijk} denote the response status of individual k in household j in EA i in stratum h , where $R_{hijk} = 1$ if individual k completed the survey, and $R_{hijk} = 0$ otherwise, and m_c denote the number of eligible individuals in adjustment cell c . Similarly, the nonresponse adjustment factor R_{hijk}^{bt} of the blood sample weights was calculated based on the individual base weight and the response status to the blood testing.

The adjustment cells involved in the calculations of nonresponse adjustment factors in both the individual base weight and the blood base weight were formed based on predictors of nonresponse according to separate LASSO and CHAID-based models.

3.7.1.3 Poststratification (undercoverage) adjustments

The PHIA's aimed 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 were further adjusted for undercoverage to a set of population projections for the country. Undercoverage adjustments were made in similar fashion to nonresponse adjustments, by creating cells within which weights were adjusted for undercoverage. Such adjustment cells were 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. Similar to the nonresponse adjustment, the nonresponse-adjusted weights were multiplied by poststratification factors that were calculated for each adjustment cell as the known/projected total divided by the sum of the nonresponse-adjusted weights for all individuals within that cell. For example, the poststratified individual weight can be calculated as:

Equation 7. Poststratified individual weight

$$W_{hijk}^{indv} = w_{hijk}^{indv} \frac{M_c}{\sum_{k=1}^{\hat{m}_c} w_{hijk}^{indv}}$$

where M_c is the known/projected population total in poststratification cell c , and \hat{m}_c denote the number of interviewed individuals with valid w_{hijk}^{indv} in adjustment cell c . The blood test weights also undergo post-stratification adjustment.

3.7.2. Using PHIA survey weight variables

Using survey weight variables correctly is essential to analyzing PHIA data properly. This section explains the available weight and weight-related variables. For additional technical details, refer to each PHIA's **Supplement**.

Weights and related variables needed for jackknife and for Taylor Series variance estimation are provided.

Examples of correct use of weights are provided in Stata, SAS and R programs that are included with requested datasets.

3.7.2.1 Survey weight variables

PHIA datasets include weight variables to support weighted analyses for each survey. For each survey weight variable, analytic variables identify each observation's eligibility and response status (see section 3.3). Such variables were critical for calculations of survey weights. Refer to each PHIA's **Supplement** for details on survey specific eligibility criteria and how these eligibility and response indicators were derived.

Figure 8. Main survey weight variables in all PHIA

Level	Survey weight variable	Values for all records with...
Household	hhwt0	hhstatus =1
Individual interview	intwt0	indstatus =1
Blood test	btwt0	bt_status =1

The final nonresponse-adjusted and poststratified sample weights were provided in each dataset and labeled accordingly (refer to the table above for variable names of survey weights for each level of analysis). Availability of survey estimation procedures varies by statistical software. Users should use the appropriate weights for the specific analysis of interest, which is generally determined by the target population of inference.

- Household weights should be used for analyses conducted at the household level, for example, distribution of households by urban/rural residence. 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 should be used for analyses conducted at the individual level for data collected for all potentially 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 test weights should 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.
- Data on sexual partners and marital relationships was collected, and couples may be a unit of analysis of interest to users (see section 3.8). As was 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.” The man’s individual sample weight was

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.

There are also weights provided on the household and roster files which were used in response rate calculations, including the household base weight (hhbwt0), individual base weight (indiv_bwt0), and trimmed person nonresponse adjusted weight (trmpnr1w0). See section 6.1 for details on the response rate calculations.

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's *Intermediary Weights* datasets.

Refer to each PHIA's **Supplement** and each PHIA's **Survey Sampling and Technical Report** for details on how weights are calculated.

3.8. Linkage Variables

3.8.1 Sexual partner linkages

Sexual and marital partnership data were collected as part of the Individual Interview (Marriage and Sexual Activity modules). To support analyses of partners, three types of partner linkage variables were provided in PHIA datasets. Note that survey weights were not provided for analyses with couples as the unit of analysis since sampling procedures do not identify couples during household listing. For couple analyses, we suggest the use of the men's individual interview or blood weight (see section 3.7.2).

3.8.1.1 Husband ID

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

3.8.1.2 Sexual partner IDs

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.

3.8.1.3 Last partner

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

3.8.1.4 Partner cluster ID

Researchers may be interested in analyzing groups of 3+ individuals in the same household who were 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 were a potential unit of interest for study. These “partnership clusters” were relevant where an individual has multiple wives and/or sexual partners in the household, thus pairs of individuals who were not themselves sexually partnered were connected indirectly through the common partner.

The variable *partnerclusterid* captured these complex partnerships by assigning a unique ID to all individuals who were 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. 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 were included. Any relationships reported outside the household were not identified.

The variable *partnerclusterid* uniquely identified each partnership cluster across the whole dataset. Partnership clusters were defined using the following rules:

1. All wives linked to their husbands using the *husid* variable were a part of the same cluster.
2. Any persons reported as sexual partners by a given person was a part of that person's cluster.
3. A person can only be in one cluster: if a person was linked to two or more other people then all of them, and anyone linked to them as a sexual or marital partner, was combined into a single larger cluster.
4. Self-reported information was assumed to be correct, even if only one side of the partnership reports the partnership.

Figure 9 below lists expected types of partnership cluster structures through six household cases. 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 have been shown.

Case 1 is a household with two pairs of partners: a husband and wife who were 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 were assigned a cluster number of 1 and the second couple was assigned to cluster 2.

Case 2 shows another relatively simple situation. Each partner has reported another sexual partner who was 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 did not change the cluster numbering. Note that the husband/wife did 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 were linked to the husband, and to each other, through the partnership cluster number.

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

Case 5 illustrated 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 were treated as correct regardless of whether the relationship was reciprocated, so these people were linked. In this example, person 501 was also married to and a reciprocal sexual partner with person 502. As a result, person 502 was 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 were two married couples, but these were also connected by an additional non-marital sexual partnership. All four persons were part of the same partnership cluster. Note that in this case person 602 was linked to 603 and 604 by the partnership cluster number, but that neither of these numbers occur on her record at all, and 603 did not occur on either her or her husband’s record. That is, persons 602 and 604 were indirectly linked through 603.

Figure 9. Examples of partner cluster structures

	<i>personid</i>	<i>gender</i>	<i>husid</i>	<i>partid1</i>	<i>partid2</i>	<i>partid3</i>	<i>partnerclusterid</i>
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							

	<i>personid</i>	<i>gender</i>	<i>husid</i>	<i>partid1</i>	<i>partid2</i>	<i>partid3</i>	<i>partnerclusterid</i>
	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

3.8.2 Parent and Child Linkage

PHIAs did not collect interview or biomarker data for children (see each PHIA's **Supplement** for exceptions) but data was captured in the household roster for children and regarding parent and child relationships. Because biomarker information was not captured for children, users in mother to child transmission should use the mother as the unit of analysis. Linkage data can be used to link children to their parents and may be of interest for researchers interested in subjects such as orphans and other vulnerable children.

For children aged 0-17 years, the identity of their mother was captured in the household roster by the identifier variable *natmomnm*, a numeric variable indicating the "line number" of the mother on the roster; this variable can be used to generate *personid* of the mother within the rostered household members. For children whose mothers were not in the household, *natmomnm* is missing. Similarly, the variable *dadmalename* was used to capture the "line number" of the child's father and was missing for children without a known father in the household.

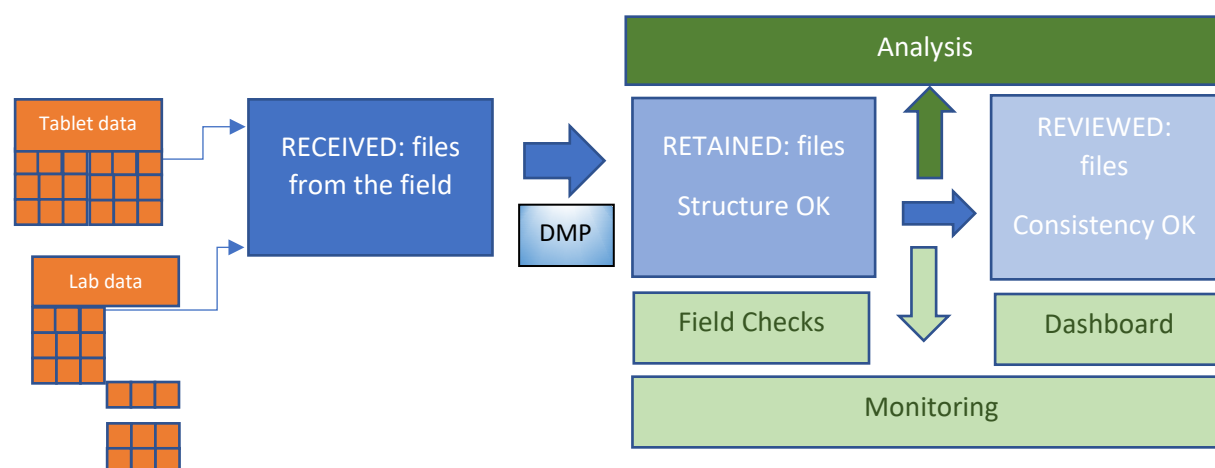
The household head reports the information on the mother and father of the child, including whether they were alive and whether they were a usual household resident. The household interview also collects information on whether the mother or father have been sick for at least 3 months in the last 12 months and their HIV status as well as demographic information on the child including education, gender, and age.

For some surveys, there may be additional information on children collected through their parents or mother. See each PHIA's **Supplement** for details.

4. Data Management and Cleaning

Although there were controls on data quality in the field, this section on data cleaning deals primarily with those processes carried out once the raw data has been transmitted to the central server. Depending upon the survey, the data management architecture may vary. However, the generalized conceptual model remains the same (Figure 10). Data was transmitted to a central server and undergoes a three-step process where data were received, retained and reviewed.

Figure 10. Flow of PHIA data from the field



Data editing focuses primarily on reviewing the structure of the data file and the consistency of the data files prior to data analysis. Data editing can be done on various platforms such as CSPro or SAS. The data editing platform depends on the capacity of the country and the specific Data Management Plan (DMP) developed for the country.

4.1 Retaining stage

Once the data file was collected on the tablet, it was transmitted to a central server. It was vital that this data has been backed up and safeguarded and remains unaltered. Retained data files undergo some basic reviews to validate the structure. This basic review ensures that the structure was complete, and the questionnaires were complete. Some data files were joined at this stage, and other relevant files were also retained including:

- Household interview
- Individual interview

- Lab data
- Roster
- Minor roster (if applicable)
- Signature (if applicable)
- Barcode data

4.2 Reviewing stage

During the reviewing stage, data undergo various checks for assessing data quality. These checks typically include the following:

- Checking the number of questionnaires in the EA match the expected responses
- Structural checks such as: completion, missing records, duplicates
- Frequency checks and reconciliation of denominators
- Outliers were identified for continuous variables
- Inter-variable consistencies were examined

During this phase, data quality issues were logged and fed-back to the field for resolution.

4.2.1 Checking questionnaires

A check that the number of household questionnaires assigned per EA matches the number of actual questionnaires administered has been done. Households that were not completed for any reason will still be accounted for in the count of questionnaires as a result code is assigned and recorded in the final dataset. Any significant differences between the numbers of households surveyed and the numbers of households in the sample design will require tracking and may require consultation with the sampling specialist and adjustment to the sample weighting.

4.2.2 Structural checks

After confirming that the correct number of individual questionnaires was completed, all individual questionnaires were checked against the roster to ensure that all expected persons and relevant questions were completed.

4.2.3 Frequency checks

Frequency checks ensure that the skip patterns of the questionnaire were followed correctly. Skip patterns occur when there are filter questions that determine the path of an interview. A respondent may not be required to answer certain questions if the filter questions do not route them to the path. In electronic data collection, these kinds of errors have been minimized but they may occur. A skipped question was a simple blank. The respondent was not considered in the denominator to that question. If for some reason, an expected response was missing then it was important to reconcile the denominator. A correction may be undertaken to code the “in-path” missing to differentiate it from a blank (not applicable). These frequency checks were designed to reconcile the denominators to assure integrity of the analysis.

4.2.4 Outliers

Generally, unusual and outlier values were found for variables that represent numeric values, such as age, animals, land area, times, and numbers of items. Unusual values often indicate that data entry errors have been made during the interviews. Any values that were outside the range of the majority of the responses, values with low or single frequencies, values that seem unusually precise, and values that seem unlikely in the context of the question are reviewed for errors. There were ranges and edit checks programmed into the tablet to minimize outliers, but checks may still be performed during and after data collection is occurring.

4.3 Data Monitoring

It was crucial to monitor the data while in the field for quality control. The exact details of survey data monitoring varied from country to country depending on the implementing partners, and typically takes place using tools such as field check tables and survey dashboards. These tools were used to give higher level management information required for executive oversight and to provide information on team performance in order to provide quick feedback into the field.

Some of the quality indicators examined in field check tables, dashboards, or both include:

- Result of household interviews
- Result of individual interviews
- Eligible men and women per household
- Response rates for the household, individual interview, and blood test
- Age displacement

These indicators were produced overall, by EA, by team, by specified geography, and over time to support survey monitoring.

4.4 Additional Data Management Issues

4.4.1 Write-in responses

Write-in responses were those responses to open-ended questions where the respondents were free to define or opt for a category not listed and they write-in their response. Write-in responses require secondary processing, involving an intensive review of all write-in responses. In many cases, the write-in responses were reclassified into existing categories. However, if there is an unforeseen response with a large number of responses, a new code may be assigned.

4.4.2 Multi-response questions

Multi-response questions must undergo a special secondary recoding process to assure that the variables were rendered useable for analysis. Each alpha coded response must become a binary response. Usually this is a yes/no response for each response category. The example provided in section 3.3.2 stores all alpha characters A-Z into one field. After the reshaping, there will be 15 variables from A to Z with each response category being a “Yes” if it was selected and “No” if it was not mentioned by the respondent.

Note that in the case that “don’t know” or “refused” is selected, the standardized categories for “don’t know” (-8) and “refused” (-9) will have to be assigned to all the possible choices. Also, the response of “Other” will require a write-in as outlined below and will be reviewed and up-coded prior to the secondary reshaping of the multiple response question.

4.4.3 Withdrawals and survey stopped

Any individual who did not complete the individual interview was classified as an incomplete interview or “stopped”. The respondent may stop and withdraw their responses. In this case, they are considered a withdrawal and their responses will not be included on the individual interview dataset. Similarly, if an individual originally consented to blood draw and later rescinded their consent, they will not be included on the biomarker dataset.

5. 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: process of identifying categories or variables containing a minimum number of cases or a minimum percent of households or individuals reporting the category or variable response; may be managed through top-coding, bottom-coding, or redaction.

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. See section 3.2 for more information.
- 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's 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's **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's **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.

Disclosure risk mitigation necessarily results in a degree of information loss. Recodes and redactions applied to PHIA data may introduce some data limitations. Therefore, it may not be possible to exactly reproduce all standard analytic variables from the variables available on the public-use datasets.

For more information about disclosure risk mitigation and specific methods applied, see each PHIA's supplement.

6. Statistical Guidance

6.1 Response rate calculations

Response rates were reported in the **Summary Sheet** and **Final Report** for each PHIA. In order to calculate household and individual response rates, the following procedure is used.

6.1.1 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 (see section 3.3).

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 was defined as $PE = (R+NR)/(R+NR+IE)$. In other words, *PE* was the eligibility rate among households with known eligibility. Then, unweighted household response rates were calculated following AAPOR's Response Rate 4 (AAPOR, 2015):

Equation 8. Household response rate

$$\text{Household response rate} = 100 \times \frac{R}{R + NR + (PE \times UE)}$$

To obtain weighted household response rates, households were weighted using the household base weight *hhbwt0* which was not adjusted for household nonresponse.

6.1.2 Individual response rates

Individual response rates were based on individual eligibility and response status (see section 3.3).

Unweighted interview response rates were calculated by dividing the number of eligible respondents (*indstatus* =1) by the total number of eligible respondents (*indstatus* =1 or 2). To obtain weighted interview response rates, individuals were weighted using the interview base weight *indiv_bwt0* which was not adjusted for individual-level nonresponse.

Unweighted blood draw response rates were 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 were weighted using the trimmed, non-response adjusted individual weight *trmpnr1w0* which is a non-poststratified

weight not adjusted for the blood draw nonresponse. See each PHIA's **Supplement** for details on these additional survey weights.

6.1.3 Overall response rates

The overall response rate was used as a summary measure of response to the PHIA's, and was calculated using the unweighted household, interview, and blood draw response rate.

Equation 9. Overall response rate

$$\begin{aligned} \text{Overall response rate} \\ &= \text{Household response rate} \times \text{Interview response rate} \\ &\times \text{Blood draw response rate} \end{aligned}$$

6.2 Jackknife variance estimation

Though multiple existing variance estimation methods can appropriately account for PHIA's complex survey design, Jackknife (JK) repeated replication was a preferred method and was typically used for the **Summary Sheet** and **Final Report** for each PHIA. JK variance estimation involves the creation of replicate weights, where one primary sampling unit (PSU) was omitted from the analysis in each replicate. This general 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 was 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.

The PHIA's 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 (Westat, 2007). Variance estimates produced by the Jackknife method reflect nonresponse and poststratification adjustments since the replicate weights were based on the original final survey weights.

In order to implement the JK2 design in the PHIA data, variance estimation strata (*varstrat*) and clusters within those strata (*varunit*) have been created. See section 3.7.2.2 for further discussion of the construction of the variance estimation strata.

Users should note two technical considerations when using JK2 variance estimation methods in their analyses: (see example code in the attachments to this **Manual**).

- **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 et al (2013).
- **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 (Burns et al, 2003). 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 overstate the precision of confidence intervals. A conservative approach is recommended for stratified analysis, where the user should override the default df calculated by software and set df=25.

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

Figure 11. Naming conventions for JK weight variables

Level	Variable name	
	Survey weight	JK replicate weights
Household	hhwt0	hhwt001-hhwt...
Individual interview	intwt0	intwt001-intwt...
Blood test	btwt0	btwt001-btwt...

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

6.3 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). See section 3.7.2 for further discussion of the assignment of varstrat and varunit. Users will need to specify the variance stratum and unit variables and base weights appropriate for the analysis of interest.

Figure 12. Taylor Series weight variables

Level	Variable name		
	Survey weight	Sampling stratum	Sampling unit
Household	hhwt0	varstrat	varunit
Individual interview	intwt0	varstrat	varunit
Blood test	btwt0	varstrat	varunit

6.4. Multi-country analysis

Users may be interested in conducting analyses where data from multiple countries is combined to produce regional estimates. The following describes how to create a single file for multi-country analyses.

Note that the methods described below can also be applied to combine data from multiple survey rounds or to combine data from multiple countries and multiple survey rounds.

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

Equation 10. Multi-country estimate of a parameter

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

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

Equation 11. Multi-country estimate of a parameter as a linear combination of country estimates

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

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.

There are four alternative methods of variance estimation that can be used for multi-country analyses, presented here. The first method involves computing the multi-country estimates and their variances for each country separately, and then appropriately combining the results. 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 by concatenating (i.e., stacking) the files of each country in one single file, which is required for the other three methods of variance estimation. Once the data are concatenated, the second method of variance estimation is to use the Taylor Series linearization method and the variance strata (varstrat) and PSUs (varunit) recorded on the country-level data files, after re-numbering to make unique within each country, to produce estimates in the standard way.

The third and fourth method involve combining the country-level replicates and computing the combined variance 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 subpopulations.

There are two options for processing the replicate weights in the combined dataset; the first involves stacking the replicate weights for all countries side-by-side, whereas the second involves concatenating the replicate weights. These correspond to the third and fourth variance estimation method and are presented below.

6.4.1 Example using the side-by-side method

An illustrative example is described below with the aim of conducting a combined analysis of the Zambia (ZAMPHIA 2016), Zimbabwe (ZIMPHIA 2015-2015) and Malawi (MPHIA 2015-2016) 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.

Figure 13. Number of replicates by country

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 `btwtj` 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 **Figure 23**.

Figure 14. Assignment of combined weights in the side-by-side method

Country	Full sample weight	New replicate weights		
		<code>btwt001-btwt253</code>	<code>btwt254-btwt501</code>	<code>btwt502-btwt751</code>
Zambia	<code>btwt0</code>	<code>btwt001-btwt253</code>	<code>btwt0</code>	<code>btwt0</code>
Zimbabwe	<code>btwt0</code>	<code>btwt0</code>	<code>btwt001-btwt248</code>	<code>btwt0</code>
Malawi	<code>btwt0</code>	<code>btwt0</code>	<code>btwt0</code>	<code>btwt001-btwt250</code>

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.

6.4.2 Example using the concatenation method

This method is an alternative for the side-by-side method described in the previous example. This method addresses the main disadvantage of the side-by-side method that the number of replicate weights in the combined file becomes too large for fast processing and production of estimates when many countries are involved. The number of replicates with the concatenation

method is fixed at the largest number of replicates across the countries in the analyses. To illustrate the concatenation method for creating the analysis file for a multi-county analysis, consider the same three countries from the previous example using the blood test datasets.

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

1. Identify the country with the largest number of replicates, in this case, Zambia with 253 replicate weights.
2. Copy these replicates weight into the combined file, with the weights in the combined files being designated as cbtwt as below. The file for the combined analysis will have a full sample weight (cbtwt0) and 253 replicate weights cbtwt1-cbtwt253 for Zambia.

Figure 15. Replicate weights in the combined file:

Zambia replicate weights assignment

Weight	Description
cbtwt0 = bwt0	Zambia Full sample weight
cbtwt1 = bwt1	Zambia Replicate weight 1
cbtwt2 = bwt2	Zambia Replicate weight 2
...	
cbtwtj = bwtj	Zambia Replicate weight <i>j</i>
...	
cbtwt253 = bwt253	Zambia Replicate weight 253

3. Append the replicate weight for the next largest country in number of replicates, Malawi, to the combined replicate cbtwt1-cbtwt253. The full sample replicate weight is cbtwt0=bwt0 for Malawi. Each of the replicate weights bwt1-bwt250 is assigned randomly (without replacement) to the one of the 253 positions corresponding to the replicates cbtwt1-cbtwt253 in the combined file, as illustrated below.

**Figure 16. Replicate weights in the combined file:
Malawi replicate weights assignment**

Weight	Description
cbtwt0 = btwt0	Malawi Full sample weight
cbtwt1 = btwt200	Malawi Replicate weight 200
cbtwt2 = btwt34	Malawi Replicate weight 34
Cbtwt3	
Cbtwt4 = btwt124	Malawi Replicate weight 124
...	
cbtwt253 = btwt2	Malawi Replicate weight 2

4. As a result of this assignment, there are three replicates without an assigned replicate weight for Malawi. These “holes” or empty replicate weights are filled out with the full sample weight btwt0 for Malawi.
5. The same process is repeated for the last country Zimbabwe with 248 replicates. After assigning randomly 248 replicates from Zimbabwe to the 253 positions, there will be five holes or empty replicate weights. These five positions are filled with the Zimbabwe full sample weight btwt0.

With any of the two methods illustrated above, and using the combined country data file, analysis with JK variance estimation can proceed as usual. 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.

Lastly, researchers conducting multi-country analyses are strongly advised to consult each PHIA’s **Supplement** to ensure that differences in question wording and response options across surveys are understood prior to pooling data. 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 multi-country analyses.

6.5 New HIV infections and annual incidence

This section summarizes the methods used by PHIA to estimate HIV incidence and the expected number of new HIV cases that will occur per year.

PHIA used blood test results to determine whether HIV positive participants became infected within a specified time period prior to the survey. A specialized estimator was used to convert the number of people infected during this specified time period prior to the survey into a standardized annual incidence rate. The population at risk was calculated as the weighted number of HIV-negative people, using the survey HIV test results. These two figures were 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 methodological details used to identify recent HIV infections and calculate annual HIV incidence from PHIA data, accounting for the complex survey design.

Refer to each PHIA's Supplement for survey-specific details. The example statistical programs included in the publicly available data contains the program code used to estimate HIV incidence and the expected number of new HIV cases per year.

6.5.1 Definitions

The following definitions were used throughout this manual to explain the properties of the recent infection testing and were 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”. The term ‘recently’ is expressed in quotation marks as it refers to recency derived from the test, rather than true recency. In other words, it acknowledges the possibility of false recent infection results.
- Cutoff time (T) – a time period that is set with regard to the recency test being used. Ideally, T is set to the minimum value while simultaneously ensuring that as few participants 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.”

6.5.2 Identification of recently infected people

To distinguish recent from long-term HIV infections, the survey used a laboratory-based testing algorithm that employed a combination of assays: an HIV-1 LAg avidity assay, viral load, and ARV detection.

The PHIA's 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*).

A person whose measured LAg-Avidity EIA ODn ≤ 1.5 using plasma was classified as recently infected. For those specimens where plasma is not available for LAg testing, comparable testing was done on DBS specimens, using a cutoff of 1.0. The measured ODn value from the assay

increases over time as an HIV infection progresses. A study by Duong et al. using specimens with known times since infection found that, on average, ODn reaches a value of 1.5 130 days after HIV infection for subtype C, with a 95% CI 118-142 days. This characteristic time was called the Mean Duration of Recent Infection (MDRI). As MDRI varies according to HIV subtype, countries with atypical subtype distributions may warrant adjustment of the MDRI value. Country-specific guidance based on HIV subtype distribution can be found in each survey's **Supplement**. See Kassanjee et al. and Longosz et al. for more details on the effect of HIV subtype on the LAg-EIA MDRI assay.

Kassanjee et al. 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 (or ≤ 1.0 for DBS) must also have a viral load measured at ≥ 1000 copies/mL to be classified as recently infected.

6.5.3 Use of ARV test results

Some people with long term infections and who tested 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. Based on these findings, PHIA used an alternative recent infection algorithm that incorporates ARV blood test results. This alternative algorithm reclassified people who were recently infected according to the LAg and VL criteria as long-term cases if they tested positive for ARVs. Although most people reclassified in this way were expected to truly have long-term infections, some false non-recents could be introduced when people have started treatment immediately 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. 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.

6.5.4 Proportion of false recents

False recent results can inflate incidence estimates, but it was not possible to directly estimate the proportion of false recents (PFR) specific to each survey. The approach taken by PHIA to address this challenge was 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.

6.5.5 Incidence estimation

PHIA uses the following approach to estimate annual HIV incidence (Voetsch et al, 2021). Because the MDRI was less than one year, adjustments were 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:

Equation 12. Instantaneous incidence

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

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. N' is the adjusted number of HIV negative people in the sample, accounting for the possibility that not all HIV-positive participants are tested for recency.

Equation 13. Adjusted number of HIV negative people in the sample

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

In the equation above, N and P are the numbers of negative and positive participants in the sample. If all HIV-positive participants were tested for recency, $N' = N$. In situations where recency results were not available for a certain proportion of HIV-positive participants, the count of HIV-negative participants in the sample was 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:

Equation 14. Simplified instantaneous incidence estimator

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

This simplified estimator effectively scales up the number of recent cases as a proportion of the population at risk by a factor of 365/130 (or replacing 130 with the MDRI if differs for the country), or approximately 2.81, to calculate the instantaneous incidence rate. The annual incidence rate is calculated from the instantaneous incidence using:

Equation 15. Annual Incidence Rate

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

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

6.5.6 Detailed steps for incidence estimation

SAS, Stata, R programs to calculate point estimates and confidence intervals for incidence from PHIA data are provided with PHIA datasets. 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 .
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 were carried out for each age group and gender-specific sub-population required. Confidence intervals were calculated using the formulae in the appendices of Kassanjee et al. (2012). The standard deviation of the MDRI was an important parameter for calculating these confidence intervals. The values of the key parameters used in the estimation are as shown in Figure 17.

Figure 17. Values of key parameters for incidence estimation

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

6.5.7 Incidence variables included in the PHIA datasets

Provided in PHIA datasets were 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 and a 2 for long-term infections. The *recentlagvl* variable has the same set of values but uses only the LAg ODn and viral load results.

6.5.8 Accounting for the PHIA sample design in incidence estimation

6.5.8.1 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 incorporate these weighted figures directly

in the estimator (i.e. 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 dividing each weight by the average weight computed for 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 HIV-positive, recent, and HIV-negative status participants estimated using our full sample design and calibrated weights, normalized so that they sum to the actual number of people tested in each cell.

6.5.8.2 Design effect adjustment

In addition to using normalized weights, we also used the survey design effect to adjust the variance in cases where the sample design underperforms the simple random sample assumed by Kassanjee et al. 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 have taken a conservative approach which aims to avoid giving overly optimistic estimates of precision.

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

6.5.9 Estimation of the annual number of new infections

The number of new infections per year was 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 was a weighted total of the HIV negative people in the survey population. PHIA blood test weights were adjusted for non-response and post-stratified, so this weighted total was calibrated according to external population estimates from the national census or official projections. The methods used to calculate and adjust PHIA weights were described in detail in each PHIA's **Supplement**.

6.5.10 Confidence interval calculation for zero cells

PHIA calculates the upper confidence limit using the Clopper-Pearson binomial confidence interval approach for situations when HIV incidence is estimated as zero due to no estimated recently infected persons. The example statistical programs included in the publicly available data contains the code used to estimate HIV incidence upper confidence limits based on the Clopper-Pearson binomial confidence interval.

The PHIA estimation method was 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 PHIAs, the total number of participants classified as recently infected was approximately

30-40, and at least one recently infected person was identified in each required age by gender incidence estimation cell. However, in some PHIA, one or more of the age group by gender estimation cells had no recently infected persons, resulting in a degenerate confidence limit when variance was 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 (UCL) for a proportion when zero successes were observed in n trials was given by:

Equation 16. Clopper-Pearson UCL for a proportion

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

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

To apply the Clopper-Pearson equation, the PHIA estimation method uses the weighted number of HIV-negative participants 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 participants in the calculation. The upper confidence limit for the number of recent infections is calculated as:

Equation 17. Clopper-Pearson UCL for the number of recent infections

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

Finally, this upper limit for the number of recent cases in the cells was substituted into the standard incidence estimation equations (from earlier in this section)

Equation 18. Simplified instantaneous incidence estimator

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

And

Equation 19. Annual Incidence Rate

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

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.

6.5.11 Application to number of new infections

The number of new infections per year was another estimate of interest reported by PHIA. It was 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 was multiplied by the relative standard error of this estimated population.

7. References

- AAPOR. The American Association for Public Opinion Research. Standard Definitions: Final Dispositions of Case Codes and Outcome Rates for Surveys. 2016.
- Brown LD, Cai TT, DasGupta A. Interval estimation for a binomial proportion. *Statist Sci.* 2001;16(2):101-133.
- Clopper CJ, Pearson ES. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika.* 1934;26(4):404-413.
- 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.
- 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.
- 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.
- 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.
- 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.
- 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.
- 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.
- 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.
- 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.

Newcombe RG. Two-sided confidence intervals for the single proportion: comparison of seven methods. *Statist Med.* 1998;17:857-872.

Parekh B. Case for including ARV in recent infection testing algorithm (RITA). Paper presented at: WHO Incidence Meeting 2018.

Rutstein SO, Kiersten J. The DHS Wealth Index. Calverton, Maryland: ORC Macro; 2004.

Valliant R, Dever JA, Kreuter F. Practical tools for designing and weighting survey samples. Vol 51. New York, NY: Springer-Verlag; 2013.

Voetsch AC, Duong YT, Stupp P, et al. HIV-1 Recent Infection Testing Algorithm With Antiretroviral Drug Detection to Improve Accuracy of Incidence Estimates. *J Acquir Immune Defic Syndr.* 2021 Aug 1;87(Suppl 1):S73-S80. doi: 10.1097/QAI.0000000000002707. PMID: 34166315; PMCID: PMC8630595.

Westat. WesVar® User's Guide. 2007.

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.

8. Attachments

8.1 Tabulation plan for standard tables

PHIA 2 Data Manual Attachment 1 - Tabulation Plan

8.2 Guide to getting started with PHIA data

PHIA 2 Data Manual Attachment 2 - Guide to getting started with PHIA data