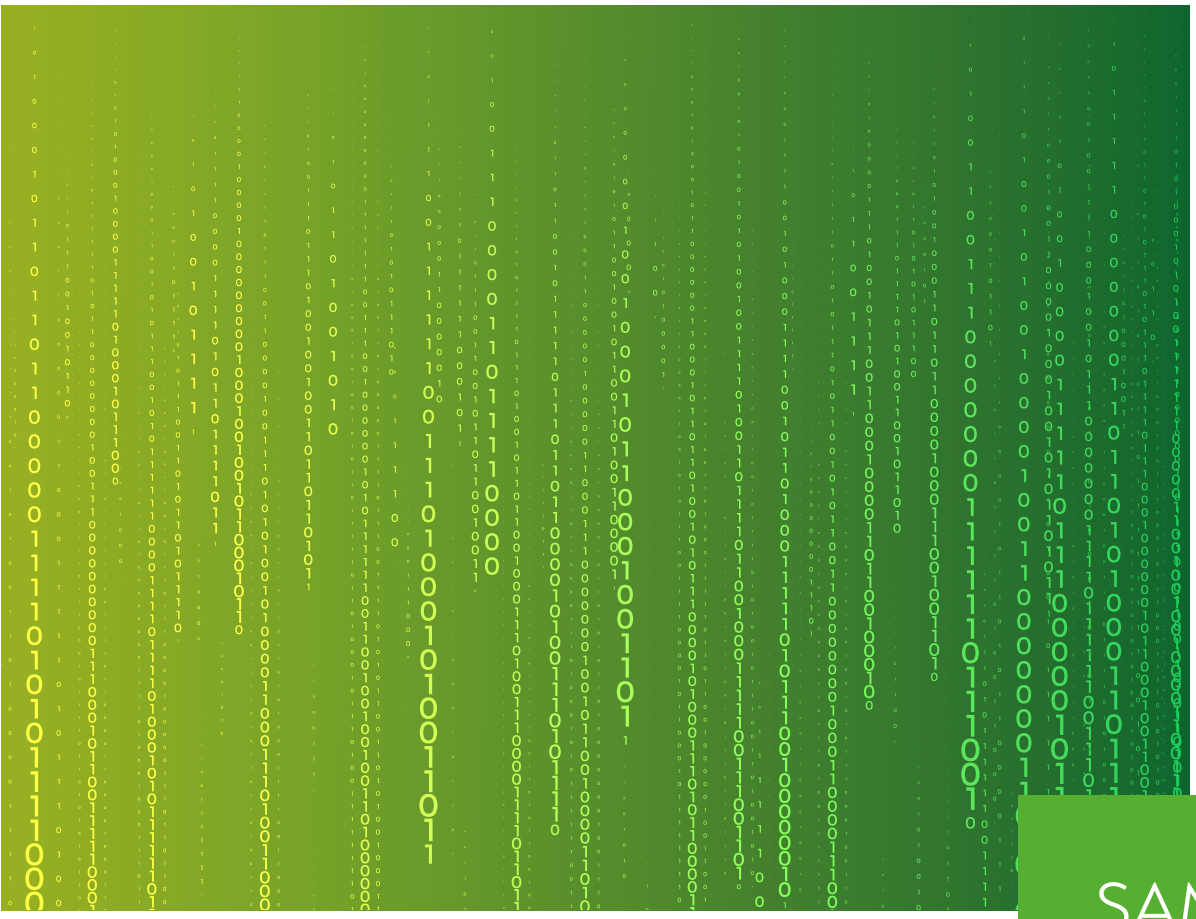
MPHIA 2020

Malawi Population-based HIV Impact Assessment



SAMPLING AND WEIGHTING TECHNICAL REPORT

Malawi Population-based HIV Impact Assessment 2020-2021

MPHIA 2020-2021

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.

Table of Contents

Contents Continued

Acronyms

CDC	Centers for Disease Control and Prevention
CHAID	Chi-square Automatic Interaction Detector
CI	Confidence Interval
CV	Coefficient of Variation
DEFF	Design Effect
DU	Dwelling Unit
EA	Enumeration Area
HH	Household
HIV	Human Immunodeficiency Virus
ICC	Intra Cluster Correlation
LASSO	Least Absolute Shrinkage and Selection Operator
MPHIA	Malawi Population-based HIV Impact Assessment
MDRI	Mean Duration of Recent Infection
MOS	Measure of Size
PHIA	Population-based HIV Impact Assessment
PSU	Primary Sampling Unit
RSE	Relative Standard Error
SAS	Statistical Analysis System
NSO	National Statistics Office
UEW	Unequal Weighting
MPHIA	Malawi Population-based HIV Impact Assessment
VLS	Viral Load Suppression
WLM	Weighted Log-linear Modeling

1. Introduction

The 2020 Malawi Population-based HIV Impact Assessment (MPHIA 2020) is a cross-sectional sample survey designed to assess the prevalence of key human immunodeficiency virus (HIV)-related health indicators among individuals 15 years or older. Data collection for the MPHIA 2020 was conducted between January 2020 and April 2021 with a temporary pause in data collection from April 2020 until March 2021 due to the SARS-CoV-2 global pandemic. The survey included approximately 26,500 interviewed individuals and over 22,600 individuals with valid blood tests in approximately 12,800 randomly-selected households. The purpose of this report is to document the procedures used to select the households and individuals for the study and the subsequent weighting of the respondent sample.

1.1 Overview of Sample Design

The sample design for the MPHIA 2020 is a stratified multistage probability sample design, with strata defined to be seven health zones within the country, first-stage sampling units defined by enumeration areas (EAs) within strata, second-stage sampling units defined by households within EAs, and finally age-eligible persons within households. Within each sampling stratum, the first-stage sampling units (also referred to as "primary sampling units" or PSUs) were selected with probabilities proportionate to updated numbers of households in the PSU derived from the 2018 Malawi Population and Housing Census. The allocation of the sample PSUs to the seven health zones 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) zonal estimates of viral load suppression (VLS) rates among HIV-positive persons 15 to 49 years old.

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

Within the 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 MPHIA data collection.

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

1.2 Overview of Weighting Process

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

The main steps of the weighting process include

- Initial checks to confirm that the probabilities of selection associated with the sampled units are computed correctly;
- Creation of jackknife replicates to be used for variance estimation;
- Calculation of PSU base weights to reflect the overall PSU probabilities of selection, and to compensate for PSU nonresponse;
- Calculation of household weights to reflect the probabilities of selecting households within PSUs, and to compensate for household nonresponse;
- Calculation of person-level interview weights to reflect the differential probabilities of selecting individuals within households, and to compensate for nonresponse to the interview;
- 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; and
- Calculation of person-level blood test weights to reflect the differential probabilities of selecting individuals within households, compensate for nonresponse to the blood test, and adjust for potential undercoverage through post-stratification.

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

2. Sample Design

2.1 Population of Inference

The population of inference for the MPHIA 2020 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 data collection. Table 2-1 summarizes estimates (projections) of the 2020 Malawi population by gender and age group.

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

Age group	Gender		Total
	Male	Female	
15 to 49 years	4,243,881	4,649,546	8,893,427
50 years or older	810,620	955,032	1,765,652
Total	5,054,501	5,604,578	10,659,079

Source: Population projections published in the 2018 Malawi Population and Housing Census report, National Statistical Office (NSO) - http://www.nsomalawi.mw.

2.2 Precision Specifications and Assumptions

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

2.2.1 Specifications

- Relative standard error (RSE) of the national estimate of HIV incidence among adults 15 to 49 years old should be 30% or less;
- 95% confidence interval (CI) bounds around the estimated VLS rate among HIV positive adults aged 15 to 49 years for each of the seven health zones 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 of ±0.025 or less.

- 95% CI bounds around the national estimate of VLS rate among all HIV positive females aged 15 to 24 years should be ±0.09 or less.

2.2.2 Statistical Assumptions

- National HIV prevalence rate of 0.10 (10.0%) for adults 15-49 years old that varies by zone (see Table 2-2), (Source: MPHIA 2015-2016);

- A national HIV prevalence rate of 0.034 (3.4%) for women aged 15 to 24 years old that varies by zone (see Table 2-2), (Source: MPHIA 2015-2016);

- Annual national incidence rate for adults aged 15-49 of p_a 0.0033 (0.33%), (Source: MPHIA 2015-2016);

- Stratum-level (zonal) incidence rates of p_{ah}, h = 1, 2, …, 7, which are obtained by adjusting the national incidence rate using the stratum-level (zonal) prevalence rates as follows:

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

  where p_h and p are the HIV prevalence rates for zone h and the country, respectively, and p_a is the annual national incidence rate (Source: MPHIA 2015-2016);

- 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 = 0.0033/2.8077 = 0.0012 (0.12%), and the corresponding stratum-level (zonal) 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 zone, which is a conservative estimate of the underlying population variance associated with VLS rate;

- Intracluster correlation (ICC) of 0.02 for VLS and 0.01 for prevalence (Source: tabulations of MPHIA 2015-2016 data);

- ICC of 0.000 for incidence (Source: analyses of prior PHIA surveys);

- Overall sex-age distributions (Source: tabulations of MPHIA 2015-2016 data); and

- Stratum-level (zonal) population distribution obtained from the 2018 Malawi Population & Housing Preliminary Report, December 2018.

2.2.3 Operational Assumptions

- Varying numbers of dwelling units/households to be sampled per PSU, resulting in an average of 35 sampled dwelling units/households per PSU;
- Occupancy rate of 89.2% for sampled dwelling units (Source: MPHIA 2015-2016);
- Household response rate of 89.4% among occupied households (Source: MPHIA 2015-2016);
- Average household size of 3.90 (*de facto)* persons per household (Source: MPHIA 2015-2016);
- Overall percentage of *de facto* persons 15-49 years of age per household of 42.7%; and an overall percentage of *de facto* persons 50+ years of age of 10.9% (Source: MPHIA 2015-2016);
- Within the responding households, a person-level interview response rate of 87.3% (Source: MPHIA 2015-2016); and
- Among persons completing the interview, a blood test response rate of 87.2% (Source: MPHIA 2015-2016). Thus, among the persons selected for MPHIA 2020, the assumed overall response rate for the blood tests is 76.1% (87.3% * 87.2%).

Based on the specifications and assumptions listed above, a sample of 438 EAs (clusters) was determined to be the minimum needed to meet the specified precision goals. The allocation of the sample to the seven health zones of Malawi is shown in Table 2-2. The expected numbers of households included in the study and the corresponding projected numbers of respondents by age group are also summarized in this table. The actual numbers of respondents achieved are presented in Sections 2.4 and 2.5 and differ from the counts in Table 2-2 because of differences between the response rates and other assumptions used to develop the sample design and those achieved during data collection. Further details about the sampling of households are given in Section 2.4.

Table 2-2 Allocation of sample clusters (EAs) and dwelling units/households and projected sample sizes (expected number of respondents) by stratum

Zone code	Zone name	HIV prevalence rate[1]		Total Number of sample clusters	Target Number of DUs/HHs to be sampled	Number of participating HHs[2]	Projected Number of respondents[3]	
		Adults 15-49	Females 15-24				Adults 15-49	Adults 50+
1	Blantyre City	0.171	0.079	26	910	726	920	235
2	Central East	0.046	0.007	63	2,205	1,758	2,230	569
3	Central West	0.054	0.013	65	2,275	1,814	2,301	587
4	Lilongwe City	0.107	0.052	30	1,050	837	1,062	271
5	Northern	0.068	0.021	45	1,575	1,256	1,593	407
6	South East	0.145	0.043	106	3,710	2,959	3,753	958
7	South West	0.158	0.061	103	3,605	2,875	3,646	931
All	Malawi	0.100	0.034	438	15,330	12,225	15,505	3,958

DU = dwelling unit; HH= household

[1] Source: MPHIA 2015-2016.

[2] Assumes occupancy rate of 89.2% and household response rate of 89.4%.

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

2.3 Selection of the Primary Sampling Units (PSUs)

2.3.1 Definition of PSUs

In MPHIA 2020 the first-stage sampling units, PSUs, were National Statistics Office (NSO) enumeration areas (EAs). The term PSU is the more general statistical term. The first-stage MPHIA 2020 sample was selected from a sampling frame of EAs that originally had been created for the 2018 Malawi Population Census, and subsequently updated by NSO sometime prior to July 2019. The EAs in the updated sampling frame were generally the same as those created for the 2014 Population Census, except that some EAs were subdivided into separate EAs. The updated sampling frame consisted of slightly over 18,463 EAs containing an estimated 3,978,558 households as of 2019.

2.3.2 Selection of the PSU Sample

A stratified sample of 438 EAs was selected from the updated EA sampling frame in accordance with the sample allocation given in Table 2-2. The following procedure was used to select the EAs for the MPHIA 2020. Within each zone, the EAs in the updated sampling frame were sorted in the

same way they had been sorted in the MPHIA 2015-2016 frame to the extent feasible; i.e., by district within zone, TA code within district and finally by EA code[1] within TA code. The sorting of EAs prior to sample selection induces an implicit geographic substratification within each zone.

Next, a systematic sample of the EAs was selected from each zone. The EAs were selected with probabilities proportionate to a measure of size (MOS) equal to the estimated number of households in the EA in 2019. To select the sample from a given zone, the cumulative MOS was determined for each EA in the ordered list of EAs, and the sample selections were designated using a random start and a sampling interval equal to the total MOS of the EAs in the zone divided by the number of EAs to be selected. The resulting sample has the property that the probability of selecting an EA within a zone is proportional to the MOS of the EA.

Of the 438 EAs selected using this method, 40 had been selected previously for the MPHIA 2015-2016. Following recommendations by NSO, none of the 40 overlapping EAs was replaced by another EA.

2.3.3 Out-of-Scope PSUs

Out-of-scope PSUs 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" PSUs. There were no out-of-scope PSUs in the MPHIA 2020 sample.

2.3.4 Non-responding PSUs and Substitution

A sampled PSU 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. One PSU was considered nonresponding due to

[1] Both TA code and EA code were identifiers defined by NSO and provided on the EA frame.

household listing being conducted outside of the EA boundary. A decision was made not to replace this nonresponding PSU.

2.3.5 Summary of the PSU Sample

As indicated in the previous sections, 438 PSUs (EAs) were selected for the MPHIA 2020. There were no out-of-scope (ineligible) PSU. Additionally, there was one PSU considered as nonresponding after the sampled households were found to be outside the PSU boundary. Table 2-3 summarizes the distribution of the sampled PSUs by zone and sampling status of the PSU.

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

Zone code	Zone name	Sampled PSUs	Nonresponding PSUs excluded from 2nd stage DU/HH selection	Ineligible PSUs excluded from 2nd stage DU/HH selection	Number of in-scope PSUs included in study
1	Blantyre City	26	-	-	26
2	Central East	63	-	-	63
3	Central West	65	1	-	64
4	Lilongwe City	30	-	-	30
5	Northern	45	-	-	45
6	South East	106	-	-	106
7	South West	103	-	-	103
All	Malawi	438	1	0	437

DU = dwelling unit; HH= household

2.4 Selection of Households

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

2.4.1 Definition of Second-Stage Sampling Units

For both sampling and analysis purposes, a household is defined to be a group of individuals who reside in a physical structure such as a house, apartment, compound, or homestead, and share in housekeeping arrangements. The physical structure in which people reside is referred to as the

"dwelling unit" which may contain more than one household meeting the above definition. Households are eligible for participation in the study if they are located within the sampled EA. For the purpose of PHIA, the sampling unit is households. When vacancy of a "dwelling unit" cannot be determined, the "dwelling unit" is included on the household sampling frame. Therefore, the sampling frame and the sample of second-stage sampling units is a mixture of households and dwelling units.

2.4.2 Listing

In essence, the listing process involves compiling complete, up-to-date, and accurate lists of all dwelling units and households for each sampled EA through a field operation using trained staff referred to as "listers." Local leaders and knowledgeable community members were consulted to assist in the listing process. Listers were provided with maps from which to delineate the boundaries of the EA, and to record the locations of the dwelling units/households found by the listers in the field. Information about the listed dwelling units/households was entered into computer tablets. The information recorded in the tablets included the address or description of the listed dwelling unit/household, the name of the head of household (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 103,700 eligible records were listed for the MPHIA 2020.

2.4.3 Determination of Eligibility for Sampling

As indicated above, all known households at the time of listing, plus vacant dwelling units that could potentially be occupied at the time of interview, were initially entered into the tablets as separate records. However, not all of these records were eligible for subsequent sampling purposes. Those records marked with the notation "discard" were data entry errors and were eliminated from the listing file. To establish eligibility for the remaining records, three key variables collected during listing were used: (1) the structure type, (2) whether the listed structure was vacant or under construction, and (3) whether anyone was living in the structure at the time of listing. Based on the values of these three variables, those records meeting the criteria specified in Appendix A were eligible for second-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 dwelling units/households (records) eligible for sampling.

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

Zone code	Zone 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	Blantyre City	6,614	214	214	6,400	6,612
2	Central East	13,458	588	584	12,869	13,447
3	Central West	16,168	457	451	15,706	16,157
4	Lilongwe City	7,386	338	335	7,048	7,357
5	Northern	7,962	369	369	7,591	7,959
6	South East	26,544	1,310	1,302	25,232	26,519
7	South West	25,662	1,171	1,170	24,482	25,652
All	Malawi	103,794	4,447	4,425	99,328	103,703

DU = dwelling unit; HH= household

[1] See Appendix A for additional details.

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

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

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

2.4.4 Selection of Dwelling Units/Households

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

The calculation of the required within-PSU sampling rates proceeded as follows. First, the target overall sampling rate for zone h = 1, 2, ..., 7, was computed as:

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

where

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

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

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

P_{hi} = probability of selecting PSU *i* in zone *h*.

The total *expected* number of listings to be selected across all seven health zones is $\sum_{h=1}^{7} T_h = 15{,}330$ (see Table 2-2). To obtain an equal probability sample within zone *h*, the required within-PSU sampling rate for PSU *i* in zone *h* was then computed as:

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

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

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

To reduce the variation in workload across the sampled PSUs, the maximum number of dwelling units/households to be selected in any PSU was capped at 70 and the minimum number was set to 15. Inspection of the values of $\mathrm{E}(n_{hi})$ indicated that the expected sample sizes for two PSUs would fall below 15, and 5 would exceed 70. The difference between the number of dwelling units/households 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 PSUs in the same stratum so as to maintain as closely as possible the desired total sample size for the stratum. The within-PSU sampling rates, f_{hi}^{within}, were therefore adjusted to account for the redistribution of the sample within the zone. The adjusted within-PSU sampling rate used to select the sample of dwelling units/households, $f_{hi}^{adj(w)}$, was calculated as:

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

where the adjustment factors, A_{hi}, were determined such that

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

L = 15 = the minimum PSU sample size,

U = 70 = the maximum PSU sample size, and

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

To achieve a geographical ordering of the listed dwelling units/households, the dwelling unit/household records in each PSU were sorted by a proximity variable that indicated the distance between the listed dwelling unit/household and the dwelling unit/household closest to the centroid of the PSU. Dwelling units/households within the EA were then selected systematically from the ordered list of records at the rates, $f_{hi}^{adj(w)}$, specified above.

2.4.5 Results of Second-Stage Sampling

Table 2-5 summarizes the numbers of dwelling units/households selected for the study and the minimum and maximum PSU sample size by zone. 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-based estimate resulting from the use of variable overall sampling rates within a zone (e.g., see Kish, 1965, page 403). With an equal-probability sample within each zone, the DEFFs would ordinarily equal 1.0. Variable sampling rates will increase the DEFF, which would arise, for example, from the capping of sample sizes that is done to control workload across EAs. However, since the extent of the capping and redistribution of the sample described previously was moderate, the corresponding increase in the variation of the overall sampling rates was small, resulting in stratum-level (zonal) UEW DEFFs that range from 1.00 to 1.02 (Table 2-5).

Table 2-5 Number of sampled dwelling units/households and expected unequal weighting DEFF by zone

Zone code	Zone name	Number of PSUs	Number of sampled DUs/HHs	Minimum number of DUs/HHs per PSU	Maximum number of DUs/HHs selected per PSU	Unequal weighting DEFF
1	Blantyre City	26	910	24	70	1.0184
2	Central East	63	2,205	17	68	1.0000
3	Central West	65	2,275	15	70	1.0049
4	Lilongwe City	30	1,050	16	53	1.0000
5	Northern	45	1,575	22	54	1.0000
6	South East	106	3,709	15	70	1.0000
7	South West	103	3,606	22	70	1.0004
All	Malawi	438	15,330	15	70	1.0502[1]

DU = dwelling unit; HH= household

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

Table 2-6 summarizes the distribution of the sampled dwelling units/households by final dwelling unit/household response status. Of the 15,330 sampled dwelling units 1,344 (8.8%) were determined during data collection to be vacant/unoccupied, 28 (0.2%) for which eligibility for the survey (i.e., occupancy status) could not be established, 1,143 (7.5%) were determined to be eligible for the study (i.e., contained eligible household members) but did not complete the household interview, and 12,815 (83.6%) completed the household interview. Excluding the ineligible cases, the overall unweighted household response rate was 91.6%.

Table 2-6 Distribution of dwelling unit/household sample by zone and response status

Zone code	Zone name	Number of sampled DUs/HHs	Number of ineligible DUs[1]	Number of DUs/HHs with unknown eligibility[2]	Number of responding households[3]	Number of eligible non-responding households[4]	Unweighted response rate[5]
1	Blantyre City	910	65	4	757	84	0.896
2	Central East	2,205	194	5	1,911	95	0.950
3	Central West	2,275	185	3	1,912	175	0.915
4	Lilongwe City	1,050	53	1	829	167	0.832
5	Northern	1,575	154	3	1,280	138	0.901
6	South East	3,709	330	4	3,132	243	0.927
7	South West	3,606	363	8	2,994	241	0.923
All	Malawi	15,330	1,344	28	12,815	1,143	0.916

DU = dwelling unit; HH= household

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

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

[3] Households completing the household interview.

[4] Occupied 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 dwelling units, and U = number of dwelling units with unknown eligibility.

2.5 Selection of Individuals

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

2.5.1 Household Rosters

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

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

Zone code	Zone name	Number of households completing interview	Rostered persons by resident status[1]				
			Usual resident/did not sleep here[2]	Usual resident/ slept here	Nonresident/ slept here	Nonresident/ did not sleep here[2]	Total rostered persons[3]
1	Blantyre City	757	150	2,844	91	48	3,133
2	Central East	1,911	347	8,687	175	274	9,483
3	Central West	1,912	326	7,856	117	294	8,593
4	Lilongwe City	829	142	3,313	131	150	3,736
5	Northern	1,280	264	5,998	107	202	6,571
6	South East	3,132	562	13,480	189	254	14,485
7	South West	2,994	511	11,937	282	290	13,020
All	Malawi	12,815	2,302	54,115	1,092	1,512	59,021

[1] Counts include persons of all ages.

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

[3] Sixteen roster entries from households that did not complete the household interview are not included in this table, five in Central East, four in South East and seven in South West.

2.5.2 Selecting Individuals for Data Collection

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

Table 2-8 Number of individuals eligible for data collection

Zone code	Zone 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	Blantyre City	1,565	61	1,626	213	8	221
2	Central East	3,877	121	3,998	767	13	780
3	Central West	3,398	83	3,481	773	11	784
4	Lilongwe City	1,728	112	1,840	200	8	208
5	Northern	2,647	67	2,714	602	17	619
6	South East	5,576	119	5,695	1,211	15	1,226
7	South West	5,303	206	5,509	1,331	36	1,367
All	Malawi	24,094	769	24,863	5,097	108	5,205

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

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

2.5.3 Distribution of Sampled Persons

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

Table 2-9 Distribution of sampled persons by age group, response status, and zone

Zone code	Zone 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	Blantyre City	1,626	1,398	1,153	221	184	147
2	Central East	3,998	3,580	3,227	780	710	624
3	Central West	3,481	3,067	2,520	784	717	585
4	Lilongwe City	1,840	1,411	1,102	208	165	134
5	Northern	2,714	2,367	2,070	619	565	494
6	South East	5,695	5,062	4,408	1,226	1,111	951
7	South West	5,509	4,930	4,215	1,367	1,252	1,032
All	Malawi	24,863	21,815	18,695	5,205	4,704	3,967

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

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

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

3. Weighting and Estimation

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

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

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

3.1 Overview of the Weighting Process

The overall weighting approach for MPHIA 2020 includes several steps.

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

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

Calculation of PSU Weights: The weighting process begins with the calculation and checking of the sample PSU (EA) base weights as the reciprocals of the overall PSU probabilities of selection. The PSU base weights are adjusted first to account for nonresponding eligible PSUs. This adjustment is generally made within the stratum in which the PSUs are located. The resulting weight is the final PSU weight.

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

Calculation of Person-Level Interview Weights: Once the household weights are determined, they 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 adjusted 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 2020 population projections.

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

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

3.2 Preparation for Weighting

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

3.2.1 Data Files for Weighting

The MPHIA 2020 survey data that are used to construct the sampling weights are contained in the following data files.

- **mw_CFF_hh_int_STAT_20210526**: A household (HH) file that contains the household data collected in the HH questionnaire.
- **mw_CFF_roster_STAT_20210526**: A file that contains the roster of household members collected in the HH questionnaire with a record for each rostered person.
- **mw_CFF_ind_int_STAT_20210526**: 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.
- **MW2Biomarker20210527**: A biomarker file containing identifying information and results for lab analyses of blood samples for individuals whose blood was drawn and analyzed in the lab.

Each of these data files except the Biomarker file contains records for all sampled or collected cases, irrespective of response and eligibility status. However, for weighting purposes, a subset of the roster file was created with only "roster eligible" cases: these are person-level records from a

responding household with a roster age of 15 or older and who were identified on the roster as having slept in the household the night before the interview. At the time of creating weight delivery files the "roster ineligible" cases were returned to the delivery files; however they have missing values for the weight variables.

3.2.2 Checks of Data Files

Prior to the start of the weighting process, the survey data files are checked and compared against information available in the sampling files. These 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 HH selection process.)
- Check counts of sampled and responding HHs against what was expected, overall and by zone.
- Adjust for substitution of EAs, if applicable. Check that guidelines have been followed and selection probabilities are consistent with guidelines.
- Set disposition codes (respondent, eligible nonrespondent, ineligible, unknown eligibility) to be used for weighting purposes based on data elements received for (a) sampled households, (b) sampled individuals, and (c) individuals selected for blood draws.

3.3 Creation of Variables for Variance Estimation

Two general methods can be used for estimating the sampling errors of survey-based estimates derived from MPHIA 2020: the jackknife replication and Taylor's Series methods. The jackknife replication variance estimation method is a widely used method for producing variance estimates using data from a complex survey. This method can correctly account for the stratification, clustering, and sample weighting, including nonresponse and poststratification weighting adjustments, from the MPHIA 2020 complex sample design. The Taylor's Series is another widely used method that uses linear approximations to calculate the variance of a sample-derived estimate.

In order to implement either method, certain variables required for variance estimation must be included in the weighted data files. In the case of jackknife replication, the required variables are a series of weights that correspond to each of the jackknife replicates. In the case of the Taylor's

Series method, the required variables are those that indicate the "variance stratum" and the "variance unit" to which each sampled respondent belongs.

3.3.1 Jackknife Replication

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

As described in Section 2.3.2, the PSUs were selected systematically from a list of PSUs that had been ordered geographically within zone. To take account of the precision benefits of implicit stratification as fully as possible, the sampled PSUs within each zone were paired off in the systematic order in which they were selected, treating each pair as a variance-estimation stratum. When there was an odd number of sampled PSUs in a zone, one of the variance-estimation strata was defined to contain three sampled PSUs. To fully reflect the sample design, the formation of the variance-estimation strata was applied to all 438 of the sampled PSUs.

For the MPHIA 2020, 217 variance-estimation strata were created. A jackknife replicate was then formed by randomly deleting a PSU from a particular variance-estimation stratum k, say, and retaining all of the PSUs in the remaining variance-estimation strata. For a variance-estimation stratum consisting of a pair of PSUs, the weight of the retained PSU within the variance-estimation stratum k was doubled. For a variance-estimation stratum consisting of three PSUs, the weight of the two retained PSUs within the variance-estimation stratum were increased by 1.5 (see Section 3.4.1). The process was repeated for all $k = 1, 2, 3, \ldots 217$ variance-estimation strata, resulting in a total of 217 jackknife replicates. Table 3-1 summarizes the number of jackknife replicates that were created for variance estimation.

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

Zone code	Zone name	Sampled PSUs[1]	Variance strata consisting of pairs	Variance strata consisting of triplets	Number of jackknife replicates
1	Blantyre City	26	13	0	13
2	Central East	63	30	1	31
3	Central West	65	31	1	32
4	Lilongwe City	30	15	0	15
5	Northern	45	21	1	22
6	South East	106	53	0	53
7	South West	103	50	1	51
All	Malawi	438	213	4	217

[1] Includes nonresponding and ineligible PSUs if applicable.

3.3.2 Taylor's Series

Even though jackknife replication is the recommended method for variance estimation, not all software packages have a replication option to produce variance estimates. Therefore, information for producing Taylor's Series estimates of variance is included in the MPHIA 2020 data files.

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

3.4 Development of Weights

3.4.1 PSU Weights

The initial weighting step after the jackknife replicates were defined was to calculate PSU base weights for the full sample and the replicates.

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

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

where P_{ghi}^{PSU} = probability of selecting PSU *i* from subgroup *g* in zone *h*. Using the PSU weights defined above, the sampled PSUs weight up to the numbers shown in the fourth column of Table 3-2.

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

Let *gh* denote the subgroup with a nonresponding PSU:

m_{gh} is the number of sample PSUs in the subgroup, and

m_{gh}^{r} is the number of responding PSUs in the subgroup.

The nonresponse-adjusted full sample PSU weight was computed as

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

where

$$A_{ghi}^{(1)} = \sum_{i=1}^{m_{gh}} W_{ghi}^{(1)} \; / \sum_{i=1}^{m_{gh}^{r}} W_{ghi}^{(1)}$$

is the PSU weight adjustment factor for subgroup *gh*. The adjustment factor is the reciprocal of the EA response rate within the subgroup. The values of $A_{ghi}^{(1)}$, equal to 1.00 except for subgroups with a nonresponding PSU, are shown in Table 3-2. The corresponding replicate-specific PSU nonresponse adjustment factor for cell *gh* were similarly computed for jackknife replicate *k* = 1, 2, ..., 217.

The adjusted PSU weights, $W_{ghi}^{(1A)}$, are passed to the household weighting process described in the next section.

As described in Section 3.3.1, 217 jackknife replicates were formed from the 438 sampled PSUs. For variance estimation, replicate-specific PSU weights, $W_{ki}^{(1)}$, $r = 1, 2, ..., 217$ were created to provide the basis for calculating the required replicate weights in subsequent stages of the weighting process. Let k denote one of the variance-estimation strata created for jackknife replication (Section 3.3.1) and let i denote the PSU within variance-estimation stratum k. For a given jackknife replicate, $r = 1, 2, ..., 217$, the corresponding replicate-specific PSU base weight where the variance-estimation strata consist of pairs was computed as

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

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

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

Table 3-2 Number of PSUs and corresponding weighted counts by zone

Zone Code	Zone Name	Number of sample EAs (PSUs)	Weighted number of EAs (PSUs) [1]	Number of in-scope PSUs in study	PSU nonresponse adjustment factor	In-scope PSUs weighted by nonresponse adjusted weights[2]	Weighted measure of size (MOS) [3]
1	Blantyre City	26	959	26	1.00	959	226,486
2	Central East	63	3,642	63	1.00	3,642	797,579
3	Central West	65	4,328	65	1;1.08 [4]	4,328	1,218,862
4	Lilongwe City	30	1,170	30	1.00	1,170	292,899
5	Northern	45	3,181	45	1.00	3,181	509,019
6	South East	106	4,037	106	1.00	4,037	1,085,049
7	South West	103	3,969	103	1.00	3,969	1,018,221
All	Malawi	438	21,286	438	-	21,286	5,148,115

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

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

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

[4] In zone 3, Central West, one subgroup has nonresponse adjustment factor of 1.08.

3.4.2 Dwelling Unit/Household Weights

3.4.2.1 Dwelling Unit Base Weights

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

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

where

$W_{ghi}^{(1A)}$ = the nonresponse-adjusted weight for PSU *i* in PSU weighting subgroup *g*

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

The corresponding weights for jackknife replicate $r = 1, 2, \ldots, 217$ were computed as:

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

where $W_{(r)ki}^{(1A)}$ is the PSU nonresponse-adjusted weight for PSU i in variance estimation stratum k 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/households by response status and zone 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 zone.

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

Response status group[1]	Description	Number of sampled dwelling units/households
1	Respondent (household with completed household interview)	12,815
2	Nonrespondent (household without a completed household interview)	1,143
3	Ineligible (dwelling units with no households)	1,344
4	Unknown eligibility (not known if dwelling unit contains household)	28
All	—	15,330

[1] See Appendix B for definitions.

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

Zone code	Zone name	Response status[1]				Total groups 1-4
		Group 1: responding household	Group 2: nonresponding household	Group 3: ineligible dwelling unit	Group 4: unknown eligibility	
1	Blantyre City	197,153	21,364	16,886	998	236,401
2	Central East	615,697	30,608	62,504	1,611	710,420
3	Central West	857,486	78,614	76,689	1,320	1,014,108
4	Lilongwe City	205,211	41,339	13,120	248	259,917
5	Northern	408,471	44,038	49,144	957	502,611
6	South East	802,464	62,224	84,561	1,024	950,273
7	South West	775,157	62,656	94,171	2,067	934,050
All	Malawi	3,861,638	340,843	397,074	8,225	4,607,780

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

3.4.2.2 Adjustment for Dwelling Unit/Household Nonresponse

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

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

Phase 1 Adjustment

In the first phase of adjustment, the weights of the dwelling units/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 PSU 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/household nonresponse adjustments are either the individual PSUs or a group of PSUs. Let n_{hi}^{DU} denote the number of sampled dwelling units/households in PSU weighting cell *i* in zone *h*. Note that n_{hi}^{DU} is the sum of the sample sizes in each of the four response status groups defined in Table 3-3, i.e.,

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

where

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

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

$n_{hi}^{(3)}$ = the number of known ineligible dwelling units (i.e., dwelling units known to contain no households) in PSU weighting cell *i* in zone *h*

$n_{hi}^{(4)}$ = the number of sampled dwelling units for which it is not known whether a household is present in PSU weighting cell *i* in zone *h* .

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

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

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

The corresponding replicate-specific first-phase dwelling units/households nonresponse adjustment factor for cell *c* were similarly computed for jackknife replicate *r* = 1, 2, ..., 217.

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

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

The corresponding replicate weights for replicate r = 1, 2, …, 217 were computed in similar fashion as:

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

where

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

Note that for the dwelling units in response-status group 4 (dwelling units of unknown eligibility), $W_{hij}^{DU1} = W_{(r)hij}^{DU1} = 0$ for r = 1, 2, …, 217.

The effect of this adjustment is to distribute the total weight of the unknown-eligibility cases (i.e., the estimated 8,225 dwelling units shown in the next-to-last column of Table 3-4) to the combined weight of the remaining three groups of sampled dwelling units/households. The resulting weighted counts using W_{hij}^{DU1} as computed above are summarized in Table 3-5.

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

Zone code	Zone name	Response status				Total households: groups 1-2
		Group 1: responding household	Group 2: nonresponding household	Group 3: ineligible dwelling unit	Total status 1-3	
1	Blantyre City	198,037	21,428	16,937	236,401	219,465
2	Central East	617,057	30,721	62,641	710,420	647,779
3	Central West	858,555	78,758	76,795	1,014,108	937,313
4	Lilongwe City	205,369	41,394	13,154	259,917	246,763
5	Northern	409,311	44,100	49,201	502,611	453,411
6	South East	803,272	62,343	84,658	950,273	865,615
7	South West	776,983	62,776	94,292	934,050	839,759
All	Malawi	3,868,583	341,520	397,677	4,607,780	4,210,103

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

Phase 2 Adjustment

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

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

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

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

The corresponding replicate-specific interview nonresponse adjustment factor for cell *i* were similarly computed for jackknife replicate *r* = 1, 2, ..., 217.

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

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

The corresponding replicate weights for replicate *r* = 1, 2, …, 217 were computed in similar fashion as:

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

where

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

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

3.4.3 Person-Level Interview Weights

In this section, we detail the calculation of person-level sampling weights to be used to analyze the individual interview responses in the MPHIA 2020 data files. First, we define the initial person-level

(interview) base weights in Section 3.4.3.1. Next, to compensate for interview nonresponse, the person base weights are adjusted within cells defined by variables available for both the responding and nonresponding individuals. Like the dwelling unit/household nonresponse adjustments described previously, this person-level nonresponse adjustment was implemented in two phases.

3.4.3.1 Person Base Weights

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

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

The corresponding replicate base weights, $W_{(r)hijk}^{(base)}$, for $r = 1, 2, \ldots, 217$ were computed in an analogous manner, with $W_{hij}^{(2A)}$ replaced by $W_{(r)hij}^{(2A)}$ in the above formula.

3.4.3.2 Adjustment of Person Weights for Interview Nonresponse

Since the final eligibility of a rostered person cannot be determined until after the actual age is confirmed during the interview, the person-level base weights were adjusted in two phases. Table 3-6 summarizes the distribution of the rostered persons by the five response-status groups specified for the first-phase adjustment. Response status groups 4 and 5 are the cases determined to be ineligible for the study because they were either under 15 years old, or because they were neither present in the household nor a usual resident of the household at the time the household roster was compiled. All of these cases are treated as "known ineligible" cases and are excluded from the first-phase adjustment. The cases in response-status group 3 are cases for which final eligibility for the study is not known 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 3-6 Distribution of rostered persons by age group and first-phase response status

First-phase response status group[1]	Resident status and age based on roster	Confirmed age based on interview	Number of rostered persons	Weighted number of rostered persons[2]
1	De facto person 15 years or older	15+	30,049	9,851,958
2	De facto person 15 years or older	Under 15	1	363
3	De facto person 15 years or older	Unknown	18	5,742
4	Non de facto persons 15 years or older	NA	3,095	1,024,104
5	Persons under 15 years	NA	25,863	8,507,480
All	—	—	59,026[3]	19,389,646[3]

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

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

[3] Of the 59,026 rostered persons, 1,512 were those that neither slept in the household nor were usual residents (see Table 2-7). On a weighted basis, these 1,512 persons account for 511,068 of the total weighted count of 19,389,646 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 3-7), the first-phase interview nonresponse adjustment factor for cell c is, $A_c^{(1)}$, was computed as

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

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

The corresponding replicate-specific first-phase interview nonresponse adjustment factors for cell c were similarly computed for jackknife replicate r = 1, 2, ..., 217.

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, ..., 217.

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 $r = 1, 2, ..., 217$ were similarly computed as

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

Phase 2 Adjustment

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

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

Sex/Age group[1]	Interview response status[2]	Unweighted sample size	Weighted count[3]
Male 15 or older	Eligible respondent	11,151	3,674,764
	Eligible nonrespondent	2,192	726,047
	All response statuses	**13,343**	**4,400,812**
Female 15 or older	Eligible respondent	15,368	5,005,159
	Eligible nonrespondent	1,338	451,729
	All response statuses	**16,706**	**5,456,888**
Total 15 years or older	Eligible respondent	26,519	8,679,924
	Eligible nonrespondent	3,530	1,177,776
	All response statuses	**30,049**	**9,857,700**

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

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

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

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

There are 45 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 45 variables used in the initial model, the LASSO identified 30 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 base weight, $W_{hijk}^{(base)}$. An output of the CHAID procedure is a tree diagram that specifies the optimum number of final weighting cells, and their definitions based on the input predictor variables. The depth limit of the tree was set to 5, and the minimum subgroup size required to allow splitting and minimum terminal node size were set to 50 observations (both respondents and nonrespondents).

To create the CHAID tree, gender (variable SEX) and an indicator of whether or not the individual was under 18 years of age (H_AGETEENYEARS) were forced into the model to make the initial splits. The reason for doing this is that males and females in the specified age groups received different questions; without forcing this variable into the model, the resulting tree would not have been created correctly. After forcing these two variables into the model, the tree was then allowed to grow freely. The CHAID algorithm identified 17 variables to create the weighting classes for nonresponse adjustment. Table 3-8 lists the variables that were included in the final CHAID models. The final trees produced by the CHAID algorithm are documented in Appendix C.1. The corresponding nonresponse-adjustment classes used to adjust the person-level base weights are given in Appendix C.2.

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

Variable number	Variable name	Description
1	DEATHS	Has Any Usual Resident Of Your Household Died Since January 1, 2018?
2	ECONSUPCOVID	Household Economic Support - Covid19
3	EMANCIPATED	Is Name Emancipated?
4	H_AGETEENYEARS	Teen Indicator: 1 – 15-17 Years Old; 2 – Otherwise; Based On Roster Age
5	H_AGEYEARS	Categorical, Based On Roster Age, Corresponding to population based control groupings
6	H_ECONSUP12_A	Household Economic Support: Nothing
7	H_HHQOWN	Does any member of your HH own: 1:bicycle; 2:working Motorcycle Or Motor Scooter; 3:working Car Or Truck; 4:a Working Boat With A Motor; 5: None Of The Above
8	H_HH_SIZE_C	Household size
9	H_OWNCHIKNNUM	Chickens: Altogether, How Many Of The Below Listed Animals Do Members Of Your Household Own?
10	H_OWNDOGNUM	Dogs: Altogether, How Many Of The Below Listed Animals Do Members Of Your Household Own?
11	H_RELATTOHH	What Is The Relationship Of Name To The Head Of The Household? 1. Head 2. Wife/Husband/Partner 3. Son Or Daughter 4. Son-In-Law/Daughter-In-Law 5. Grandchild 6. Parent 7. Parent-In-Law 8. Brother/Sister 9. Co-Wife 10.Other
12	H_ROOMSLEEP	How Many Rooms Are Used For Sleeping?
13	LIVEHERE	Does Name Usually Live Here?
14	SEX	Is Name Male Or Female?
15	SICK_HOUSEHOLD	Sickhouse Flag
16	STRATA	Sampling Stratum Code - Assigned By Stat Team
17	URBAN_RURAL	1=urban, 2=rural

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 is $A_m^{(int)}$, 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, ..., 217 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, ..., 217.

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, ..., 217 were similarly computed as

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

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

Table 3-9 Summary of the interview nonresponse adjustment process

Characteristic	Total sample
Number of variables in initial model	45
Number of variables selected by LASSO	30
Number of variables selected by CHAID	17
Number of final nonresponse-adjustment cells	56
Number of interview respondents	26,519
Minimum adjustment factor	1.00
Maximum adjustment	3.02
Weighted count of respondents before adjustment[1]	8,679,924
Weighted count of respondents after adjustment[2]	9,857,700

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

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

3.4.3.3 Poststratification Adjustment

The final step in computing the individual interview weights was to adjust the nonresponse-adjusted interview weights using a procedure called poststratification (Kalton and Kasprzyk, 1986). The

primary goal of poststratification is to mitigate noncoverage biases that result when some persons in the study population do not have a chance to be sampled and interviewed. For example, undercoverage can occur:

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

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

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

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

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

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

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

for the $r = 1, 2, \ldots, 217$ jackknife replicates.

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

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

and the corresponding poststratified replicate weights were computed as:

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

for $r = 1, 2, \ldots, 217$.

Table 3-10 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 3-10 2020 Malawi population projections and weighted counts before 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	1,003,449	890,412	1.127	1,077,303	902,690	1.193	2,080,752	1,793,103	1.160
20-24	839,923	739,242	1.136	921,840	961,494	0.959	1,761,763	1,700,736	1.036
25-29	697,827	532,079	1.312	773,244	760,610	1.017	1,471,071	1,292,689	1.138
30-34	568,066	432,941	1.312	633,213	569,570	1.112	1,201,279	1,002,511	1.198
35-39	459,440	402,672	1.141	514,554	571,007	0.901	973,994	973,679	1.000
40-44	374,117	343,280	1.090	408,429	406,590	1.005	782,546	749,871	1.044
45-49	301,059	292,120	1.031	320,963	331,246	0.969	622,022	623,366	0.998
50-54	229,937	197,147	1.166	245,182	222,786	1.101	475,119	419,933	1.131
55-59	173,116	148,405	1.167	187,685	189,013	0.993	360,801	337,418	1.069
60-64	131,597	125,916	1.045	150,417	167,085	0.900	282,014	293,002	0.962
65+	275,970	283,017	0.975	371,748	388,374	0.957	647,718	671,391	0.965
Total 15+	5,054,501	4,387,233	1.152	5,604,578	5,470,467	1.025	10,659,079	9,857,700	1.081

[1] Source: National Statistics Office (NSO).

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

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

3.4.4 Person-Level Blood Test Weights

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

3.4.4.1 Initial Weights

The starting point for the construction of the blood test weights is the set of final full-sample nonresponse-adjusted interview weights and corresponding replicate weights described in Section 3.4.3.2. These weights are given by $W_{hijk}^{(int)}$ and $W_{(r)hijk}^{(int)}$ (for replicate $r = 1, 2, \ldots, 217$), respectively, where k denotes the interview respondent, h denotes the zone, i denotes the PSU, and j denotes the household. These weights have been adjusted for interview nonresponse, and thus act as the "base" weights for developing nonresponse adjustments for the blood test weights. Table 3-11 summarizes the counts of individuals by sex, age group and blood test response status, and the corresponding weighted counts using the nonresponse person-level interview weights, $W_{hijk}^{(int)}$.

Table 3-11 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	7,846	3,108,710
		Eligible nonrespondent	1,267	524,037
	Female	Eligible respondent	10,840	3,817,929
		Eligible nonrespondent	1,847	685,280
50 years or older	Male	Eligible respondent	1,749	657,428
		Eligible nonrespondent	252	97,059
	Female	Eligible respondent	2,227	786,625
		Eligible nonrespondent	491	180,633
15 years or older	Male	Eligible respondent	9,595	3,766,137
		Eligible nonrespondent	1,519	621,096
	Female	Eligible respondent	13,067	4,604,554
		Eligible nonrespondent	2,338	865,912

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

3.4.4.2 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 PSU characteristics such as zone 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, 77 potential predictor variables were available for initial selection. For females, 80 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 41 significant variables for males and 57 significant variables for females. The selected variables were then input into the CHAID program to create the final weighting cells for nonresponse adjustment.

The CHAID algorithm identified 16 variables for males and 17 variables for females that were then used to create weighting classes for nonresponse adjustment. Table 3-12 lists the variables that were included in the final CHAID models. The final trees produced by the CHAID algorithm are documented in Appendix C.1. The corresponding nonresponse-adjustment classes used to adjust the person-level base weights are given in Appendix C.2.

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

Sex	Variable number	Variable name	Description
Male	1	ANXIETY	Tb And Other Health Issues: Over The Past Two Weeks, How Often Have You Felt Nervous, Anxious Or On Edge?
	2	AT_FIRSTSXAGE	Age Of First Sexual Activity
	3	AT_LIFETIMESEX	In Total, With How Many Different People Have You Had Sex In Your Lifetime?
	4	CURMAR	Marriage: What Is Your Marital Status Now: Are You Married, Living Together With Someone As If Married, Widowed, Divorced, Or Separated/Single?
	5	HFHIVTSTOFFER	HIV Testing: During Any Of Your Visits To The Health Facility In The Last 12 Months, Did A Doctor, Clinical Officer Or Nurse Offer You An HIV Test?
	6	HIVPOSPROV	HIV Testing: Has A Health Care Provider Ever Told You That You Have HIV?
	7	HIVSELFTST	HIV Testing: Have You Ever Tested Yourself For HIV Using A Self-Test Kit?
	8	MATEXWALLS	HH Characteristics: Main Material Of Exterior Walls
	9	MCPLANS	Male Circumcision: Are You Planning To Get Circumcised Within The Next 6 Months?
	10	OUTREGIONTYPE	Background: Just Before You Moved Here, Did You Live In A City, In A Town (Boma, Big Trading Centre) Or In A Rural Area?
	11	SCHLCUR	Background: Are You Currently Enrolled In School?
	12	SICK3MO	HH Roster: Has Name Been Very Sick For At Least 3 Months During The Past 12 Months, That Is Name Was Too Sick To Work Or Do Normal Activities?
	13	SICK_HOUSEHOLD	Calc - Sickhouse Flag
	14	STRATA	Sampling Stratum code - assigned by STAT team
	15	URBAN_RURAL	1=Urban, 2=Rural
	16	WORKIND	Background: What Is Your Occupation? That Is, What Kind Of Work Do You Mainly Do?
Female	17	ANXIETY	Tb And Other Health Issues: Over The Past Two Weeks, How Often Have You Felt Nervous, Anxious Or On Edge?
	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_LIVEB	How Many Times Have You Had A Pregnancy That Resulted In A Live Birth?
	21	AVOIDPREG	Reproduction: Are You Or Your Partner Currently Doing Something Or Using Any Method To Delay Or Avoid Getting Pregnant?
	22	CHTSTHIVBIRTH1	Reproduction: After Childname Was Born, Was He/She Tested For HIV?
	23	CURMAR	Marriage: What Is Your Marital Status Now: Are You Married, Living Together With Someone As If Married, Widowed, Divorced, Or Separated/Single?
	24	HFHIVTSTOFFER	HIV Testing: During Any Of Your Visits To The Health Facility In The Last 12 Months, Did A Doctor, Clinical Officer Or Nurse Offer You An HIV Test?

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

Sex	Variable number	Variable name	Description
Female	25	KNOWN_HIV_STATUS_R	Categorical Known HIV Status
	26	PARTLASTCNDM1	Sexual Activity: The Last Time You Had Sex With Initials, Was A Condom Used?
	27	PARTLASTETOH1	Sexual Activity: The Last Time You Had Sex With Initials, Did Either Of You Drink Alcohol Beforehand?
	28	SCHCOM	Background: What Is The Highest Class/Form/Year You Have Completed?
	29	SICK_HOUSEHOLD	Calc - Sickhouse Flag
	30	STRATA	Sampling Stratum code - assigned by STAT team
	31	TOILETTYPE	HH Characteristics: What Kind Of Toilet Facility Do Members Of Your Household Usually Use?
	32	URBAN_RURAL	1=Urban, 2=Rural
	33	WORRY	Tb And Other Health Issues: Over The Past Two Weeks, How Often Have You Not Been Able To Stop Or Control Worrying?

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 for blood-test nonresponse adjustment factor for cell *m*, $A_m^{(BT)}$, was computed as

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

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, ..., 217.

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, ..., 217.

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 $r = 1, 2, ..., 217$ were similarly computed as

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

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

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

Characteristic	Male	Female
Number of variables in initial model	77	80
Number of variables selected by LASSO	41	57
Number of variables selected by CHAID	16	17
Number of final nonresponse-adjustment cells	30	39
Number of interview respondents	9,595	13,067
Minimum adjustment factor	1.00	1.00
Maximum adjustment	1.65	1.92
Weighted count of respondents before adjustment[1]	3,766,137	4,604,554
Weighted count of respondents after adjustment[2]	4,387,233	5,470,467

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

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

3.4.4.3 Poststratification Adjustment

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

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

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

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

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

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

for the *r* = 1, 2, …, 217 jackknife replicates.

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

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

and the corresponding poststratified replicate weights were computed as:

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

for *r* = 1, 2, …, 217.

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

Table 3-14 2020 Malawi 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	1,003,449	897,904	1.118	1,077,303	907,199	1.188	2,080,752	1,805,103	1.153
20-24	839,923	729,668	1.151	921,840	955,446	0.965	1,761,763	1,685,114	1.045
25-29	697,827	514,026	1.358	773,244	752,603	1.027	1,471,071	1,266,628	1.161
30-34	568,066	425,413	1.335	633,213	576,477	1.098	1,201,279	1,001,890	1.199
35-39	459,440	409,422	1.122	514,554	571,938	0.900	973,994	981,360	0.992
40-44	374,117	346,262	1.080	408,429	412,357	0.990	782,546	758,619	1.032
45-49	301,059	292,308	1.030	320,963	340,944	0.941	622,022	633,253	0.982
50-54	229,937	205,570	1.119	245,182	227,356	1.078	475,119	432,926	1.097
55-59	173,116	153,565	1.127	187,685	191,019	0.983	360,801	344,584	1.047
60-64	131,597	132,765	0.991	150,417	173,482	0.867	282,014	306,247	0.921
65+	275,970	280,329	0.984	371,748	361,646	1.028	647,718	641,975	1.009
Total 15+	5,054,501	4,387,233	1.152	5,604,578	5,470,467	1.025	10,659,079	9,857,700	1.081

[1] Source: National Statistics Office (NSO).

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

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

References

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

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

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

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

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

2018 Malawi Population and Housing Census Main Report, National Statistical Office (NSO) - http://www.nsomalawi.mw.

Appendix A

Definition of Eligibility for Dwelling Unit/Household Sampling

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

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

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

Of the 103,794 dwelling unit/household records in the listing file, 1,091 were classified as ineligible for sampling based on the structure type, vacancy status, and residential status. Thus, a total of 102,703 records in the Master Listing File were eligible for household sampling.

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

Structure type (STOBS_D)	Vac/Constr. Status (STVAC_D)	Resid. Status (RESYN_D)	ELIG_FLAG	Total in master file	Eligible
Cases with no GPS information			N	19	
1 = Single House / compound of houses	1 = Not Vacant and not under construction	1 = Yes	Y	89,160	89,160
1 = Single House / compound of houses	1 = Not Vacant and not under construction	2 = No	Y	278	278
1 = Single House / compound of houses	2 = Vacant	1 = Yes	Y	150	150
1 = Single House / compound of houses	2 = Vacant	2 = No	Y	1,665	1,665
1 = Single House / compound of houses	3 = Under Construction	1 = Yes	Y	885	885
1 = Single House / compound of houses	3 = Under Construction	2 = No	Y	1,534	1,534
2 = Flat/Block/Apartment building	1 = Not Vacant and not under construction		Y	3	3
2 = Flat/Block/Apartment building	1 = Not Vacant and not under construction	1 = Yes	Y	8,816	8,816
2 = Flat/Block/Apartment building	1 = Not Vacant and not under construction	2 = No	Y	204	204
2 = Flat/Block/Apartment building	1 = Vacant	1 = Yes	Y	4	4
2 = Flat/Block/Apartment building	2 = Vacant	2 = No	Y	33	33
2 = Flat/Block/Apartment building	3 = Under Construction		Y	7	7
2 = Flat/Block/Apartment building	3 = Under Construction	1 = Yes	Y	29	29
2 = Flat/Block/Apartment building	3 = Under Construction	2 = No	Y	106	106
3 = Church/Mosque/Temple	1 = Not Vacant and not under construction	1 = Yes	Y	21	21
3 = Church/Mosque/Temple	1 = Not Vacant and not under construction	2 = No	N	1	
3 = Church/Mosque/Temple	2 = Vacant	1 = Yes	Y		
3 = Church/Mosque/Temple	2 = Vacant	2 = No	N	4	
3 = Church/Mosque/Temple	3 = Under Construction	1 = Yes	Y		
3 = Church/Mosque/Temple	3 = Under Construction	2 = No	N	1	
4 = Shop/office/bus. cntr/comm. bldg.	1 = Not Vacant and not under construction	1 = Yes	Y	697	697
4 = Shop/office/bus. cntr/comm. bldg.	1 = Not Vacant and not under construction	2 = No	N	46	
4 = Shop/office/bus. cntr/comm. bldg.	2 = Vacant	1 = Yes	Y	4	4
4 = Shop/office/bus. cntr/comm. bldg.	2 = Vacant	2 = No	N	12	
4 = Shop/office/bus. cntr/comm. bldg.	3 = Under Construction	1 = Yes	Y	3	3
4 = Shop/office/bus. cntr/comm. bldg.	3 = Under Construction	2 = No	N	5	
5 = School/University	1 = Not Vacant and not under construction	1 = Yes	Y	89	89
5 = School/University	1 = Not Vacant and not under construction	2 = No	Y	3	3
5 = School/University	2 = Vacant	1 = Yes	Y		
5 = School/University	2 = Vacant	2 = No	Y	5	5
5 = School/University	3 = Under Construction	1 = Yes	Y		
5 = School/University	3 = Under Construction	2 = No	N		
6 = Clinic/hospital/Doctors office	1 = Not Vacant and not under construction	1 = Yes	Y	5	5
6 = Clinic/hospital/Doctors office	1 = Not Vacant and not under construction	2 = No	N		
6 = Clinic/hospital/Doctors office	2 = Vacant	1 = Yes	Y		

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

Structure type (STOBS_D)	Vac/Constr. Status (STVAC_D)	Resid. Status (RESYN_D)	ELIG_FLAG	Total in master file	Eligible
6 = Clinic/hospital/Doctors office	2 = Vacant	2 = No	N		
6 = Clinic/hospital/Doctors office	3 = Under Construction	1 = Yes	Y		
6 = Clinic/hospital/Doctors office	3 = Under Construction	2 = No	N		
7 = Community Center/CBO	1 = Not Vacant and not under construction	1 = Yes	Y	1	1
7 = Community Center/CBO	1 = Not Vacant and not under construction	2 = No	N	2	
7 = Community Center/CBO	2 = Vacant	1 = Yes	Y		
7 = Community Center/CBO	2 = Vacant	2 = No	N		
7 = Community Center/CBO	3 = Under Construction	1 = Yes	Y		
7 = Community Center/CBO	3 = Under Construction	2 = No	N		
96 = Other	1 = Not Vacant and not under construction	1 = Yes	Y	1	1
96 = Other	1 = Not Vacant and not under construction	2 = No	N	1	
96 = Other	2 = Vacant	1 = Yes	Y		
96 = Other	2 = Vacant	2 = No	N		
96 = Other	3 = Under Construction	1 = Yes	Y		
96 = Other	3 = Under Construction	2 = No	N		
				103,794	103,703

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

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

Appendix B

Definition of Household, Interview, and Blood Test Response Status

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

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

B.1 Survey Status for Household: HH_STATUS

B.1.1 Summary

HH_STATUS is defined for all sampled 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
1	Responding household (completed household interview)
2	Nonresponding in-scope household
3	Household not in scope for the survey
4	Household whose survey eligibility could not be determined

B.1.2 SAS code defining HH_STATUS

```
HH_STATUS = UPCODE_STAT_HH;
  IF EA_HHID_FIXED IN ('MW309050230190', 'MW308020090137', 'MW310010600081',
                       'MW202050090120', 'MW312010140114')
        THEN HH_STATUS = 2;
```

Note: These five households had a roster but not a completed household survey, so their HH_STATUS was recoded to '2' in the weighting program.

Definition for responding household:

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

ELSE assign UPCODE_STAT_HH to 2, 3 or 4 using rules shown below.

UPCODE_STAT_HH = 2 if

- RESULTNDT OR UPCODE_RESLTNDT = 1 or 2 or 7 or 8 or 9
- If RESULTNDT=NULL, then
  - If HHELIG = 2 OR
  - (HHCONSTAT = 2 or 3) or
  - HHELIG = 1 AND HHCONSTAT=NULL OR
  - STARTINT = 4 and ROSTER_MENU is NULL

UPCODE_STAT_HH = 3 if

  - RESULTNDT OR UPCODE_RESLTNDT = 3 or 4 or 6

UPCODE_STAT_HH = 4 if

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

Tables showing upcoding scheme for RESULTNDT = '96' cases used in Malawi

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
		UPCODE_RESLTNDT	
	OTHER	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	7	2
SHE LOST HER BOYFRIEND WHO WAS BURIED LAST SUNDAY. HE DIED OF LIVER PROBLEMS IN SOUTH AFRICA		
FUNERAL AT THE HOUSEHOLD		
GRIEVING.SHE RECENTLY LOST A SON AND MOURNERS ARE STILL GATHERED		
NOT IN AN EMOTIONAL STATE TO PARTICIPATE, HH MISSING, DEATH OF A GRANDCHILD AND BIRTH OF CHILD		
CLOSE RELATIVE (DAUGHTER-IN-LAW) TO THE DECEASED BURIAL SCHEDULED		
Participant/Household Head unable to do survey (incapacitated, language barrier, under age)	8	2
HOUSEHOLD HEAD INCAPACITATED MENTALLY CHALLENGED		
THE PARTICIPANT IS INCAPACITATED -DEAF		
SINGLE HOUSEHOLD MEMBER WHO IS TOO OLD AND INCAPACITATED		
HH IS 14 YEARS OLD SO PARTICIPANT IS INELIGIBLE		

Table of examples for RESULTNDOTH upcoding (continued)

RESULTNDOTH	UPCODE_ RESLTNDT	UPCODE_ STAT_HH
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.2 INDIV_STATUS

B.2.1 Summary

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

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

B.2.2 SAS Code for INDIV_STATUS

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

```
label roster_elig = "Flag for roster eligible";

if hh_status ^= 1 then roster_elig = 2;
else
  if sleephere = 1 and
    ageyears => 15 then roster_elig = 1;
  else
    roster_elig = 0;
```

Next, combine Roster_Elig with endmsg1 and Confagey to create INDIV_STATUS (endmsg1 = 'A' indicates a completed Individual questionnaire)

```
label INDIV_STATUS = "Individual Response Status";

if roster_elig = 2 then indiv_status = 6;
else
  if roster_elig = 0 then do;

    if sleephere = 2 and
       livehere  = 1 and
       ageyears >= 15 then indiv_status = 9;
    else
      indiv_status = 5;
end;

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

B.3 BT_STATUS

B.3.1 Summary

BT_STATUS is only defined for cases where INDIV_STATUS = 1. It is based on information from the Biomarker data set.

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

B.3.2 SAS Code for BT_STATUS

```
ATTRIB BT_STATUS LABEL="Blood test disposition code: 1 = Valid lab results, 2 = No valid
lab results or didn't do BT;
        IF HIV1statusfinalsurvey IN ("Positive" "Negative") THEN BT_STATUS=1;
        ELSE BT_STATUS=2;
```

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 3.4.3.2) and the blood test weights (Section 3.4.4.2). The names of the PDF files containing the CHAID trees are listed below. Each tree indicates diagrammatically how the final weighting cells were created by successively partitioning the sample into heterogeneous subsets with respect to response propensity. The final cells (prior to collapsing, if done to control variation in weights) are indicated by the number underneath the box defining the cell.

Individual Interview

AD_INDIV_STATUS.pdf (Persons 15+ years)

Blood Test

AM_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

MW_AD_INDIV.xlsx (Persons 15+ years)

Blood Test

MW_AM_BT.xlsx (Males 15+ years)

MW_AF_BT.xlsx (Females 15+ years)