

# Zimbabwe Population-based HIV Impact Assessment 2020

## ZIMPHIA 2020



## SAMPLING AND WEIGHTING TECHNICAL REPORT



The mark "CDC" is owned by the US Dept. of Health and Human Services and is used with permission. Use of this logo is not an endorsement by HHS or CDC of any particular product, service or enterprise.

# Zimbabwe Population-based HIV Impact Assessment 2020

## ZIMPHIA 2020

This project is supported by the US President's Emergency Plan for AIDS Relief (PEPFAR) through CDC under the terms of cooperative agreement #U2GGH002173. The findings and conclusions are those of the authors and do not necessarily represent the official position of the funding agencies.





## Table of Contents

---

1. Introduction .....	SA5-5
1.1 Overview of Sample Design .....	SA5-5
1.2 Overview of Weighting Process.....	SA5-5
2. Sample Design .....	SA5-7
2.1 Population of Inference .....	SA5-7
2.2 Precision Specifications and Assumptions .....	SA5-7
2.3 Selection of the Primary Sampling Units (PSUs) .....	SA5-9
2.3.1 Definition of PSUs .....	SA5-9
2.3.2 Selection of the PSU Sample .....	SA5-10
2.3.3 Out-of-Scope PSUs .....	SA5-10
2.3.4 Substitution .....	SA5-11
2.3.5 Segmentation .....	SA5-11
2.3.6 Summary of the PSU Sample .....	SA5-11
2.4 Selection of Households .....	SA5-12
2.4.1 Definition of Second-Stage Sampling Units .....	SA5-12
2.4.2 Listing .....	SA5-12
2.4.3 Determination of Eligibility for Sampling .....	SA5-12
2.4.4 Selection of Dwelling Units .....	SA5-14
2.4.5 Results of Second-Stage Sampling .....	SA5-16
2.5 Selection of Individuals .....	SA5-19
2.5.1 Household Rosters .....	SA5-19
2.5.2 Selecting Individuals for Data Collection .....	SA5-20
2.5.3 Distribution of Person Samples.....	SA5-21
3. Weighting and Estimation .....	SA5-23
3.1 Overview of the Weighting Process .....	SA5-23
3.2 Preparation for Weighting .....	SA5-24
3.2.1 Data Files for Weighting .....	SA5-24
3.2.2 Checks of Data Files .....	SA5-25
3.3 Creation of Variables for Variance Estimation.....	SA5-25
3.3.1 Jackknife Replication .....	SA5-26
3.3.2 Taylor's Series .....	SA5-27
3.4 Development of Weights .....	SA5-27
3.4.1 PSU Weights .....	SA5-27
3.4.2 Dwelling Unit/Household Weights .....	SA5-29
3.4.3 Person-Level Interview Weights .....	SA5-34
3.4.4 Person-Level Blood Test Weights.....	SA5-44

References .....	SA5-52
Appendix A - Definition of Eligibility for Dwelling Unit/Household Sampling .....	SA5-54
Appendix B - Definition of Household, Interview, and Blood Test Response Status ..	SA5-58
Appendix C - CHAID Trees and Definition of Final Nonresponse-Adjustment Weighting Cells .....	SA5-66

## Acronyms

---

CDC	US Centers for Disease Control and Prevention
CHAID	Chi-square Automatic Interaction Detector
CI	Confidence Interval
CV	Coefficient of Variation
DEFF	Design Effect
DHS	Demographic and Health Survey
DU	Dwelling Unit
EA	Enumeration Area
FTP	File Transfer Protocol
HH	Household
HIV	Human Immunodeficiency Virus
ICC	Intra Cluster Correlation
LASSO	Least Absolute Shrinkage and Selection Operator
MDRI	Mean Duration of Recent Infection
MOS	Measure of Size
PHIA	Population-based HIV Impact Assessment
PEPFAR	President's Emergency Plan for AIDS Relief
PSU	Primary Sampling Unit
RSE	Relative Standard Error
SAS	Statistical Analysis System
UEW	Unequal Weighting
UNAIDS	Joint United Nations Programme on HIV and AIDS
USAID	United States Agency for International Development
VLS	Viral Load Suppression
WHO	World Health Organization
WLM	Weighted Log linear Modeling
ZIMPHIA	Zimbabwe Population-based HIV Impact Assessment
ZIMSTAT	Zimbabwe National Statistics Agency

## 1. Introduction

---

The 2020 Zimbabwe Population-based HIV Impact Assessment (ZIMPHIA 2020) is a cross-sectional sample survey designed to assess the prevalence of key human immunodeficiency virus (HIV)-related health indicators among individuals 15 years or older. Data collection for the ZIMPHIA 2020 was conducted between November 2019 and March 2020 with almost 21,000 interviewed individuals including 19,500 individuals with valid blood tests in approximately 10,500 randomly-selected households. The purpose of this report is to document the procedures used to select the households and individuals for the study and the subsequent weighting of the respondent sample.

### 1.1 Overview of Sample Design

The sample design for the ZIMPHIA 2020 is a stratified multistage probability sample design, with strata defined to be the 10 provinces of the country, first-stage sampling units defined by enumeration areas (EAs) within strata, second-stage sampling units defined by households within EAs, and finally age-eligible persons within households. Within each sampling stratum, the first-stage sampling units (also referred to as “primary sampling units” or PSUs) were selected with probabilities proportionate to the estimated number of households in the PSU based on updated information available for 2017. The allocation of the sample PSUs to the 10 provinces was made in a manner designed to achieve specified precision levels for (a) national estimate of HIV incidence among persons 15-49 years of age; and (b) provincial estimates of viral load suppression (VLS) rates among HIV-positive persons 15-49 years of age.

The second-stage sampling units were selected from lists of dwelling units/households compiled by trained staff for each of the sampled PSUs. Upon completion of the listing process, random samples of specified numbers of dwelling units/households were selected from each PSU.

Within the sampled households, all eligible persons 15 years of age and older who were present in the household on the night prior to the interview were included in the study sample for PHIA data collection.

Details of sample design employed for the ZIMPHIA 2020 are provided in Section 2.

### 1.2 Overview of Weighting Process

The purpose of weighting survey data from a complex sample design is to (1) compensate for variable probabilities of selection, (2) account for differential nonresponse rates across relevant subsets of the sample, and (3) adjust for possible undercoverage of certain population groups. Weighting is accomplished by assigning an appropriate sampling weight to each responding sampled unit (e.g., a household or person), and using that weight to calculate weighted estimates from the sample.

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 probabilities of selecting individuals within households, compensate for nonresponse to the blood test, and adjust for potential undercoverage through poststratification.

Technical details of the weighting procedures employed for the ZIMPHIA 2020 are provided in Section 3.



## 2. Sample Design

### 2.1 Population of Inference

The population of inference for the ZIMPHIA 2020 is comprised of individuals 15 years of age and older who were present in households (i.e., “slept in the household”) on the night prior to the date of interview. This population is referred to as the de facto population. In contrast, those individuals who are usual residents of the household regardless of whether they were present in the household during the previous night comprise the de jure population. Individuals belonging to either the de facto or de jure populations were included on the rosters compiled for sampling purposes; however, only members of the de facto population were eligible for data collection. Table 2-1 summarizes estimates (projections) of the de facto population in Zimbabwe in 2020 by gender and age group.

**Table 2-1 2020 population estimates for Zimbabwe by gender and age group**

Age group	Gender		Total
	Male	Female	
15-49 years	3,756,582	4,079,456	7,836,038
50 years or older	699,884	959,700	1,659,584
Total	4,456,466	5,039,156	9,495,622

Source: Updated 2020 population projections provided by Zimbabwe National Statistics Agency (ZIMSTAT)

### 2.2 Precision Specifications and Assumptions

The following specifications and assumptions were used to develop the sample design for the ZIMPHIA 2020.

#### Specifications

- Relative standard error (RSE) of the national estimate of HIV incidence among persons 15-49 years old should be 30% or less.
- 95% confidence interval (CI) bounds around the estimated VLS rate among HIV positive adults aged 15-49 years for each of the 10 provinces should be  $\pm 0.08$  or less.

#### Statistical Assumptions

- A national HIV prevalence rate of 0.134 (13.4%) for adults 15-49 years old that varies by province (e.g., see Table 2-2). Source: 2015-16 Zimbabwe PHIA (ZIMPHIA 2015-2016).
- An annual national incidence rate for adults aged 15-49 of  $p_a = 0.0044$  (0.44%). Source: ZIMPHIA 2015-2016.
- Stratum-level (provincial) incidence rates of  $p_{ah}$ ,  $h = 0, 2, \dots, 9$ , which are obtained by adjusting the national incidence rate using the provincial prevalence rates as follows:  
$$p_{ah} = (p_h/p) p_a,$$
where  $p_h$  and  $p$  are the HIV prevalence rates for province  $h$  and the country, respectively, and  $p_a$  is the annual national incidence rate obtained from ZIMPHIA 2015-2016.
- A mean duration of recent infection (MDRI) of 130 days, yielding an annualization rate of  $365/130 = 2.8077$ .

- Hence, an estimated incidence rate for MDRI=130 days of  $p_m = 0.0044/2.8077 = 0.0016$  (0.16%). The corresponding provincial estimates are obtained by  $p_{mh} = p_{ah}/2.8077$ .
- A viral load suppression rate among HIV positive adults aged 15-49 of  $p_{VLS} = 0.50$  (50%) in each province. This assumption provides a conservative estimate of the underlying population variance associated with VLS rate.
- An intracluster correlation (ICC) of 0.05 for VLS and 0.01 for prevalence. Source: tabulations of ZIMPHIA 2015-2016 data.
- An intracluster correlation (ICC) of 0.000 for incidence. Source: analyses of prior PHIA surveys.
- Overall sex-age distributions derived from the ZIMPHIA 2015-2016.
- Stratum-level (provincial) population projections for 2020 obtained from the 2015 ZIMSTAT Population Projections Thematic Report.

### Operational Assumptions

- Varying number of dwelling units to be sampled per PSU, resulting in an average of 35 sampled dwelling units per PSU.
- An overall occupancy rate of 93.2% for the sampled dwelling units (source: ZIMPHIA 2015-2016).
- A household response rate of 83.7% among occupied dwelling units (source: ZIMPHIA 2015-2016).
- An average household size of 3.95 (de facto) persons per household (source: ZIMPHIA 2015-2016). The de facto population consists of persons of all ages who were present in the household during the night prior to the interview.
- An average of 1.86 de facto persons 15-49 years of age per household (source: ZIMPHIA 2015-2016).
- An average of 0.49 de facto persons 50+ years of age per household (source: ZIMPHIA 2015-2016).
- Within the responding households, a person-level interview response rate of 88.9% (source: ZIMPHIA 2015-2016).
- Among de facto persons 15+ years of age completing the interview, a blood test response rate of 91.4%. Thus, the overall response rate for the blood tests is  $88.9\% * 91.4\% = 81.2\%$  (source: ZIMPHIA 2015-2016).

Based on the specifications and assumptions listed above, a sample of 356 EAs (clusters) was determined to be the minimum needed to meet the specified precision goals. The allocation of the sample to the 10 provinces of Zimbabwe is shown in Table 2-2. The expected numbers of households included in the study and the corresponding projected numbers of respondents by age group are also summarized in this table. The actual numbers of respondents achieved are presented in Sections 2.4 and 2.5 and differ from the counts in Table 2-2 because of differences between the response rates and other assumptions used to develop the sample design and those achieved during data collection. Further details about the sampling of households are given in Section 2.4.

**Table 2-2 Allocation of sample clusters (EAs) and dwelling units and projected sample sizes (expected number of respondents) by province**

Province code	Province name	HIV+ prevalence rate <sup>[1]</sup>	Total no sample clusters	Target no. of DUs to be sampled	Projected no. of participating households <sup>[2]</sup>	Projected no. of respondents <sup>[3]</sup>	
						Adults 15-49	Adults 50+
0	Bulawayo	0.165	31	1,085	846	1,279	340
1	Manicaland	0.102	44	1,540	1,201	1,816	483
2	Mashonaland Central	0.129	37	1,295	1,010	1,527	406
3	Mashonaland East	0.129	37	1,295	1,010	1,527	406
4	Mashonaland West	0.119	40	1,400	1,092	1,651	439
5	Matabeleland North	0.188	29	1,015	792	1,197	318
6	Matabeleland South	0.204	27	945	737	1,114	296
7	Midlands	0.127	38	1,330	1,038	1,568	417
8	Masvingo	0.137	36	1,260	983	1,486	395
9	Harare	0.13	37	1,295	1,010	1,527	406
All		0.134	356	12,460	9,720	14,690	3,906

[1] Source: 2015-16 Zimbabwe PHIA.

[2] Assumes occupancy rate of 93.2% and household response rate of 83.7%.

[3] Projected numbers of individuals providing valid blood draw based on assumptions used to develop the sample design.

## 2.3 Selection of the Primary Sampling Units (PSUs)

### 2.3.1 Definition of PSUs

The first-stage or primary sampling units (PSUs) for the ZIMPHIA 2020 were selected from a sampling frame of enumeration areas (EAs) that originally had been created for the 2012 Zimbabwe Population Census, and subsequently updated by the Zimbabwe National Statistics Agency (ZIMSTAT) in 2017. The enumeration areas in the updated sampling frame were generally the same as those created for the 2012 Population Census, except that some EAs that had grown appreciably in population by 2017 were subdivided into two or more separate EAs. In addition, a small number of EAs in Manicaland province (accounting for an estimated 0.50 % of the households in the province) that had been devastated by Cyclone Idai in 2019 were deleted from the sampling frame. The updated sampling frame consisted of slightly over 30,600 EAs containing an estimated 3.1 million households as of 2017.

### **2.3.2 Selection of the PSU Sample**

A stratified sample of 356 EAs was selected from the EA sampling frame in accordance with the sample allocation given in Table 2-2. To avoid re-selecting the same EAs that had been selected for the ZIMPHIA 2015-2016, the following procedure was used to select EAs for the ZIMPHIA 2020. Within each province, the EAs in the updated sampling were sorted in the same way they had been sorted in the ZIMPHIA 2015-2016 frame to the extent feasible; i.e., by urban/rural status, district within urban/rural status, and finally by ward within district. Since the EAs in the updated frame were defined somewhat differently from those in the original frame, the resulting ordering of the EAs approximated (but did not replicate exactly) the ordering that was used to select the EA sample for ZIMPHIA 2015-2016. The sorting of EAs prior to sample selection induces an implicit geographic substratification within each province.

Next, a systematic sample of the same number of EAs selected for the ZIMPHIA 2015-2016 was selected from the given province using a random starting point that was offset by a specified amount to minimize selecting EAs that had been selected for the ZIMPHIA 2015-2016, and an adjusted sampling interval that reflected the change in measure of size (number of households) between the original and updated sampling frames. The EAs were selected with probabilities proportionate to a measure of size (MOS) equal to the estimated number of households in the EA in 2017. To select the sample from a given province, the cumulative MOS was determined for each EA in the ordered list of EAs, and the sample selections were designated using the specified random start and a sampling interval equal to the total MOS of the EAs in the province divided by the number of EAs to be selected. The resulting sample has the property that the probability of selecting an EA within a province is proportional to the MOS of the EA.

Since the number of EAs required for the ZIMPHIA 2020 (see Table 2-2) was less than that specified for the ZIMPHIA 2015-2016 for every province, the final step was to select an equal-probability systematic sample of the desired number of EAs from the set of initially-selected EAs. Of the 356 sampled EAs, only three had been selected previously for the ZIMPHIA 2015-2016. Each of the three EAs was replaced by another EA of roughly the same size using guidelines developed for PHIA.

### **2.3.3 Out-of-Scope PSUs**

Out-of-scope PSUs are defined to be those EAs with no dwelling units (e.g., EAs that are no longer occupied due to flooding or other natural disasters, or where all residents have been permanently relocated). These are also sometimes referred to as “empty” PSUs. There were no out-of-scope PSUs in the ZIMPHIA 2020 sample.

#### 2.3.4 Substitution

One sampled PSU in Masvingo province that was confirmed to contain eligible dwelling units could not be entered for security reasons. This PSU was replaced by a PSU in the same general area following guidelines developed for PHIA.

#### 2.3.5 Segmentation

Of the 356 sampled PSUs that were included in data collection, four were considered to be too large for subsequent listing activities (see Section 2.4.2). These were generally (but not always) EAs with 300 or more households, where the size cutoff for segmentation could vary depending on local conditions such as the land area of the EA. Thus, these four EAs underwent another stage of sampling in which (a) the EA was subdivided into a specified number of segments of manageable size, (b) a rough measure of size was assigned to each defined segment, and (c) one segment was randomly selected with probability proportionate to the rough measure of size. The segmentation procedures are described in the listing manual developed for the ZIMPHIA 2020.

#### 2.3.6 Summary of the PSU Sample

As indicated in the previous sections, 356 PSUs (EAs) were selected for the ZIMPHIA 2020. Of these, three were found to have been selected previously for the ZIMPHIA 2015-2016, and were replaced to avoid going back to the same PSUs that had been surveyed earlier. Of the 356 PSUs included in the ZIMPHIA 2020 data collection, one eligible PSU was replaced for security reasons, and four were segmented because they were too large to be canvassed efficiently. There were no out-of-scope (ineligible) PSUs. Table 2-3 summarizes the distribution of the sampled PSUs by province and sampling status of the PSU.

**Table 2-3 Distribution of sample PSUs by province and PSU sampling status**

Province code	Province name	Sample PSUs	PSUs replaced due to overlap with ZIMPHIA 2015-2016	Eligible PSUs replaced for other reasons	Number of ineligible PSUs	Number of segmented PSUs	Number of in-scope PSUs included in study <sup>[1]</sup>
0	Bulawayo	31	0	0	0	1	31
1	Manicaland	44	0	0	0	0	44
2	Mashonaland Central	37	0	0	0	0	37
3	Mashonaland East	37	1	0	0	2	37
4	Mashonaland West	40	1	0	0	1	40
5	Matabeleland North	29	1	0	0	0	29
6	Matabeleland South	27	0	0	0	0	27
7	Midlands	38	0	0	0	0	38

8	Masvingo	36	0	1	0	0	36
9	Harare	37	0	0	0	0	37
All	Zimbabwe	356	3	1	0	4	356

[1] Includes a PSU in Harare for which households were listed and sampled, but for which data collection was not conducted (see Section 3.4.2.2).

## 2.4 Selection of Households

The selection of households for the ZIMPHIA 2020 involved the following steps: (1) listing all potentially eligible dwelling units/households within the sampled EAs, (2) assigning eligibility codes to the listed dwelling unit/household records based on characteristics of the listed units, and (3) selecting the sample of dwelling units/households from those records determined to be eligible for selection.

### 2.4.1 Definition of Second-Stage Sampling Units

For both sampling and analysis purposes, a household is defined to be a group of individuals who reside in a physical structure such as a house, apartment, compound, or homestead, and share in housekeeping arrangements. The physical structure in which people reside is referred to as the “dwelling unit” which may contain more than one household meeting the above definition. Households are eligible for participation in the study if they are located within the sampled enumeration area (EA).

### 2.4.2 Listing

In essence, the listing process involves compiling complete, up-to-date, and accurate lists of all dwelling units and households for each sampled EA through a field operation using trained staff referred to as “listers.” Local leaders and knowledgeable community members were consulted to assist in the listing process. Listers were provided with maps from which to delineate the boundaries of the EA, and to record the locations of the dwelling units/households found by the listers in the field. Information about the listed dwelling units/households was entered into computer tablets. The information recorded in the tablets included the address or description of the listed dwelling unit/household, the name of the head of household, the type of structure (house, apartment, compound, etc.), occupancy status, and GPS coordinates. Vacant structures were listed along with households in occupied dwelling units. Slightly over 75,000 eligible dwelling units/households were listed for the ZIMPHIA 2020.

### 2.4.3 Determination of Eligibility for Sampling

As indicated above, all known households at the time of listing, plus vacant dwelling units that could potentially be occupied at the time of interview, were initially entered into the tablets as separate records. However, not all of these records were eligible for subsequent sampling purposes. Those records marked with the notation “discard” were data entry errors and were eliminated from the final listing file. To establish

eligibility for the remaining records, three key variables collected during listing were used: (1) the structure type, (2) whether the listed structure was vacant or under construction, and (3) whether anyone was living in the structure at the time of listing. Based on the values of these three variables, those records meeting the criteria specified in Appendix A were eligible for household sampling. Table 2-4 summarizes the total number of records entered into the tablets, the numbers of unoccupied and occupied dwelling units eligible for sampling, and the total number of dwelling units/households (records) eligible for sampling.

Table 2-4 Distribution of records in listing file by type of record, eligibility status, and province

Province code	Province name	Number of inscope PSUs included in study <sup>[1]</sup>	Number of records (DUs) in listing file <sup>[2]</sup>	Number of unoccupied DUs eligible for sampling <sup>[3]</sup>	Number of occupied DUs eligible for sampling <sup>[4]</sup>	Total number of DUs/households eligible for sampling
0	Bulawayo	31	3,247	50	3,197	3,247
1	Manicaland	44	5,403	83	5,319	5,402
2	Mashonaland Central	37	3,902	101	3,801	3,902
3	Mashonaland East	37	4,057	200	3,857	4,057
4	Mashonaland West	40	4,651	54	4,597	4,651
5	Matabeleland North	29	3,204	127	3,077	3,204
6	Matabeleland South	27	3,118	70	3,047	3,117
7	Midlands	38	4,181	87	4,094	4,181
8	Masvingo	36	3,959	165	3,794	3,959
9	Harare	37	4,220	33	4,187	4,220
All	Zimbabwe	356	39,942	970	38,970	39,940

[1] Includes a PSU in Harare for which households were listed and sampled, but data collection was not conducted (see Section 3.4.2.2).

[2] See Appendix A for additional details.

[3] Records coded as vacant, under construction, or with no residents at time of listing.

[4] All records not coded as vacant, under construction, or with no residents at the time of listing.

#### 2.4.4 Selection of Dwelling Units

In order to achieve equal-probability samples of dwelling units within each of the five sampling strata, the sampling rates required to select dwelling units within a PSU (i.e., EA or segment) will depend on the difference between the size measure used in sampling (i.e., the estimated number of households in the PSU based on the most recent census projections) and the actual number of dwelling units/households found at the time of listing in late 2019. Thus, application of these within-PSU sampling rates can yield more than the desired number households in PSUs that have experienced growth in population since the latest census projections, and fewer than the desired number of households in PSUs that have declined in population.

The calculation of the required within-PSU sampling rates proceeded as follows. First, the target overall sampling rate for province  $h = 0, 2, \dots, 9$ , was computed as:

$$F_h^{overall} = T_h / \sum_{i=1}^{m_h} (N_{hi} / P_{hi}),$$

where

$T_h$  = target sample size for province  $h$  given in Table 2-2 ;



$m_h$	=	number of sample PSUs in province h ;
$N_{hi}$	=	number of eligible dwelling units in PSU i in province h based on listing counts;
$P_{hi}$	=	probability of selecting PSU i in province h .

Note that for those PSUs in which the segmentation process described in Section 2.3.5 was implemented,  $P_{hi}$  is equal to the overall probability of selecting the segment (cluster) within the province, i.e., the product of the probability of selecting the EA and the conditional probability of selecting the segment within the EA.

The total expected number of listings to be selected across all 10 provinces is  $\sum_{h=0}^9 T_h = 12,460$  (see Table 2-2). To obtain an equal probability sample within province h, the required within-PSU sampling rate for PSU i in province h was then computed as:

$$f_{hi}^{within} = F_h^{overall} / P_{hi}.$$

and the corresponding expected sample size for PSU i in stratum h was computed as:

$$E(n_{hi}) = N_{hi} f_{hi}^{within}.$$

Inspection of the values of  $E(n_{hi})$  indicated that there would be unduly large workloads in some PSUs and very small workloads in others. To reduce the variation in workload across the sampled PSUs, the maximum number of dwelling units to be selected in any PSU was capped at 70 except for one PSU where the difference between the sampling measure of size and the actual confirmed listing count was so great that the sample size for this PSU was set to 210. In addition, the minimum number of dwelling units to be selected in any PSU was set to a value equal to the lesser of 15 and the number of listed units in the PSU. The difference between the number of dwelling units that would have been selected using the rates,  $f_{hi}^{within}$ , and the specified maximum and minimum numbers was then re-distributed to the other PSUs in the same province so as to maintain as closely as possible the desired total sample size for the province. The within-PSU sampling rates,  $f_{hi}^{within}$ , were therefore adjusted to reflect the redistribution of the sample within the stratum. The adjusted within-PSU sampling rate used to select the sample of dwelling units,  $f_{hi}^{adj(w)}$ , was calculated as:

$$f_{hi}^{adj(w)} = A_{hi} f_{hi}^{within},$$

where the adjustment factors,  $A_{hi}$ , were determined such that  $L \leq N_{hi} A_{hi} f_{hi}^{within} \leq U$ ,  $L$  = the minimum PSU sample size,  $U$  = the maximum PSU sample size, and  $\sum_{i=1}^{m_h} A_{hi} f_{hi}^{within} = T_h$ .

To achieve a geographical ordering of the listed dwelling units, the dwelling unit records in each PSU were sorted by a proximity variable that indicated the distance between the listed dwelling unit and the dwelling

unit closest to the centroid of the PSU. Dwelling units/households within the EA were then selected systematically from the ordered list of records at the rates,  $f_{hi}^{adj(w)}$ , specified above.

#### **2.4.5 Results of Second-Stage Sampling**

Table 2-5 summarizes the number of PSUs and dwelling units/households selected for the study, the minimum and maximum PSU sample size, and the weighted count of the sampled DUs/households by province. The last column shows the unequal weighting (UEW) design effects (DEFF) to be expected for the selected sample. The UEW design effect provides a measure of the increase in the variance of a sample-based estimate resulting from the use of variable sampling fractions within a province (e.g., see Kish, 1965, page 403). With an equal probability sample within each province, the design effects would ordinarily equal 1.0. However, with the capping and redistribution of the sample described previously, the overall sampling rates (and, hence, household weights) varied to some extent within a province. As indicated in Table 2-5, this variation in sampling rates is expected to result in UEW design effects exceeding 1.00 for three provinces.

**Table 2-5 Number of sampled dwelling units/households and expected unequal weighting design effects by province**

Province code	Province name	Number of PSUs <sup>[1]</sup>	Number of sampled DUs/households	Minimum number of DUs selected per PSU	Maximum number of DUs selected per PSU	Weighted count of sampled DUs/households <sup>[2]</sup>	Unequal weighting design effect
0	Bulawayo	31	1,085	27	70	183,218	1.00
1	Manicaland	44	1,540	17	60	488,837	1.00
2	Mashonaland Central	37	1,295	23	70	305,492	1.00
3	Mashonaland East	37	1,295	15	210	529,882	1.09
4	Mashonaland West	40	1,400	15	70	430,949	1.03
5	Matabeleland North	29	1,015	15	57	176,432	1.00
6	Matabeleland South	27	945	22	59	173,875	1.00
7	Midlands	38	1,330	22	50	388,645	1.00
8	Masvingo	36	1,260	15	63	348,703	1.01
9	Harare	37	1,295	20	70	540,276	1.00
All	Zimbabwe	356	12,460	15	210	3,566,309	1.11 <sup>[3]</sup>

[1] The number of eligible PSUs that were fielded for listing. Includes a PSU in Harare for which data collection was not conducted (see Section 3.4.2.2).

[2] Weight is the reciprocal of the product of the PSU selection probability and the within-PSU sampling rate used to select DUs/households.

[3] Overall DEFF reflects total variation in weights within and across provinces.

Table 2-6 summarizes the distribution of the sampled dwelling units/households by final household response status. Of the 12,460 sampled dwelling units 675 (5.4%) were determined during data collection to be vacant/unoccupied, 78 (0.6%) for which eligibility for the survey (i.e., occupancy status) could not be established, 1,208 (9.7%) were determined to be eligible for the study (i.e., contained household members) but did not complete the household interview, and 10,499 (84.3%) completed the household interview. The overall unweighted household response rate was 89.1%.

**Table 2-6 Distribution of dwelling unit sample by province and response status**

Province code	Province name	Number of sampled DUs	Number of ineligible DUs/households <sup>[1]</sup>	Number of DUs with unknown eligibility <sup>[2]</sup>	Number of responding households <sup>[3]</sup>	Number of eligible non-responding households <sup>[4]</sup>	Unweighted response rate <sup>[5]</sup>
0	Bulawayo	1,085	44	0	917	124	0.881
1	Manicaland	1,540	76	3	1,303	158	0.890
2	Mashonaland Central	1,295	70	4	1,118	103	0.913
3	Mashonaland East	1,295	80	7	1,081	127	0.890
4	Mashonaland West	1,400	82	4	1,218	96	0.924
5	Matabeleland North	1,015	48	1	867	99	0.897
6	Matabeleland South	945	41	4	822	78	0.909
7	Midlands	1,330	76	3	1,138	113	0.908
8	Masvingo	1,260	64	9	1,066	121	0.892
9	Harare	1,295	94	43	969	189	0.809
All	Zimbabwe	12,460	675	78	10,499	1,208	0.891

[1] Vacant dwelling units or nonresidential units as determined during data collection.

[2] Unoccupied dwelling units for which eligibility for PHIA could not be ascertained.

[3] Households completing the household interview. Excludes a PSU in Harare for which no household interviews were obtained (see Section 3.4.2.2).

[4] Occupied dwelling units that did not complete the household interview.

[5] Computed as  $R / [R + N + U \cdot \{(R + N) / (R + N + I)\}]$ , where R = number of households completing interview; N = number of eligible nonresponding households; I = number of ineligible DUs, and U = number of DUs with unknown eligibility.

## **2.5 Selection of Individuals**

The selection of individuals for the ZIMPHIA 2020 involved the following steps: (1) compiling a list of all individuals known to reside in the household or who slept in the household during the night prior to data collection; (2) identifying those rostered individuals who are eligible for data collection; and (3) selecting for the study those individuals meeting the age and residency requirements of the study. As noted below, only those individuals who were present (i.e., slept) in the household on the night prior to the time the household roster was compiled (i.e., the de facto population) were eligible for data collection and retained for subsequent weighting and analysis.

### **2.5.1 Household Rosters**

A comprehensive list (roster) of all household members was compiled during the administration of the household interview. Included on the roster were all persons who were present in the household during the night prior to the interview, along with other individuals who are usual residents of the household but were not present during that time. The information recorded for each rostered individual included sex, age, relationship to head of household, residency status (i.e., whether a usual resident), and physical presence in household (i.e., slept in household the night prior to interview). Table 2-7 summarizes the number of households completing the roster and the corresponding number of rostered individuals by province and resident status.

**Table 2-7 Distribution of households completing rosters and corresponding numbers of rostered persons by resident status and province**

Province code	Province name	Number of households completing interview	Rostered persons by resident status <sup>[1]</sup>			
			Usual resident/did not sleep here	Usual resident/slept here	Nonresident/slept here	Total rostered persons
0	Bulawayo	917	354	2,826	71	3,288
1	Manicaland	1,303	520	4,919	128	5,712
2	Mashonaland Central	1,118	436	4,166	129	4,794
3	Mashonaland East	1,081	395	3,610	61	4,152
4	Mashonaland West	1,218	584	4,258	120	5,047
5	Matabeleland North	867	351	3,651	106	4,182
6	Matabeleland South	822	290	2,960	89	3,384
7	Midlands	1,138	442	4,390	67	4,978
8	Masvingo	1,066	448	3,890	136	4,591
9	Harare	969	415	2,750	49	3,275
All	Zimbabwe	10,499	4,235	37,420	956	43,403

[1] Counts include rostered persons of all ages in the 10,499 responding households. There were two sampled households that provided roster information for 12 individuals but for which the household questionnaire was not completed. These households and associated individuals are not included in this table and will be excluded from the nonresponse adjustment weighting process described in Section 3.

## 2.5.2 Selecting Individuals for Data Collection

All individuals listed in the household rosters who were 15 years of age and older and slept in the household on the night prior to the household interview were eligible for data collection. Table 2-8 summarizes the number of individuals eligible for data collection by province, age group, and resident status.

**Table 2-8**      **Number of individuals eligible for data collection in responding households**

Provin ce code	Province name	Persons 15-49 years <sup>[1]</sup>			Persons 50 years or older <sup>[1]</sup>		
		Usual residen t/slept here	Nonreside nt/slept here	Total sampled persons <sup>[2]</sup>	Usual residen t/slept here	Nonreside nt/slept here	Total sampled persons <sup>[2]</sup>
0	Bulawayo	1,530	49	1,579	362	13	375
1	Manicaland	2,102	72	2,174	653	13	666
2	Mashonaland Central	1,897	69	1,966	487	13	500
3	Mashonaland East	1,624	39	1,663	492	12	504
4	Mashonaland West	1,973	65	2,038	579	9	588
5	Matabeleland North	1,505	48	1,553	528	11	539
6	Matabeleland South	1,285	54	1,339	526	12	538
7	Midlands	1,967	49	2,016	575	6	581
8	Masvingo	1,645	81	1,726	606	13	619
9	Harare	1,487	40	1,527	267	4	271
All	Zimbabwe	17,015	566	17,581	5,075	106	5,181

[1] Age recorded in roster. In a small number of cases, the actual age at interview may be different.

[2] Eligible persons selected for data collection based on information reported in roster.

### 2.5.3      **Distribution of Person Samples**

Table 2-9 summarizes the number of individuals selected for data collection and the corresponding numbers completing the interview and blood test by age group and province. Note that the age classification in this table is based on rostered age. Interview respondents are those persons who met the criteria for completing the individual interview. Among the interview respondents, the blood test respondents are those persons with a final HIV status determination. The criteria used to define the interview and blood test respondents are given in Appendix B.

Table 2-9 Distribution of sampled persons in responding households by age group, response status, and province

Province code	Province name	Persons 15-49 years <sup>[1]</sup>			Persons 50 years or older <sup>[1]</sup>		
		Selected for data collection	Interview respondents <sup>[2]</sup>	Blood test respondent <sup>[3]</sup>	Selected for data collection	Interview respondents <sup>[2]</sup>	Blood test respondent <sup>[3]</sup>
0	Bulawayo	1,579	1,438	1,361	375	333	299
1	Manicaland	2,174	1,956	1,811	666	597	549
2	Mashonaland Central	1,966	1,811	1,685	500	463	442
3	Mashonaland East	1,663	1,510	1,375	504	469	432
4	Mashonaland West	2,038	1,927	1,834	588	554	530
5	Matabeleland North	1,553	1,356	1,282	539	490	471
6	Matabeleland South	1,339	1,220	1,118	538	505	481
7	Midlands	2,016	1,807	1,747	581	541	521
8	Masvingo	1,726	1,611	1,536	619	567	550
9	Harare	1,527	1,390	1,285	271	248	226
All	Zimbabwe	17,581	16,026	15,034	5,181	4,767	4,501

[1] Age recorded in household roster. In a small number of instances, the actual confirmed age at interview may be different.

[2] Persons who completed all relevant modules of the individual interview (see Appendix B.2).

[3] Subset of interview respondents with confirmed results of blood tests (see Appendix B.3).



### 3. Weighting and Estimation

---

In general, the purpose of weighting survey data from a complex sample design is to (1) compensate for variable probabilities of selection, (2) account for differential nonresponse rates within relevant subsets of the sample, and (3) adjust for possible undercoverage of certain population groups. Weighting is accomplished by computing an appropriate sampling weight for each responding sampled unit (e.g., a household or person), and using that weight to calculate weighted estimates from the sample. The critical component of the sampling weight is the base weight which is defined to be the reciprocal of the probability of including a household or person in the sample. The base weights are used to inflate the responses of the sampled units to population levels and are generally unbiased (or consistent) if there is no nonresponse or noncoverage in the sample (e.g., see Kish, 1965, p. 67). When nonresponse or noncoverage occurs in the survey, weighting adjustments are applied to the base weights to compensate for both types of sample omissions.

Nonresponse is unavoidable in virtually all surveys of human populations. For the ZIMPHIA 2020, nonresponse can occur at different stages of data collection, for example, (1) before the enumeration of individuals in the household, (2) after household enumeration and selection of persons but before completion of the individual interview, and (3) after completion of the interview but before collection of a usable blood sample. The procedures used to compensate for nonresponse at each of the relevant stages of data collection are described in Section 3.4.

Noncoverage arises when some members of the survey population have no chance of being selected for the sample. For example, noncoverage can occur if the field operations fail to enumerate all dwelling units during the listing process, or if certain household members are omitted from the household rosters. To compensate for such omissions, the poststratification procedures described in Sections 3.4.3.3 and 3.4.4.3 are used to calibrate the weighted sample counts to available population projections.

#### 3.1 Overview of the Weighting Process

The overall weighting approach for ZIMPHIA 2020 includes several steps.

**Initial checks:** Checks of the data files are carried out as part of the survey and data quality control, and the probabilities of selection for PSUs and households are calculated and checked.

**Creation of Jackknife Replicates:** The variables needed to create the jackknife replicates for variance estimation are established at this point. This step can be implemented immediately after the PSU sample has been selected. All of the subsequent weighting steps described below are applied to the full sample, and to each of the jackknife replicates.

**Calculation of PSU Base Weights:** The weighting process begins with the calculation and checking of the sample PSU (EA) base weights as the reciprocals of the overall PSU probabilities of selection.

**Calculation of Household Weights:** The next step is to calculate household weights. The household base weights are calculated as the nonresponse adjusted PSU weights times the reciprocal of the within-EA household selection probabilities. The household base weights are adjusted first to account for dwelling units for which it could not be determined whether the dwelling unit contained an eligible household (see Table 2-6) and then the responding households have their weights adjusted to account for nonresponding eligible households. These adjustments are generally made within the EA in which the households are located. The resulting weight is the final household weight.

**Calculation of Person-Level Interview Weights:** Once the household weights are determined, they are used to calculate the individual base weights. The individual base weights are then adjusted for nonresponse among the eligible individuals, with a final adjustment for the individual weights to compensate for undercoverage in the sampling process by weighting up to 2020 population projections.

**Calculation of Person-Level Blood Test Weights:** The individual weights adjusted for nonresponse are in turn the base weights for the blood data sample, with a further adjustment for nonresponse to the blood draw, and a final poststratification adjustment to compensate for undercoverage.

**Application of Weighting Adjustments to Jackknife Replicates:** All of the adjustment processes are applied to the full sample and the replicate samples so that the final set of full sample and replicate weights can be used for variance estimation that takes into account the complex sample design and every step of the weighting process.

## **3.2 Preparation for Weighting**

Four basic data files are used as input to the weighting process. In this section we discuss these files from the perspective of the weighting process.

### **3.2.1 Data Files for Weighting**

The ZIMPHIA 2020 survey data that are used to construct the sampling weights are contained in the following data files. These are work files created and used during the weighting process and are not included in the data package for dissemination.

- **zw\_CFF\_hh\_int\_STAT\_20200416:** A household (HH) file that contains the household data collected in the HH questionnaire.

- **zw\_CFF\_roster\_STAT\_20200416:** A file that contains the roster of household members collected in the HH questionnaire with a record for each rostered person.
- **zw\_CFF\_ind\_int\_STAT\_20200416:** An individual level file that includes data collected on individual questionnaire tablets. This file contains data from the appropriate questionnaire modules for each person, with “null” values for those modules that do not apply to that person.
- **ZW2Biomarker20200417:** A biomarker file containing identifying information and results for lab analyses of blood samples for individuals whose blood was drawn and analyzed in the lab.

Each of these data files except the ZW2Biomarker file contains records for all sampled or collected cases, irrespective of response and eligibility status. However, for weighting purposes, a subset of the roster file was created with only “roster eligible” cases: these are person-level records from a responding household with a roster age of 15 or older and who were identified on the roster as having slept in the household the night before the interview. At the time of creating weight delivery files the “roster ineligible” cases were returned to the delivery files; however they have missing values for the weight variables.

### 3.2.2 Checks of Data Files

Prior to the start of the weighting process, the survey data files are checked and compared against information available in the sampling files. These checks include:

- Checking IDs, merging household survey files with sampling files, and accounting for records found in one file and not the other. (This type of check for the EAs occurs as part of the HH selection process.)
- Check counts of sampled and responding HHs against what was expected, overall and by province.
- Adjust for substitution of EAs, if applicable. Check that guidelines have been followed and selection probabilities are consistent with guidelines.
- Set disposition codes (respondent, eligible nonrespondent, ineligible, unknown eligibility) to be used for weighting purposes based on data elements received for (a) sampled households, (b) sampled individuals, and (c) individuals selected for blood draws.

### 3.3 Creation of Variables for Variance Estimation

Two general methods can be used for estimating the sampling errors of survey-based estimates derived from ZIMPHIA 2020: the jackknife replication and Taylor’s Series methods. The jackknife replication variance estimation method is a widely used method for producing variance estimates using data from a complex survey. This method can correctly account for the stratification, clustering, and sample weighting, including nonresponse and poststratification weighting adjustments, from the ZIMPHIA 2020 complex

sample design. The Taylor's Series is another widely used method that uses linear approximations to calculate the variance of a sample-derived estimate.

In order to implement either method, certain variables required for variance estimation must be included in the weighted data files. In the case of jackknife replication, the required variables are a series of weights that correspond to each of the jackknife replicates. In the case of the Taylor's Series method, the required variables are those that indicate the "variance stratum" and the "variance unit" to which each sampled respondent belongs.

### **3.3.1 Jackknife Replication**

To permit the calculation of variance estimates from the survey data, a series of weights, referred to as jackknife replicate weights, are attached to each record in the data file, along with the corresponding final full-sample weight. Calculation of the replicate weights first requires the construction of a set of subsamples of the full sample referred to as "jackknife replicates." Since these replicates depend only on the selected PSUs, they can be created immediately after the selection of PSUs.

As described in Section 2.3, the PSUs were selected systematically from a list of PSUs that had been ordered by EA within province. To take account of the precision benefits of implicit stratification as fully as possible, the sampled PSUs within each province were paired off in the systematic order in which they were selected, treating each pair as a variance-estimation stratum. When there was an odd number of sampled PSUs in a province, one of the variance-estimation strata was defined to contain three sampled PSUs. To fully reflect the sample design, the formation of the variance-estimation strata was applied to all 356 of the sampled PSUs, including nonresponding and out of scope PSUs if any (see Table 2-3).

For the ZIMPHIA 2020, a total of 175 variance-estimation strata were created. A jackknife replicate was then formed by randomly deleting a PSU from a particular variance-estimation stratum  $k$ , say, and retaining all of the PSUs in the remaining variance-estimation strata. For a variance-estimation stratum consisting of a pair of PSUs, the weight of the retained PSU within the variance-estimation stratum  $k$  was doubled. For a variance-estimation stratum consisting of three PSUs, the weight of the two retained PSUs within the variance-estimation stratum were increased by 1.5 (see Section 3.4.1). This process was repeated for all  $r = 1, 2, \dots, 175$  variance-estimation strata, resulting in a total of 175 jackknife replicates. Table 3-1 summarizes the number of jackknife replicates that were created for variance estimation.

**Table 3-1      Number of PSUs and variance-estimation strata constructed for variance estimation**

Province code	Province name	Sampled PSUs <sup>[1]</sup>	Variance strata consisting of pairs	Variance strata consisting of triplets	Number of jackknife replicates
0	Bulawayo	31	14	1	15
1	Manicaland	44	22	0	22
2	Mashonaland Central	37	17	1	18
3	Mashonaland East	37	17	1	18
4	Mashonaland West	40	20	0	20
5	Matabeleland North	29	13	1	14
6	Matabeleland South	27	12	1	13
7	Midlands	38	19	0	19
8	Masvingo	36	18	0	18
9	Harare	37	17	1	18
All	Zimbabwe	356	169	6	175

[1] Includes nonresponding and ineligible PSUs if applicable.

### 3.3.2 Taylor's Series

Even though jackknife replication is the recommended method for variance estimation, not all software packages have a replication option to produce variance estimates. For example, SPSS has built-in options for estimating variance using Taylor's Series methods, but the end user has to write a program within SPSS to produce replicate estimates of variance. Therefore, information for producing Taylor's Series estimates of variance is included in the ZIMPHIA 2020 data files.

The full-sample weight (see Section 3.4) is used as the weight to compute Taylor's Series variance estimates. The variable **varstrat** indicates the variance-estimation stratum and the variable **varunit** indicates the primary sampling unit (PSU) or cluster within the variance-estimation stratum. This pair of variables allows the analyst to produce variance estimates if their software does not easily accommodate replication methods, but does have a Taylor's Series capability.

## 3.4 Development of Weights

### 3.4.1 PSU Weights

The initial weighting step after the jackknife replicates were defined was to calculate PSU weights for the full sample and the replicates. Note that for convenience, we use the term PSU (primary sampling unit) to refer to either the originally-sampled EA, or the selected segment within the EA if the segmentation process was applied to the PSU.

The full-sample PSU weight was computed from the formula:

$$W_{hi}^{(1)} = 1/P_{hi}^{PSU},$$

where  $P_{hi}^{PSU}$  = probability of selecting PSU  $i$  from province  $h$ . Note that if the PSU was segmented, then  $P_{hi}^{PSU}$  is the product of the probability of selecting the EA and the conditional probability of selecting the segment within the EA. Using the PSU weights defined above, the sampled PSUs (i.e., whole EAs or segments) weight up to the numbers shown in the last column of Table 3-2.

As described in Section 3.3.1, 175 jackknife replicates were formed from the 356 sampled PSUs. For variance estimation, replicate-specific PSU weights,  $W_{(r)hi}^{(1)}$ ,  $r = 1, 2, \dots, 175$  were created to provide the basis for calculating the required replicate weights in subsequent stages of the weighting process. Let  $h$  denote one of the variance-estimation strata created for jackknife replication (Section 3.3.1) and let  $i$  denote the PSU within variance-estimation stratum  $h$ . For a given jackknife replicate,  $r = 1, 2, \dots, 175$ , the corresponding replicate-specific PSU base weight was computed as

$$\begin{aligned} W_{(r)hi}^{(1)} &= a W_{hi}^{(1)} && \text{if } h = r \text{ and PSU } i \text{ in variance-estimation stratum } h \text{ is included in} \\ &&& \text{replicate } r \\ &= 0 && \text{if } h = r \text{ and PSU } i \text{ in variance-estimation stratum } h \text{ is not included in} \\ &&& \text{replicate } r \\ &= W_{hi}^{(1)} && \text{if } h \neq r \end{aligned}$$

where the coefficient  $a = 2$  or  $1.5$  depending on whether the variance-estimation stratum consisted of 2 or 3 PSUs, respectively.

**Table 3-2 Number of PSUs and weighted number of PSUs by province**

Province code	Province name	Sampled PSUs <sup>[1]</sup>	Weighted number of PSUs <sup>[2]</sup>
0	Bulawayo	31	1,778.20
1	Manicaland	44	4,103.46
2	Mashonaland Central	37	2,965.22
3	Mashonaland East	37	4,243.38
4	Mashonaland West	40	3,777.93
5	Matabeleland North	29	1,614.19
6	Matabeleland South	27	1,525.09
7	Midlands	38	3,576.87
8	Masvingo	36	3,218.45
9	Harare	37	4,902.43
All	Zimbabwe	356	31,705.22

[1] Includes all sampled PSUs, including nonresponding and ineligible PSUs if applicable

[2] Weights are the PSU base weights,  $W_{hi}^{(1)}$ .

### 3.4.2 Dwelling Unit/Household Weights

#### 3.4.2.1 Dwelling Unit Base Weights

The household weighting process starts by calculating the dwelling unit-level base weights. These are the product of the PSU weight (described in Section 3.4.1) and the reciprocal of the within-PSU dwelling unit (DU) selection probability; i.e., the dwelling unit base weight for sampled dwelling unit  $j$  in PSU  $i$  in province  $h$  was computed as:

$$W_{hij}^{(2)} = W_{hi}^{(1)} / P_{j|hi}^{DU}$$

where

$W_{hi}^{(1)}$  = the reciprocal of the probability of selection for PSU  $i$  in province  $h$

$P_{j|hi}^{DU}$  = the conditional probability of selecting dwelling unit  $j$  in PSU  $i$  in province  $h$ .

The corresponding weights for jackknife replicate  $r = 1, 2, \dots, 175$  were computed as:

$$W_{(r)hij}^{(2)} = W_{(r)hi}^{(1)} / P_{j|hi}^{DU},$$

where  $W_{(r)hi}^{(1)}$  is the PSU base weight for PSU  $i$  in province  $h$  in replicate  $r$  described in Section 3.4.1.

Next, the sampled dwelling units were assigned to one of the four response status groups specified in Table 3-3. Note that by definition, a dwelling unit containing a household is classified as a “responding household” if a completed household interview was obtained. The specific rules used to classify dwelling

units into the response status groups are given in Appendix B.1. In Table 3-4, we show the weighted counts of dwelling units/households by response status and province using the dwelling unit base weights described above. The characteristics of the dwelling unit base weights were checked by examining statistical summaries of the weights such as the mean weight, CV (coefficient of variation) of the weights, sum of the weights, and the minimum and maximum values of the weights, both overall and by province.

**Table 3-3 Distribution of sampled dwelling units/households by response status**

Response status group <sup>[1]</sup>	Description	Number of sampled dwelling units/households
1	Respondent (household with completed household interview)	10,499
2	Nonrespondent (household without a completed household interview)	1,208
3	Ineligible (dwelling units with no households)	675
4	Unknown eligibility (not known if dwelling unit contains household)	78
All	---	12,460

[1] See Appendix B.1 for definitions.

**Table 3-4 Weighted counts of dwelling unit/household base weights by response status and province**

Province code	Province name	Response status <sup>[1]</sup>				Total groups 1-4
		Group 1: responding household	Group 2: nonresponding household	Group 3: ineligible dwelling unit	Group 4: unknown eligibility	
0	Bulawayo	155,115	20,701	7,402	0	183,218
1	Manicaland	413,607	50,153	24,124	952	488,837
2	Mashonaland Central	263,929	24,232	16,397	935	305,492
3	Mashonaland East	435,574	56,867	34,132	3,309	529,882
4	Mashonaland West	374,786	29,343	25,626	1,194	430,949
5	Matabeleland North	150,748	17,130	8,380	175	176,432
6	Matabeleland South	151,244	14,352	7,544	736	173,875
7	Midlands	332,540	33,020	22,208	877	388,645
8	Masvingo	294,955	33,571	17,666	2,511	348,703
9	Harare	404,891	78,578	39,018	17,789	540,276
All	Zimbabwe	2,977,387	357,948	202,497	28,477	3,566,309

[1] See Table 3.3. Counts given in table are weighted counts using the dwelling unit base weights,  $W_{hi}^{(2)}$ . Counts may not add to totals due to rounding.



### 3.4.2.2 Adjustment for Dwelling Unit Nonresponse

The general approach for handling dwelling unit nonresponse was to increase the weights of responding dwelling units so that they represent the nonresponding dwelling units in the same PSU or group of PSUs. Because such nonresponse could occur before establishing whether or not a sampled dwelling unit is eligible for the study (i.e., whether or not the associated household contains persons eligible for ZIMPHIA 2020), the nonresponse adjustment was implemented in two phases. In the first phase of adjustment, the base weights were adjusted to compensate for sampled dwelling units for which eligibility for the survey (e.g., occupancy status) was not ascertained. In the second phase of adjustment, the first-phase adjusted weights were further adjusted to compensate for the nonresponding dwelling units among those dwelling units known to be eligible for the study.

To account for variation in response rates across different types of PSUs, the dwelling unit nonresponse adjustments were made within weighting cells defined by the individual PSUs whenever possible. In a small number of instances, the adjustment was made within a group of two or more PSUs because either (a) the household response rate within the EA was so low that it would have resulted in unduly large sampling weights, or (b) data collection was not initiated for any of the sampled households in the EA. Both types of situations occurred in the ZIMPHIA 2020. In the first situation, the households in the EA (PSU 324) with the low response rate were combined with households in an adjacent EA (PSU 325) in the same district. In the second situation, the households in the EA (PSU 340) where data collection was not conducted due to a data collection omission error were treated as nonresponding households and combined with the sampled households in two adjacent EAs (PSUs 339; and 341) in the same district. To compensate for the omission of the households in PSU 340, a household-level nonresponse adjustment was made within the combined group of PSUs indicated above. Details of the procedures used to compute the nonresponse-adjusted dwelling unit/household weights are provided below.

#### ***Phase 1 Adjustment***

As indicated above, the weighting cells for the dwelling unit nonresponse adjustments are either the individual PSUs or a group of PSUs. Let  $n_{hi}^{DU}$  denote the number of sampled dwelling units in PSU  $i$  in province  $h$ . Note that  $n_{hi}^{DU}$  is the sum of the sample sizes in each of the four response status groups defined in Table 3-3, i.e.,

$$n_{hi}^{DU} = n_{hi}^{(1)} + n_{hi}^{(2)} + n_{hi}^{(3)} + n_{hi}^{(4)}$$

where

$$n_{hi}^{(1)} = \text{the number of responding households (i.e., households with a completed household interview) in PSU weighting cell } i \text{ in province } h$$

- $n_{hi}^{(2)}$  = the number of eligible nonresponding households (i.e., households without a completed household interview) in PSU weighting cell  $i$  in province  $h$
- $n_{hi}^{(3)}$  = the number of known ineligible dwelling units (i.e., dwelling units known to contain no households) in PSU weighting cell  $i$  in province  $h$
- $n_{hi}^{(4)}$  = the number of sampled dwelling units for which it is not known whether a household is present in PSU weighting cell  $i$  in province  $h$ .

The first-phase nonresponse adjustment factor for PSU weighting cell  $i$  in province  $h$  was computed as the ratio:

$$A_{hi}^{(DU1)} = \sum_{j=1}^{n_{hi}^{DU}} W_{hij}^{(2)} / \sum_{j=1}^{n_{hi}^{(1)} + n_{hi}^{(2)} + n_{hi}^{(3)}} W_{hij}^{(2)}$$

where  $W_{hi}^{(2)}$  is the base weight for dwelling unit/household  $j$  in PSU weighting cell  $i$  in province  $h$ , and where the sum in the numerator extends over the entire sample of dwelling units/households in PSU weighting cell  $i$  in province  $h$ , while the sum in the denominator extends over the first three groups of dwelling units.

For the sampled dwelling units/households in response-status groups 1, 2 or 3, the first-phase adjusted weight for dwelling unit/household  $j$  in PSU weighting cell  $i$  in province  $h$  was then computed as:

$$W_{hij}^{DU1} = A_{hi}^{(DU1)} W_{hi}^{(2)}$$

The corresponding replicate weights for replicate  $r = 1, 2, \dots, 175$  were computed in similar fashion as:

$$W_{(r)hij}^{DU1} = A_{(r)hi}^{(DU1)} W_{(r)hij}^{(2)},$$

where

$$A_{(r)hi}^{(DU1)} = \sum_{j=1}^{n_{(r)hi}^{DU}} W_{(r)hi}^{(2)} / \sum_{j=1}^{n_{(r)hi}^{(1)} + n_{(r)hi}^{(2)} + n_{(r)hi}^{(3)}} W_{(r)hij}^{(2)}.$$

Note that for the dwelling units in response-status group 4 (dwelling units of unknown eligibility),  $W_{hij}^{DU1} = W_{(r)hij}^{DU1} = 0$  for  $r = 1, 2, \dots, 175$ .

The effect of this adjustment is to distribute the total weight of the unknown-eligibility cases (i.e., the estimated 28,477 dwelling units shown in the next-to-last column of Table 3-4) to the combined weight of the remaining three groups of sampled dwelling units/households. The resulting weighted counts using  $W_{hij}^{DU1}$  as computed above are summarized in Table 3-5.

**Table 3-5 Weighted counts of dwelling units/households adjusted for unknown eligibility**

Province code	Province name	Response status				Total households: groups 1-2
		Group 1: responding household	Group 2: nonresponding household	Group 3: ineligible dwelling unit	Total status 1-3	
0	Bulawayo	155,115	20,701	7,402	183,218	175,816
1	Manicaland	414,403	50,250	24,185	488,837	464,652
2	Mashonaland Central	264,761	24,284	16,448	305,492	289,044
3	Mashonaland East	438,187	57,322	34,373	529,882	495,509
4	Mashonaland West	375,795	29,398	25,756	430,949	405,194
5	Matabeleland North	150,893	17,155	8,384	176,432	168,048
6	Matabeleland South	151,886	14,403	7,586	173,875	166,290
7	Midlands	333,271	33,102	22,273	388,645	366,372
8	Masvingo	297,126	33,723	17,855	348,703	330,849
9	Harare	418,515	80,977	40,784	540,276	499,492
All	Zimbabwe	2,999,951	361,315	205,043	3,566,309	3,361,266

Note: Counts in table are weighted counts using first-phase adjusted household weights,  $W_{hij}^{DU1}$ . Counts may not add to totals due to rounding.

### Phase 2 Adjustment

In the second phase of adjustment, the weights of the responding households (response status group 1) were inflated by the inverse of the (weighted) response rate in the PSU weighting cell after eliminating the known ineligible dwelling units (i.e., response-status group 3). The second-phase household nonresponse adjustment factor for PSU weighting cell  $i$  in province  $h$  was computed as the ratio:

$$A_{hi}^{(HH2)} = \frac{\sum_{j=1}^{n_{hi}^{(1)} + n_{hi}^{(2)}} W_{hij}^{DU1}}{\sum_{j=1}^{n_{hi}^{(1)}} W_{hij}^{DU1}}$$

where  $W_{hij}^{DU1}$  is the first-phase adjusted weight for dwelling unit/household  $j$  in PSU weighting cell  $i$  in province  $h$ , and where the sum in the numerator extends over the sample of responding and nonresponding households in PSU weighting cell  $i$  in province  $h$ , while the sum in the denominator extends over the responding households.

The final nonresponse-adjusted weight for *responding* household  $j$  in PSU weighting cell  $i$  in province  $h$  was then computed as:

$$W_{hij}^{(2A)} = A_{hi}^{(HH2)} W_{hij}^{DU1}.$$

The corresponding replicate weights for replicate  $r = 1, 2, \dots, 175$  were computed in similar fashion as:

$$W_{(r)hij}^{(2A)} = A_{(r)hi}^{(HH2)} W_{(r)hij}^{DU1},$$

where

$$A_{(r)hi}^{(HH2)} = \sum_{j=1}^{n_{(r)hi}^{(1)} + n_{(r)hi}^{(2)}} W_{(r)hij}^{DU1} / \sum_{j=1}^{n_{(r)hi}^{(1)}} W_{(r)hij}^{DU1}.$$

The sum of the final nonresponse-adjusted household weights,  $W_{hij}^{(2A)}$ , summed across the responding households (response status group 1), is equal to the weighted count shown in the last column of Table 3-5.

### 3.4.3 Person-Level Interview Weights

In this section, we detail the calculation of person-level sampling weights to be used to analyze the individual interview responses in the ZIMPHIA 2020 data files. First we define the initial person-level (interview) base weights in Section 3.4.3.1. Next, to compensate for interview nonresponse, the person base weights are adjusted within cells defined by variables available for both the responding and nonresponding individuals. Like the dwelling unit/household nonresponse adjustments described previously, this person-level nonresponse adjustment was implemented in two phases.

#### 3.4.3.1 Person Base Weights

All persons included on the rosters provided by responding households initially receive a person-level base weight equal to the final nonresponse-adjusted household weight,  $W_{hij}^{(2A)}$ . That is, the base weight for rostered person  $k$  in household  $j$  in PSU  $i$  in province  $h$  was computed from the formula

$$W_{hijk}^{(base)} = W_{hij}^{(2A)}.$$

The corresponding replicate base weights,  $W_{(r)hijk}^{(base)}$ , for  $r = 1, 2, \dots, 175$  were computed in an analogous manner, with  $W_{hij}^{(2A)}$  replaced by  $W_{(r)hij}^{(2A)}$  in the above formula.

#### 3.4.3.2 Adjustment of Person Weights for Interview Nonresponse

Since the final eligibility of a rostered person cannot be determined until after the actual age is confirmed during the interview, the person-level base weights were adjusted in two phases. Table 3-6 summarizes the distribution of the rostered persons by the five response-status groups specified for the first-phase adjustment. Response status groups 4 and 5 are the cases determined to be ineligible for the study because they are either under 15 years old or because they were not present in the household at the time they were rostered (i.e., “non *de facto*”). All of these cases are treated as “known ineligible” cases and are excluded from the first-phase adjustment. The cases in response-status group 3 are cases for which final eligibility for the study is not known. The combined weight of these individuals was distributed to the

cases in response-status groups 1 and 2 within weighting classes defined by sex and age group as described below.

**Table 3-6 Distribution of rostered persons in responding households by response status for first-phase nonresponse adjustment**

First-phase response status group <sup>[1]</sup>	Resident status and age based on roster	Confirmed age based on interview	Number of rostered persons	Weighted number of rostered persons <sup>[2]</sup>
1	<i>De facto</i> person 15 years or older	15+	22,751	7,135,805
2	<i>De facto</i> person 15 years or older	Under 15	0	0
3	<i>De facto</i> person 15 years or older	Unknown	11	3,876
4	Non <i>de facto</i> persons 15 years or older	NA	4,271	1,390,203
5	Persons under 15 years	NA	16,370	5,109,112
All	---	---	43,403	13,638,997

[1] See Appendix B for definitions of response status categories.

[2] Weighted by the person-level base weight,  $W_{hijk}^{(base)}$ .

### ***First Phase Adjustment***

The procedure for computing the first-phase adjustment was as follows. For each of the sex-age weighting classes specified for the adjustment, the weighted full-sample first-phase response rate,  $R_c^{(1)}$ , was computed as

$$R_c^{(1)} = ( \sum_{k=1}^{n_c^{(1)}} W_{ck}^{(base)} + \sum_{i=1}^{n_c^{(2)}} W_{ck}^{(base)} ) / ( \sum_{i=1}^{n_c^{(1)}} W_{ck}^{(base)} + \sum_{i=1}^{n_c^{(2)}} W_{ck}^{(base)} + \sum_{i=1}^{n_c^{(3)}} W_{ck}^{(base)} )$$

where  $c$  denotes the first-phase adjustment cell,  $W_{ck}^{(base)}$  is the base weight for person  $k$  in cell  $c$ , and  $n_c^{(a)}$  = the number of cases in response-status group  $a = 1, 2, 3$  in weighting class  $c$ .

The corresponding replicate-specific weighted response rates were similarly computed for jackknife replicate  $r = 1, 2, \dots, 175$  as

$$R_{(r)m}^{(1)} = ( \sum_{k=1}^{n_{(r)c}^{(1)}} W_{(r)ck}^{(base)} + \sum_{i=1}^{n_{(r)c}^{(2)}} W_{(r)ck}^{(base)} ) / ( \sum_{i=1}^{n_{(r)c}^{(1)}} W_{(r)ck}^{(base)} + \sum_{i=1}^{n_{(r)c}^{(2)}} W_{(r)ck}^{(base)} + \sum_{i=1}^{n_{(r)c}^{(3)}} W_{(r)ck}^{(base)} )$$

The first-phase interview nonresponse adjustment factor for cell  $c$  is  $A_c^{(1)} = 1/R_c^{(1)}$  for the full sample, and  $A_{(r)c}^{(1)} = 1/R_{(r)c}^{(1)}$  for jackknife replicate  $r = 1, 2, \dots, 175$ .

The full-sample first-phase nonresponse-adjusted weight for person  $k$  in cell  $c$  was then computed as

$$W_{ck}^{(3)} = A_c^{(1)} W_{ck}^{(base)},$$

and the corresponding jackknife replicate weights for replicate  $r = 1, 2, \dots, 175$  were similarly computed as

$$W_{(r)ck}^{(3)} = A_{(r)c}^{(1)} W_{(r)ck}^{(base)}.$$

### Second Phase Adjustment

Table 3-7 summarizes the unweighted and weighted counts of eligible sample persons by sex and interview response status. The weights used to derive the weighted counts in this table are the first-phase person-level nonresponse-adjusted weights,  $W_{ck}^{(3)}$ . To compensate for interview nonresponse, the first-phase nonresponse-adjusted weights,  $W_{ck}^{(3)}$ , were further adjusted within cells defined by variables available for both the responding and nonresponding individuals. These variables included data from the household roster and other information collected in the household questionnaire, and selected PSU characteristics such as province and urban/rural status. The age and sex variables used to make the nonresponse adjustments are those reported in the household roster and not the interview-reported age and sex, because the latter values are not known for the nonrespondents. The Least Absolute Shrinkage and Selection Operator (LASSO) was used for initial variable selection, and the Chi-Square Automatic Interaction Detector (CHAID) was used to form the final weighting cells for nonresponse adjustment.

**Table 3-7 Unweighted and weighted counts of eligible sample persons by sex and interview response status**

Sex/Age group <sup>[1]</sup>	Interview response status <sup>[2]</sup>	Unweighted sample size	Weighted count <sup>[3]</sup>
Male 15 or older	Eligible respondent	8,271	2,560,300
	Eligible nonrespondent	1,241	396,012
	<b>All response statuses</b>	<b>9,512</b>	<b>2,956,312</b>
Female 15 or older	Eligible respondent	12,522	3,953,222
	Eligible nonrespondent	717	230,147
	<b>All response statuses</b>	<b>13,239</b>	<b>4,183,369</b>
Total 15 years or older	Eligible respondent	20,793	6,513,523
	Eligible nonrespondent	1,958	626,159
	<b>All response statuses</b>	<b>22,751</b>	<b>7,139,681</b>

[1] Age reported in roster which may differ from the confirmed age in the interview.

[2] See Appendix B for definitions of the interview response status categories.

[3] Weighted by the first-phase nonresponse adjusted person weight,  $W_{hijk}^{(3)}$ .

### The Least Absolute Shrinkage and Selection Operator (LASSO) for Initial Variable Selection

There are 47 variables from the household questionnaire and EA sampling frame that could potentially be used for nonresponse adjustment. The LASSO regression was used to reduce the number of variables to

a manageable subset of the most important and relevant predictors that would ultimately be entered into the CHAID algorithm to define the final nonresponse adjustment weighting cells. The LASSO is a restrictive procedure similar to linear regression that shrinks regression coefficient estimates to zero. In other words, predictors that are found to be nonsignificant have their regression coefficients set to 0 (Hastie, Tibshirani, and Friedman, 2009).

In the final model produced by the LASSO, only the most significant variables predictive of the response variable were identified and kept. The HPGENSELECT procedure (Johnston and Rodriguez, 2015) with selection method=lasso in SAS 9.4 was used to select the variables, with the weight set to the person-level base weight,  $W_{hijk}^{(base)}$ . The final model was selected on the basis of cross validation with observations in the input data set partitioned into disjoint subsets, reserving 25% for training, 50% for validation, and 25% for testing. As there is some randomness in how the LASSO selects the variables, we set the seed to a known constant value to remove the randomness so that if the program had to be re-run, the same results would be reproduced. Of the 47 variables used in the initial model, the LASSO identified 33 variables as the significant predictors of response.

### ***The Chi-Square Automatic Interaction Detector (CHAID) for Cell Formation***

The next step was to apply the CHAID algorithm (Magidson, 2005) to the variables selected by the LASSO procedure. CHAID classifies the sampled individuals (i.e., the respondents and nonrespondents) into “cells” based on information available for all sample persons. The cells are formed in such a way that persons belonging to the same cell are expected to have similar propensities to participate in the study. Using the variables selected by the LASSO as input, CHAID uses a weighted log-linear modeling (WLM) algorithm for the computation of chi-square statistics associated with each predictor, where the weight is the person base weight,  $W_{hijk}^{(base)}$ . An output of the CHAID procedure is a tree diagram that specifies the optimum number of final weighting cells, and their definitions based on the input predictor variables. The depth limit of the tree was set to 5 (not including any variables that are forced into the model), and the minimum subgroup size required to allow splitting and minimum terminal node size were set to 50 observations (both respondents and nonrespondents).

To create the CHAID tree, gender (SEX) and a variable indicating whether the sampled person was 15-17 years of age or 18 or older (H\_AGETEENYEARS) were forced into the model to make the initial splits. The reason for doing this was because the subgroups defined by these variables received different questions; without forcing these variable into the model, the resulting tree would not have been created correctly. After forcing the two variable in the model, the tree was then allowed to grow freely. The CHAID algorithm identified 27 variables that were used to create the weighting classes for nonresponse adjustment. Table 3-8 lists the variables that were included in the final CHAID models. The final trees

produced by the CHAID algorithm are documented in Appendix C.1. The corresponding nonresponse-adjustment classes used to adjust the person-level base weights are given in Appendix C.2.



**Table 3-8 Variables selected by CHAID to produce classes for interview nonresponse adjustment**

Variable number	Variable name	Description
1	COOKINGFUEL	HH Characteristics: What Type Of Fuel Does Your Household Mainly Use For Cooking?
2	DADALIVE	HH Minors: Is [dispname]'s Natural Father Alive?
3	FEMGUARDHHM	HH Minors: Does [name] Have A Female Guardian Who Usually Lives In This Household Or Was A Guest Last Night?
4	HH_ECONSUP12_H	HH Economic suppot: Income Generation Support In Cash Or Kind (E.G. Agrigultural Inputs)
5	H_AGETEENYEARS	TEEN INDICATOR: 1 – 15-17 YEARS OLD; 2 – OTHERWISE; BASED ON AGEYEARS (ROSTER)
6	H_AGEYEARS	AGE (CATEGORICAL), BASED ON ROSTER AGE. MATCHES POSTSTRATIFICATION CELLS
7	H_HHQITEMS	1-Electricity; 2-Working Radio; 3-Working Television; 4-Working Telephone/Mobile Telephone; 5-Working Refrigerator; 6-None Of The Above
8	H_HHQOWN	1-Bicycle; 2-Working Motorcycle Or Motor Scooter; 3-Working Car Or Truck; 4-A Working Boat With A Motor; 5-None Of The Above
9	H_HH_SIZE_C	1-9, where 9 includes all HHs with 9 or more people
10	H_OWNCCHIKNUM	Chickens: Altogether, How Many Of The Below Listed Animals Do Members Of Your Household Own?
11	H_OWNCOWNUM	Cows: How Many Of The Below Listed Animals Do Members Of Your Household Own?
12	H_OWNDOGNUM	Dogs: Altogether, How Many Of The Below Listed Animals Do Members Of Your Household Own?
13	H_OWNGOATNUM	Goats/Sheep: Hh Characteristics: Altogether, How Many Of The Below Listed Animals Do Members Of Your Household Own?
14	H_OWNHORSENUM	Work Animals (Camels, Horses, Donkeys): Hh Characteristics: Altogether, How Many Of The Below Listed Animals Do Members Of Your Household Own?
15	H_RELATTOHH	1-Head, 2-Wife/Husband/Partner, 3-Son Or Daughter, 4-Son-In-Law/Daughter-In-Law, 5-Grandchild, 6-Parent, 7-Parent-In-Law, 8-Brother/Sister, 9-Co-Wife, 10-Other
16	H_ROOMSLEEP	How Many Rooms Are Used For Sleeping?
17	H_WATERSOURCE	11-Piped to Dwelling, 12-Piped To Yrd/Plt, 13-Public Tap/Standpipe, 21-Tube Well Or Borehole, 31-Protected Well, 32-Unprotected Well, 41-Protected Spring, 42-Unprotected Spring, 51-Rainwater, 81-Surface Water (River/Dam/Lake/Pond/Stream/Canal), 96-other
18	MATEXWALLS	HH Characteristics: Main Material Of Exterior Walls
19	MATFLO	HH Characteristics: Main Material Of Floor
20	MATROOF	HH Characteristics: Main Material Of Roof
21	MOMSICK	HH Sickness: Has [dispname]'s Natural Mother Been Very Sick For At Least 3 Months During The Past 12 Months, That Is She Was Too Sick To Work Or Do Normal Activities?
22	SEX	HH Roster: Is [name] Male Or Female?
23	SICK_HOUSEHOLD	HH Sickhouse Flag: Any Member Of The Household Has Answered That They Are Sick On last 3 months
24	STRATA	Numeric code for EA sampling stratum
25	TOILETSARE	HH Characteristics: Do You Share This Toilet Facility With Other Households?

26	TOILETTYPE	HH Characteristics: What Kind Of Toilet Facility Do Members Of Your Household Usually Use?
27	URBAN_RURAL	1 = Urban, 2 = Rural

### **Calculation of Second-Phase Nonresponse-Adjusted Person Weights**

The general approach for computing the second-phase nonresponse-adjusted person-level interview weights was as follows. Within each of the final adjustment cells specified in Appendix C.2, the full-sample weighted response rate,  $R_m^{(int)}$ , was computed as

$$R_m^{(int)} = \sum_{k=1}^{n_m^{resp}} W_{mk}^{(3)} / \left( \sum_{i=1}^{n_m^{resp}} W_{mk}^{(3)} + \sum_{i=1}^{n_m^{nr}} W_{mk}^{(3)} \right),$$

where  $m$  denotes the adjustment cell,  $W_{mk}^{(3)}$  is the first-phase nonresponse-adjusted weight for person  $k$  in cell  $m$ ,  $n_m^{resp}$  = the number of responding persons in cell  $m$ , and  $n_m^{nr}$  = the number of eligible nonresponding persons in cell  $m$ .

The corresponding replicate-specific weighted response rates were similarly computed for jackknife replicate  $r = 1, 2, \dots, 175$  as

$$R_{(r)m}^{(int)} = \sum_{k=1}^{n_{(r)m}^{resp}} W_{(r)mk}^{(3)} / \left( \sum_{i=1}^{n_{(r)m}^{resp}} W_{(r)mk}^{(3)} + \sum_{i=1}^{n_{(r)m}^{nr}} W_{(r)mk}^{(3)} \right).$$

The interview nonresponse adjustment factor for cell  $m$  is  $A_m^{(int)} = 1/R_m^{(int)}$  for the full sample, and  $A_{(r)m}^{(int)} = 1/R_{(r)m}^{(int)}$  for jackknife replicate  $r = 1, 2, \dots, 175$ .

The full-sample nonresponse-adjusted interview weight for responding person  $k$  in cell  $m$  was then computed as

$$W_{mk}^{(int)} = A_m^{(int)} W_{mk}^{(3)},$$

and the corresponding jackknife replicate weights for replicate  $r = 1, 2, \dots, 175$  were similarly computed as

$$W_{(r)mk}^{(int)} = A_{(r)m}^{(int)} W_{(r)mk}^{(3)}.$$

A summary of selected features of the nonresponse adjustment process is given in Table 3-9.

**Table 3-9 Summary of the interview nonresponse adjustment process**

Characteristic	Total sample
Number of variables in initial model	47
Number of variables selected by LASSO	33

Number of variables selected by CHAID	27
Number of final nonresponse-adjustment cells	82
Number of interview respondents	20,793
Minimum adjustment factor	1.00
Maximum adjustment <sup>[1]</sup>	1.91
Weighted count of respondents before adjustment <sup>[2]</sup>	6,513,523
Weighted count of respondents after adjustment <sup>[3]</sup>	7,139,681

[1] Maximum adjustment after collapsing CHAID cells (see Appendix C.2).

[2] Weight is the first-phase nonresponse-adjusted person weight,  $W_{mk}^{(3)}$ .

[3] Weight is the second-phase nonresponse-adjusted person weight,  $W_{mk}^{(int)}$ .

### 3.4.3.3 Poststratification Adjustment

The final step in computing the individual interview weights was to adjust the nonresponse-adjusted interview weights using a procedure called poststratification (Kalton and Kasprzyk, 1986). The primary goal of poststratification is to mitigate noncoverage biases that result when some persons in the study population do not have a chance to be sampled and interviewed. For example, undercoverage can occur:

- At the dwelling unit (DU) level if field operations fail to include all eligible dwelling units during the implementation of the listing procedures.
- At the household level if all households within multi-family dwelling units are not accounted for in sampling.
- At the person level where under- or overcoverage can occur if errors are made in the enumeration of household members.

To compensate for the types of coverage problems indicated above, the nonresponse-adjusted person weights were ratio-adjusted so that the resulting weighted sample counts match the population control totals indicated in Table 3-10. The population control totals given in this table are projected 2020 national population projections by gender and five-year age groups provided by the Zimbabwe National Institute of Statistics (ZIMSTAT). The poststratified interview weights were computed as follows.

Let  $N_{ga}^{2020}$  denote the 2020 Zimbabwe population control total for gender  $g$  and (five-year) age group  $a$  as given in Table 3-10. The poststratification ratio adjustment factor for gender  $g$  and age group  $a$  was then computed as:

$$T_{ga}^{2020} = N_{ga}^{2020} / \sum_{k=1}^{n_{ga}^{resp}} W_{gak}^{(int)},$$

where  $W_{gak}^{(int)}$  is the nonresponse-adjusted interview weight for respondent  $k$  in gender group  $g$  and age group  $a$ .

The corresponding replicate-specific adjustment factors were computed in a similar way as:

$$T_{(r)ga}^{2020} = N_{ga}^{2020} / \sum_{k=1}^{n_{(r)ga}^{resp}} W_{(r)gak}^{(int)}$$

for the  $r = 1, 2, \dots, 175$  jackknife replicates.

The full-sample poststratified interview weight was then computed as:

$$W_{gak}^{(ps-int)} = T_{ga}^{2020} W_{gak}^{(int)},$$

and the corresponding poststratified replicate weights were computed as:

$$W_{(r)gak}^{(ps-int)} = T_{ga}^{2020} W_{(r)gak}^{(int)}$$

for  $r = 1, 2, \dots, 175$ .

Weighted counts of the interview respondents before and after poststratification are summarized in Table 3-10.

**Table 3-10 2020 Zimbabwe population projections and weighted counts before and after poststratification**

Age group	Male			Female			Total		
	Population control total <sup>[1]</sup>	Wtd. count before post-stratification <sup>[2]</sup>	Poststratification ratio <sup>[3]</sup>	Population control total <sup>[1]</sup>	Wtd. count before post-stratification <sup>[2]</sup>	Poststratification ratio <sup>[3]</sup>	Population control total <sup>[1]</sup>	Wtd. count before post-stratification <sup>[2]</sup>	Poststratification ratio <sup>[3]</sup>
15-19	875,183	553,356	1.582	871,128	600,787	1.450	1,746,311	1,154,143	1.513
20-24	745,086	412,512	1.806	750,035	578,593	1.296	1,495,121	991,105	1.509
25-29	594,652	319,783	1.860	664,852	525,534	1.265	1,259,504	845,317	1.490
30-34	470,910	300,702	1.566	595,670	473,673	1.258	1,066,580	774,375	1.377
35-39	422,277	305,116	1.384	501,577	453,010	1.107	923,854	758,126	1.219
40-44	372,980	232,421	1.605	402,977	330,826	1.218	775,957	563,247	1.378
45-49	275,494	216,142	1.275	293,217	292,508	1.002	568,711	508,651	1.118
50-54	185,193	132,928	1.393	196,941	196,334	1.003	382,134	329,262	1.161
55-59	143,251	103,612	1.383	182,650	199,840	0.914	325,901	303,452	1.074
60-64	110,685	108,551	1.020	176,094	177,072	0.994	286,779	285,622	1.004
65+	260,755	252,313	1.033	404,015	374,067	1.080	664,770	626,381	1.061
15+	4,456,466	2,937,436	1.517	5,039,156	4,202,245	1.199	9,495,622	7,139,681	1.330

[1] 2020 population projections provided by Zimbabwe National Statistics Agency (ZIMSTAT)

[2] Weighted count of interview respondents using nonresponse-adjusted interview weight,  $W_{gak}^{(int)}$ .

[3] Ratio of population control total to weighted count of interview respondents using nonresponse-adjusted interview weight,  $W_{gak}^{(int)}$ .

### 3.4.4 Person-Level Blood Test Weights

Not every interview respondent provided a useable blood sample. Thus, a separate set of weights is required for analysis of the blood test results. Similar to the construction of the interview weights described previously, development of the final blood test weights involves adjustments for nonresponse and poststratification to 2020 population control totals.

#### 3.4.4.1 Initial Weights

The starting point for the construction of the blood test weights is the set of final full-sample nonresponse-adjusted interview weights and corresponding replicate weights described in Section 3.4.3.2. These weights are given by  $W_{hijk}^{(int)}$  and  $W_{(r)hijk}^{(int)}$  (for  $r = 1, 2, \dots, 175$ ), respectively, where  $k$  denotes the interview respondent,  $h$  denotes the province,  $i$  denotes the PSU, and  $j$  denotes the household. These weights have been adjusted for interview nonresponse, and thus act as the “base” weights for developing nonresponse adjustments for the blood test weights. Table 3-11 summarizes the counts of individuals by gender, age group and blood test response status, and the corresponding weighted counts using the person-level interview weights,  $W_{hijk}^{(int)}$ .

**Table 3-11 Unweighted and weighted distributions of sample persons completing the blood test by age group, sex, and response status**

Age group <sup>[1]</sup>	Sex	Blood test response status <sup>[2]</sup>	Unweighted sample size	Weighted count <sup>[3]</sup>
15-49 years	Male	Eligible respondent	5,946	2,163,515
		Eligible nonrespondent	458	176,517
	Female	Eligible respondent	9,088	3,064,868
		Eligible nonrespondent	535	190,064
50 years or older	Male	Eligible respondent	1,718	562,370
		Eligible nonrespondent	98	35,035
	Female	Eligible respondent	2,783	891,817
		Eligible nonrespondent	167	55,496
15 years or older	Male	Eligible respondent	7,664	2,725,885
		Eligible nonrespondent	556	211,551
	Female	Eligible respondent	11,871	3,956,685
		Eligible nonrespondent	702	245,560

[1] Age reported in the interview, which may differ from the age reported on the roster.

[2] Status among the interview respondents. See Appendix B for definitions of the response status groups.

[3] Weighted count of interview respondents using final nonresponse-adjusted interview weight,  $W_{gak}^{(int)}$ .

#### **3.4.4.2 Nonresponse Adjustment of Blood Test Weights**

To compensate for blood test nonresponse, the nonresponse-adjusted interview weights were further adjusted within cells defined by variables available for both the responding and nonresponding individuals (i.e., individuals completing the interview who may or may not have a final HIV status determination). These variables included data from the household roster and other information collected in the household questionnaire, selected PSU characteristics such as province and urban/rural status, and the individual interview. The age and sex variables used to make the nonresponse adjustments are those reported in the interview.

For males, 111 potential predictor variables were available for initial selection. For females, 118 potential predictor variables were available for initial selection. The LASSO procedure was used to identify a reduced set of predictor variables to be used in the CHAID algorithm. From these initial sets of variables, the LASSO regression identified 33 significant variables for males and 61 significant variables for females. The selected variables were then input into the CHAID program to create the final weighting cells for nonresponse adjustment.

The CHAID algorithm identified 17 variables for males and 12 variables for females that were then used to create weighting classes for nonresponse adjustment. Table 3-12 lists the variables that were included in the final CHAID models. The final trees produced by the CHAID algorithm are documented in Appendix C.1. The corresponding nonresponse-adjustment classes used to adjust the person-level base weights are given in Appendix C.2.

**Table 3-12 Variables selected by CHAID to produce classes for blood test nonresponse adjustment**

Sex	Variable number	Variable name	Description
Male	1	ADDISHIV	Prevention Intervention: Have You Ever Discussed Hiv With Your Parents Or Guardian?
	2	AT_BESTAGE_C	CATEGORICAL AGE BASED ON INTERVIEW AGE (CONFAGEY)
	3	AT_FIRSTSXAGE	AGE OF FIRST SEXUAL ACTIVITY
	4	AT_SCHCOM	WHAT IS THE HIGHEST LEVEL YOU HAVE COMPLETED?
	5	COOKINGFUEL	HH Characteristics: What Type Of Fuel Does Your Household Mainly Use For Cooking?
	6	H_HHQITEMS	1-Electricity; 2-Working Radio; 3-Working Television; 4-Working Telephone/Mobile Telephone; 5-Working Refrigerator; 6-None Of The Above
	7	H_HH_SIZE_C	1-9, where 9 includes all HHs with 9 or more people
	8	H_OWNCNIKNNUM	Chickens: Altogether, How Many Of The Below Listed Animals Do Members Of Your Household Own?
	9	H_OWNDOGNUM	Dogs: Altogether, How Many Of The Below Listed Animals Do Members Of Your Household Own?
	10	H_RELATTOHH	1-Head, 2-Wife/Husband/Partner, 3-Son Or Daughter, 4-Son-In-Law/Daughter-In-Law, 5-Grandchild, 6-Parent, 7-Parent-In-Law, 8-Brother/Sister, 9-Co-Wife, 10-Other
	11	LITTLEINTEREST	Tb And Other Health Issues: Over The Past Two Weeks, How Often Have You Been Bothered By Having Little Interest In Doing Things?
	12	MONTHOUTEVER	Background: Have You Ever Lived Away From Home For More Than 1 Month At A Time?
	13	NORMWORK	Background: Where Do You Normally Work? In Your Home Community, Elsewhere In Region/Country, Or Outside The Country?
	14	PARTLASTCNDM1	Sexual Activity: The Last Time You Had Sex With [partinit], Was A Condom Used?
	15	PREPWDTK	HIV Testing: Would You Take Prep To Help Prevent Hiv?
	16	SICK3MO	HH Roster: Has [name] Been Very Sick For At Least 3 Months During The Past 12 Months, That Is [name] Was Too Sick To Work Or Do Normal Activities?
	17	STRATA	Numeric code for EA sampling stratum
Female	18	AT_FIRSTSXAGE	AGE OF FIRST SEXUAL ACTIVITY
	19	AT_LIVEB	HOW MANY TIMES HAVE YOU HAD A PREGNANCY THAT RESULTED IN A LIVE BIRTH?
	20	CERNCNRSLT	Tb And Other Health Issues: What Was The Result Of Your Last Test For Cervical Cancer?
	21	HIVTPRG	Reproduction: Were You Tested For Hiv Anytime During Pregnancy Or Delivery With [childlast]?
	22	HIVTSAD	Reproduction: Were You Tested For Hiv At Any Time After Delivery Of Your Last Pregnancy With [childlast]?
	23	H_OWNCNIKNNUM	Chickens: Altogether, How Many Of The Below Listed Animals Do Members Of Your Household Own?
	24	H_RELATTOHH	1-Head, 2-Wife/Husband/Partner, 3-Son Or Daughter, 4-Son-In-Law/Daughter-In-Law, 5-Grandchild, 6-Parent, 7-Parent-In-Law, 8-Brother/Sister, 9-Co-Wife, 10-Other
	25	H_ROOMSLEEP	How Many Rooms Are Used For Sleeping?



Sex	Variable number	Variable name	Description
	26	H_WATERSOURCE	11-Piped to Dwelling, 12-Piped To Yrd/Plt, 13-Public Tap/Standpipe, 21-Tube Well Or Borehole, 31-Protected Well, 32-Unprotected Well, 41-Protected Spring, 42-Unprotected Spring, 51-Rainwater, 81-Surface Water (River/Dam/Lake/Pond/Stream/Canal), 96-other
	27	MATROOF	HH Characteristics: Main Material Of Roof
	28	PREPWDTK	HIV Testing: Would You Take Prep To Help Prevent Hiv?
	29	WORKIND	Background: What Is Your Occupation? That Is, What Kind Of Work Do You Mainly Do?

### Calculation of Nonresponse-Adjusted Blood Test Weights

The general approach for computing the nonresponse-adjusted person-level blood test weights was as follows. Within each of the final adjustment cells specified in Appendix B.2 for blood-test weighting, the full-sample weighted response rate,  $R_m^{(BT)}$ , was computed as

$$R_m^{(BT)} = \sum_{k=1}^{n_m^{BT}} W_{mk}^{(int)} / \left( \sum_{i=1}^{n_m^{BT}} W_{mk}^{(int)} + \sum_{i=1}^{n_m^{NBT}} W_{mk}^{(int)} \right),$$

where  $m$  denotes the adjustment cell,  $W_{mk}^{(int)}$  is the final nonresponse-adjusted interview weight for interview respondent  $k$  in cell  $m$ ,  $n_m^{BT}$  is the number of interview respondents in cell  $m$  with a final HIV status determination, and  $n_m^{NBT}$  is the number of interview respondents in cell  $m$  who did not have a final HIV status determination.

The corresponding replicate-specific weighted response rates were similarly computed for jackknife replicate  $r = 1, 2, \dots, 175$  as

$$R_{(r)m}^{(BT)} = \sum_{k=1}^{n_{(r)m}^{BT}} W_{(r)mk}^{(int)} / \left( \sum_{i=1}^{n_{(r)m}^{BT}} W_{(r)mk}^{(int)} + \sum_{i=1}^{n_{(r)m}^{NBT}} W_{(r)mk}^{(int)} \right).$$

The blood test nonresponse adjustment factor for cell  $m$  is  $A_m^{(BT)} = 1/R_m^{(BT)}$  for the full sample, and  $A_{(r)m}^{(BT)} = 1/R_{(r)m}^{(BT)}$  for jackknife replicate  $r = 1, 2, \dots, 175$ .

The full-sample nonresponse-adjusted blood test weight for interview respondent  $k$  in cell  $m$  was then computed as

$$W_{mk}^{(BT)} = A_m^{(BT)} W_{mk}^{(int)}$$

and the corresponding jackknife replicate weights for replicate  $r = 1, 2, \dots, 175$  were similarly computed as

$$W_{(r)mk}^{(BT)} = A_{(r)m}^{(BT)} W_{(r)mk}^{(int)}.$$

A summary of selected features of the blood-test nonresponse adjustment process is given in Table 3-13.

**Table 3-13 Summary of the blood test nonresponse adjustment process**

Characteristic	Male	Female
Number of variables in initial model	111	118
Number of variables selected by LASSO	33	61
Number of variables selected by CHAID	17	12
Number of final nonresponse-adjustment cells	40	16
Number of blood test respondents	7,664	11,871

Minimum adjustment factor	1.00	1.00
Maximum adjustment	1.69	1.57
Weighted count of respondents before adjustment <sup>[1]</sup>	2,725,885	3,956,685
Weighted count of respondents after adjustment <sup>[2]</sup>	2,937,436	4,202,245

[1] Weight is person interview weight,  $W_{mk}^{(int)}$ .

[2] Weight is nonresponse-adjusted blood test weight,  $W_{mk}^{(BT)}$ .

### 3.4.4.3 Poststratification Adjustment

Like the nonresponse-adjusted interview weights described previously, the nonresponse-adjusted blood test weights were poststratified to projected 2020 population counts within classes defined by gender and five-year age group.

Let  $N_{ga}^{2020}$  denote the 2020 Zimbabwe population control total for gender  $g$  and (five-year) age group  $a$  as given in Table 3-14. The poststratification ratio adjustment factor used to adjust the blood test weights for gender  $g$  and age group  $a$  was computed as:

$$T_{ga}^{2020} = N_{ga}^{2020} / \sum_{k=1}^{n_{ga}^{BT}} W_{gak}^{(BT)},$$

where  $W_{gak}^{(BT)}$  is the nonresponse-adjusted blood test weight for blood test respondent  $k$  in gender group  $g$  and age group  $a$ .

The corresponding replicate-specific adjustment factors were computed in a similar way as:

$$T_{(r)ga}^{2020} = N_{ga}^{2020} / \sum_{k=1}^{n_{(r)ga}^{BT}} W_{(r)gak}^{(BT)}$$

for the  $r = 1, 2, \dots, 175$  jackknife replicates.

The full-sample poststratified blood test weight was then computed as:

$$W_{gak}^{(ps-BT)} = T_{ga}^{2020} W_{gak}^{(BT)},$$

and the corresponding poststratified replicate weights were computed as:

$$W_{(r)gak}^{(ps-B)} = T_{ga}^{2020} W_{(r)gak}^{(BT)}$$

for  $r = 1, 2, \dots, 175$

Weighted counts of the blood test respondents before and after poststratification are summarized in Table 3-14.

**Table 3-14 2020 Zimbabwe population projections and weighted counts of blood test respondents before and after poststratification**

Age group	Male			Female			Total		
	Population control total <sup>[1]</sup>	Wtd. count before post-stratification <sup>[2]</sup>	Poststratification ratio <sup>[3]</sup>	Population control total <sup>[1]</sup>	Wtd. count before post-stratification <sup>[2]</sup>	Poststratification ratio <sup>[3]</sup>	Population control total <sup>[1]</sup>	Wtd. count before post-stratification <sup>[2]</sup>	Poststratification ratio <sup>[3]</sup>
15-19	875,183	557,006	1.571	871,128	595,560	1.463	1,746,311	1,152,566	1.515
20-24	745,086	408,254	1.825	750,035	577,716	1.298	1,495,121	985,970	1.516
25-29	594,652	318,208	1.869	664,852	518,992	1.281	1,259,504	837,200	1.504
30-34	470,910	300,141	1.569	595,670	471,012	1.265	1,066,580	771,153	1.383
35-39	422,277	301,676	1.400	501,577	453,949	1.105	923,854	755,625	1.223
40-44	372,980	229,556	1.625	402,977	326,238	1.235	775,957	555,794	1.396
45-49	275,494	215,662	1.277	293,217	295,851	0.991	568,711	511,513	1.112
50-54	185,193	135,072	1.371	196,941	199,918	0.985	382,134	334,990	1.141
55-59	143,251	104,896	1.366	182,650	205,516	0.889	325,901	310,411	1.050
60-64	110,685	109,691	1.009	176,094	176,765	0.996	286,779	286,456	1.001
65+	260,755	257,184	1.014	404,015	380,820	1.061	664,770	638,003	1.042
15+	4,456,466	2,937,346	1.517	5,039,156	4,202,335	1.199	9,495,622	7,139,681	1.330

[1] 2020 population projections provided by Zimbabwe National Statistics Agency (ZIMSTAT).

[2] Weighted count of blood test respondents using nonresponse-adjusted blood test weight,  $W_{gak}^{(BT)}$ .

[3] Ratio of population control total to weighted count of blood test respondents using nonresponse-adjusted blood test weight,  $W_{gak}^{(BT)}$ .

## References

---

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning*. Springer Series in Statistics.

Johnston, G. and Rodriguez, R (2015). Introducing the HPGENSELECT Procedure: Model Selection for Generalized Linear Models and More. Paper SAS1742-2015.

<https://support.sas.com/resources/papers/proceedings15/SAS1742-2015.pdf>

Kalton, G., and Kasprzyk, D. (1986). The treatment of missing survey data. *Survey Methodology* 12, 1-16.

Kish, L. (1965). *Survey Sampling*. New York, NY: John Wiley & Sons.

Magidson, J. (2005) SI-CHAID Users Guide. Statistical Innovations.

<https://www.statisticalinnovations.com/wp-content/uploads/SICHAIDUsersguide.pdf>

## **Appendix A**

### **Definition of Eligibility for Dwelling Unit/Household Sampling**

## **Appendix A - Definition of Eligibility for Dwelling Unit/Household Sampling**

---

The listing process was implemented by trained field staff using computer tablets. The aim in establishing eligibility was to make sure that all potentially eligible dwelling units (e.g., including vacant units or buildings under construction) are given appropriate chances of selection for the study. Based on three variables recorded for each listing in the computer tablets (the structure type, whether the structure was vacant or under construction, and whether the structure was occupied or not), an eligibility flag (ELIG\_FLAG) was assigned to each combination of values of the three variable as either being eligible for the study (ELIG\_FLAG = Y) or not (ELIG\_FLAG = N).

Table A-1 shows all possible combinations of the three relevant variables used to define eligibility status and the corresponding counts of records in the Master Listing File. Table A-2 contains a detailed description of the three variables.

Of the 39,942 dwelling unit/household records in the listing file, 2 were classified as ineligible for sampling based on the structure type, vacancy status, and residential status. Thus, a total of 39,940 records in the Master Listing File were eligible for household sampling.



**Table A-1 Definition of eligibility and number of records by eligibility status**

Structure type (STOBS)	Vacant/Constr. Status (STVAC)	Resid. Status (RESYN_D)	ELIG_FLAG	Total in master file	Eligible
Cases with no GPS information			N	0	0
Missing	1. Occupied	1. Residential	Y	1	1
1. Single House/Compound of Houses	1. Occupied	1. Residential	Y	35,312	35,312
1. Single House/Compound of Houses	1. Occupied	2. Non_Residential	Y	8	8
1. Single House/Compound of Houses	2. Temporarily Absent	1. Residential	Y	707	707
1. Single House/Compound of Houses	3. Vacant/Unoccupied	1. Residential	Y	99	99
1. Single House/Compound of Houses	3. Vacant/Unoccupied	2. Non_Residential	Y	1	1
1. Single House/Compound of Houses	4. Absent during survey	1. Residential	Y	79	79
1. Single House/Compound of Houses	5. Short term occupation	1. Residential	Y	12	12
1. Single House/Compound of Houses	6. Destroyed/Abandoned	1. Residential	Y	5	5
2. Flat/Block/Apartment Building	1. Occupied	1. Residential	Y	927	927
2. Flat/Block/Apartment Building	2. Temporarily Absent	1. Residential	Y	15	15
2. Flat/Block/Apartment Building	3. Vacant/Unoccupied	1. Residential	Y	1	1
2. Flat/Block/Apartment Building	4. Absent during survey	1. Residential	Y	9	9
3. ChurchMosque/Temple	1. Occupied	1. Residential	Y	20	20
3. ChurchMosque/Temple	2. Temporarily Absent	1. Residential	Y	1	1
4. Shop/Office/Bus. Centre/Comm. Bldg.	1. Occupied	1. Residential	Y	351	351
4. Shop/Office/Bus. Centre/Comm. Bldg.	1. Occupied	2. Non_Residential	N	2	0
4. Shop/Office/Bus. Centre/Comm. Bldg.	2. Temporarily Absent	1. Residential	Y	7	7
5. School/University	1. Occupied	1. Residential	Y	261	261
5. School/University	2. Temporarily Absent	1. Residential	Y	28	28
6. Clinic/Hospital/Doctor's Office	1. Occupied	1. Residential	Y	85	85
7. Community Centre/CBO	1. Occupied	1. Residential	Y	5	5
8. Semi-Detached	1. Occupied	1. Residential	Y	1,988	1,988
8. Semi-Detached	2. Temporarily Absent	1. Residential	Y	15	15
8. Semi-Detached	3. Vacant/Unoccupied	1. Residential	Y	3	3
Total				39,942	39,940

**Table A-2      Definition of variables used to define eligibility status**

<b>Structure Type (STOBS)</b>	
1	-Single House/Compound of Houses
2	-Flat/Block/Apartment Building
3	-Church Mosque/Temple
4	-Shop/Office/Business Centre/Commercial Building
5	-School/University
6	-Clinic/Hospital/Doctor's Office
7	-Community Centre/CBO
8	-Semi-Detached
<b>Structure vacant or under construction? (STVAC)</b>	
1	-Occupied
2	-Temporarily Absent
3	-Vacant/Unoccupied
4	-Absent during survey
5	-Short term occupation
6	-Destroyed/Abandoned
<b>Anyone living in the structure? (RESYN_D)</b>	
1	-Residential
2	-Non-Residential

## **Appendix B**

### **Definition of Household, Interview, and Blood Test Response Status**

## Appendix B - Definition of Household, Interview, and Blood Test Response Status

The response status variables required for weighting as previously described in Section 3.4.2.1 (household weights), Section 3.4.3.1 (interview weights), and Section 3.4.4.1 (blood test weights) were created using the SAS program code given below. In general, a response code of 1 is assigned to respondents, 2 to (eligible) nonrespondents, 3 to ineligible/out-of-scope cases, and 4 to cases for which eligibility is unknown.

### B.1 HH\_STATUS

#### B.1.1 Summary

HH\_STATUS is defined for all sampled DUs. First, the variable UPCODE\_RESULTNDT is derived using RESULTNDTOTHRR. Next, the questionnaire completion variable and the upcoded RESULTNDT are used to calculate UPCODE\_STAT\_HH. Lastly, HH\_STATUS is set equal to UPCODE\_STAT\_HH when the Data Lock files are delivered.

HH_STATUS	Description
1	Responding Household (Questionnaire data)
2	Eligible Household, NonRespondent (no questionnaire data)
3	Ineligible
4	Unknown eligibility Status

#### B.1.2 SAS code defining HH\_STATUS

HH\_STATUS = UPCODE\_STAT\_HH;

#### Definition for household with completed questionnaire:

UPCODE\_STAT\_HH = 1 if:

- RESULTNDT is NULL and (STARTINT = 1 AND HHELIG = 1 AND HHCONSTAT = 1 AND HHQDTHSINS is NOT NULL AND ROSTER\_MENU is NOT NULL AND HHQINSHH is NOT NULL AND HHQASSIGN\_INST is NOT NULL) OR
- RESULTNDT is NULL and (STARTINT = 4 and ROSTER\_MENU is NOT NULL)

The table below shows the values for RESULTNDT on the data file:

CANNOT COLLECT CSPRO CODE (RESULTNDT)	Map to UPCODE_STAT_HH
1 = HH NOT AVAILABLE AT ALL VISIT ATTEMPTS	2 = NONRESPONDING HH
2 = REFUSED	2 = NONRESPONDING HH
3= DWELLING VACANT OR ADDRESS NOT A DWELLING	3 = INELIGIBLE HH
4= DWELLING DESTROYED	3 = INELIGIBLE HH
5= DWELLING NOT FOUND	4 = UNKNOWN STATUS HH

6= HOUSEHOLD ABSENT FOR EXTENDED PERIOD OF TIME	3 = INELIGIBLE HH
96 = OTHER	Will be upcoded to UPCODE_RSLTNDT

**Definitions for household without completed questionnaire:**

**ELSE** assign UPCODE\_STAT\_HH to 2, 3 or 4 using rules shown below.

UPCODE\_STAT\_HH = 2 if

- RESULTNDT OR UPCODE\_RESLTNDT = 1 or 2 or 7 or 8 or 9
- If RESULTNDT=NULL, then
  - If HHELIG = 2 OR
  - (HHCONSTAT = 2 or 3) or
  - HHELIG = 1 AND HHCONSTAT=NULL OR
  - STARTINT = 4 and ROSTER\_MENU is NULL

UPCODE\_STAT\_HH = 3 if

- RESULTNDT OR UPCODE\_RESLTNDT = 3 or 4 or 6

UPCODE\_STAT\_HH = 4 if

- (RESULTNDT OR UPCODE\_RESLTNDT = 5 or 99) or
- the record does not meet the criteria for 1, 2, or 3

Tables showing upcoding scheme for RESULTNDT = '96' cases

RESULTNDT	Value label		UPCODE_STAT_HH
1	HOUSEHOLD NOT AVAILABLE AT ALL VISIT ATTEMPTS		2
2	REFUSED		2
3	DWELLING VACANT OR ADDRESS NOT A DWELLING		3
4	DWELLING DESTROYED		3
5	DWELLING NOT FOUND		4
6	HOUSEHOLD ABSENT FOR EXTENDED PERIOD OF TIME		3
96	OTHER	UPCODE_RESLTNDT	
		Suggested add'l codes	
		Bereavement related	7
		No capable Head of Household available to do survey	8
		Dwelling inaccessible	9
		Recorded in another HH or tablet (discrepant record)	99
			2
			2
			2
			4

UPCODE_STAT_HH	Value label	
1	RESPONDING HH	Use when HH_INT has completed questionnaire.
2	NONRESPONDING HH	Based on RESULTNDT or UPCODE_RESULTNDT
3	INELIGIBLE HH	Based on RESULTNDT or UPCODE_RESULTNDT
4	UNKNOWN STATUS HH	RESULTNDT or UPCODE_RESLTNDT = 5 OR RESULTNDOTH cannot be upcoded OR unresolved discrepant record

Table of examples for RESULTNDOTH upcoding

RESULTNDOTH	UPCODE_ RESLTNDT	UPCODE_ STAT_HH
<b>Not available at three occasions</b>	1	2
HOUSEHOLD HEAD TOO BUSY TO ACCOMODATE SURVEY		
HOUSEHOLD HEAD NOT AVAILABLE FOR AN EXTENDED PERIOD OF TIME		
HOUSEHOLD HEAD IS AWAY IN SOUTH AFRICA AND WIFE IS NOT ABLE TO MAKE DECISIONS OR GIVE PERMISSION		
HHH IS AN ARTISAN MINOR HE COMES BACK AROUND 10 PM AND GOES VERY EARLY IN THE MORNING AROUND 4 AM		
KEPT GIVING APPOINTMENTS BUT WAS NOWHERE TO BE FOUND ON LAST DAY		
PARTICIPANT 'S WORK SHIFTS COULD NOT ACCOMMODATE SURVEY ACTIVITIES TO BE CONDUCTED.		
<b>Refusing Behavior</b>	2	2
COULD NOT ACCOMODATE SURVEY DUE TO RELIGIOUS AFFILIATION.THEY ARE FROM THE JOHANNE MARANGE CHURCH		
DATA CANNOT BE COLLECTED DUE TO STRONG RELIGIOUS BELIEF		
HEAD OF HOUSE STATED THAT IF THERE ARE NO MONETARY BENEFITS HIS HOUSEHOLD SHOULD NOT BE INCLUDED		
PARTICIPANT REFUSED TO PARTICIPATE IN THE SURVEY AND THE REASON BEING DOMESTIC ISSUES.		
THE FAMILY WAS RECENTLY ATTACHED AND ROBBED BY ARMED ROBBERS AT GUN POINT. WRONG TIMING		
HH HEAD LISTED AGREED HOWEVER THE SON IS NOT ALLOWING THE PROCEDURES TO BE DONE		
<b>Death/Funeral</b>	7	2
SHE LOST HER BOYFRIEND WHO WAS BURIED LAST SUNDAY. HE DIED OF LIVER PROBLEMS IN SOUTH AFRICA		
FUNERAL AT THE HOUSEHOLD		
GRIEVING.SHE RECENTLY LOST A SON AND MOURNERS ARE STILL GATHERED.		
NOT IN AN EMOTIONAL STATE TO PARTICIPATE, HH MISSING, DEATH OF A GRANDCHILD AND BIRTH OF CHILD		
CLOSE RELATIVE (DAUGHTER IN LAW) TO THE DECEASED BURIAL SCHEDULED ON 01/12/19		

Table of examples for RESULTNDOTH upcoding - continued

RESULTNDOTH	UPCODE_ RESLTNDT	UPCODE_ STAT_HH
<b>Participant/Household Head unable to do survey (incapacitated, language barrier, under age)</b>	<b>8</b>	<b>2</b>
HOUSEHOLD HEAD INCAPACITATED MENTALLY CHALLENGED		
THE PARTICIPANT IS INCAPACITATED -DEAF		
SINGLE HOUSEHOLD MEMBER WHO IS TOO OLD AND INCAPACITATED.		
HH IS 14 YEARS OLD SO PARTICIPANT IS INELIGIBLE		
HOUSEHOLD HEAD UNABLE TO SPEAK ANY OF THE SURVEY LANGUAGES.		
THE HOUSEHOLD HEAD PASSED ON IN BULAWAYO ON THE 3RD DAY VISIT. NO ONE TO CONSENT FOR THE HOUSEHOLD		
HOUSEHOLD HEAD INVOLVED IN A CAR ACCIDENT THEREFORE CANNOT ACCOMODATE AN INTERVIEW		
<b>Dwelling inaccessible</b>	<b>9</b>	<b>2</b>
DWELLING CANT BE REACHED ROADS SLIPPERY DUE TO RAINS AND BAD TERRAIN		
HOUSEHOLD INACCESSIBLE BECAUSE OF A FLOODED STREAM FOR TWO DAYS		
<b>Vacant or not a dwelling</b>	<b>3</b>	<b>3</b>
STRUCTURE UNDER CONSTRUCTION STILL AT FOUNDATION LEVEL		
NO ONE SLEEPS AT THE HOUSE		
HOUSEHOLD HEAD DECEASED. DWELLING VACANT		
VACANT		
DWELLING IS A BOTTLESTORE		
<b>Household absent for extended period of time</b>	<b>6</b>	<b>3</b>
MEMBERS OF THE HOUSEHOLD HAVE TRAVELLED FOR A LONG PERIOD OF TIME		
THE INDIVIDUAL STAYS ALONE AND HE HAS TRAVELLED TO ARGENTINA AND THERE IS NOONE STAYING AT THE HOUSE		



## B.2 INDIV\_STATUS

### B.2.1 Summary

INDIV\_STATUS is defined for all final roster records. This variable is derived when the Data Lock files are delivered.

INDIV_STATUS	Description
1	Respondent
2	Eligible non-Respondent
3	Roster eligible but confirmed age <15
4	Roster eligible but no confirmed age
5	Roster ineligible (roster age < 15 or SLEEPHERE=2, except cases in status 9)
6	Rostered case from household with no questionnaire data
9	DeJure ineligible (SLEEPHERE = 2, LIVEHERE = 1 and roster age >=15)

### B.2.2 SAS Code for INDIV\_STATUS

First create a variable to designate whether the case is survey eligible based on the roster:

```
label roster_elig = "Flag for roster eligible";
if hh_status ^= 1 then roster_elig = 2;
else
  if sleephere = 1 and
    ageyears => 15 then roster_elig = 1;
  else
    roster_elig = 0;
```

Next, combine Roster\_Elig with endmsg1 and Confagey to create INDIV\_STATUS

(endmsg1 = 'A' indicates a completed Individual questionnaire)

```
label INDIV_STATUS = "Individual Response Status";
if roster_elig = 2 then indiv_status = 6;
else
  if roster_elig = 0 then do;
    If sleephere = 2 and
      livehere = 1 and
      ageyears >= 15 then indiv_status = 9;
  else
    indiv_status = 5;
end;
```

```

else
  if confagey => 15 and
    endmsg1 = "A" then indiv_status = 1;
  else
    if confagey => 15 and
      endmsg1 = " " then indiv_status = 2;
    else
      if confagey ^= . and
        confagey < 15 then indiv_status = 3;
      else
        if confagey = . then indiv_status = 4;
      end;
    end;
  end;
run;

```

### B.3 BT\_STATUS

#### B.3.1 Summary

BT\_STATUS is only defined for cases where INDIV\_STATUS = 1. It is based on information from the Biomarker data set.

BT_STATUS	Description
1	Blood test respondent (Interview respondent with valid HIV lab result)
2	Blood test nonrespondent (Interview respondent with no valid HIV lab result)

#### B.3.2 SAS Code for BT\_STATUS

ATTRIB BT\_STATUS LABEL="Blood test disposition code: 1 = Valid lab results, 2 = No valid lab results or didn't do BT;

```
IF HIV1statusfinalsurvey IN ("Positive" "Negative") THEN BT_STATUS=1;
```

```
ELSE BT_STATUS=2;
```

## **Appendix C**

### **CHAID Trees and Definition of Final Nonresponse-Adjustment Weighting Cells**

## **Appendix C - CHAID Trees and Definition of Final Nonresponse-Adjustment Weighting Cells**

---

### **C.1 Final CHAID Trees**

The final CHAID trees used to construct the weighting cells for nonresponse adjustment are documented in PDF files in the zipped file Appendix\_C.zip. There are three PDF files corresponding to the groups for which the CHAID analysis was conducted for adjustment of the interview weights (Section 3.4.3.2) and the blood test weights (Section 3.4.4.2). The names of the PDF files containing the CHAID trees are listed below. Each tree indicates diagrammatically how the cells were created by successively partitioning the sample into subsets with similar response propensities. The final cells (prior to collapsing, if done to control variation in weights) are indicated by the number underneath the box defining the cell.

#### **Individual Interview**

AD\_INDIV\_STATUS.pdf (Persons 15+ years)

#### **Blood Test**

AM\_BTEST.pdf (Males 15+ years)

AF\_BTEST.pdf (Females 15+ years)

### **C.2 Final Nonresponse-Adjustment Weighting Cells**

The final nonresponse-adjustment weighting cells are documented in Excel files in the zipped file Appendix\_C.zip. There are three Excel files corresponding to the groups for which the nonresponse adjustments were made. The names of the Excel files are listed below. Each row of the Excel file corresponds to a weighting cell, and shows the variables and the corresponding values used to define the weighting cell, the numbers of responding and nonresponding cases in the cell, the weighted counts of the responding and nonresponding cases, the weighted response rate, and the nonresponse weight adjustment factor (which is defined to be the reciprocal of the weighted response rate). In some cases, cells were combined to control the variation in weights. The combined cells have the same final adjustment cell number and are highlighted in the tables.

#### **Individual Interview**

zim\_AD\_INDIV.xlsx (Persons 15+ years)

**Blood Test**

zim\_AM\_BT.xlsx (Males 15+ years)

zim\_AF\_BT.xlsx (Females 15+ years)