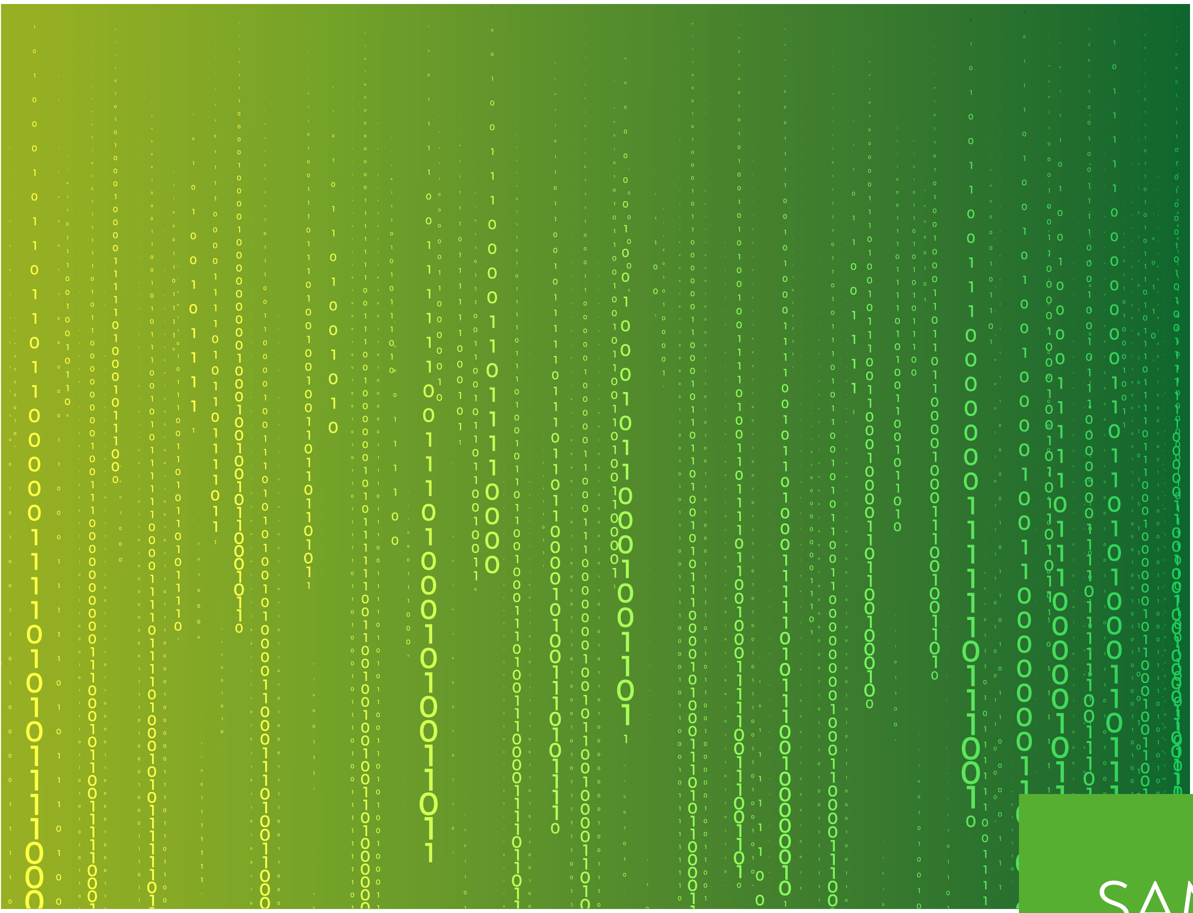
LePHIA 2020

LESOTHO
POPULATION-BASED
HIV IMPACT ASSESSMENT



SAMPLING AND WEIGHTING TECHNICAL REPORT

BOPHELO
MINISTRY OF HEALTH

ΣBOS Bureau of Statistics

PEPFAR
U.S. President's Emergency Plan for AIDS Relief

CDC
CENTERS FOR DISEASE CONTROL AND PREVENTION

Westat

icap Global Health

Lesotho Population-based HIV Impact Assessment 2020

LePHIA 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

Section **Page**

Contents continued

Section **Page**

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
LBOS	Lesotho Bureau of Statistics
LePHIA	Lesotho Population-based HIV Impact Assessment
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

1. Introduction

The 2020 Lesotho Population-based HIV Impact Assessment (LePHIA 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 LePHIA 2020 was conducted between December 2019 and March 2020 with approximately 16,500 interviewed individuals and 15,300 individuals with valid blood tests in approximately 9,700 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 LePHIA 2020 is a stratified multistage probability sample design, with strata defined to be the 10 districts 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 updated numbers of households in the PSU derived from the 2016 Lesotho Population and Housing Census. The allocation of the sample PSUs to the 10 districts was made in a manner designed to achieve specified precision levels for (a) national estimate of HIV incidence among persons 15 years of age and older; and (b) provincial estimates of viral load suppression (VLS) rates among HIV-positive persons 15 years of age and older.

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 the sample design employed for the LePHIA 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 differential 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 LePHIA 2020 are provided in Section 3.

2. Sample Design

2.1 Population of Inference

The population of inference for the LePHIA 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 2020 Lesotho population by gender and age group.

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

Age group	Gender		Total
	Male	Female	
15 to 49 years	565,477	539,568	1,105,046
50 years or older	128,522	192,268	320,791
Total	693,999	731,836	1,425,837

Source: Population projections provided by Lesotho Bureau of Statistics (LBOS)

2.2 Precision Specifications and Assumptions

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

2.2.1 Specifications

- Relative standard error (RSE) of the national estimate of HIV incidence among adults 15 to 49 years old should be 30% or less.

- 95% confidence interval (CI) bounds around the proportion virally suppressed among HIV positive adults aged 15 to 49 years for each of the 10 districts of the country should be ±0.10 or less.

- 95% confidence interval (CI) bounds around the national estimate of the proportion virally suppressed (i.e., VLS rate) among all HIV positive adults aged 15 to 49 years should be ±0.02 or less.

- 95% confidence interval (CI) bounds around the national estimate of VLS rate among all HIV positive females aged 15 to 24 years should be ±0.06 or less.

2.2.2 Statistical Assumptions

- A national HIV prevalence rate of 0.243 (24.3%) for adults 15-49 years old that varies by province (e.g., see Table 2-2). Source: LePHIA 2016-2017 (e.g., see Lesotho Population-based HIV Impact Assessment (LePHIA) 2016-2017: Final Report).

- A national HIV prevalence rate of 0.111 (11.1%) for women aged 15 to 24 years old that varies by province (see Table 2-2). Source: LePHIA 2016-2017.

- An annual national incidence rate for adults aged 15-49 of p_a = 0.0119 (1.19%). Source: LePHIA 2016-2017.

- District-level incidence rates of p_{ah}, *h* = 1, 2, …, 10, which are obtained by adjusting the national incidence rate using the district-level prevalence rates as follows:

  $$p_{ah} \ = \ (p_h/p)\, p_a \ ,$$

  where p_h and p are the HIV prevalence rates for district h and the country, respectively, and p_a is the annual national incidence rate obtained from the LePHIA 2016-2017.

- 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 = 0.0119/2.8077 = 0.0042 (0.42%). The corresponding district-level 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 district. This assumption provides a conservative estimate of the underlying population variance associated with VLS rate.

- An intracluster correlation (ICC) of 0.015 for VLS and 0.005 for prevalence. Source: tabulations of LePHIA 2016-2017 data.

- An intracluster correlation (ICC) of 0.000 for incidence. Source: analyses of prior PHIA surveys.

- Overall sex-age distributions derived from the LePHIA 2016-2017.

- District-level population estimates obtained from the 2016 Lesotho Population and Housing Census.

2.2.3 Operational Assumptions

- Varying numbers of households to be sampled per PSU, resulting in an average of 35 sampled households per PSU.

- An occupancy rate of 90.9% for sampled dwelling units based on the LePHIA 2016-2017.

- A household response rate of 89.1% among occupied dwelling units based on the LePHIA 2016-2017.

- An average household size of 3.05 (*de facto)* persons per household (LePHIA 2016-2017). The *de facto* population consists of persons of all ages who were present in the household during the night prior to the interview.

- An overall percentage of *de facto* persons 15-49 years of age per household of 47.0%; and an overall percentage of *de facto* persons 50+ years of age of 17.1% (LePHIA 2016-2017).

- Within the responding households, a person-level interview response rate of 91.4% (LePHIA 2016-2017).

- Among persons completing the interview, a blood test response rate of 90.3% (LePHIA 2016-2017). Thus, among the persons selected for the PHIA sample, the overall response rate for the blood tests is 82.5% (91.4% * 90.3%).

Based on the specifications and assumptions listed above, a sample of 342 EAs (clusters) was determined to be the minimum needed to meet the specified precision goals. The allocation of the sample to the 10 districts of Lesotho 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 stratum

District code	District name	HIV prevalence rate[1]		Total no sample clusters	Target no. of DUs to be sampled	No. of participating HHs[2]	Projected no. of respondents[3]	
		Adults 15-49	Females 15-24				Adults 15-49	Adults 50+
1	Botha-Bothe	0.1640	0.0720	22	770	624	738	269
2	Leribe	0.2290	0.1060	55	1,925	1,559	1,845	672
3	Berea	0.2160	0.1000	41	1,435	1,162	1,376	501
4	Maseru	0.2670	0.1320	93	3,255	2,636	3,121	1,137
5	Mafeteng	0.2480	0.1140	30	1,050	850	1,007	367
6	Mohale's Hoek	0.2770	0.1080	27	945	765	906	330
7	Quthing	0.2440	0.1160	20	700	567	671	244
8	Qacha's Nek	0.2490	0.0750	16	560	454	537	196
9	Mokhotlong	0.2500	0.1080	17	595	482	570	208
10	Thaba Tseka	0.2450	0.0900	21	735	595	705	257
All	Lesotho	0.2430	0.1110	342	11,970	9,695	11,476	4,180

[1] Source: LePHIA 2016-2017.

[2] Assumes occupancy rate of 90.9% and household response rate of 89.1%.

[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 LePHIA 2020 were selected from a sampling frame of enumeration areas (EAs) that originally had been created for the 2016 Lesotho Population Census, and subsequently updated by the Lesotho Bureau of Statistics in 2017. The enumeration areas in the updated sampling frame were generally the same as those created for the 2016 Population Census, except that the measures of size of EAs had been updated to reflect current information. The updated sampling frame consisted of slightly over 5,600 EAs containing an estimated 540,000 households as of 2017.

2.3.2 Selection of the PSU Sample

A stratified sample of 342 EAs was selected from the updated 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 LePHIA 2016-2017, the following procedure was used to select the EAs for the LePHIA 2020. Within each district, the EAs in the updated sampling were sorted in the same way they had been sorted in the LePHIA 2016-2017 frame to the extent feasible; i.e., by constituency within district, community council within constituency, zone within community council, settlement code (urban, periurban, rural) within zone, and finally by EA within settlement code. The sorting of EAs prior to sample selection induces an implicit geographic substratification within each district.

Next, a systematic sample of the same number of EAs selected for the LePHIA 2016-2017 was selected from the each district using a random starting point that was offset by a specified amount to minimize selecting EAs that had been selected for the LePHIA 2016-2017, 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 district, 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 district divided by the number of EAs to be selected. The resulting sample has the property that the probability of selecting an EA within a district is proportional to the MOS of the EA.

Since the number of EAs required for the LePHIA 2020 (see Table 2-2) was less than that specified for the LePHIA 2016-2017 for every district, 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 342 sampled EAs, 21 had been selected previously for the LePHIA 2016-2017. Each of the 21 overlapping 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 LePHIA 2020 sample.

2.3.4 Substitution

Under the general procedures established for PHIA, a sampled PSU that contains eligible dwelling units can be replaced if it cannot be entered (e.g., roads/bridges or other means of entry are temporarily closed, access points are flooded, the area contains army barracks or government facilities for which entry is prohibited), military conflict or other dangerous conditions, or failure to receive permission to visit sampled areas when such approval is needed. There were no such inaccessible PSUs in the LEPHIA 2020 sample.

2.3.5 Segmentation

Of the 342 sampled PSUs, 41 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 41 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 LePHIA 2020.

2.3.6 Summary of the PSU Sample

As indicated in the previous sections, 342 PSUs (EAs) were selected for the LePHIA 2020. Of these, 21 were found to have been selected previously for the LePHIA 2016-2017, and were replaced to avoid going back to the same PSUs that had been surveyed earlier. Of the 342 PSUs included in the LePHIA 2020, 41 were segmented because they were too large to be canvassed in their entirety. There were no out-of-scope (ineligible) or nonresponding eligible PSUs. Table 2-3 summarizes the distribution of the sampled PSUs by district and sampling status of the PSU.

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

District code	District name	Sample PSUs	PSUs replaced due to overlap with LEPHIA 2016	Number of segmented PSUs	Number of inscope PSUs included in study
1	Botha-Bothe	22	3	4	22
2	Leribe	55	6	6	55
3	Berea	41	4	7	41
4	Maseru	93	2	7	93
5	Mafeteng	30	1	6	30
6	Mohale's Hoek	27	0	3	27
7	Quthing	20	0	1	20
8	Qacha's Nek	16	5	2	16
9	Mokhotlong	17	0	1	17
10	Thaba Tseka	21	0	4	21
All	Lesotho	342	21	41	342

2.4 Selection of Households

The selection of households for the LePHIA 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 38,500 eligible dwelling units/households were listed for the LePHIA 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 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 district

District code	District name	Number of records (DUs) in listing file[1]	Number of unoccupied DUs[2]	Number of unoccupied DUs eligible for sampling[3]	Number of occupied DUs eligible for sampling[4]	Total number of DUs/households eligible for sampling
1	Botha-Bothe	2,585	94	93	2,491	2,584
2	Leribe	7,109	370	366	6,733	7,099
3	Berea	4,297	248	248	4,049	4,297
4	Maseru	10,492	280	278	10,211	10,489
5	Mafeteng	3,321	149	148	3,170	3,318
6	Mohale's Hoek	2,844	160	160	2,683	2,843
7	Quthing	2,124	162	162	1,962	2,124
8	Qacha's Nek	1,659	68	68	1,591	1,659
9	Mokhotlong	2,103	44	44	2,057	2,101
10	Thaba Tseka	2,011	95	95	1,916	2,011
All	Lesotho	38,545	1,670	1,662	36,863	38,525

[1] See Appendix A for additional details.

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

[3] Subset of the unoccupied DUs that could potentially become residential units by the time of data collection.

[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 10 sampling strata (districts), 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 time of 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 district $h = 1, 2, ..., 10$, was computed as:

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

where

T_h = target sample size for district *h* given in Table 2-2 ;

m_h = number of sample PSUs in district *h* ;

N_{hi} = number of eligible dwelling units in PSU *i* in district *h* based on listing counts;

P_{hi} = probability of selecting PSU *i* in district *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 district, 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 districts is $\sum_{h=1}^{10} T_h$ = 11,970 (see Table 2-2). To obtain an equal probability sample within district *h*, the required within-PSU sampling rate for PSU *i* in district *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:

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

Inspection of the values of $\mathrm{E}(n_{hi})$ indicated that the expected sample sizes for three PSUs would fall below 15, and none would exceed 70. For the three PSUs with an expected sample size below 15, the sample size was set to a value 15 to ensure a minimum acceptable workload 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 minimum number was then re-distributed to the other PSUs in the same stratum so as to maintain as closely as possible the desired total sample size for the stratum. The within-PSU sampling rates, f_{hi}^{within}, were therefore adjusted to account for 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 = 15 = the minimum PSU sample size, U = 70 = 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 numbers of dwelling units/households selected for the study and the minimum and maximum PSU sample size by district. 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 overall sampling rates within a district (e.g., see Kish, 1965, page 403). With an equal-probability sample within each district, the design effects would ordinarily equal 1.0. Variable sampling rates will increase the design effect, which would arise, for example, from the capping of sample sizes that is done to control workload across EAs. However, since the extent of the capping and redistribution of the sample described previously was minimal, the corresponding increase in the variation of the overall sampling rates was also minimal, resulting in district-level UEW design effects that are for all practical purposes equal to 1.00 (Table 2-5).

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

District code	District name	Number of PSUs	Number of sampled DUs/house-holds	Minimum number of DUs selected per PSU	Maximum number of DUs selected per PSU	Unequal weighting design effect
1	Botha-Bothe	22	770	25	52	1.00
2	Leribe	55	1,924	26	57	1.00
3	Berea	41	1,435	15	47	1.00
4	Maseru	93	3,256	15	61	1.00
5	Mafeteng	30	1,050	15	55	1.00
6	Mohale's Hoek	27	945	26	59	1.00
7	Quthing	20	700	18	58	1.00
8	Qacha's Nek	16	559	25	66	1.00
9	Mokhotlong	17	594	22	59	1.00
10	Thaba Tseka	21	735	28	50	1.00
All	Lesotho	342	11,968	15	66	1.01[1]

[1] Overall DEFF reflects total variation in weights within and across districts.

Table 2-6 summarizes the distribution of the sampled dwelling units/households by final household response status. Of the 11,968 sampled dwelling units 1,583 (13.2%) were determined during data collection to be vacant/unoccupied, 103 (0.9%) for which eligibility for the survey (i.e., occupancy status) could not be established, 617 (5.2%) were determined to be eligible for the study (i.e., contained eligible household members) but did not complete the household interview, and 9,665 (80.8%) completed the household interview. Excluding the ineligible cases, the overall unweighted household response rate was 93.2%.

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

District code	District name	Number of sampled DUs	Number of ineligible DUs/ households[1]	Number of DUs with unknown eligibility[2]	Number of responding households[3]	Number of eligible non-responding households[4]	Unweighted response rate[5]
1	Botha-Bothe	770	75	7	678	10	0.977
2	Leribe	1,924	237	10	1,616	61	0.959
3	Berea	1,435	214	11	1,099	111	0.901
4	Maseru	3,256	346	42	2,613	255	0.899
5	Mafeteng	1,050	152	6	867	25	0.966
6	Mohale's Hoek	945	166	12	728	39	0.937
7	Quthing	700	148	3	508	41	0.921
8	Qacha's Nek	559	88	2	456	13	0.969
9	Mokhotlong	594	48	3	504	39	0.923
10	Thaba Tseka	735	109	7	596	23	0.954
All	Lesotho	11,968	1,583	103	9,665	617	0.932

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

[2] Sampled dwelling units for which existence of eligible households could not be ascertained.

[3] Households completing the household interview.

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

[5] Computed as R/ [R + N + U*{(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 LePHIA 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 district and resident status.

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

District code	District name	Number of households completing interview	Rostered persons by resident status[1]				
			Usual resident/did not sleep here[2]	Usual resident/ slept here	Nonresident/ slept here	Nonresident/ did not sleep here[2]	Total rostered persons
1	Botha-Bothe	678	215	1,867	108	239	2,429
2	Leribe	1,616	596	4,073	166	381	5,216
3	Berea	1,099	476	2,944	257	405	4,082
4	Maseru	2,613	692	6,435	213	443	7,783
5	Mafeteng	867	349	2,239	229	362	3,179
6	Mohale's Hoek	728	238	1,915	119	296	2,568
7	Quthing	508	239	1,407	80	218	1,944
8	Qacha's Nek	456	261	1,332	78	180	1,851
9	Mokhotlong	504	217	1,512	100	182	2,011
10	Thaba Tseka	596	199	1,732	81	207	2,219
All	Lesotho	9,665	3,482	25,456	1,431	2,913	33,282

[1] Counts include persons of all ages.

[2] Not eligible to be surveyed for LePHIA 2020.

2.5.2 Selecting Individuals for Data Collection

All individuals listed in the household rosters who were 15 years of age and older and were present (slept in the household) on the night prior to the household interview were eligible for data collection. Excluded are usual residents and any rostered nonresidents who were not present in the household on the night prior to the interview. Table 2-8 summarizes the number of individuals eligible for data collection by district, age group, and resident status.

Table 2-8 Number of individuals eligible for data collection

		Persons 15-49 years[1]			Persons 50 years or older[1]		
District code	District name	Usual resident/ slept here	Nonresident/ slept here	Total sampled persons[2]	Usual resident/ slept here	Nonresident/ slept here	Total sampled persons[2]
1	Botha-Bothe	875	70	945	326	20	346
2	Leribe	1,991	111	2,102	690	27	717
3	Berea	1,366	138	1,504	541	19	560
4	Maseru	3,502	156	3,658	958	27	985
5	Mafeteng	1,014	118	1,132	447	14	461
6	Mohale's Hoek	812	74	886	389	23	412
7	Quthing	543	46	589	325	5	330
8	Qacha's Nek	529	44	573	259	7	266
9	Mokhotlong	667	82	749	228	6	234
10	Thaba Tseka	784	49	833	305	19	324
All	Lesotho	12,083	888	12,971	4,468	167	4,635

[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 Sampled Persons

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 district. 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 who provided analyzable blood test results (i.e., had 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 by age group, response status, and district

		Persons 15-49 years[1]			Persons 50 years or older[1]		
District code	District name	Selected for data collection	Interview respondents[2]	Blood test respondent[3]	Selected for data collection	Interview respondents[2]	Blood test respondent[3]
1	Botha-Bothe	945	899	850	346	335	324
2	Leribe	2,102	1,993	1,894	717	685	654
3	Berea	1,504	1,346	1,179	560	530	499
4	Maseru	3,658	3,381	3,034	985	930	867
5	Mafeteng	1,132	1,052	1,007	461	433	409
6	Mohale's Hoek	886	822	769	412	399	389
7	Quthing	589	540	489	330	317	308
8	Qacha's Nek	573	533	499	266	250	240
9	Mokhotlong	749	678	631	234	218	210
10	Thaba Tseka	833	808	781	324	319	316
All	Lesotho	12,971	12,052	11,133	4,635	4,416	4,216

[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 LePHIA 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 and 3.4.4 are used to calibrate the weighted sample counts to available population projections.

3.1 Overview of the Weighting Process

The overall weighting approach for LePHIA 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 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. This adjustment is 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 LEPHIA 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.

- **Ls_CFF_hh_int_STAT_20200602**: A household (HH) file that contains the household data collected in the HH questionnaire.
- **Ls_CFF_roster_STAT_20200602**: A file that contains the roster of household members collected in the HH questionnaire with a record for each rostered person.
- **Ls_CFF_ind_int_STAT_20200602**: 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.
- **LS2Biomarker20200603**: 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 Biomarker 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 (b) 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 LePHIA 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 LePHIA 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 2.3.2, the PSUs were selected systematically from a list of PSUs that had been ordered geographically within district. To take account of the precision benefits of implicit stratification as fully as possible, the sampled PSUs within each district 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 district, 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 342 of the sampled PSUs.

For the LePHIA 2020, 168 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 weights 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 $r = 1, 2, ..., 168$ variance-estimation strata, resulting in a total of 168 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

District code	District name	Sampled PSUs	Variance strata consisting of pairs	Variance strata consisting of triplets	Number of jackknife replicates
1	Botha-Bothe	22	11	0	11
2	Leribe	55	26	1	27
3	Berea	41	19	1	20
4	Maseru	93	45	1	46
5	Mafeteng	30	15	0	15
6	Mohale's Hoek	27	12	1	13
7	Quthing	20	10	0	10
8	Qacha's Nek	16	8	0	8
9	Mokhotlong	17	7	1	8
10	Thaba Tseka	21	9	1	10
All	Lesotho	342	162	6	168

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 LePHIA 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 district *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, 168 jackknife replicates were formed from the 356 sampled PSUs. For variance estimation, replicate-specific PSU weights, $W_{(r)hi}^{(1)}$, *r* = 1, 2, ..., 168 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, ..., 168, the corresponding replicate-specific PSU base weight was computed as

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

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 corresponding weighted counts by district

District code	District name	Sampled PSUs	Weighted number of PSUs[1]	Weighted measure of size (MOS)[2]
1	Botha-Bothe	22	269	30,127
2	Leribe	55	856	90,386
3	Berea	41	736	70,192
4	Maseru	93	1,762	158,678
5	Mafeteng	30	492	46,526
6	Mohale's Hoek	27	507	40,813
7	Quthing	20	324	26,297
8	Qacha's Nek	16	202	17,586
9	Mokhotlong	17	249	24,362
10	Thaba Tseka	21	393	33,510
All	Lesotho	342	5,790	538,477

[1] Weights are the PSU base weights, $W_{hi}^{(1)}$. The weighted count provides an estimate of the number of PSUs in the sampling frame.

[2] The measure of size used to select the sample of PSUs; i.e., the updated 2017 count of households in the PSU.

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 district *h* was computed as:

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

where

$W_{hi}^{(1)}$ = the weight for PSU *i* in district *h*

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

The corresponding weights for jackknife replicate *r* = 1, 2, …, 168 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 district *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. In Table 3-4, we show the weighted counts of dwelling units/households by response status and district 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 district.

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

Response status group[1]	Description	Number of sampled dwelling units/households
1	Respondent (household with completed household interview)	9,665
2	Nonrespondent (household without a completed household interview)	617
3	Ineligible (dwelling units with no households)	1,583
4	Unknown eligibility (not known if dwelling unit contains household)	103
All	—	11,968

[1] See Appendix B for definitions.

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

District code	District name	Response status[1]				Total groups 1-4
		Group 1: responding household	Group 2: nonresponding household	Group 3: ineligible dwelling unit	Group 4: unknown eligibility	
1	Botha-Bothe	32,198	475	3,562	332	36,567
2	Leribe	95,866	3,619	14,060	593	114,138
3	Berea	65,042	6,570	12,667	651	84,931
4	Maseru	159,542	15,567	21,131	2,565	198,805
5	Mafeteng	51,502	1,486	9,030	357	62,374
6	Mohale's Hoek	42,651	2,285	9,725	703	55,365
7	Quthing	22,361	1,805	6,515	132	30,813
8	Qacha's Nek	19,191	547	3,704	84	23,526
9	Mokhotlong	25,582	1,980	2,436	152	30,150
10	Thaba Tseka	31,212	1,204	5,708	367	38,491
All	Lesotho	545,147	35,538	88,537	5,937	675,158

[1] See Table 3.3. Counts given in table are weighted counts using the dwelling unit base weights, $W_{hij}^{(2)}$ described in Section 3.4.2.1.

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. 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 LePHIA 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 or group of PSUs. The procedures used to compute the nonresponse-adjusted dwelling unit/household weights are described 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 district *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

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

$n_{hi}^{(2)}$ = the number of eligible nonresponding households (i.e., households without a completed household interview) in PSU weighting cell *i* in district *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 district *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 district *h* .

The first-phase nonresponse adjustment factor for PSU weighting cell *i* in district *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 $W_{hij}^{(2)}$ is the base weight for dwelling unit/household *j* in PSU weighting cell *i* in district *h*, and where the sum in the numerator extends over the entire sample of dwelling units/households in PSU weighting cell *i* in district *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 district *h* was then computed as:

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

The corresponding replicate weights for replicate r = 1, 2, …, 168 were computed in similar fashion as:

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

where

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

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, …, 168.

The effect of this adjustment is to distribute the total weight of the unknown-eligibility cases (i.e., the estimated 5,937 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

District code	District 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	
1	Botha-Bothe	32,489	487	3,590	36,567	32,976
2	Leribe	96,363	3,634	14,141	114,138	99,997
3	Berea	65,529	6,643	12,758	84,931	72,172
4	Maseru	161,498	15,771	21,536	198,805	177,269
5	Mafeteng	51,821	1,492	9,060	62,374	53,314
6	Mohale's Hoek	43,235	2,313	9,817	55,365	45,548
7	Quthing	22,462	1,819	6,532	30,813	24,281
8	Qacha's Nek	19,262	553	3,711	23,526	19,815
9	Mokhotlong	25,707	1,987	2,455	30,150	27,694
10	Thaba Tseka	31,528	1,216	5,747	38,491	32,744
All	Lesotho	549,895	35,915	89,348	675,158	585,810

Note: Counts in table are weighted counts using first-phase adjusted household weights, W_{hij}^{DU1}.

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 district *h* was computed as the ratio:

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

where W_{hij}^{DU1} is the first-phase adjusted weight for dwelling unit/household *j* in PSU weighting cell *i* in district *h*, and where the sum in the numerator extends over the sample of responding and nonresponding households in PSU weighting cell *i* in district *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 district *h* was then computed as:

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

The corresponding replicate weights for replicate $r = 1, 2, \ldots, 168$ 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 LePHIA 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 district *h* was computed from the formula

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

The corresponding replicate base weights, $W_{(r)hijk}^{(base)}$, for $r = 1, 2, \ldots, 168$ 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 were either under 15 years old, or because they were neither present in the household nor a usual resident of the household at the time the household roster was compiled. 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 by age group and first-phase response status

First-phase response status group[1]	Resident status and age based on roster	Confirmed age based on interview	Number of rostered persons	Weighted number of rostered persons[2]
1	*De facto* person 15 years or older	15+	17,590	1,064,872
2	*De facto* person 15 years or older	Under 15	0	0
3	*De facto* person 15 years or older	Unknown	16	996
4	Non *de facto* persons 15 years or older	NA	5,561	328,739
5	Persons under 15 years	NA	10,115	602,447
All	—	—	33,282[3]	1,997,055[3]

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

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

[3] Of the 33,282 rostered persons, 2,913 were those that neither slept in the household nor were usual residents (see Table 2-7). On a weighted basis, these 2,913 persons account for 171,231 of the total weighted count of 1,997,055 rostered persons.

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

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

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, ..., 168 as

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

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, ..., 168.

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, ..., 168 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 district 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[1]	Interview response status[2]	Unweighted sample size	Weighted count[3]
Male 15 or older	Eligible respondent	6,745	406,754
	Eligible nonrespondent	736	45,722
	All response statuses	***7,481***	***452,477***
Female 15 or older	Eligible respondent	9,723	589,234
	Eligible nonrespondent	386	24,157
	All response statuses	***10,109***	***613,392***
Total 15 years or older	Eligible respondent	16,468	995,989
	Eligible nonrespondent	1,122	69,880
	All response statuses	***17,590***	***1,065,868***

[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 adjusted person weight, $W_{hijk}^{(3)}$.

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

There are 50 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 subsequently 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 not significant 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 produced. Of the 50 variables used in the initial model, the LASSO identified 18 variables as 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 for responding to 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, 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 (variable SEX) and an indicator of whether or not the individual was under 18 years of age (H_AGETEENYEARS) were forced into the model to make the initial splits. The reason for doing this was because males and females in the specified age groups received different questions; without forcing this variable into the model, the resulting tree would not have been created correctly. After forcing these two variable in the model, the tree was then allowed to grow freely. The CHAID algorithm identified 17 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	HHQHUNGFRQ	Hh Food Support: How Often Did This Happen In The Past 4 Weeks?
2	H_AGETEENYEARS	Teen Indicator: 1 – 15-17 Years Old; 2 – Otherwise; Based On Ageyears (Roster)
3	H_AGEYEARS	Age (Categorical), Based On Roster Age. Matches Poststratification Cells
4	H_ECONSUP12_B	B. Cash Transfer
5	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
6	H_HH_SIZE_C	1-9, Where 9 Includes All Hhs With 9 Or More People
7	H_MATEXWALLS	Main Material Of Exterior Walls
8	H_OWNDOGNUM	Altogether, How Many Of The Below Listed Animals Do Members Of Your Household Own?
9	H_OWNHORSENUM	Work Animals (Camels, Horses, Donkeys): Hh Characteristics: Altogether, How Many Of The Below Listed Animals Do Members Of Your Household Own?
10	H_OWNPIGNUM	Pigs: How Many Of The Below Listed Animals Do Members Of Your Household Own?
11	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
12	H_WATERSOURCE	What Is The Main Source Of Drinking Water For Members Of Your Household?
13	SEX	Is [name] Male Or Female?
14	SICK3MO	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?
15	SICK_HOUSEHOLD	Any Member Of The Household Has Answered That They Are Sick
16	STRATA	District Code (Sampling Stratum Code)
17	URBAN_RURAL	1=urban, 2=peri-Urban, 3=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, ..., 168 as

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

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, ..., 168.

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, ..., 168 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	50
Number of variables selected by LASSO	18
Number of variables selected by CHAID	17
Number of final nonresponse-adjustment cells	64
Number of interview respondents	16,468
Minimum adjustment factor	1.00
Maximum adjustment	1.71
Weighted count of respondents before adjustment[1]	995,989
Weighted count of respondents after adjustment[2]	1,065,868

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

[2] 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 Lesotho Bureau of Statistics (LBOS). The poststratified interview weights were computed as follows.

Let N_{ga}^{2020} denote the 2020 Lesotho 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 $r = 1, 2, \ldots, 168$ 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 $r = 1, 2, \ldots, 168$.

Weighted counts of the interview respondents before and after poststratification (viz., the population control totals) are summarized in Table 3-10.

.

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

	Male			Female			Total		
Age group	Population control total[1]	Wtd. count before post-stratification[2]	Poststrat-ification ratio[3]	Population control total[1]	Wtd. count before post-stratification[2]	Poststrat-ification ratio[3]	Population control total[1]	Wtd. count before post-stratification[2]	Poststrat-ification ratio[3]
15-19	101,567	63,595	1.597	98,021	79,414	1.234	199,588	143,009	1.396
20-24	96,596	63,351	1.525	97,515	83,139	1.173	194,111	146,490	1.325
25-29	101,102	54,580	1.852	97,521	72,852	1.339	198,623	127,432	1.559
30-34	98,439	51,977	1.894	90,599	69,107	1.311	189,038	121,085	1.561
35-39	77,985	44,612	1.748	68,294	55,069	1.240	146,279	99,682	1.467
40-44	52,463	38,656	1.357	49,256	46,497	1.059	101,719	85,153	1.195
45-49	37,325	29,042	1.285	38,362	37,100	1.034	75,687	66,142	1.144
50-54	33,043	22,531	1.467	40,364	32,246	1.252	73,407	54,777	1.340
55-59	26,813	21,357	1.255	37,138	28,971	1.282	63,951	50,328	1.271
60-64	22,509	17,701	1.272	31,690	29,097	1.089	54,199	46,798	1.158
65+	46,157	42,487	1.086	83,076	82,486	1.007	129,233	124,973	1.034
Total 15+	693,999	449,890	1.543	731,836	615,978	1.188	1,425,835	1,065,868	1.338

[1] Source: Lesotho Bureau of Statistics (LBOS)

[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 replicate *r* = 1, 2, …, 168), respectively, where *k* denotes the interview respondent, *h* denotes the district, *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 sex, age group and blood test response status, and the corresponding weighted counts using the person-level interview weights, $W_{hijk}^{(int)}$.

Table 3-11 Distribution of sample persons completing the blood test by age group and response status

Age group[1]	Sex	Blood test response status[2]	Unweighted sample size	Weighted count[3]
15 to 49 years	Male	Eligible respondent	4,661	316,259
		Eligible nonrespondent	420	29,554
	Female	Eligible respondent	6,466	409,301
		Eligible nonrespondent	498	33,878
50 years or older	Male	Eligible respondent	1,539	98,018
		Eligible nonrespondent	87	6,059
	Female	Eligible respondent	2,683	165,294
		Eligible nonrespondent	114	7,506
15 years or older	Male	Eligible respondent	6,200	414,277
		Eligible nonrespondent	507	35,613
	Female	Eligible respondent	9,149	574,594
		Eligible nonrespondent	612	41,384

[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, 88 potential predictor variables were available for initial selection. For females, 105 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 41 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 16 variables for males and 15 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	ALCFREQ	(901) Alcohol And Drug Use: How Often Do You Have A Drink Containing Alcohol?
	2	AT_BESTAGE_C	CATEGORICAL AGE BASED ON INTERVIEW AGE (CONFAGEY)
	3	AT_FIRSTSXAGE	AGE OF FIRST SEXUAL ACTIVITY - 1=LESS THAN 20; 2=20 OR MORE
	4	AVOIDPREG	(330) Reproduction: Are You Or Your Partner Currently Doing Something Or Using Any Method To Delay Or Avoid Getting Pregnant?
	5	DEATHS	(40) Hh Deaths: Has Any Usual Resident Of Your Household Died Since January 1, 2017?
	6	H_AGETEENYEARS	TEEN INDICATOR: 1 – 15-17 YEARS OLD; 2 – OTHERWISE; BASED ON AGEYEARS (ROSTER)
	7	KNOWN_HIV_STATUS_R	CATEGORICAL KNOWN HIV STATUS
	8	MATFLOOR	(51) Hh Characteristics: Main Material Of Floor
	9	PREPWDTK	(618) Hiv Testing: Would You Take Prep To Help Prevent Hiv?
	10	SCHCOM	(103) Background: What Is The Highest Level You Have Completed?
	11	STRATA	District Code (Sampling Stratum code)
	12	TOILETTYPE	(47) Hh Characteristics: What Kind Of Toilet Facility Do Members Of Your Household Usually Use?
	13	WORK12MO	(112) Background: Have You Done Any Work In The Last 12 Months For Which You Received Cash Or Goods As Payment?
	14	WORK7DAYS	(113) Background: Have You Done Any Work In The Last Seven Days For Which You Received Cash Or Goods As Payment?
	15	WORKIND	(114) Background: What Is Your Occupation? That Is, What Kind Of Work Do You Mainly Do?
	16	WORRY	(815) Tb And Other Health Issues: Over The Past Two Weeks, How Often Have You Not Been Able To Stop Or Control Worrying?
Female	17	ALCNUMDAY	(902) Alcohol And Drug Use: How Many Drinks Containing Alcohol Do You Have On A Typical Day?
	18	AT_FIRSTSXAGE	AGE OF FIRST SEXUAL ACTIVITY - 1=LESS THAN 20; 2=20 OR MORE
	19	AT_LIFETIMESEX	IN TOTAL, WITH HOW MANY DIFFERENT PEOPLE HAVE YOU HAD SEX IN YOUR LIFETIME? - 1=LESS THAN 10; 2=10 OR MORE
	20	AT_LIVEB	HOW MANY TIMES HAVE YOU HAD A PREGNANCY THAT RESULTED IN A LIVE BIRTH?
	21	COOKINGFUEL	(50) Hh Characteristics: What Type Of Fuel Does Your Household Mainly Use For Cooking?
	22	H_AGETEENYEARS	TEEN INDICATOR: 1 – 15-17 YEARS OLD; 2 – OTHERWISE; BASED ON AGEYEARS (ROSTER)
	23	NORMWORK	(115) Background: Where Do You Normally Work? In Your Home Community, Elsewhere In Region/Country, Or Outside The Country?

Sex	Variable number	Variable name	Description
	24	OUTREGIONTYPE	(105) Background: Just Before You Moved Here, Did You Live In A City, In A Town, Or In A Rural Area?
	25	OUTREGIONWHR	(106) Background: Before You Moved Here, Which Province Did You Live In? If You Lived Outside Of Zimbabwe, Which Country Did You Live In?
	26	PREPWDTK	(618) Hiv Testing: Would You Take Prep To Help Prevent Hiv?
	27	SCHCOM	(103) Background: What Is The Highest Level You Have Completed?
	28	STRATA	District Code (Sampling Stratum code)
	29	TBCLINHIVTST	(802) Tb And Other Health Issues: When You Visited A Tb Clinic In The Last 12 Months, Were You Tested For Hiv?
	30	TOILETTYPE	(47) Hh Characteristics: What Kind Of Toilet Facility Do Members Of Your Household Usually Use?
	31	URBAN_RURAL	1=Urban, 2=Peri-urban, 3=Rural

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

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

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} = the number of interview respondents in cell *m* who provided a useable blood sample, and n_m^{NBT} = the number of interview respondents in cell *m* who did not provide a useable blood sample.

The corresponding replicate-specific weighted response rates were similarly computed for jackknife replicate *r* = 1, 2, ..., 168 as

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

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, ..., 168.

The full-sample nonresponse-adjusted blood test weight for 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, ..., 168 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	88	105
Number of variables selected by LASSO	33	41
Number of variables selected by CHAID	16	15
Number of final nonresponse-adjustment cells	29	31
Number of interview respondents	6,200	9,149
Minimum adjustment factor	1.00	1.00
Maximum adjustment	1.42	1.38
Weighted count of respondents before adjustment[1]	414,277	574,594
Weighted count of respondents after adjustment[2]	449,890	615,978

[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 Lesotho 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, …, 168 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:

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

for *r* = 1, 2, …, 168.

Weighted counts of the blood test respondents before and after poststratification (viz., the population control totals) are summarized in Table 3-14.

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

	Male			Female			Total		
Age group	Population control total[1]	Wtd. count before post-stratification[2]	Poststrat-ification ratio[3]	Population control total[1]	Wtd. count before post-stratification[2]	Poststrat-ification ratio[3]	Population control total[1]	Wtd. count before post-stratification[2]	Poststrat-ification ratio[3]
15-19	101,567	64,447	1.576	98,021	79,540	1.232	199,588	143,987	1.386
20-24	96,596	64,050	1.508	97,515	84,919	1.148	194,111	148,970	1.303
25-29	101,102	55,174	1.832	97,521	72,980	1.336	198,623	128,154	1.550
30-34	98,439	50,421	1.952	90,599	68,582	1.321	189,038	119,003	1.589
35-39	77,985	44,855	1.739	68,294	53,686	1.272	146,279	98,541	1.484
40-44	52,463	38,032	1.379	49,256	45,727	1.077	101,719	83,759	1.214
45-49	37,325	28,839	1.294	38,362	36,104	1.063	75,687	64,943	1.165
50-54	33,043	22,074	1.497	40,364	32,421	1.245	73,407	54,495	1.347
55-59	26,813	21,732	1.234	37,138	29,211	1.271	63,951	50,943	1.255
60-64	22,509	17,062	1.319	31,690	29,130	1.088	54,199	46,192	1.173
65+	46,157	43,203	1.068	83,076	83,679	0.993	129,233	126,882	1.019
Total 15+	693,999	449,890	1.543	731,836	615,978	1.188	1,425,835	1,065,868	1.338

[1] Source: Lesotho Bureau of Statistics (LBOS)

[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

Ministry of Health, Lesotho, Centers for Disease Control and Prevention (CDC), and ICAP at Columbia University. Lesotho Population-based HIV Impact Assessment (LePHIA) 2016-2017: Final Report. Maseru, Lesotho, Atlanta, Georgia, and New York, New York, USA: Ministry of Health, CDC, and ICAP, September 2019.

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 vacants 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 38,545 dwelling unit/household records in the listing file, 20 were classified as ineligible for sampling based on the structure type, vacancy status, and residential status. Thus, a total of 38,525 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)	Vac/Constr. Status (STVAC)	Resid. Status (RESYN_D)	ELIG_FLAG	Records in master file	Eligible records
Cases with no GPS information			N	0	0
1 = Single House / compound of houses	1 = Not Vacant and not under constr.	1 = Yes	Y	26,827	26,827
1 = Single House / compound of houses	1 = Not Vacant and not under constr.	2 = No	Y	878	878
1 = Single House / compound of houses	2 = Vacant	1 = Yes	Y	50	50
1 = Single House / compound of houses	2 = Vacant	2 = No	Y	976	976
1 = Single House / compound of houses	3 = Under Construction	1 = Yes	Y	74	74
1 = Single House / compound of houses	3 = Under Construction	2 = No	Y	408	408
2 = Flat/Block/Apartment building	1 = Not Vacant and not under constr.	Missing	Y	4	4
2 = Flat/Block/Apartment building	1 = Not Vacant and not under constr.	1 = Yes	Y	8,123	8,123
2 = Flat/Block/Apartment building	1 = Not Vacant and not under constr.	2 = No	Y	829	829
2 = Flat/Block/Apartment building	1 = Vacant	1 = Yes	Y	4	4
2 = Flat/Block/Apartment building	2 = Vacant	2 = No	Y	49	49
2 = Flat/Block/Apartment building	3 = Under Construction	Missing	Y	15	15
2 = Flat/Block/Apartment building	3 = Under Construction	1 = Yes	Y	30	30
2 = Flat/Block/Apartment building	3 = Under Construction	2 = No	Y	54	54
3 = Church/Mosque/Temple	1 = Not Vacant and not under constr.	1 = Yes	Y	17	17
3 = Church/Mosque/Temple	1 = Not Vacant and not under constr.	2 = No	N	1	0
3 = Church/Mosque/Temple	2 = Vacant	1 = Yes	Y	0	0
3 = Church/Mosque/Temple	2 = Vacant	2 = No	N	1	0
3 = Church/Mosque/Temple	3 = Under Construction	1 = Yes	Y	0	0
3 = Church/Mosque/Temple	3 = Under Construction	2 = No	N	0	0
4 = Shop/office/bus. cntr/commercial bldg.	1 = Not Vacant and not under constr.	1 = Yes	Y	133	133
4 = Shop/office/bus. cntr/commercial bldg.	1 = Not Vacant and not under constr.	2 = No	N	7	0
4 = Shop/office/bus. cntr/commercial bldg.	2 = Vacant	1 = Yes	Y	2	2
4 = Shop/office/bus. cntr/commercial bldg.	2 = Vacant	2 = No	N	6	0
4 = Shop/office/bus. cntr/commercial bldg.	3 = Under Construction	1 = Yes	Y	0	0
4 = Shop/office/bus. cntr/commercial bldg.	3 = Under Construction	2 = No	N	0	0
5 = School/University	1 = Not Vacant and not under constr.	1 = Yes	Y	31	31
5 = School/University	1 = Not Vacant and not under constr.	2 = No	Y	1	1
5 = School/University	2 = Vacant	1 = Yes	Y	0	0
5 = School/University	2 = Vacant	2 = No	Y	0	0
5 = School/University	3 = Under Construction	1 = Yes	Y	0	0
5 = School/University	3 = Under Construction	2 = No	N	0	0
6 = Clinic/hospital/Doctors office	1 = Not Vacant and not under constr.	1 = Yes	Y	18	18
6 = Clinic/hospital/Doctors office	1 = Not Vacant and not under constr.	2 = No	N	4	0

Structure type (STOBS)	Vac/Constr. Status (STVAC)	Resid. Status (RESYN_D)	ELIG_FLAG	Records in master file	Eligible records
6 = Clinic/hospital/Doctors office	2 = Vacant	1 = Yes	Y	0	0
6 = Clinic/hospital/Doctors office	2 = Vacant	2 = No	N	0	0
6 = Clinic/hospital/Doctors office	3 = Under Construction	1 = Yes	Y	0	0
6 = Clinic/hospital/Doctors office	3 = Under Construction	2 = No	N	0	0
7 = Community Center/CBO	1 = Not Vacant and not under constr.	1 = Yes	Y	0	0
7 = Community Center/CBO	1 = Not Vacant and not under constr.	2 = No	N	0	0
7 = Community Center/CBO	2 = Vacant	1 = Yes	Y	0	0
7 = Community Center/CBO	2 = Vacant	2 = No	N	0	0
7 = Community Center/CBO	3 = Under Construction	1 = Yes	Y	0	0
7 = Community Center/CBO	3 = Under Construction	2 = No	N	0	0
96 = Other	1 = Not Vacant and not under constr.	1 = Yes	Y	2	2
96 = Other	1 = Not Vacant and not under constr.	2 = No	N	0	0
96 = Other	2 = Vacant	1 = Yes	Y	0	0
96 = Other	2 = Vacant	2 = No	N	0	0
96 = Other	3 = Under Construction	1 = Yes	Y	0	0
96 = Other	3 = Under Construction	2 = No	N	1	0
TOTAL				38,545	38,525

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

Structure type (STOBS_D)
1 - Single House / compound of houses
2 - Flat/Block/Apartment building
3 - Church/Mosque/Temple
4 - Shop/office/business cntr/commercial bldg.
5 - School/University
6 - Clinic/hospital/Doctors office
7 - Community Center/CBO
96 – Other
Structure vacant or under construction? (STVAC_D)
1 – Not Vacant and not under construction
2 – Vacant
3 – Under construction
Anyone living in the structure? (RESYN_D)
1 – Yes
2 – No

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 Survey Status for Household: HH_STATUS

B.1.1 Summary

HH_STATUS is defined for all sampled DUs. First, the variable UPCODE_RESULTNDT is derived using RESULTNDTOTHR. 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
	OTHER	**UPCODE_RESLTNDT** **Additional codes**	
96	Bereavement related	7	2
	No capable Head of Household available to do survey	8	2
	Dwelling inaccessible	9	2
	Recorded in another HH or tablet (discrepant record)	99	4

UPCODE_STAT_HH	Value label	Conditions
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
Not available at three occasions	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.		
Refusing Behavior	2	2
COULD NOT ACCOMODATE SURVEY DUE TO RELIGIOUS AFFILIATION.THEY ARE FROM THE JOHANNE MARANGE CHURCH		
DATA CANNOT BE COLLECTED DUE TO STRONG RELIGOUS 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		
Death/Funeral	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		
Participant/Household Head unable to do survey (incapacitated, language barrier, under age)	8	2
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		
Dwelling inaccessible	9	2
DWELLING CANT BE REACHED ROADS SLIPPERY DUE TO RAINS AND BAD TERRAIN		
HOUSEHOLD INACCESSIBLE BECAUSE OF A FLOODED STREAM FOR TWO DAYS		

RESULTNDOTH	UPCODE_ RESLTNDT	UPCODE_ STAT_HH
Vacant or not a dwelling	3	3
STRUCTURE UNDER CONSTRUCTION STILL AT FOUNDATION LEVEL		
NO ONE SLEEPS AT THE HOUSE		
HOUSEHOLD HEAD DECEASED. DWELLING VACANT		
VACANT		
DWELLING IS A BOTTLESTORE		
Household absent for extended period of time	6	3
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[1]	Rostered case from household with no questionnaire data
9	DeJure ineligible (SLEEPHERE = 2, LIVEHERE = 1 and roster age >=15)

[1] This code is defined for PHIA but the situation did not occur in LePHIA 2020.

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;
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 final weighting cells were created by successively partitioning the sample into heterogeneous subsets with respect to response propensity. 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).

Individual Interview

LES_AD_INDIV.xlsx (Persons 15+ years)

Blood Test

LES_AM_BT.xlsx (Males 15+ years)

LES_AF_BT.xlsx (Females 15+ years)