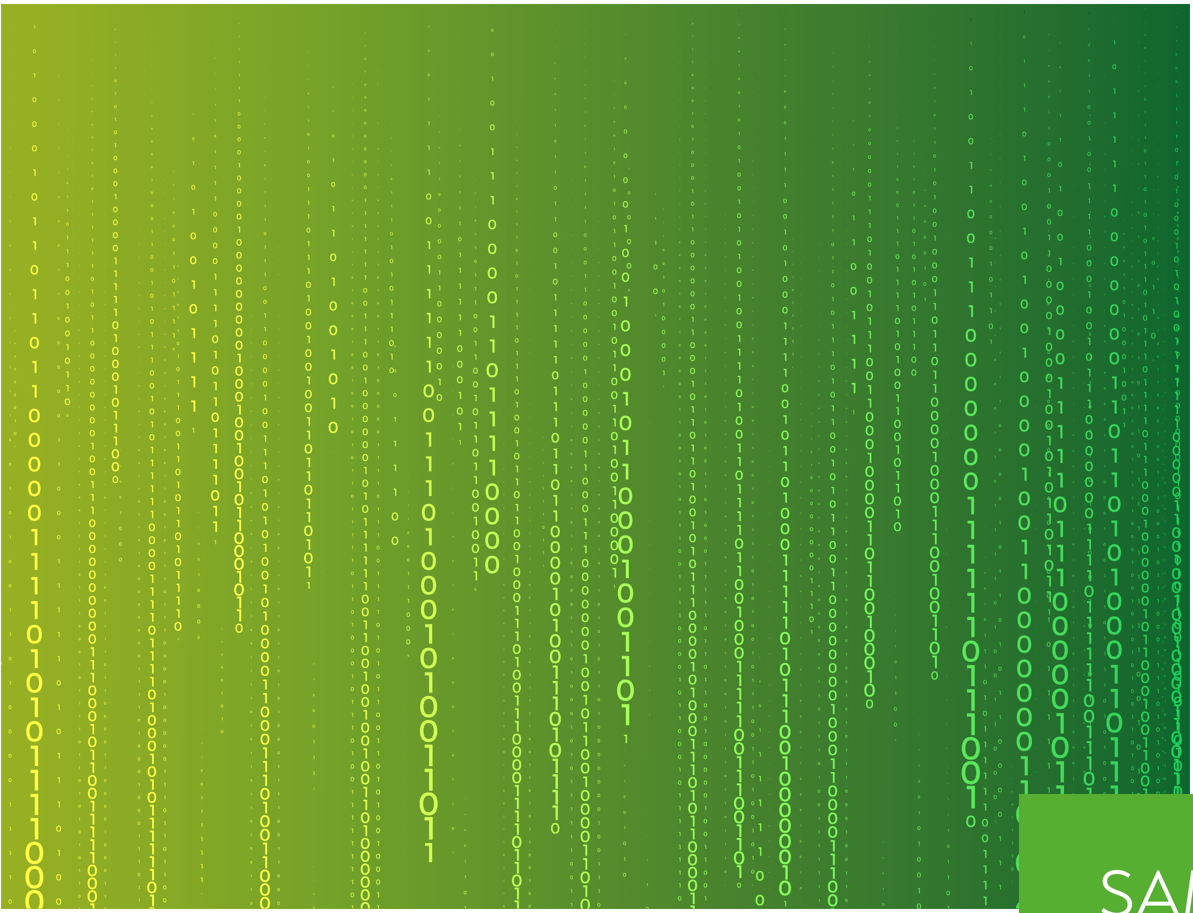
INSIDA

Mozambique Population-based HIV Impact Assessment



SAMPLING AND WEIGHTING TECHNICAL REPORT

INSTITUTO NACIONAL DE SAÚDE MOÇAMBIQUE

MINISTÉRIO DA SAÚDE

Conselho Nacional de Combate ao HIV/SIDA CNCS

INSTITUTO NACIONAL DE ESTATÍSTICA

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

CDC Centers for Disease Control and Prevention

Westat

icap Global Health

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

Mozambique Population-based HIV Impact Assessment 2021

INSIDA 2021

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.

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

Table of Contents

Contents Continued

Section **Page**

Acronyms

CDC	Centers for Disease Control and Prevention
CHAID	Chi-square Automatic Interaction Detector
CA	Control Area
CI	Confidence Interval
CV	Coefficient of Variation
DEFF	Design Effect
DU	Dwelling Unit
EA	Enumeration Area
HH	Household
HIV	Human Immunodeficiency Virus
ICC	Intracluster Correlation
LASSO	Least Absolute Shrinkage and Selection Operator
INE	Instituto Nacional de Estatística
INSIDA	National Survey on Prevalence, Behavioral Risk Factors and Information about HIV and AIDS
IMASIDA	Mozambique Immunization, Malaria and HIV/AIDS Indicator Survey
MDRI	Mean Duration of Recent Infection
MOS	Measure of Size
PHIA	Population-based HIV Impact Assessment
PSU	Primary Sampling Unit
SSU	Secondary Sampling Unit
RSE	Relative Standard Error
SAS	Statistical Analysis System
UEW	Unequal Weighting
VLS	Viral Load Suppression
WLM	Weighted Log-linear Modeling

1. Introduction

The 2021 Mozambique Population-based HIV Impact Assessment (INSIDA 2021) 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. Household listing for INSIDA 2021 occurred late January through early March of 2020. Data collection for INSIDA 2021 was conducted between April 2021 and February 2022 with a temporary pause in data collection from July through mid-September 2021 for training. The survey included approximately 17,100 interviewed individuals and over 14,400 individuals with valid blood tests in approximately 8,700 randomly-selected households from approximately 9,000 randomly-selected responding dwelling units. 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 INSIDA 2021 is a stratified four-stage probability sample design, with strata defined to be eleven provinces within the country, first-stage sampling units defined by control areas (CAs) within strata, the second stage sampling units defined by enumeration areas (EAs) within control areas, third-stage sampling units defined by dwelling units within EAs, fourth-stage sampling units defined by households within dwelling units, and finally age-eligible persons within households. Within each sampling stratum, the first-stage sampling units or CAs, also referred to as "primary sampling units" (PSUs) were selected from a nationally-representative master sample maintained by the Instituto Nacional de Estatística (INE) with probabilities proportionate to size. The measure of size (MOS) is the number of households in the CA derived from the 2017 Mozambique Population and Housing Census. In the second-stage sampling, one "secondary sampling unit" (SSU), also referred to as EA, was randomly selected with equal probability from each of the CAs selected for INSIDA. The allocation of the sample PSUs and SSUs to the eleven strata was made in a manner designed to achieve specified precision levels for (a) national estimate of HIV incidence among persons 15 to 49 years old; and (b) provincial estimates of viral load suppression (VLS) rates among HIV-positive persons 15 to 49 years old.

A list of households was compiled by trained staff for each of the sampled SSUs. This list was converted to a dwelling unit list by combining all households in the same structure into a dwelling unit record. The third-stage sample was selected from the dwelling unit list as a systematic random sample of dwelling units within stratum. The sample was selected with probabilities proportionate to the number of households within a dwelling unit.

For the fourth stage of sampling, one household was selected at random from each responding dwelling unit. Within the responding 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 INSIDA 2021.[1]

Details of the sample design employed for INSIDA 2021 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 SSU base weights to reflect the overall SSU probabilities of selection, and to compensate for SSU nonresponse;

[1] Usual members aged 15 and older who were not present in the household on the night before the household interview were eligible for data collection but were not part of the survey (not eligible for weighting and analysis).

- Calculation of dwelling unit weights to reflect the probabilities of selecting dwelling units within SSUs, and to compensate for dwelling unit nonresponse;

- Calculation of household weights to reflect the probabilities of selecting households within dwelling units, and to compensate for household nonresponse;

- Calculation of person-level interview weights to compensate for nonresponse to the interview;

- Post-stratification 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 compensate for nonresponse to the blood test; and

- Post-stratification of the person-level blood test weights to adjust for potential undercoverage of the population.

Technical details of the weighting procedures employed for INSIDA 2021 are provided in Section 3.

2. Sample Design

2.1 Population of Inference

The population of inference for INSIDA 2021 is comprised of the *de facto* population of individuals 15 years of age and older. The *de facto* population is comprised of all individuals who were present in households (i.e., "slept in the household") on the night prior to the date of interview. 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 the weighting and analysis. Table 2-1 summarizes estimates (projections) of the 2021 Mozambique population by gender and age group.

Table 2-1 2021 population estimates for Mozambique by gender and age group

	Gender		
Age group	Male	Female	Total
15 to 49 years	6,593,576	7,393,475	13,987,051
50 years or older	1,293,979	1,544,303	2,838,282
Total	7,887,555	8,937,778	16,825,333

Source: Population Projections from INE website http://www.ine.gov.mz/

2.2 Precision Specifications and Assumptions

The following specifications and assumptions were used to develop the sample design for INSIDA 2021.

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 provincial-level estimates of VLS rate among HIV positive adults aged 15 to 49 years should be ±0.10 or less.

- 95% CI bounds around the national estimate of VLS rate among all HIV positive adults aged 15 to 49 years should be ±0.03 or less.

- 95% 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

- National HIV prevalence rate of 0.128 (12.8%) for adults 15-49 years old that varies by province (see Table 2-2), (Sources: Instituto Nacional de Estatística (INE), 2018[2], and Instituto Nacional de Estatística (INE) (1996-2022);
- National HIV prevalence rate of 9.8% for women aged 15 to 24 years old that varies by province (see Table 2-2), (Source: Instituto Nacional de Estatistica (INE), 2018, with adjustments for sex ratio based on population projections obtained from and U.S. Census Bureau's International Database to account for lower male response.);
- Annual national incidence rate for adults aged 15-49 of p_a = 0.0046 (0.46%), (Source: Instituto Nacional de Estatística (INE), 2018. .);
- Stratum-level (provincial) incidence rates of p_{ah}, h = 1, 2, …, 11, which are obtained by adjusting the national incidence rate using the provincial prevalence rates as follows:

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

  where p_h and p are the HIV prevalence rates for province h and the country, respectively, and p_a is the annual national incidence rate obtained from the 2015 IMASIDA;
- Mean duration of recent infection (MDRI) of 130 days, yielding an annualization rate of 365/130= 2.8077;
- Estimated incidence rate for MDRI = 130 days of p_m = 0.0046/2.8077 = 0.00164 (0.164%), and the corresponding stratum-level (provincial) estimates obtained by p_{mh} = p_{ah}/2.8077;
- Viral load suppression rate among HIV positive adults aged 15-49 of p_{VLS} = 0.50 (50%) in each province, which yields a conservative estimate of the underlying population variance associated with VLS rate;
- Intracluster correlation (ICC) of 0.069 for VLS and 0.039 for prevalence (Source: tabulations of 2015 IMASIDA data);
- ICC of 0.000 for incidence (Source: analyses of prior PHIA surveys);
- Overall sex-age distributions based on 2015 IMASIDA; and

[2] The name of this survey from INE is IMASIDA (2015).

- Population distributions by province based on published 2017 Mozambique census projections.

2.2.3 Operational Assumptions

- Varying numbers of dwelling units to be sampled per SSU, resulting in an average of 35 sampled dwelling units per SSU;[3]
- Occupancy rate of 99.6% for sampled dwelling units (Source: Instituto Nacional de Estatística (INE), 2018);
- Household response rate of 97.6% among occupied households (Source: Instituto Nacional de Estatística (INE), 2018);
- Average household size of 4.41 (*de facto)* persons per household (Source: Instituto Nacional de Estatística (INE), 2018);
- Overall percentage of *de facto* persons 15-49 years of age per household of 38.6%; and an overall percentage of *de facto* persons 50+ years of age of 10.4% (Source: Instituto Nacional de Estatística (INE), 2018);
- Within the responding households, a person-level interview response rate of 90.7% (Source: Instituto Nacional de Estatística (INE), 2018); and
- Among persons completing the interview, a blood test response rate of 87.8% (Source: Instituto Nacional de Estatística (INE), 2018). Thus, among the persons selected for INSIDA 2021, the assumed overall response rate for the blood tests is 79.6% (90.7% * 87.8%).

Based on the specifications and assumptions listed above, a sample of 324 SSUs within sampled PSUs was determined to be the minimum needed to meet the specified precision goals. The allocation of the PSU and SSU sample to the eleven strata of Mozambique is shown in Table 2-2. Twelve (11 in Cabo Delgado province, 1 in Sofala province) out of 324 sampled EAs were excluded for security reasons. The target numbers of dwelling units to be sampled, the expected numbers of households to be 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

[3] In Cabo Delgado, the average number of DUs to be sampled was increased to 58 to account for the 11 excluded SSUs.

achieved during data collection. Further details about the sampling of dwelling units and households are given in Section 2.4.

Table 2-2 Allocation of Secondary Sampling Units (SSUs), dwelling units, households and projected number of respondents by stratum

Stratum code	Stratum name	HIV prevalence rate[1]		Total Number of PSUs/SSUs	Number of SSUs not listed for security or other reasons	Number of eligible SSUs for 3rd stage sampling	Target Number of DUs to be sampled	Number of participating HHs[3]	Projected Number of respondents providing valid blood draw[4]	
		Adults 15-49	Females 15-24[2]						Adults 15-49	Adults 50+
1	Niassa	0.074	0.052	35		35	1,225	1,191	1,613	434
2	Cabo Delgado	0.135	0.119	29	11	18	1,050	1,021	1,383	372
3	Nampula	0.058	0.034	43	1*	42	1,505	1,463	1,982	533
4	Zambézia	0.148	0.143	30		30	1,050	1,021	1,383	372
5	Tete	0.05	0.019	48		48	1,680	1,633	2,212	595
6	Manica	0.131	0.099	24		24	840	817	1,106	297
7	Sofala	0.16	0.116	22	1	21	770	749	1,014	273
8	Inhambane	0.135	0.115	24		24	840	817	1,106	297
9	Gaza	0.238	0.159	19		19	665	646	876	235
10	Maputo Provincia	0.221	0.157	29		29	1,015	987	1,337	359
11	Maputo Cidade	0.163	0.111	21		21	735	714	968	260
All	Mozambique	0.128	0.098	324	13	311	11,375	11,058	14,980	4,026

DU = dwelling unit; HH= household

[*] SSU was empty. No eligible households found in SSU.

[1] Source: Instituto Nacional de Estatística (INE), 2018 .

[2] 2015 IMASIDA (Instituto Nacional de Estatística (INE), 2018) prevalence rates adjusted by use of a sex-ratio based on population projections obtained from U.S. Census Bureau's International Database, to account for lower male response.

[3] Within each responding dwelling unit, one household was selected at random for the survey. Assumes occupancy rate of 99.6% and household response rate of 97.6%.

[4] 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) and Secondary Sampling Units (SSUs)

2.3.1 Definition and Selection of PSUs

In INSIDA 2021, the first-stage sampling units were INE control areas (CAs). The term primary sampling unit (PSU) is the more general statistical term. The first-stage INSIDA 2021 sample was selected by INE from a nationally-representative master sample of CAs, maintained and updated by INE. The INSIDA 2021 master sample has over 1,800 CAs, with each CA consisting of about 3 to 4 EAs and their number of households in the 2017 census. Following the sample allocation given in Table 2-3, a stratified sample of 324 CAs was then selected from the master sample.

2.3.2 Selection of the SSU Sample

For the second stage of sampling, an equal probability random sample of one EA was selected within sampled CAs. A more general statistical term is secondary sampling unit (SSU). Since CAs were selected with probabilities proportionate to an MOS equal to the number of households in 2017, the EAs were therefore selected with probabilities proportionate to an MOS equal to the estimated number of households in the EA in 2017.

2.3.3 Out-of-Scope SSUs

Out-of-scope SSUs are defined to be those EAs with no households (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" SSUs. There was one out-of-scope SSU in the INSIDA 2021 sample. Following standard statistical practice, this SSU was not replaced in the SSU sample.

2.3.4 Non-responding SSUs and Substitution

A sampled SSU that contains eligible households is considered nonresponding 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), is subject to military conflict or other dangerous conditions, or if permission to visit sampled areas is not received when such approval is needed. Of the 323 in-scope sampled EAs, 12 were excluded from INSIDA 2021 due to security concerns. Following recommendations by INE, none of the 12 EAs excluded for security reasons were replaced by another EA.

2.3.5 Summary of the PSU and SSU Samples

As indicated in the previous sections, 324 SSUs (EAs) were selected from 324 sampled PSUs (CAs) for INSIDA 2021. There was one out-of-scope (ineligible) SSU. Additionally, there were 12 SSUs considered as nonresponding since listing staff could not enter the EA due to security concerns. Table 2-3 summarizes the distribution of the sampled PSUs and SSUs by province and response status of the SSUs.

Table 2-3 Distribution of sampled PSUs and SSUs by stratum and SSU response status

Stratum code	Stratum name	Sampled PSU/SSUs	Nonresponding SSUs excluded from 3rd stage DU selection	Ineligible SSUs excluded from 3rd stage DU selection	Number of in-scope SSUs included in 3rd stage DU selection
1	Niassa	35	0	0	35
2	Cabo Delgado	29	11	0	18
3	Nampula	43	0	1	42
4	Zambézia	30	0	0	30
5	Tete	48	0	0	48
6	Manica	24	0	0	24
7	Sofala	22	1	0	21
8	Inhambane	24	0	0	24
9	Gaza	19	0	0	19
10	Maputo Provincia	29	0	0	29
11	Maputo Cidade	21	0	0	21
All	Mozambique	324	12	1	311

DU = dwelling unit

2.4 Selection of Dwelling Units and Households

The selection of dwelling units for INSIDA 2021 involved the following steps: (1) listing all potentially eligible dwelling units and households within the sampled EAs, (2) assigning eligibility codes to the listed dwelling units and household records based on characteristics of the listed units, (3) creating a dwelling unit sampling frame with a dwelling unit identified as a collection of eligible households sharing the same housing structure, and (4) selecting the sample of dwelling units with probabilities proportionate to the number of eligible households in each dwelling unit.

The selection of households for INSIDA 2021 involved the following steps: (1) re-listing the households within the responding selected dwelling units and (2) randomly selecting one household from each responding dwelling unit.

2.4.1 Definition of Third- and Fourth-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 a sampled EA. For the purpose of INSIDA, the third-stage sampling units are dwelling units. All dwelling units, including those appearing to be vacant, are included on the dwelling unit sampling frame.

During data collection, households within responding dwelling units were listed prior to fourth-stage sample selection. In the fourth stage, one household was randomly selected from the list of households within each responding dwelling unit.

2.4.2 Dwelling Unit 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." Listers were to list households where they were able, and dwelling units if the dwelling unit was vacant or the household list could not be obtained. Since the list includes both dwelling units and households, we refer to it as the dwelling unit/household list.

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 (where available), the type of structure (house, apartment, compound, etc.), occupancy status, and GPS coordinates. Vacant structures were listed along with occupied households. Slightly over 33,700 eligible records were listed for INSIDA 2021.

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 third-stage sampling. Table 2-4 summarizes the total number of records entered into the tablets, the numbers of unoccupied dwelling units, households eligible for sampling, and the total number of records eligible for sampling.

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

Stratum code	Stratum name	Number of records (DUs/HHs) in listing file[1]	Number of unoccupied DUs[2]	Number of unoccupied DUs eligible for sampling[3]	Number of occupied HHs eligible for sampling[4]	Total number of DUs/HHs eligible for sampling
1	Niassa	3,498	248	248	3,250	3,498
2	Cabo Delgado	2,203	61	60	2,142	2,202
3	Nampula	4,803	119	119	4,684	4,803
4	Zambezia	3,087	90	90	2,997	3,087
5	Tete	4,853	44	44	4,809	4,853
6	Manica	2,649	26	26	2,623	2,649
7	Sofala	2,291	61	60	2,230	2,290
8	Inhambane	2,546	108	108	2,438	2,546
9	Gaza	1,949	29	29	1,920	1,949
10	Maputo Provincia	3,483	278	278	3,205	3,483
11	Maputo Cidade	2,366	94	94	2,272	2,366
All	Mozambique	33,728	1,158	1,156	32,570	33,726

DU = dwelling unit; HH= household

[1] The listing file includes all households listed individually within DUs, and vacant/under construction DUs. 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 and Households

The listing of households and dwelling units served as the basis for creating the dwelling unit sampling frame. Household level records on the listing were summarized to the dwelling unit level. A variable indicating the number of household level records included in one dwelling unit record was created, with a value of one for the records on the listing which were already at the dwelling unit level or with only one household listed for the dwelling unit. Dwelling units were sampled with probabilities-proportional-to-size where the measure of size was the number of households in a dwelling unit. This measure of size was used in order to achieve an equal probability sample of households at the fourth stage of sampling.

Table 2-5 Distribution of listed households, dwelling units, the minimum and the maximum dwelling unit measure of sizes by province

Stratum Name	Number of eligible DUs/HHs listed	Total Number distinct DUs on sampling frame	Minimum DU measure of size	Maximum DU measure of size
Niassa	3,498	3,470	1	3
Cabo Delgado	2,202	2,195	1	4
Nampula	4,803	4,802	1	2
Zambezia	3,087	3,064	1	3
Tete	4,853	4,809	1	5
Manica	2,649	2,596	1	7
Sofala	2,290	2,191	1	7
Inhambane	2,546	2,533	1	3
Gaza	1,949	1,937	1	4
Maputo Provincia	3,483	3,145	1	14
Maputo Cidade	2,366	1,990	1	7
Mozambique	**33,726**	**32,732**	**1**	**14**

The calculation of the required within-SSU sampling rates proceeded as follows. First, the target overall sampling rate for stratum *h* = 1, 2, ..., 11, was computed as:

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

where

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

m_h = number of sampled SSUs in stratum *h*;

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

P_{hi} = probability of selecting SSU *i* in stratum *h*.

The total *expected* number of listings to be selected across all eleven strata is $\sum_{h=1}^{11} T_h$ = 11,375 (see Table 2-2). To obtain an equal probability sample within stratum *h*, the required within-SSU sampling rate for SSU *i* in stratum *h* was then computed as:

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

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

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

To reduce the variation in workload across the sampled SSUs, the maximum number of dwelling units to be selected in any SSU was capped at 70 (except for Cabo Delgado province where the sample size was capped at 120) and the minimum number was set to 15. Inspection of the values of E(n_{hi}) indicated that the expected sample sizes for three SSUs would fall below 15. The difference between the number of dwelling units that would have been selected using the rates, f_{hi}^{within}, and the specified maximum and minimum number was then re-distributed to the other SSUs in the same stratum so as to maintain as closely as possible the desired total sample size for the stratum. The within-SSU sampling rates, f_{hi}^{within}, were therefore adjusted to account for the redistribution of the sample within the stratum. The adjusted within-SSU 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

$$\mathrm{L} \leq N_{hi}\, A_{hi}\, f_{hi}^{within} \leq \mathrm{U},$$

L = 15 = the minimum SSU sample size,

U = 70 = the maximum SSU sample size, and

$\sum_{i=1}^{m_h} \; A_{hi} f_{hi}^{within} = T_h.$

The adjusted expected sample size for SSU i in stratum *h* was computed as:

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

To achieve a geographical ordering of the listed dwelling units, the dwelling unit records in each SSU 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 SSU. Dwelling units within the EA were then selected systematically (with probability proportional to size, the number of households per dwelling unit) from the ordered list of dwelling units to achieve the adjusted expected sample size specified above.

Certainty dwelling units were identified as dwelling units that would be selected at least once because their measure of size (e.g., number of households) are larger than the sampling interval. Certainty dwelling units were selected and removed from the dwelling unit frame while simultaneously reducing the number of dwelling units to sample from the reduced sampling frame.

Table 2-6 shows DU distribution by certainty and non-certainty status in the sampling frame by stratum.

Table 2-6 Distribution of Certainty and Non-Certainty DUs in Sampling Frame by stratum

Stratum Name	Number of eligible DUs/HHs listed	Total Number distinct DUs on sampling frame	Certainty DUs	Non-Certainty DUs	Target sampled DUs	Number to sample from Non-Certainty DUs
Niassa	3,498	3,470	5	3,465	1,225	1,220
Cabo Delgado	2,202	2,195	1	2,194	1,050	1,049
Nampula	4,803	4,802	46	4,756	1,505	1,459
Zambezia	3,087	3,064	2	3,062	1,050	1,048
Tete	4,853	4,809	12	4,797	1,680	1,668
Manica	2,649	2,596	10	2,586	840	830
Sofala	2,290	2,191	99	2,092	770	671
Inhambane	2,546	2,533	1	2,532	840	839
Gaza	1,949	1,937	2	1,935	665	663
Maputo Provincia	3,483	3,145	45	3,100	1,015	970
Maputo Cidade	2,366	1,990	95	1,895	735	640
Mozambique	**33,726**	**32,732**	**318**	**32,414**	**11,375**	**11,057**

The fourth stage of sampling was the selection of households within each selected dwelling unit. This stage occurred during data collection. For each responding dwelling unit a new household listing was created at the time of interview. This process was referred to as the mini-listing. One household was randomly selected within each responding dwelling unit's mini-listing.

The sampling goal was an equal-probability sample of households. Achieving this goal depends on the comparison of housing unit counts at three time periods - the housing unit count used in sampling SSUs (i.e., the estimated number of households in the SSU based on the most recent census projections), the actual number of households/dwelling units found at the time of listing, and the number of housing units found during the mini-listing. Application of the within-SSU sampling rates based on the size measure used in SSU sampling can yield more than the desired number of dwelling units in SSUs that have experienced growth in population since the time of the latest census projections, and fewer than the desired number of dwelling units in SSUs that have declined in population. Since there was a gap in time from the time of dwelling unit listing and data collection, conducting the mini-listing within sampled dwelling units captured growth or decline in household counts within sampled dwelling units. If there was growth or decline in the housing unit count across the three time periods, the change will be reflected in the sampling rates and will affect the variation in the sampling weights. Any decrease or increase in the number of dwelling units in a given region between the point of dwelling unit listing and household mini-listing was not affected because dwelling units were not relisted.

2.4.5 Results of Third and Fourth-Stage Sampling

Table 2-7 summarizes the numbers of dwelling units selected for the study and the minimum and maximum SSU sample size by stratum. The last column shows the unequal weighting (UEW) design effects (DEFF) to be expected for the selected sample. The UEW DEFF provides a measure of the increase in the variance of a sample estimate resulting from the use of variable overall sampling rates within a stratum (e.g., see Kish, 1965, page 403). With an equal-probability sample within each stratum, the UEW DEFFs would equal 1.0. Variable sampling rates increase the DEFF and would arise, for example, from the capping of sample sizes to control workload across EAs. However, since the extent of the capping and redistribution of the sample described previously was moderate, the corresponding increase in the variation of the overall sampling rates was small, resulting in stratum-level (provincial) UEW DEFFs that range from 1.00 to 1.05 (Table 2-7).

Table 2-7 Number of sampled dwelling units and expected unequal weighting DEFF by stratum

Stratum code	Stratum name	Number of SSUs for 3rd stage sampling	Number of sampled DUs	Minimum number of DUs per SSU	Maximum number of DUs selected per SSU	Unequal weighting DEFF
1	Niassa	35	1,225	15	70	1.00
2	Cabo Delgado	18	1,050	25	101	1.00
3	Nampula	42	1,505	15	70	1.00
4	Zambezia	30	1,050	15	70	1.00
5	Tete	48	1,680	15	70	1.00
6	Manica	24	840	18	63	1.00
7	Sofala	21	770	7	70	1.05
8	Inhambane	24	840	15	50	1.00
9	Gaza	19	665	15	52	1.00
10	Maputo Provincia	29	1,015	15	70	1.00
11	Maputo Cidade	21	735	15	63	1.01
All	**Mozambique**	**311**	**11,375**	**7**	**101**	**1.06**[1]

DU = dwelling unit; HH= household

[1] Overall DEFF reflects total variation in weights within and across stratum (provinces).

Table 2-8 summarizes the distribution of the sampled dwelling units by final dwelling unit response status. Of the 11,375 sampled dwelling units, 1,087 (9.56%)[4] were determined during data collection to be vacant/unoccupied, 435 (3.82%) for which eligibility for the survey (i.e., occupancy status) could not be established, 838 (7.37%) were determined to be eligible for the study (i.e., contained eligible households) but did not complete the household mini-listing, and 9,015 (79.25%) completed the household mini-listing. Excluding the ineligible cases, the overall unweighted dwelling unit response rate was 88.0%.

Table 2-8 Distribution of dwelling unit sample by stratum and response status

Stratum code	Stratum name	Number of sampled DUs	Number of ineligible DUs[1]	Number of DUs with unknown eligibility[2]	Number of responding DUs[3]	Number of eligible non-responding DUs[4]	Unweighted response rate[5]
1	Niassa	1,225	116	129	868	112	0.792
2	Cabo Delgado	1,050	90	78	697	185	0.732
3	Nampula	1,505	148	39	1,261	57	0.932
4	Zambezia	1,050	182	49	711	108	0.828
5	Tete	1,680	231	82	1,316	51	0.916
6	Manica	840	72	11	693	64	0.903
7	Sofala	770	48	15	660	47	0.915
8	Inhambane	840	69	7	685	79	0.889
9	Gaza	665	62	16	550	37	0.914
10	Maputo Provincia	1,015	45	2	912	56	0.940
11	Maputo Cidade	735	24	7	662	42	0.931
All	Mozambique	11,375	1,087	435	9,015	838	0.880

DU = dwelling unit

[1] Vacant dwelling units, nonresidential units, and units located outside the sampled SSU, as determined during data collection.

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

[3] Dwelling units that completed the mini-listing of households.

[4] Dwelling units with eligible households, but did not complete the mini-listing of households.

[5] Computed as R/ [R + N + U*{(R + N)/(R + N + I)}], where R = number of responding dwelling units; N = number of eligible nonresponding dwelling units; I = number of ineligible dwelling units, and U = number of dwelling units with unknown eligibility.

Table 2-9 summarizes the distribution of the sampled households from the 9,015 responding dwelling units by final household response status. Of the 9,015 sampled households 13 (0.1%) were

[4] The proportion of unoccupied dwelling units of 9.56% was noticeably higher than was included in the assumptions (0.4%).

determined during data collection to be vacant/unoccupied, four (<0.00%) for which eligibility for the survey (i.e., occupancy status) could not be established, 308 (3.42%) were determined to be eligible for the study (i.e., contained eligible household members) but did not complete the household questionnaire, and 8,690 (96.4%) completed the household questionnaire. Excluding the 13 ineligible cases, the overall unweighted household response rate was 96.5%.

Table 2-9 Distribution of household sample by stratum and response status

Stratum code	Stratum name	Number of sampled HHs	Number of ineligible HHs[1]	Number of HHs with unknown eligibility[2]	Number of responding HHs[3]	Number of eligible non-responding HHs[4]	Unweighted response rate[5]
1	Niassa	868	5	1	747	115	0.866
2	Cabo Delgado	697	0	0	678	19	0.973
3	Nampula	1,261	0	1	1,250	10	0.991
4	Zambezia	711	4	0	661	46	0.935
5	Tete	1,316	0	0	1,299	17	0.987
6	Manica	693	0	0	681	12	0.983
7	Sofala	660	0	0	650	10	0.985
8	Inhambane	685	0	0	668	17	0.975
9	Gaza	550	0	0	541	9	0.984
10	Maputo Provincia	912	3	1	878	30	0.966
11	Maputo Cidade	662	1	1	637	23	0.964
All	**Mozambique**	**9,015**	**13**	**4**	**8,690**	**308**	**0.965**

HH= household

[1] Households not eligible for the survey, for example absent for a prolonged period of time as determined during data collection.

[2] Sampled dwelling units/households for which existence of eligible households was not ascertained during data collection.

[3] Households completing the household interview.

[4] Eligible households 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 households, and U = number of households with unknown eligibility.

2.5 Selection of Individuals

The selection of individuals for INSIDA 2021 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 retained for subsequent weighting and analysis; individuals who did not sleep in the household on the night prior to the time the household roster was compiled but who were usual residents of the household (*de jure*) were eligible for data collection but not for 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 2-10 summarizes the number of households completing the roster and the corresponding number of rostered individuals by stratum and resident status.

Table 2-10 Distribution of households completing questionnaires and corresponding numbers of rostered persons by resident status and stratum

Stratum code	Stratum 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[3]	Total rostered persons[4]
1	Niassa	747	103	3,127	76	70	3,376
2	Cabo Delgado	678	88	2,831	19	28	2,966
3	Nampula	1,250	104	4,728	132	87	5,051
4	Zambezia	661	99	2,432	59	99	2,689
5	Tete	1,299	227	5,143	40	61	5,471
6	Manica	681	183	3,472	45	85	3,785
7	Sofala	650	88	2,902	28	53	3,071
8	Inhambane	668	205	2,641	74	106	3,026
9	Gaza	541	235	2,402	31	168	2,836
10	Maputo Provincia	878	215	3,323	53	83	3,674
11	Maputo Cidade	637	158	2,607	62	53	2,880
All	**Mozambique**	**8,690**	**1,705**	**35,608**	**619**	**893**	**38,825**

[1] Counts include persons of all ages.

[2] Not eligible for INSIDA 2021 weighting and analysis, but eligible for data collection

[3] Not eligible for INSIDA 2021 weighting and analysis or data collection

[4] Three roster entries from households that did not complete the household interview are not included in this table.

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 INSIDA 2021. Excluded are usual residents and any rostered nonresidents who were not present in the household on the night prior to the interview. Usual residents who were not present on the night prior to the household interview were eligible for data collection but not considered part of the weighting and analysis sample. Table 2-11 summarizes the number of individuals eligible for INSIDA 2021 sample by stratum, age group, and resident status.

Table 2-11 Number of individuals ages 15 or older who were eligible for INSIDA 2021 sampling

Stratum code	Stratum name	Persons 15-49 years[1]			Persons 50 years or older[1]		
		Usual resident/ slept here	Nonresident/ slept here	Total sampled persons[2]	Usual resident/ slept here	Nonresident/ slept here	Total sampled persons[2]
1	Niassa	1,360	57	1,417	234	13	247
2	Cabo Delgado	1,205	8	1,213	220	3	223
3	Nampula	1,962	93	2,055	445	22	467
4	Zambezia	1,036	38	1,074	241	9	250
5	Tete	2,120	22	2,142	481	7	488
6	Manica	1,474	32	1,506	330	4	334
7	Sofala	1,323	19	1,342	361	3	364
8	Inhambane	1,086	39	1,125	452	10	462
9	Gaza	986	16	1,002	341	5	346
10	Maputo Provincia	1,740	31	1,771	299	3	302
11	Maputo Cidade	1,432	32	1,464	395	9	404
All	**Mozambique**	**15,724**	**387**	**16,111**	**3,799**	**88**	**3,887**

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

[2] Eligible persons selected for the INSIDA 2021 sample based on information reported in roster

2.5.3 Distribution of Sampled Persons

Table 2-12 summarizes the number of individuals sampled for the INSIDA 2021 survey and the corresponding numbers completing the interview and blood test by age group and stratum. 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 2-12 Distribution of sampled persons by age group, response status, and stratum

Stratum code	Stratum name	Persons 15-49 years[1]			Persons 50 years or older[1]		
		Selected for data collection	Interview respondents[2]	Blood test respondent[3]	Selected for data collection	Interview respondents[2]	Blood test respondent[3]
1	Niassa	1,417	1,095	930	247	219	196
2	Cabo Delgado	1,213	973	730	223	201	161
3	Nampula	2,055	1,918	1,712	467	448	408
4	Zambezia	1,074	923	814	250	234	208
5	Tete	2,142	1,835	1,460	488	458	370
6	Manica	1,506	1,224	1,038	334	292	239
7	Sofala	1,342	1,159	1,064	364	335	310
8	Inhambane	1,125	929	752	462	420	357
9	Gaza	1,002	813	670	346	319	280
10	Maputo Provincia	1,771	1,507	1,281	302	267	242
11	Maputo Cidade	1,464	1,197	987	404	339	279
All	**Mozambique**	**16,111**	**13,573**	**11,438**[4]	**3,887**	**3,532**	**3,050**[4]

[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).

[4] Actual counts of respondents were lower than projected, as shown in Table 2-2.

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., dwelling unit, 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 control area, enumeration area, dwelling unit, 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 INSIDA 2021, nonresponse can occur at different stages of data collection, for example, (1) after the selection of the sample of SSUs at the second stage, (2) after the enumeration of DUs in each EA at the third stage, (3) after the enumeration of households in each DU at the fourth stage, (4) after household enumeration and selection of persons but before completion of the individual interview, and (5) 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 3.4.4.3 and 3.4.5.3 are used to calibrate the weighted sample counts to available population projections.

3.1 Overview of the Weighting Process

The overall weighting approach for INSIDA 2021 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 SSUs, dwelling units 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 SSU 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 Weights: The weighting process begins with the calculation and checking of the sample PSU base weights as the reciprocals of the overall PSU probabilities of selection.

Calculation of SSU Weights: The next step is the calculation and checking of the sample SSU base weights. The SSU base weights are the product of the PSU base weights and the reciprocal of the within-PSU SSU selection probabilities. The SSU base weights are adjusted to account for nonresponding eligible SSUs. This adjustment is made within the stratum in which the SSUs are located. The resulting weight is the final SSU weight.

Calculation of Dwelling Unit Weights: The next step is to calculate dwelling unit weights. The dwelling unit base weights are calculated as the product of the nonresponse adjusted SSU weights and the reciprocal of the within-SSU dwelling unit selection probabilities. The dwelling unit base weights are adjusted first to account for nonresponding dwelling units for which it could not be determined whether the dwelling unit is in-scope (unknown eligibility) (see Table 2-8). These adjusted dwelling unit weights are then adjusted for nonresponse amongst the eligible dwelling units. Adjustments are made within the SSU (or sometimes group of SSUs if collapsing is needed) in which the dwelling unit is located. The resulting weight is the final dwelling unit weight.

Calculation of Household Weights: The next step is to calculate household weights. The household base weights are calculated as the product of the nonresponse adjusted dwelling unit weights and the adjusted number of households within the dwelling unit. The household base weights are adjusted first to account for households for which it could not be determined whether the household is in-scope (unknown eligibility) (see Table 2-9). These adjusted household weights are then adjusted for nonresponse amongst the eligible households. This adjustment is made within the SSU or groups of SSUs 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 become the individual base weights for individuals found from the household roster to be eligible for the survey. Similar to the household weights, the first phase of individual weight adjustment is for any individuals whose eligibility is unknown. Eligibility is unknown when age was not confirmed at the interview stage. These individual weights are then adjusted for nonresponse among the eligible individuals, with a final poststratification adjustment for the individual weights to compensate for undercoverage in the sampling process by adjusting the weighted frequencies to correspond to 2021 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 INSIDA 2021 survey data that are used to construct the sampling weights are contained in the following data files.

- **mz_CFF_hh_int_STAT_20220421**: A household (HH) file that contains the dwelling unit (DU) and household data collected from DU sampling and in the HH questionnaire.

- **mz_CFF_roster_STAT_20220421**: A file that contains the roster of household members collected in the HH questionnaire with a record for each rostered person.

- **mz_CFF_ind_int_STAT_20220421**: 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.

- **MZ2Biomarker20220426**: 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. The "roster ineligible" cases were included in the final weighting files with 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 steps include:

- Check identification variables, 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 DU and HH selection process.

- Check counts of sampled and responding DUs and HHs against what was expected, overall and by sampling stratum.

- Adjust for segmentation and/or 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) enumeration areas, (b) dwelling units, (c) sampled households, (d) sampled individuals, and (e) 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 estimates derived from INSIDA 2021: the jackknife replication and Taylor 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 INSIDA 2021 complex sample design. Taylor Series is another widely used method that uses linear approximations to calculate the variance of a sample 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 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 SSUs, they can be created immediately after the selection of SSUs.

As described in Section 2.3.2, the SSU sample was obtained by randomly selecting one EA from each sampled PSU. The sampled SSUs were ordered geographically within sampling stratum and paired off in systematic order, treating each pair as a variance-estimation stratum. When there was an odd number of sampled SSUs in a sampling stratum, one of the variance-estimation strata was defined to contain three sampled SSUs. The formation of the variance-estimation strata was applied to all 324 of the sampled SSUs.

For INSIDA 2021, 159 variance-estimation strata were created. A jackknife replicate was then formed by randomly deleting an SSU from a particular variance-estimation stratum *k*, say, and retaining all of the SSUs in the remaining variance-estimation strata. For a variance-estimation stratum consisting of a pair of SSUs, the weight of the retained SSU within the variance-estimation stratum *k* was doubled. For a variance-estimation stratum consisting of three SSUs, the weight of the retained SSUs within the variance-estimation stratum were multiplied by 1.5 (see Section 3.4.1). The process was repeated for all *k* = 1, 2, 3, …, 159 variance-estimation strata, resulting in a total of 159 jackknife replicates. Table 3-1 summarizes the number of jackknife replicates that were created for variance estimation.

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

Sampling stratum code	Sampling stratum name	Sampled SSUs[1]	Variance strata consisting of pairs	Variance strata consisting of triplets	Number of jackknife replicates
1	Niassa	35	16	1	17
2	Cabo Delgado	29	13	1	14
3	Nampula	43	20	1	21
4	Zambezia	30	15	0	15
5	Tete	48	24	0	24
6	Manica	24	12	0	12
7	Sofala	22	11	0	11
8	Inhambane	24	12	0	12
9	Gaza	19	8	1	9
10	Maputo Provincia	29	13	1	14
11	Maputo Cidade	21	9	1	10
All	**Mozambique**	**324**	**153**	**6**	**159**

[1] Includes nonresponding and ineligible SSUs if applicable.

3.3.2 Taylor Series

Even though jackknife replication is the recommended method for variance estimation, not all software packages have a replication option to produce variance estimates. Therefore, information for producing Taylor Series estimates of variance is included in the INSIDA 2021 delivery files.

The full-sample weight (see Section 3.4) is used as the weight to compute Taylor Series variance estimates. The variable **VarStrat** indicates the variance-estimation stratum and the variable **VarUnit** indicates the SSU 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 Series capability.

3.4 Development of Weights

3.4.1 PSU and SSU Weights

The initial weighting step was to calculate PSU base weights for the full sample.

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

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

where P_{ha}^{PSU} = probability of selecting PSU *a* from sampling stratum *h*.

The conditional probability of selecting SSU *i* within PSU *a* in sampling stratum *h* was computed from the formula:

$$P_{i|ha}^{SSU} = 1/n_{ha},$$

where n_{ha}= the count of SSUs within the sampled PSU *a* in sampling stratum *h*.

The first step of computing the SSU weight is calculating the SSU level base weights as the product of the PSU weight and the reciprocal of the within-PSU SSU selection probability. The SSU base weight for sampled SSU *i* in PSU *a* in sampling stratum *h* was computed as:

$$W_{hai}^{(1)} = W_{ha}^{(1)} \ / \ P_{i|ha}^{SSU}$$

where

$W_{ha}^{(1)}$ = the base weight for PSU *a* in PSU sampling stratum *h*

$P_{i|ha}^{SSU}$ = the conditional probability of selecting SSU *i* in PSU *a* in sampling stratum *h*.

Using the SSU weights defined above, the sampled SSUs weight up to the numbers shown in the fourth column of Table 3-2.

SSU nonresponse occurs when no survey data are collected from an otherwise in-scope (eligible) SSU.

To compensate for the dwelling units from the nonresponding SSUs, the weights of the responding SSUs were inflated by the inverse of the (weighted) response rate in the SSU weighting cell after eliminating the known ineligible ("out of scope") SSUs (i.e., response-status group 3). The weighting cells for the SSU nonresponse adjustments are groups of SSUs within administrative boundaries inside each sampling stratum.

Let *hg* denote the substratum *g* within stratum *h* with a nonresponding SSU:

m_{hg} is the number of sample SSUs in substratum g in stratum h, and

m_{hg}^{r} is the number of responding SSUs in substratum g in stratum h.

The nonresponse-adjusted full sample SSU weight for SSU *i* in PSU *a* in sampling stratum *h* was computed as

$$W_{hai}^{(1A)} = A_{hgai}^{(1)} W_{hai}^{(1)},$$

where

$$A_{hgai}^{(1)} = \sum_{i=1}^{m_{hg}} W_{hai}^{(1)} \ / \sum_{i=1}^{m_{hg}^{r}} W_{hai}^{(1)}$$

is the SSU weight adjustment factor for substratum *hg*. The adjustment factor is the reciprocal of the SSU weighted response rate within the substratum. The values of $A_{hgai}^{(1)}$, equal to 1.00 except for substrata with a nonresponding SSU, are shown in Table 3-2. The corresponding replicate-specific SSU nonresponse adjustment factor for cell *hg* were similarly computed for jackknife replicate *r* = 1, 2, ..., 159.

The adjusted SSU weights, $W_{hai}^{(1A)}$, are passed to the dwelling unit weighting process described in the next section.

As described in Section 3.3.1, 159 jackknife replicates were formed from the 324 sampled SSUs. For variance estimation, replicate-specific SSU weights, $W_{(r)bi}^{(1)}$, $r = 1, 2, ..., 159$ were created to provide the basis for calculating the required replicate weights in subsequent stages of the weighting process. Let *b* denote one of the variance-estimation strata within the sampling stratum created for jackknife replication (Section 3.3.1) and let *i* denote the SSU within variance-estimation stratum *b*. For a given jackknife replicate, $r = 1, 2, ..., 159$, the corresponding replicate-specific SSU base weight where the variance-estimation strata consist of pairs was computed as

$$
\begin{array}{lcll}
W_{(r)bi}^{(1)} & = & a\,W_{hai}^{(1)} & \text{if } b = r \text{ and SSU } i \text{ in variance-estimation stratum } b \text{ is included in replicate } r \\
 & = & 0 & \text{if } b = r \text{ and SSU } i \text{ in variance-estimation stratum } b \text{ is not included in replicate } r \\
 & = & W_{hai}^{(1)} & \text{if } b \neq r
\end{array}
$$

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

The adjustment for SSU nonresponse was applied to the replicate weights as well as the full sample weights.

Table 3-2 Number of SSUs and corresponding weighted counts by sampling stratum

Sampling stratum Code	Sampling stratum Name	Number of sample EAs (SSUs)	Weighted number of EAs (SSUs) [1]	Number of in-scope SSUs in study	SSU nonresponse adjustment factor	In-scope SSUs weighted by nonresponse adjusted weights[2]	Weighted measure of size (MOS) [3]
1	Niassa	35	3,857	35	1.00	3,857	353,827
2	Cabo Delgado	29	5,366	29	1.376; 1.735 [4]	5,366	516,081
3	Nampula	43	12,663	42	1.00	12,322	1,193,099
4	Zambezia	30	10,276	30	1.00	10,276	1,100,106
5	Tete	48	5,927	48	1.00	5,927	531,289
6	Manica	24	3,612	24	1.00	3,612	384,507
7	Sofala	22	5,281	22	1; 1.029 [4]	5,281	563,895
8	Inhambane	24	3,260	24	1.00	3,260	299,695
9	Gaza	19	2,826	19	1.00	2,826	244,973
10	Maputo Provincia	29	4,817	29	1.00	4,817	421,060
11	Maputo Cidade	21	2,379	21	1.00	2,379	228,962
All	**Mozambique**	**324**	**60,262**	**323**	**-**	**59,922**	**5,837,493**

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

[2] Weights are the adjusted SSU weights, $W_{hai}^{(1A)}$.

[3] The measure of size used to select the sample of SSUs; the SSU Measure of Size (MOS) equals the number of households in the SSU frame. Weights are the adjusted SSU weights, $W_{hai}^{(1A)}$. Only in-scope SSUs are included.

[4] In sampling stratum 2, Cabo Delgado, one substratum has nonresponse adjustment factor of 1.376 and another 1.735. In sampling stratum 7, Sofala, one substratum has nonresponse adjustment factor of 1.029.

3.4.2 Dwelling Unit Weights

3.4.2.1 Dwelling Unit Base Weights

The first step of computing the dwelling unit weight is calculating the dwelling unit base weights as the product of the SSU weight adjusted for nonresponse (described in Section 3.4.1) and the reciprocal of the within-SSU dwelling unit selection probability.

The conditional probability of selecting dwelling unit j within SSU i in sampling stratum h was computed from the formula:

$$P_{j|hi}^{DU} = m_{hi}\, n_{hij} / n_{hi},$$

where

m_{hi} = the targeted sample of DUs within SSU i,

n_{hij} = the count of listed HHs in DU j in SSU i in sampling stratum h, and

n_{hi} = the count of listed HHs within the sampled SSU i in sampling stratum h.

The dwelling unit base weight for sampled dwelling unit j in SSU i in sampling stratum h was computed as:

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

where

$W_{hi}^{(1A)}$ = the nonresponse-adjusted weight for SSU i in SSU sampling stratum h

$P^{DU}_{j|hi}$ = the conditional probability of selecting dwelling unit *j* in SSU *i*[5] in sampling stratum *h*.

The corresponding weights for jackknife replicate *r* = 1, 2, …, 159 were computed as:

$$W^{(2)}_{(r)bij} = W^{(1A)}_{(r)bi} \;/\; P^{DU}_{j|hi} \,,$$

where $W^{(1A)}_{(r)bi}$ is the SSU nonresponse-adjusted weight for SSU *i* in variance estimation stratum *b* 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. 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 by response status and sampling stratum 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 sampling stratum.

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

Response status group[1]	Description	Number of sampled dwelling units
1	Respondent (Dwelling Unit listed and selected household)	9,015
2	Nonrespondent (Eligible Dwelling Unit with no selected household)	838
3	Ineligible (Dwelling Unit with no households: vacant, destroyed, or other)	1,087
4	Unknown eligibility (not known if Dwelling Unit contains household)	435
All	**—**	**11,375**

[1] See Appendix B for dwelling unit status definitions.

[5] For simplicity, the subscript *a* for PSU has been removed from subsequent equations in the remainder of the document.

Table 3-4 Weighted counts of dwelling unit base weights by response status and sampling stratum

Sampling stratum code	Sampling stratum name	Response status[1]				Total groups 1-4
		Group 1: responding dwelling unit	Group 2: nonresponding dwelling unit	Group 3: ineligible dwelling unit	Group 4: unknown eligibility	
1	Niassa	267,110	35,015	35,814	37,945	375,884
2	Cabo Delgado	391,375	106,607	50,112	46,505	594,599
3	Nampula	1,078,5.90	47,345	128,110	33,797	1,287,842
4	Zambezia	687,812	101,241	173,801	45,047	1,007,902
5	Tete	449,958	17,550	79,336	28,595	575,439
6	Manica	304,391	28,844	32,240	4,943	370,418
7	Sofala	430,135	27,220	31,340	9,121	497,817
8	Inhambane	273,204	31,461	27,830	2,801	335,296
9	Gaza	228,822	15,499	26,295	6,551	277,168
10	Maputo Provincia	479,298	28,464	22,567	1,157	531,486
11	Maputo Cidade	203,235	11,516	7,397	934	223,082
All	**Mozambique**	**4,793,931**	**450,761**	**614,843**	**217,397**	**6,076,932**

Note: 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 SSU. Because such nonresponse could occur before establishing whether or not a sampled dwelling unit is eligible for the study, 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 SSUs, the dwelling unit nonresponse adjustments were made within weighting cells defined by the individual SSUs or group of SSUs. The procedures used to compute the nonresponse-adjusted dwelling unit weights are described below.

Phase 1 Adjustment

To account for those dwelling units in which eligibility status is unknown, in the first phase of adjustment, the weights of the dwelling units where eligibility status is known (response status groups 1, 2, and 3) were inflated by the inverse of the (weighted) rate of known eligibility status in the SSU weighting cell after eliminating the dwelling units with eligibility status unknown (i.e., response-status group 4). As indicated above, the weighting cells for the dwelling unit nonresponse adjustments are either the individual SSUs or a group of SSUs. Let n_{hf}^{DU} denote the number of sampled dwelling units in SSU weighting cell *f* in sampling stratum *h*. Note that n_{hf}^{DU} is the sum of the sample sizes in each of the four response status groups defined in Table 3-3, i.e.,

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

where

$n_{hf}^{(1)}$ = the number of responding dwelling units (i.e., dwelling units with a completed household listing) in SSU weighting cell *f* in sampling stratum *h*

$n_{hf}^{(2)}$ = the number of eligible nonresponding dwelling units (i.e., eligible dwelling units without a completed household listing) in SSU weighting cell *f* in sampling stratum *h*

$n_{hf}^{(3)}$ = the number of known ineligible dwelling units (i.e., dwelling units known to contain no households) in SSU weighting cell *f* in sampling stratum *h*

$n_{hf}^{(4)}$ = the number of sampled dwelling units for which eligibility is unknown in SSU weighting cell *f* in sampling stratum *h*.

The first-phase nonresponse adjustment factor for SSU weighting cell *f* in sampling stratum *h* was computed as the ratio:

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

where $W_{hij}^{(2)}$ is the base weight for dwelling unit *j* in SSU *i* in SSU weighting cell *f* in sampling stratum *h*, and where the sum in the numerator extends over the entire sample of dwelling units in

SSU weighting cell *f* in sampling stratum *h*, while the sum in the denominator extends over the first three response status groups of dwelling units.

The corresponding replicate-specific first-phase dwelling units nonresponse adjustment factor for cell *f* were similarly computed for jackknife replicate *r* = 1, 2, ..., 159.

For the sampled dwelling units in response-status groups 1, 2 or 3, the first-phase adjusted weight for dwelling unit *j* in SSU *i* in SSU weighting cell *f* in sampling stratum *h* was then computed as:

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

The corresponding replicate weights for replicate r = 1, 2, …, 159 were computed in similar fashion as:

$$W_{(r)bij}^{DU1} = A_{(r)hf}^{(DU1)} \, W_{(r)bij}^{(2)},$$

where

$$A_{(r)hf}^{(DU1)} = \sum\nolimits_{j=1}^{n_{(r)hf}^{DU}} W_{(r)bij}^{(2)} \Big/ \sum\nolimits_{j=1}^{n_{(r)hf}^{(1)}+n_{(r)hf}^{(2)}+n_{(r)hf}^{(3)}} W_{(r)bij}^{(2)} \,.$$

Note that for the dwelling units in response-status group 4 (dwelling units of unknown eligibility), $W_{hij}^{DU1} = W_{(r)bij}^{DU1} = 0$ for r = 1, 2, …, 159.

The effect of this adjustment is to distribute the total weight of the unknown-eligibility cases (i.e., the estimated 217,397 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. 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 adjusted for unknown eligibility

Sampling stratum code	Sampling stratum name	Response status				Total dwelling units: groups 1-2
		Group 1: responding dwelling unit	Group 2: nonresponding dwelling unit	Group 3: ineligible dwelling unit	Total status 1-3	
1	Niassa	298,801	38,778	38,305	375,884	337,580
2	Cabo Delgado	429,642	113,237	51,720	594,599	542,879
3	Nampula	1,110,179	47,959	129,703	1,287,842	1,158,139

4	Zambezia	718,186	105,321	184,394	1,007,902	823,507
5	Tete	470,413	19,166	85,860	575,439	489,579
6	Manica	308,737	29,223	32,458	370,418	337,960
7	Sofala	436,539	29,161	32,117	497,817	465,700
8	Inhambane	275,270	32,008	28,018	335,296	307,278
9	Gaza	234,484	15,676	27,008	277,168	250,160
10	Maputo Provincia	480,416	28,490	22,581	531,486	508,906
11	Maputo Cidade	204,084	11,542	7,456	223,082	215,626
All	**Mozambique**	**4,966,751**	**470,561**	**639,619**	**6,076,932**	**5,437,313**

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

Phase 2 Adjustment

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

$$A_{hf}^{(DU2)} = \sum_{j=1}^{n_{hf}^{(1)}+n_{hf}^{(2)}} W_{hij}^{DU1} \; / \; \sum_{j=1}^{n_{hf}^{(1)}} W_{hij}^{DU1}$$

where W_{hij}^{DU1} is the first-phase adjusted weight for dwelling unit *j* in SSU *i* in SSU weighting cell *f* in sampling stratum *h*, and where the sum in the numerator extends over the sample of responding and nonresponding dwelling units in SSU weighting cell *f* in sampling stratum *h*, while the sum in the denominator extends over the responding dwelling units.

The weighted dwelling unit response rate for cell *f* is $R_{hf}^{(DU2)} = 1/A_{hf}^{(DU2)}$.

The corresponding replicate-specific dwelling unit nonresponse adjustment factor for cell *f* were similarly computed for jackknife replicate *r* = 1, 2, ..., 159.

The nonresponse-adjusted weight for responding dwelling unit *j* in SSU *i* in SSU weighting cell *f* in sampling stratum *h* was then computed as:

$$W_{hij}^{(2A)} = A_{hf}^{(DU2)} \, W_{hij}^{DU1}.$$

The corresponding replicate weights for replicate *r* = 1, 2, …, 159 were computed in similar fashion as:

$$W^{(2A)}_{(r)bij} = A^{(DU2)}_{(r)hf}\, W^{DU1}_{(r)bij},$$

where

$$A^{(DU2)}_{(r)hf} = \sum_{j=1}^{n^{(1)}_{(r)hf}+n^{(2)}_{(r)hf}} W^{DU1}_{(r)bij} \;/\; \sum_{j=1}^{n^{(1)}_{(r)hf}} W^{DU1}_{(r)bij}\,.$$

Final dwelling unit weights were examined for outliers based on a threshold of 4.8 times the median of the dwelling unit nonresponse-adjusted full sample weights within each sampling stratum. None were identified.

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

3.4.3 Household Weights

3.4.3.1 Household Base Weights

The first step in computing the household weights is calculating the household level base weights, which is the product of the dwelling unit weight adjusted for nonresponse (described in Section 3.4.2) and the reciprocal of the within-dwelling unit probability of selection. The within-dwelling unit probability of selection is the reciprocal of the number of households listed within the dwelling unit. To reduce the effect of extreme weights, the number of households within the dwelling unit was capped at a maximum value of four for the purpose of computing the within-dwelling unit probability of selection.

The conditional probability of selecting household *l* in dwelling unit *j* within SSU *i* in sampling stratum *h* was computed from the formula:

$$P^{HH}_{l|hij} = 1/\ min(\tilde{n}_{hij},4),$$

where $\tilde{n}_{hij}$ = the count of mini-listed HHs in DU *j* in SSU *i* in sampling stratum *h*.

The household base weight for sampled household *l* in dwelling unit *j* in SSU *i* in sampling stratum *h* was computed as:

$$W_{hijl}^{(3)} = W_{hij}^{(2A)} / P_{l|hij}^{HH}$$

where

$W_{hij}^{(2A)}$ = the nonresponse-adjusted weight for dwelling unit *j* in SSU *i* in sampling stratum *h*

$P_{l|hij}^{HH}$ = the conditional probability of selecting household *l* in dwelling unit *j* in SSU *i* in sampling stratum *h*.

The corresponding weights for jackknife replicate *r* = 1, 2, …, 159 were computed as:

$$W_{(r)bijl}^{(3)} = W_{(r)bij}^{(2A)} / P_{l|hij}^{HH}$$

where $W_{(r)bij}^{(2A)}$ is the SSU nonresponse-adjusted weight for dwelling unit *j* in SSU *i* in variance estimation stratum *b* described in Section 3.4.1.

Next, the sampled households were assigned to one of the four response status groups specified in Table 3-6. The specific rules used to classify households into the response status groups are given in Appendix B. In Table 3-7, we show the weighted counts of households by response status and sampling stratum using the household base weights described above. The characteristics of the household base weights were checked by examining statistical summaries of the weights such as the mean weight, CV of the weights, sum of the weights, and the minimum and maximum values of the weights, both overall and by sampling stratum.

Table 3-6 Distribution of sampled households by response status

Response status group[1]	Description	Number of sampled households
1	Respondent (household with completed household interview)	8,690
2	Nonrespondent (household without a completed household interview)	308
3	Ineligible (household selected but not eligible for interview)	13
4	Unknown eligibility (selected household with unknown eligibility, e.g. not approached for security reasons)	4
All	**Selected households from responding dwelling units**	**9,015[2]**

[1] See Appendix B for definitions of household status.

[2] Sum equals the total number of responding DUs since only one household was selected per responding DU.

Table 3-7 Weighted counts of household base weights by response status and sampling stratum

Sampling stratum code	Sampling stratum name	Response status[1]				Total groups 1-4
		Group 1: responding household	Group 2: nonresponding household	Group 3: ineligible household	Group 4: unknown eligibility	
1	Niassa	294,553	47,659	1,717	357	344,286
2	Cabo Delgado	537,442	15,939	0	0	553,381
3	Nampula	1,169,950	8,553	0	930	1,179,432
4	Zambezia	800,623	52,893	4,110	0	857,626
5	Tete	534,145	7,034	0	0	541,179
6	Manica	356,771	11,749	0	0	368,520
7	Sofala	489,748	6,894	0	0	496,642
8	Inhambane	305,255	6,806	0	0	312,061
9	Gaza	254,792	4,180	0	0	258,973
10	Maputo Provincia	550,736	21,249	1,388	608	573,980
11	Maputo Cidade	255,059	9,808	564	389	265,819
All	**Mozambique**	**5,549,073**	**192,763**	**7,779**	**2,283**	**5,751,898**

Note: Counts given in table are weighted counts using the household base weights $W_{hijl}^{(3)}$ described in Section 3.4.3.1.

3.4.3.2 Adjustment for Household Nonresponse

The general approach for handling household nonresponse was to increase the weights of responding households so that they represent the nonresponding households in the same SSU. Because such nonresponse could occur before establishing whether or not a sampled household is eligible for the study, the nonresponse adjustment was implemented in two phases. In the first phase of adjustment, the base weights were adjusted to compensate for sampled households for which eligibility for the survey was not ascertained. In the second phase of adjustment, the first-phase adjusted weights were further adjusted to compensate for the nonresponding households among those households known to be eligible for the study.

To account for variation in response rates across different types of SSUs, the nonresponse adjustments were made within weighting cells defined by the individual SSUs or group of SSUs. The procedures used to compute the nonresponse-adjusted household weights are described below.

Phase 1 Adjustment

In the first phase of adjustment, to account for those households in which eligibility status is unknown the weights of the households where eligibility status is known (response status groups 1, 2, and 3) were inflated by the inverse of the (weighted) rate of known eligibility status in the SSU weighting cell after eliminating the households with eligibility status unknown (i.e., response-status group 4). As indicated above, the weighting cells for the household nonresponse adjustments are either the individual SSUs or a group of SSUs. Let n_{he}^{HH} denote the number of sampled households in SSU weighting cell *e* in sampling stratum *h*. Note that n_{he}^{HH} is the sum of the sample sizes in each of the four response status groups defined in Table 3-6, i.e.,

$$n_{he}^{HH} = n_{he}^{(HH,1)} + n_{he}^{(HH,2)} + n_{he}^{(HH,3)} + n_{he}^{(HH,4)}$$

where

$n_{he}^{(HH,1)} =$ the number of responding households (i.e., households with a completed household interview) in SSU weighting cell *e* in sampling stratum *h*

$n_{he}^{(HH,2)} =$ the number of eligible nonresponding households (i.e., eligible households without a completed household interview) in SSU weighting cell *e* in sampling stratum *h*

$n_{he}^{(HH,3)} =$ the number of known ineligible households in SSU weighting cell *e* in sampling stratum *h*

$n_{he}^{(HH,4)} =$ the number of sampled households for which it is not known whether the household is eligible in SSU weighting cell *e* in sampling stratum *h*.

The first-phase nonresponse adjustment factor for SSU weighting cell *e* in sampling stratum *h* was computed as the ratio:

$$A_{he}^{HH1} = \sum_{j=1}^{n_{he}^{HH}} W_{hijl}^{(3)} \bigg/ \sum_{j=1}^{n_{he}^{(HH,1)}+n_{he}^{(HH,2)}+n_{he}^{(HH,3)}} W_{hijl}^{(3)}$$

where $W_{hijl}^{(3)}$ is the base weight for household l in dwelling unit j in SSU i in sampling stratum h, and where the sum in the numerator extends over the entire sample of households in SSU weighting cell e in sampling stratum h, while the sum in the denominator extends over the first three response status groups of households.

The corresponding replicate-specific first-phase households nonresponse adjustment factor for cell e were similarly computed for jackknife replicate $r = 1, 2, ..., 159$.

For the sampled households in response-status groups 1, 2 or 3, the first-phase adjusted weight for household l in dwelling unit j in SSU i in SSU weighting cell e in sampling stratum h was then computed as:

$$W_{hijl}^{HH1} = A_{he}^{(HH1)} W_{hijl}^{(3)}$$

The corresponding replicate weights for replicate r = 1, 2, …, 159 were computed in similar fashion as:

$$W_{(r)bijl}^{HH1} = A_{(r)he}^{(HH1)} W_{(r)bijl}^{(3)},$$

where

$$A_{(r)he}^{HH1} = \sum_{j=1}^{n_{(r)he}^{HH}} W_{(r)bijl}^{(3)} \bigg/ \sum_{j=1}^{n_{(r)he}^{(HH,1)}+n_{(r)he}^{(HH,2)}+n_{(r)he}^{(HH,3)}} W_{(r)bijl}^{(3)}$$

Note that for the households in response-status group 4 (households of unknown eligibility), $W_{hijl}^{HH1} = W_{(r)bijl}^{HH1} = 0$ for r = 1, 2, …, 159.

The effect of this adjustment is to distribute the total weight of the unknown-eligibility cases (i.e., the estimated 2,283 households shown in the next-to-last column of Table 3-7) to the combined weight of the remaining three groups of sampled households. The resulting weighted counts using W_{hijl}^{HH1} as computed above are summarized in Table 3-8.

Table 3-8 Weighted counts of households adjusted for unknown eligibility

Sampling stratum code	Sampling stratum name	Response status				Total households: groups 1-2
		Group 1: responding household	Group 2: nonresponding household	Group 3: ineligible household	Total status 1-3	
1	Niassa	294,833	47,722	1,731	344,286	342,555
2	Cabo Delgado	537,442	15,939	0	553,381	553,381
3	Nampula	1,170,879	8,553	0	1,179,432	1,179,432
4	Zambezia	800,623	52,893	4,110	857,626	853,516
5	Tete	534,145	7,034	0	541,179	541,179
6	Manica	356,771	11,749	0	368,520	368,520
7	Sofala	489,748	6,894	0	496,642	496,642
8	Inhambane	305,255	6,806	0	312,061	312,061
9	Gaza	254,792	4,180	0	258,973	258,973
10	Maputo Provincia	551,328	21,264	1,388	573,980	572,592
11	Maputo Cidade	255,408	9,847	564	265,819	265,255
All	**Mozambique**	**5,551,225**	**192,881**	**7,793**	**5,751,898**	**5,744,106**

Note: Counts in table are weighted counts using first-phase adjusted household weights W_{hijl}^{HH1}.

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 SSU weighting cell after eliminating the known ineligible households (i.e., response-status group 3). The second-phase household nonresponse adjustment factor for SSU weighting cell *e* in sampling stratum *h* was computed as the ratio:

$$A_{he}^{HH2} = \sum_{j=1}^{n_{he}^{(HH,1)}+n_{he}^{(HH,2)}} W_{hijl}^{HH1} \Big/ \sum_{j=1}^{n_{he}^{(HH,1)}} W_{hijl}^{HH1}$$

where W_{hijl}^{HH1} is the first-phase adjusted weight for household *l* in dwelling unit *j* in SSU *i* in sampling stratum *h*, and where the sum in the numerator extends over the sample of responding and nonresponding households in SSU weighting cell *e* in sampling stratum *h*, while the sum in the denominator extends over the responding households.

The weighted household interview response rate for cell *e* is $R_{he}^{(HH2)} = 1/A_{he}^{(HH2)}$.

The corresponding replicate-specific interview nonresponse adjustment factor for cell *e* were similarly computed for jackknife replicate *r* = 1, 2, ..., 159

The final nonresponse-adjusted weight for responding household *l* in dwelling unit *j* in SSU *i* in SSU weighting cell *e* in sampling stratum *h* was then computed as:

$$W_{hijl}^{(3A)} = A_{he}^{(HH2)} W_{hijl}^{HH1}$$

The corresponding replicate weights for replicate *r* = 1, 2, …, 159 were computed in similar fashion as:

$$W_{(r)bijl}^{(3A)} = A_{(r)he}^{(HH2)} W_{(r)bijl}^{HH1}$$

where

$$A_{(r)he}^{HH2} = \sum_{j=1}^{n_{(r)he}^{(HH,1)}+n_{(r)he}^{(HH,2)}} W_{(r)bijl}^{HH1} \Big/ \sum_{j=1}^{n_{(r)he}^{(HH,1)}} W_{(r)bijl}^{HH1}$$

Final household weights were examined for outliers based on a threshold of 4.8 times the median of the household unit nonresponse-adjusted full sample weights within each sampling stratum. None were identified.

The sum of the final nonresponse-adjusted household weights, $W_{hijl}^{(3A)}$, summed across the responding households (response status group 1), is equal to the weighted count shown in the last column of Table 3-8.

3.4.4 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 INSIDA 2021 data files. First, we define the initial person-level (interview) base weights in Section 3.4.4.1. 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 and household nonresponse adjustments described previously, this person-level nonresponse adjustment was implemented in two phases.

3.4.4.1 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, $W_{hijl}^{(3A)}$. That is, the base weight for rostered person *k* in household *l* in dwelling unit *j* in SSU *i* in sampling stratum *h* was computed from the formula

$$W_{hijlk}^{(base)} = W_{hijl}^{(3A)} .$$

The corresponding replicate base weights,$W_{(r)bijlk}^{(base)}$, for *r* = 1, 2, …, 159 were computed in an analogous manner, with $W_{hijl}^{(3A)}$ replaced by $W_{(r)bijl}^{(3A)}$ in the above formula.

3.4.4.2 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 3-9 summarizes the distribution of the rostered persons in responding households 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 because actual age was not obtained. 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 3-9 Distribution of rostered persons in responding households 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+	19,912	12,721,957
2	De facto person 15 years or older	Under 15	79	49,948
3	De facto person 15 years or older	Unknown	7	3,539
4	Non de facto persons 15 years or older	NA	2,010	1,196,814
5	Persons under 15 years	NA	16,817	11,038,993
All	**—**	**—**	**38,825[3]**	**25,011,251[3]**

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

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

[3] Of the 38,825 rostered persons, 893 were those that neither slept in the household nor were usual residents (see Table 2-10). These 893 persons account for 577,658 of the total weighted count of 25,011,251 rostered persons.

Phase 1 Adjustment

The procedure for computing the first phase adjustment was as follows. For each of the sex-age weighting classes specified for the adjustment (see Table 3-10), the first-phase interview nonresponse adjustment factor for cell *c*, $A_c^{(1)}$, was computed as

$$A_c^{(1)} = (\sum_{i=1}^{n_c^{(1)}} W_{ck}^{(base)} + \sum_{i=1}^{n_c^{(2)}} W_{ck}^{(base)} + \sum_{i=1}^{n_c^{(3)}} W_{ck}^{(base)} + \sum_{i=1}^{n_c^{(4)}} W_{ck}^{(base)}) / (\sum_{k=1}^{n_c^{(1)}} W_{ck}^{(base)} + \sum_{i=1}^{n_c^{(2)}} W_{ck}^{(base)} + \sum_{i=1}^{n_c^{(3)}} W_{ck}^{(base)})$$

where *c* denotes the first-phase adjustment cell, $W_{ck}^{(base)}$ is the base weight for person *k* in cell *c*, and $n_c^{(a)}$= the number of cases in response-status group $a = 1, 2, 3, 4$ in weighting class *c*.

The corresponding replicate-specific first-phase interview nonresponse adjustment factors for cell *c* were similarly computed for jackknife replicate $r = 1, 2, ..., 159$.

The first-phase weighted interview response rate for cell *c* is $R_c^{(1)} = 1/A_c^{(1)}$ for the full sample, and $R_{(r)c}^{(1)} = 1/A_{(r)c}^{(1)}$ for jackknife replicate $r = 1, 2, ..., 159$.

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 $r = 1, 2, ..., 159$ were similarly computed as

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

Phase 2 Adjustment

Table 3-10 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 SSU characteristics such as sampling stratum 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 3-10 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	7,304	4,807,733
	Eligible nonrespondent	1,522	896,137
	All response statuses	**8,826**	**5,703,870**
Female 15 or older	Eligible respondent	9,801	6,228,180
	Eligible nonrespondent	1,285	793,398
	All response statuses	**11,086**	**7,021,578**
Total 15 years or older	Eligible respondent	17,105	11,035,913
	Eligible nonrespondent	2,807	1,689,535
	All response statuses	**19,912**	**12,725,448**

[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, $W_{ck}^{(3)}$

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

There are 51 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 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 zero (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 base weight adjusted for unknown eligibility, $W_{ck}^{(3)}$. 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 so that if the program had to be re-run, the same results would be produced. Of the 51 variables used in the initial model, the LASSO identified 32 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 weighting cells based on information available for all sampled 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 first-phase nonresponse-adjusted weight, $W_{ck}^{(3)}$. 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).

There are four different populations of inference at the individual level – males and females divided by age group (under and over 18 years of age). In the individual questionnaire, males and females in the specified age groups received different questions. 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. By forcing gender and age group as the first variables in the CHAID tree, the tree model essentially creates four distinct CHAID trees. After forcing these two variables into the model, the tree was then allowed to grow freely. The CHAID algorithm identified 19 variables to create the weighting classes for nonresponse adjustment. Table 3-11 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 3-11 Variables selected by CHAID to produce classes for interview nonresponse adjustment

Variable number	Variable name	Description
1	DADHHM	Natural Father Usually Live In This Household Or Was A Guest Last Night?
2	DEATHCOUNT	How Many Usual Household Residents Died Since January 1, 2019?
3	HHELITER	HH Head Eligibility: Literacy
4	H_AGETEENYEARS	Teen Indicator: 1 – 15-17 Years Old; 2 – Otherwise; Based On Ageyears (Roster)
5	H_AGEYEARS	Age (Categorical), Based On Roster Age. Matches Poststratification Cells
6	H_ECONSUP12_A	Economic support? Nothing
7	H_HHQITEMS	1-Electricity; 2-Working Radio; 3-Working Television; 4-Working Telephone/Mobile Telephone; 5-Working Refrigerator; 6-None Of The Above
8	H_HH_SIZE_C	1-9, Where 9 Includes All Hhs With 9 Or More People
9	H_OWNCHIKNNUM	Altogether, How Many Of The Below Listed Animals Do Members Of Your Household Own?
10	H_OWNCOWNUM	How Many Of The Below Listed Animals Do Members Of Your Household Own?
11	H_ROOMSLEEP	How Many Rooms Are Used For Sleeping?
12	LIVEHERE	Usually Live Here?
13	MATEXWALLS	HH Characteristics: Main Material Of Exterior Walls
14	RELATTOHH	What Is The Relationship Of Name To The Head Of The Household?
15	SEX	Male Or Female?
16	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?
17	STRATA	Sampling Stratum Code
18	TOILETTYPE	What Kind Of Toilet Facility Do Members Of Your Household Usually Use?
19	URBAN_RURAL	Urban = 1; Rural = 2

Calculation of Phase 2 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 interview nonresponse adjustment factor for cell *m*, $A_m^{(int)}$, was computed as

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

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 interview nonresponse adjustment factor for cell *m* were similarly computed for jackknife replicate *r* = 1, 2, ..., 159 as

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

The weighted interview response rate for cell *m* is $R_m^{(int)} = 1/A_m^{(int)}$ for the full sample, and $R_{(r)m}^{(int)} = 1/A_{(r)m}^{(int)}$ for jackknife replicate *r* = 1, 2, ..., 159.

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, ..., 159 were similarly computed as

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

Outlier checks for the interview nonresponse-adjusted weights were done using a threshold of 4.8 times the median of the interview nonresponse-adjusted full sample weights within each sampling stratum. There were a total of 12 outlier observations. The weights for the 12 outliers were trimmed and then re-distributed such that the total weights equal the nonresponse-adjusted interview weights in Table 3-12.

A summary of selected features of the nonresponse adjustment process is given in Table 3-12.

Table 3-12 Summary of the interview nonresponse adjustment process

Characteristic	Total sample
Number of variables in initial model	51
Number of variables selected by LASSO	32
Number of variables selected by CHAID	19
Number of final nonresponse-adjustment cells	67
Number of interview respondents	17,105
Minimum adjustment factor	1.00
Maximum adjustment	2.00
Weighted count of respondents before adjustment[1]	11,035,913
Weighted count of respondents after adjustment[2]	12,725,448

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

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

3.4.4.3 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 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 3-13. The population control totals given in this table are projected 2021 national population projections by gender and five-year age groups provided by the INE. The poststratified interview weights were computed as follows.

Let N_{ga}^{2021} denote the 2021 Mozambique population control total for gender *g* and (five-year) age group *a* as given in Table 3-13. The poststratification ratio adjustment factor for gender *g* and age group *a* was then computed as:

$$T_{ga}^{2021} = N_{ga}^{2021} / \sum_{k=1}^{n_{ga}^{resp}} W_{gak}^{(int)},$$

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:

$$T_{(r)ga}^{2021} = N_{ga}^{2021} / \sum_{k=1}^{n_{(r)ga}^{resp}} W_{(r)gak}^{(int)}$$

for the $r = 1, 2, \ldots, 159$ jackknife replicates.

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

$$W_{gak}^{(ps-int)} = T_{ga}^{2021} W_{gak}^{(int)},$$

and the corresponding poststratified replicate weights were computed as:

$$W_{(r)gak}^{(ps-int)} = T_{(r)ga}^{2021} W_{(r)gak}^{(int)}$$

for $r = 1, 2, \ldots, 159$.

Table 3-13 provides the population control totals, weighted counts of the respondents before poststratification, and the ratio of the control totals to the nonresponse adjusted weights (poststratification adjustment factor) by age and gender.

Table 3-13 2021 Mozambique population projections and weighted counts before poststratification

	Male			Female			Total		
Age group	Population control total[1]	Weighted count before post-stratification[2]	Poststrat-ification ratio[3]	Population control total[1]	Weighted count before post-stratification[2]	Poststrat-ification ratio[3]	Population control total[1]	Weighted count before post-stratification[2]	Poststrat-ification ratio[3]
15-19	1,668,015	975,806	1.709	1,671,982	1,070,985	1.561	3,339,997	2,046,791	1.632
20-24	1,318,946	903,229	1.460	1,502,812	1,173,556	1.281	2,821,758	2,076,785	1.359
25-29	1,038,112	754,459	1.376	1,284,370	1,029,389	1.248	2,322,482	1,783,847	1.302
30-34	828,163	568,406	1.457	942,711	783,792	1.203	1,770,874	1,352,198	1.310
35-39	682,042	562,018	1.214	782,345	642,739	1.217	1,464,387	1,204,756	1.216
40-44	588,712	444,337	1.325	687,033	533,728	1.287	1,275,745	978,065	1.304
45-49	469,586	419,317	1.120	522,222	454,838	1.148	991,808	874,155	1.135
50-54	356,018	241,367	1.475	413,053	382,415	1.080	769,071	623,783	1.233
55-59	271,696	225,536	1.205	316,203	283,494	1.115	587,899	509,030	1.155
60-64	207,958	204,199	1.018	235,244	238,221	0.988	443,202	442,419	1.002
65+	458,307	392,431	1.168	579,803	441,187	1.314	1,038,110	833,618	1.245
Total 15+	7,887,555	5,691,104	1.386	8,937,778	7,034,344	1.271	16,825,333	12,725,448	1.322

[1] Source: Population Projections from INE website http://www.ine.gov.mz/

[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, $W_{gak}^{(int)}$.

3.4.5 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 2021 population control totals.

3.4.5.1 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 3.4.4.2. These weights are given by $W_{hijlk}^{(int)}$ and $W_{(r)hijlk}^{(int)}$ (for replicate *r* = 1, 2, …, 159), respectively, where *k* denotes the interview respondent, *h* denotes the sampling stratum, *i* denotes the SSU, *j* denotes the dwelling unit, and *l* 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 3-14 summarizes the counts of individuals by sex, age group and blood test response status, and the corresponding weighted counts using the nonresponse person-level interview weights, $W_{hijlk}^{(int)}$.

Table 3-14 Distribution of sample persons completing the blood test by sex, 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	4,907	3,931,414
		Eligible nonrespondent	901	696,157
	Female	Eligible respondent	6,536	4,820,984
		Eligible nonrespondent	1,235	868,043
50 years or older	Male	Eligible respondent	1,326	957,765
		Eligible nonrespondent	158	105,768
	Female	Eligible respondent	1,719	1,142,510
		Eligible nonrespondent	323	202,807
15 years or older	Male	Eligible respondent	6,233	4,889,179
		Eligible nonrespondent	1,059	801,925
	Female	Eligible respondent	8,255	5,963,494
		Eligible nonrespondent	1,558	1,070,849

[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 person-level interview weight, $W_{hijlk}^{(int)}$.

3.4.5.2 Nonresponse Adjustment of Blood Test Weights

To compensate for blood test nonresponse, the nonresponse-adjusted person-level 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 SSU characteristics such as sampling stratum 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, 81 potential predictor variables were available for initial selection. For females, 84 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 36 significant variables for males and 43 significant variables for females. The selected variables were then input into the SI-CHAID program to create the final weighting cells for nonresponse adjustment.

The CHAID algorithm identified 16 variables for males and 23 variables for females that were then used to create weighting classes for nonresponse adjustment. Table 3-15 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 blood test base weights are given in Appendix C.2.

Table 3-15 Variables selected by CHAID to produce classes for blood test nonresponse adjustment

Sex	Variable number	Variable name	Description
Male	1	ALCNUMDAY	Alcohol And Drug Use: How Many Drinks Containing Alcohol Do You Have On A Typical Day?
	2	AT_FIRSTSXAGE	AGE OF FIRST SEXUAL ACTIVITY - TRUNCATED AT 20 AND COLLAPSE 0-11 TO 11
	3	AT_LIFETIMESEX	IN TOTAL, WITH HOW MANY DIFFERENT PEOPLE HAVE YOU HAD SEX IN YOUR LIFETIME?
	4	AT_LIGHTINGFUEL	WHAT TYPE OF FUEL DOES YOUR HOUSEHOLD MAINLY USE FOR LIGHTING?
	5	AT_MATEXWALLS	MAIN MATERIAL OF EXTERIOR WALLS
	6	AT_PART12MONUM	HOW MANY DIFFERENT PEOPLE HAVE YOU HAD SEX WITH IN THE LAST 12 MONTHS?
	7	AT_WATERSOURCE	WHAT IS THE MAIN SOURCE OF DRINKING WATER FOR MEMBERS OF YOUR HOUSEHOLD?
	8	EVERMAR	Marriage: Have You Ever Been Married Or Lived Together With A Man As If Married?
	9	HFLAST12MO	HIV Testing: Have You Seen A Doctor, Clinical Officer, Nurse Or A Lay Counselor In A Health Facility In The Last 12 Months?
	10	KNOWN_HIV_STATUS_R	CATEGORICAL KNOWN HIV STATUS
	11	MCSTATUS	Male Circumcision: Some Men Are Uncomfortable Talking About Circumcision, But It Is Important For Us To Have This Information. Some Men Are Circumcised. Are You Circumcised?
	12	NUMWIF	Marriage: Altogether, How Many Wives Or Live-In Partners Do You Have Who Live With You Here In This Household?
	13	PARTLASTCNDM1	Sexual Activity: The Last Time You Had Sex With Partner, Was A Condom Used?
	14	PARTLIVEW1	Sexual Activity: Is The Person That You Had Sex With A Spouse Or A Partner Who Lives In This Household?
	15	STRATA	Sampling Stratum Code
	16	URBAN_RURAL	Urban = 1; Rural = 2

Table 3-15 Variables selected by CHAID to produce classes for blood test nonresponse adjustment (continued)

Sex	Variable number	Variable name	Description
Female	17	ALCSIXMORE	Alcohol And Drug Use: How Often Do You Have Six Or More Drinks On One Occasion?
	18	AT_BESTAGE_C	CATEGORICAL AGE BASED ON INTERVIEW AGE (CONFAGEY)
	19	AT_LIFETIMESEX	IN TOTAL, WITH HOW MANY DIFFERENT PEOPLE HAVE YOU HAD SEX IN YOUR LIFETIME?
	20	AT_LIGHTINGFUEL	WHAT TYPE OF FUEL DOES YOUR HOUSEHOLD MAINLY USE FOR LIGHTING?
	21	AT_LIVEB	HOW MANY TIMES HAVE YOU HAD A PREGNANCY THAT RESULTED IN A LIVE BIRTH?
	22	AT_MATEXWALLS	MAIN MATERIAL OF EXTERIOR WALLS
	23	AT_TOILETSHARENUM	HOW MANY HOUSEHOLDS SHARE THIS TOILET FACILITY?
	24	AT_WATERSOURCE	WHAT IS THE MAIN SOURCE OF DRINKING WATER FOR MEMBERS OF YOUR HOUSEHOLD?
	25	CONDOMGET	Prevention Intervention: If You Wanted A Condom, Would It Be Easy For You To Get One?
	26	CURMAR	Marriage: What Is Your Marital Status Now: Are You Married, Living Together With Someone As If Married, Widowed, Divorced, Or Separated/Single?
	27	DEPRESSED	TB And Other Health Issues: Over The Past Two Weeks, How Often Have You Felt Down, Depressed Or Hopeless?
	28	HFLAST12MO	HIV Testing: Have You Seen A Doctor, Clinical Officer, Nurse Or A Lay Counselor In A Health Facility In The Last 12 Months?
	29	HHELITER	HH Head Eligibility: Literacy
	30	KNOWN_HIV_STATUS_R	CATEGORICAL KNOWN HIV STATUS
	31	LITTLEINTEREST	TB And Other Health Issues: Over The Past Two Weeks, How Often Have You Been Bothered By Having Little Interest In Doing Things?
	32	PARTLASTCNDM1	Sexual Activity: The Last Time You Had Sex With Partner, Was A Condom Used?
	33	PARTLASTETOH1	Sexual Activity: The Last Time You Had Sex With Partner, Did Either Of You Drink Alcohol Beforehand?
	34	SCHCOM	Background: What Is The Highest Level Of School You Attended?
	35	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?
	36	STRATA	Sampling Stratum Code
	37	TOILETTYPE	What Kind Of Toilet Facility Do Members Of Your Household Usually Use?
	38	URBAN_RURAL	Urban = 1; Rural = 2
	39	WORK12MO	Background: Have You Done Any Work In The Last 12 Months For Which You Received Cash Or Goods As Payment? This Includes Work On The Family Farm

Calculation of Nonresponse-Adjusted Blood Test Weights

The general approach for computing the nonresponse-adjusted blood test weights was as follows. Within each of the final adjustment cells specified in Appendix C.2 the blood-test nonresponse adjustment factor for cell *m*, $A_m^{(BT)}$, was computed as

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

where *m* denotes the adjustment cell, $W_{mk}^{(int)}$ is the final nonresponse-adjusted person-level 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 nonresponse adjustment factor for cell *m* were similarly computed for jackknife replicate *r* = 1, 2, ..., 159.

The weighted blood test response rate for cell *m* is $R_m^{(BT)} = 1/A_m^{(BT)}$ for the full sample, and $R_{(r)m}^{(BT)} = 1/A_{(r)m}^{(BT)}$ for jackknife replicate *r* = 1, 2, ..., 159.

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, ..., 159 were similarly computed as

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

Outlier checks for the blood test nonresponse-adjusted weights were done using a threshold of 4.8 times the median of the blood test nonresponse-adjusted full sample weights within each sampling stratum. There were a total of 9 outlier weights, with ratios of the weight to the threshold varying from 1.00 to 1.28. These 9 outliers were not trimmed.

A summary of selected features of the blood-test nonresponse adjustment process is given in Table 3-16.

Table 3-16 Summary of the blood test nonresponse adjustment process

Characteristic	Male	Female
Number of variables in initial model	81	84
Number of variables selected by LASSO	36	43
Number of variables selected by CHAID	16	23
Number of final nonresponse-adjustment cells	39	59
Number of biomarker respondents	6,233	8,255
Minimum adjustment factor	1.00	1.00
Maximum adjustment	1.65	1.83
Weighted count of respondents before adjustment[1]	4,889,179	5,963,494
Weighted count of respondents after adjustment[2]	5,691,104	7,034,344

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

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

3.4.5.3 Poststratification Adjustment

Like the nonresponse-adjusted interview weights described previously, the nonresponse-adjusted blood test weights were poststratified to projected 2021 Mozambique population counts within classes defined by gender and five-year age group.

Let N_{ga}^{2021} denote the 2021 Mozambique population control total for gender *g* and (five-year) age group *a* as given in Table 3-17. The poststratification ratio adjustment factor used to adjust the blood test weights for gender *g* and age group *a* was computed as:

$$T_{ga}^{2021} = N_{ga}^{2021} / \sum_{k=1}^{n_{ga}^{BT}} W_{gak}^{(BT)},$$

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:

$$T_{(r)ga}^{2021} = N_{ga}^{2021} / \sum_{k=1}^{n_{(r)ga}^{BT}} W_{(r)gak}^{(BT)}$$

for the *r* = 1, 2, …, 159 jackknife replicates.

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

$$W_{gak}^{(ps-BT)} = T_{ga}^{2021} \, W_{gak}^{(BT)},$$

and the corresponding poststratified replicate weights were computed as:

$$W_{(r)gak}^{(ps-BT)} = T_{(r)ga}^{2021} \; W_{(r)gak}^{(BT)}$$

for $r = 1, 2, \ldots, 159$.

Weighted counts of the blood test respondents before and after poststratification (namely, the population control totals) are summarized in Table 3-17.

Table 3-17 2021 Mozambique population projections and weighted counts of blood test respondents before and after poststratification

Age group	Male			Female			Total		
	Population control total[1]	Weighted count before post-stratification[2]	Poststrat-ification ratio[3]	Population control total[1]	Weighted count before post-stratification[2]	Poststrat-ification ratio[3]	Population control total[1]	Weighted count before post-stratification[2]	Poststrat-ification ratio[3]
15-19	1,668,015	979,563	1.703	1,671,982	1,083,408	1.543	3,339,997	2,062,972	1.619
20-24	1,318,946	906,567	1.455	1,502,812	1,178,585	1.275	2,821,758	2,085,153	1.353
25-29	1,038,112	740,603	1.402	1,284,370	1,032,788	1.244	2,322,482	1,773,390	1.310
30-34	828,163	550,769	1.504	942,711	767,994	1.227	1,770,874	1,318,764	1.343
35-39	682,042	557,391	1.224	782,345	625,005	1.252	1,464,387	1,182,395	1.238
40-44	588,712	439,048	1.341	687,033	543,405	1.264	1,275,745	982,453	1.299
45-49	469,586	412,571	1.138	522,222	458,803	1.138	991,808	871,374	1.138
50-54	356,018	249,381	1.428	413,053	389,217	1.061	769,071	638,598	1.204
55-59	271,696	234,678	1.158	316,203	278,878	1.134	587,899	513,557	1.145
60-64	207,958	210,002	0.990	235,244	243,311	0.967	443,202	453,313	0.978
65+	458,307	410,530	1.116	579,803	432,950	1.339	1,038,110	843,479	1.231
Total 15+	7,887,555	5,691,104	1.386	8,937,778	7,034,344	1.271	16,825,333	12,725,448	1.322

[1] Source: Population Projections from INE: Instituto Nacional de Estatística (1996-2022), http://www.ine.gov.mz/iv-rgph-2017/projeccoes-da-populacao-2017-2050

[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

Instituto Nacional de Estatística (INE) (2018). Inquérito de Indicadores de Imunização, Malária e HIV/SIDA em Moçambique (IMASIDA) 2015 – Relatório Final.

Instituto Nacional de Estatística (INE) (1996-2022). Projecções da População 2017 – 2050, http://www.ine.gov.mz/iv-rgph-2017/projeccoes-da-populacao-2017-2050.

U.S. Census Bureau International Database, https://www.census.gov/programs-surveys/international-programs/about/idb.html

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

Johnston, G. and Rodriguez, R (2015). Introducing the HPGENSELECT Procedure: Model Selection for Generalized Linear Models and More. Paper SAS1742-2015. https://support.sas.com/resources/papers/proceedings15/SAS1742-2015.pdf

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

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

Magidson, J. (2005). SI-CHAID Users Guide. Statistical Innovations. https://www.statisticalinnovations.com/wp-content/uploads/SICHAIDusersguide.pdf

-

Appendix A

Definition of Eligibility for Dwelling Unit/Household Sampling

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

The dwelling unit 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 33,728 dwelling unit/household records in the listing file, only 2 were classified as ineligible for sampling based on the structure type, vacancy status, and residential status. Thus, a total of 33,726 household records in the Master Listing File were eligible for sampling.

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

Structure type (STOBS_D)	Vac/Constr. Status (STVAC_D)	Resid. Status (RESYN_D)	ELIG_FLAG	Total in master file	Eligible Listed household records
Cases with no GPS information			N		
1 = Single House / compound of houses	1 = Not Vacant and not under const.	1 = Yes	Y	31,533	31,533
1 = Single House / compound of houses	1 = Not Vacant and not under const.	2 = No	Y	328	328
1 = Single House / compound of houses	2 = Vacant	1 = Yes	Y	110	110
1 = Single House / compound of houses	2 = Vacant	2 = No	Y	567	567
1 = Single House / compound of houses	3 = Under Construction	1 = Yes	Y	320	320
1 = Single House / compound of houses	3 = Under Construction	2 = No	Y	95	95
2 = Flat/Block/Apartment building	1 = Not Vacant and not under const.	1 = Yes	Y	625	625
2 = Flat/Block/Apartment building	1 = Not Vacant and not under const.	2 = No	Y	30	30
2 = Flat/Block/Apartment building	2 = Vacant	1 = Yes	Y	1	1
2 = Flat/Block/Apartment building	2 = Vacant	2 = No	Y	35	35
2 = Flat/Block/Apartment building	3 = Under Construction	1 = Yes	Y	4	4
2 = Flat/Block/Apartment building	3 = Under Construction	2 = No	Y	22	22
3 = Church/Mosque/Temple	1 = Not Vacant and not under const.	1 = Yes	Y	16	16
3 = Church/Mosque/Temple	2 = Vacant	2 = No	N	1	
3 = Church/Mosque/Temple	3 = Under Construction	1 = Yes	Y	1	1
4 = Shop/office/business cntr/commercial bldg.	1 = Not Vacant and not under const.	1 = Yes	Y	30	30
4 = Shop/office/business cntr/commercial bldg.	3 = Under Construction	1 = Yes	Y	1	1
5 = School/University	1 = Not Vacant and not under const.	1 = Yes	Y	6	6
6 = Clinic/hospital/Doctors office	1 = Not Vacant and not under const.	1 = Yes	Y	1	1
96 = Other	1 = Not Vacant and not under const.	1 = Yes	Y	1	1
96 = Other	2 = Vacant	2 = No	N	1	
				33,728	33,726

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 Dwelling Unit, Household, Interview, and Blood Test Response Status

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

The response status variables required for weighting as previously described in Section 3.4.2 (dwelling unit weights), Section 3.4.3 (household weights), Section 3.4.4 (interview weights), and Section 3.4.5 (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 Dwelling Unit: DU_STATUS

B.1.1 Summary

DU_STATUS is defined for all sampled dwelling units. First, the variable UPCODE_STARTDWELL is derived from STARTDWELLOT. Next, STARTDWELL and UPCODE_STARTDWELL are used to calculate UPCODE_STAT_DWELL. Lastly, DU_STATUS is set equal to UPCODE_STAT_DWELL when the Data Lock files are delivered.

DU_STATUS	Description
1	Responding Dwelling Unit (Household Selected)
2	Eligible Dwelling Unit, NonRespondent (no Household selected)
3	Ineligible Dwelling Unit
4	Dwelling Unit with Unknown eligibility Status

B.1.2 SAS code defining DU_STATUS

SAS code defining DU_STATUS:

```
   DU_STATUS = UPCODE_STAT_DU;
```

UPCODE_STAT_DWELL MOZ

Variables on DU/HH record – STARTDWELL and STARTDWELLOT (text field for STARTDWELL = 29). DM Calc variables – UPCODE_STARTDWELL and UPCODE_STAT_DWELL

STARTDWELL - CAN DATA BE COLLECTED FOR THE DWELLING?

CODE	LABEL
10	YES
21	NO - ALL HOUSEHOLDS IN DWELLING NOT AVAILABLE AT ALL VISIT ATTEMPTS
23	NO - DWELLING VACANT OR ADDRESS NOT A DWELLING
24	NO - DWELLING DESTROYED
25	NO - DWELLING NOT FOUND
29	NO - OTHER, SPECIFY

STARTDWELLOT is upcoded to **UPCODE_STARTDWELL** using same or similar specs to those used to upcode RESULTNDT in previous PHIA2 household cleaning. This table shows the UPCODE_STARTDWELL values and their mapping to UPCODE_STAT_DWELL

UPCODE_STARTDWELL	UPCODE_STAT_DWELL
21-ALL HOUSEHOLDS IN DWELLING NOT AVAILABLE AT ALL VISIT ATTEMPTS	2
22-REFUSED	2
23-DWELLING VACANT OR ADDRESS NOT A DWELLING	3
24-DWELLING DESTROYED	3
25-DWELLING NOT FOUND	4
26-DWELLING INACCESSIBLE	2
27-BEREAVEMENT RELATED	2
28-NO CAPABLE HH AVAILABLE TO DO SURVEY	2
89-TEAM IN IQ	4
90-COMMUNITY REFUSAL	4
91-OUT OF SCOPE (DU not in EA map boundaries)	3
92-COVID QUARANTINE - UNKNOWN ELIGIBILITY	4
93-UNKNOWN ELIGIBILITY	4
94-COVID DELAY – UNKNOWN ELIGIBILITY	4
95-CANNOT TRACE	4
97 – Inaccessible for Security Reasons	4
99-RECORDED IN ANOTHER HH OR TABLET (DISCREPANT RECORD)	3

UPCODE_STAT_DWELL definition. (See programmer spec below, also table above. All three should match).

1 = Did mini-listing and chose a household

STARTDWELL = 10 AND HHNUM not NULL

2 = Found, appears to have households, but no HH chosen, breakoff

STARTDWELLOT indicates refusal

3 = Out of scope listing (vacant, destroyed)

STARTDWELL =23, 24; or STARTDWELLOT indicates out of scope (91)

4 = No household chosen, cannot determine whether DU is in-scope

STARTDWELL = 29 and STARTDWELLOT can't be assigned to 2 or 3.

Or some other situation where DU eligibility can't be determined.

B.2 Survey Status for Household: HH_STATUS

B.2.1 Summary

HH_STATUS is defined for all responding dwelling units. First, the variable UPCODE_RESLTNDT 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
'.'	Sampled DU with no household (unresponding DU)
1	Responding Household (Questionnaire data)
2	Eligible Household, NonRespondent (no questionnaire data)
3	Ineligible Household
4	Household Unknown eligibility Status

B.2.2 SAS code defining HH_STATUS

Note: HH_STATUS is assigned only to cases with DU_STATUS = 1

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)

Definitions for household without completed questionnaire:

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

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 used in Mozambique

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
	Out of Scope	91	3
	COVID Delay – Unknown Eligibility	94	4
	Cannot Trace	95	4
	Recorded in another HH or tablet (discrepant record)	99	4

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		
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		
Death/Funeral		
SHE LOST HER BOYFRIEND WHO WAS BURIED LAST SUNDAY. HE DIED OF LIVER PROBLEMS IN SOUTH AFRICA	7	2
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		

Table of examples for RESULTNDOTH upcoding (continued)

RESULTNDOTH	UPCODE_ RESLTNDT	UPCODE_ STAT_HH
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		
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		
Out of Scope	91	3
COVID Delay – Unknown Eligibility	94	4
Cannot Trace	95	4
Recorded in another HH or tablet (discrepant record)	99	4

B.3 INDIV_STATUS

B.3.1 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 ineligible for weighting (roster age 15+ but confirmed age <15)
4	Roster eligible but no confirmed age
5	Roster ineligible (roster age < 15 or SLEEPHERE=2, except cases in status 9)
6	Rostered case from household with no questionnaire data
9	DeJure ineligible (roster age >= 15 but SLEEPHERE = 2)

B.3.2 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;
```

Rename CONFAGEY to CONFAGEY_ORIG when reading in the individual data set.

Then define a temporary value for CONFAGEY to use in the weighting process. (Note: In the delivery file this value will be the value for a new variable CONFAGEY_WEIGHT; the value of CONFAGEY_ORIG will be restored to the variable CONFAGEY.)

In the survey questionnaire, if the initial questions determine that an individual is ineligible, the confirmed age as of that determination is kept as CONFAGEY_INELIG and the survey is terminated. For these cases the later variables will have the value NULL. CONFAGEY_SOURCE is the later variable that denotes how the age was confirmed, so if that variable is NULL the best source for a confirmed age is actually CONFAGEY_INELIG.

This code creates confagey for weighting using the above information:

```
if confageysource = . then confagey = confagey_inelig;
else
  confagey = confagey_orig;
```

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

Note: BT_STATUS is assigned only to cases with INDIV_STATUS = 1

```
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;
```

Note: BT_STATUS = 2 is used for cases with no blood sample taken and also for cases where the blood sample did not result in a definite outcome.

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 3.4.4.2) and the blood test weights (Section 3.4.5.2). 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_BT_STATUS.pdf (Males 15+ years)

AF_BT_STATUS.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

MOZ_AD_INDIV.xlsx (Persons 15+ years)

Blood Test

MOZ_AM_BT.xlsx (Males 15+ years)

MOZ_AF_BT.xlsx (Females 15+ years)