

Guide to getting started with PHIA Data

The purpose of this guide is to assist users who are new to PHIA data and to orient the user on how to get started with PHIA data. Before beginning any analyses, users are encouraged to review the available documentation, summary sheet, and final report for the country of interest. A brief description of the publicly available documentation is provided below.

- The **PHIA Data Manual** applies to PHIAs conducted in 2020 and after. It includes information such as the structure of PHIA data files, the types of variables included on the files, and statistical guidance specific to analyzing PHIA data.
- Each survey has its own survey-specific **Data Manual Supplement** with survey-specific details, including:
 - Questionnaires
 - Codebook
 - Flow diagrams for analytical variables
 - HIV testing methodology diagram
 - Sampling and Weighting Technical Report

The sections below discuss specific analytical tools and approaches to consider as one begins using the PHIA data.

1. Review sample code

Example code in SAS, Stata, and R are available as standalone statistical programs. The programs include code that supports the topic areas of:

Topic 1. Preparing the data for analysis:

1. Set up program environment
2. Read PHIA datasets
3. Recode, label, and format variables

Topic 2. Example analyses:

4. Estimate HIV prevalence overall and by gender using Jackknife variance estimation
5. Estimate HIV prevalence for subpopulations using Taylor Series variance estimation
6. Estimate the 90-90-90 cascade
7. Estimate annual incidence

Topic 3. Other data tips:

8. Merge datasets and using the survey status indicator flags

2. Replicate key findings

After receiving data for a PHIA, users should replicate key findings to ensure that their programs are properly reading and analyzing the data. Matching known numbers from the codebook, summary sheet, and final report ensures that programs are subsetting data properly and that weights are being used correctly. Matching numerators, denominators and weighted percentages is a good way to become familiar with the data and documentation.

For example, a user was interested in HIV prevalence among men and women aged 15 years and older in Zimbabwe from the [ZIMPHIA 2020 Summary Sheet](#).

Figure 1. Key Findings from the ZIMPHIA 2020 Summary Sheet

KEY FINDINGS						
HIV Indicator	Women	95% CI	Men	95% CI	Total	95% CI
Annual incidence (%)						
15-49 years	0.67	0.34-0.99	0.23	0.01-0.44	0.45	0.24-0.65
15 years and older	0.54	0.28-0.81	0.20	0.02-0.37	0.38	0.20-0.55
Prevalence (%)						
15-49 years	14.8	13.9-15.7	8.6	7.8-9.3	11.8	11.1-12.5
15 years and older	15.3	14.4-16.1	10.2	9.5-11.0	12.9	12.3-13.5
Viral load suppression (%)						
15-49 years	76.8	74.2-79.3	68.1	63.6-72.7	73.8	71.4-76.2
15 years and older	79.8	77.7-81.9	73.0	69.5-76.4	77.3	75.3-79.2

Viral load suppression is defined as HIV RNA <1,000 copies per milliliter among all HIV-positive adults.

A user who can match published results can confidently move to additional analyses knowing that their basic programming is correct.

First, the user should consult the tabulation plan attached to this **Manual** to identify the variables used in the calculation. HIV prevalence for adults aged 15+ is covered by Table 6.3 “HIV prevalence by age” in the tabulation plan. The datasets and variables used for table calculations are provided in the tabulation plan, directly under the table shell, as shown in Figure 2 below. Upon consulting the tabulation plan, the user will find that the variables used to calculate HIV prevalence for adults aged 15+ are: *hivstatusfinal*, *bt_status*, and *age* from the *adult biomarker* file (*adultbio*).

Figure 2. Datasets and variables used for Table 6.3

Datasets and variables used	
Dataset	adultbio
Subset	bt_status = 1
Analytic Variables	hivstatusfinal
Row stratification variables	agegroup age15_24, age15_49, age50_up, age15_up (age)
Column stratification variables	gender
Weight variables	btwt0, btwt001, ..., btwt[MAX]

Next, the user should review the survey-specific codebook with frequencies in the attachments of the **Supplement**. The codebook with frequencies (Figure 3) contains variable definitions and frequencies for each variable in the dataset. From the codebook we see that HIV positive individuals have *hivstatusfinal* = 1 and HIV negative individuals have *hivstatusfinal* = 2.

The user can consult the frequencies and percentages in this codebook for the variables of interest and check this against the dataset using the statistical software of their choice. This is a good way to make sure that one is using the correct dataset and that it has been read in accurately. For example, in Figure 4, the frequency for *hivstatusfinal* has been confirmed using PROC FREQ in SAS.

Figure 3. ZIMPHIA2020 Adult Biomarker Codebook

Variable Order	Variable Name	Variable Category	Question Text/ Variable Description	Variable Label	Variable Type and Width	Coding Values and Labels	Frequency	Percent
1	country	ID variable	Country Name	Country Name	Text (20)	Zimbabwe	19,535	100.0
2	year	ID variable	Survey Year	Survey Year	Numeric	2020	19,535	100.0
...
18	<u>hivstatusfinal</u>	Derived	Final HIV status determination after data cleaning and QC; can be missing because respondent did not have blood drawn, problems with the sample, or other issues	Final HIV status determination	Numeric	1 - HIV Positive 2 - HIV Negative 99 - Missing	2,958 16,577 .	15.1 84.9 .

Figure 4. SAS Code and Output: Unweighted frequency of hivstatusfinal on zimphia2020adultbio

```
proc freq data = zimphia2020adultbio;
  tables hivstatusfinal / list missing;
run;
```

The SAS System					
The FREQ Procedure					
Final HIV status determination					
hivstatusfinal	Frequency	Percent	Cumulative Frequency	Cumulative Percent	
1	2958	15.14	2958	15.14	
2	16577	84.86	19535	100.00	

Now that one has matched the unweighted counts and percentages from the data documentation, the next step is to take in to account the weights and sample design. Official PHIA reports use jackknife variance estimation under a JK2 design with a prespecified 25 degrees of freedom. Figure 5 shows sample code and output in SAS to generate the weighted HIV prevalence overall and by gender, with 95% confidence limits. Comparing back to Figure 1, you will see that the estimates overall and by gender match the summary sheet.

Figure 5. SAS Code and Output: Weighted HIV Prevalence and 95% Confidence Limits

```

proc surveymeans data = zimphia2020adultbio
    varmethod = jackknife mean clm nobs;
where bt_status = 1;
domain gender;
repweight btwt001-btwt175 / jkcoefs = 1 df = 25;
weight btwt0;
class hivstatusfinal;
var hivstatusfinal;
run;

```

The SAS System			
The SURVEYMEANS Procedure			
Data Summary			
Number of Observations		19535	
Sum of Weights		9495622	

Class Level Information			
Variable	Label	Levels	Values
hivstatusfinal	Final HIV status determination	2	1 2

Variance Estimation	
Method	Jackknife
Replicate Weights	ZIMPHIA2020ADULTBIO
Number of Replicates	175

Statistics							
Variable	Level	Label	N	Mean	Std Error of Mean	95% CL for Mean	
hivstatusfinal	1	Final HIV status determination	2958	0.129027	0.003095	0.12265172	0.13540159
	2	Final HIV status determination	16577	0.870973	0.003095	0.86459841	0.87734828

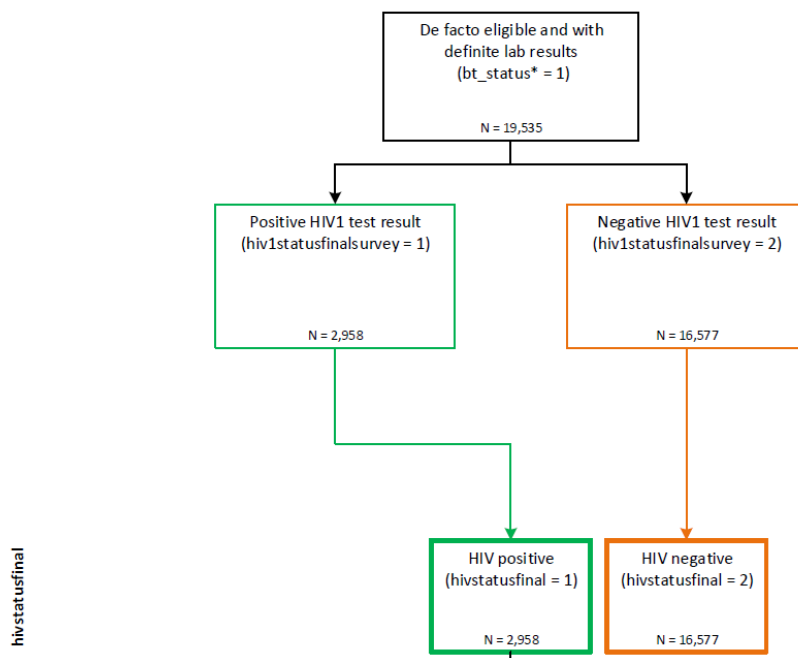
The SAS System							
The SURVEYMEANS Procedure							
Statistics for gender Domains							
gender	Variable	Level	Label	N	Mean	Std Error of Mean	95% CL for Mean
1	hivstatusfinal	1	Final HIV status determination	906	0.102303	0.003526	0.09504062 0.10956630
		2	Final HIV status determination	6758	0.897697	0.003526	0.89043370 0.90495938
2	hivstatusfinal	1	Final HIV status determination	2052	0.152660	0.003983	0.14445765 0.16086190
		2	Final HIV status determination	9819	0.847340	0.003983	0.83913810 0.85554235

Flow diagrams are useful to understand the construction of analytic variables. The user can consult the flow diagrams for selected analytic variables in the attachments of the **Supplement**. See Figure 6 for an example of *hivstatusfinal*.

Figure 6. ZIMPHIA 2020 flow diagram for *hivstatusfinal* analytic variable

Variables: *hivstatusfinal*, *vls*
Found in ZIMPHIA 2020 datasets:
Biomarker

SA3-4



The user should also review the individual or household questionnaire as needed. The questionnaire shows the order of survey questions, the name of the variable, the question text and coding labels, as well as skip patterns for determining what subsequent questions were asked based on the individual's response.

Finally, the user should consult this **Manual** for guidance on the use of survey weights. In this example, to calculate HIV prevalence using *hivstatusfinal* from the *adult biomarker* dataset, the user should use the biomarker weights.