ZAMPHIA
ZAMBIA POPULATION-BASED HIV IMPACT ASSESSMENT
A DROP THAT COUNTS

SAMPLING AND WEIGHTING TECHNICAL REPORT
ZAMPHIA 2016

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

This project is supported by the U.S. President's Emergency Plan for AIDS Relief (PEPFAR) through CDC under the terms of cooperative agreement #U2GGH001226.
The contents of this document do not necessarily represent the official position of the funding agencies.

Table of Contents

Contents Continued

Acronyms

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

1 Introduction

The 2016 Zambia Population-based HIV Impact Assessment (ZAMPHIA) is a cross-sectional sample survey designed to assess the prevalence of key human immunodeficiency virus (HIV)-related health indicators. Data collection for the ZAMPHIA was conducted between March 2016 and September 2016, and included over 31,000 individuals in approximately 11,000 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 ZAMPHIA is a stratified multistage probability sample design, with strata defined by the 10 provinces of Zambia, first-stage sampling units defined by enumeration areas (EAs) within strata, second-stage sampling units defined by households within EAs, and finally eligible persons 0-59 years of age within households.

The first-stage sampling units (also referred to as the "primary sampling units" or PSUs) were stratified by the 10 provinces of the country, and then within each province were selected with probabilities proportionate to the number of households in the PSU based on the 2010 census. The allocation of the sample PSUs to the 10 provinces was made in a manner designed to achieve specified precision levels for a national estimate of HIV incidence rate, and province-level estimates of viral load suppression (VLS) rates.

The second-stage sampling units were selected from lists of dwelling units/households complied by trained staff for each of the sampled PSUs. Upon completion of the listing process, a random systematic sample of dwelling units/households was selected from each PSU at rates designed to yield a self-weighting (i.e., equal probability) sample within each province to the extent feasible.

Within the sampled households, all eligible adults 15 to 59 years of age were included in the study sample for data collection. All eligible children 0-14 years of age in a randomly designated subset of one-half of the selected households were included in the study for data collection.

Details of sample design employed for the ZAMPHIA 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 within 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 are:

- Initial checks to confirm that the probabilities of selection associated with the sampled units are computed correctly.
- Creation of jackknife replicates to be used for variance estimation.
- Calculation of PSU base weights to reflect the overall PSU probabilities of selection.
- Calculation of household weights to reflect the probabilities of selecting households within PSUs, and to compensate for household nonresponse.
- Calculation of person-level interview weights to reflect the differential probabilities of selecting individual within households, and to compensate for nonresponse to the interview.
- Poststratification of the person-level interview weights to calibrate the weighted counts of persons completing the interview so that they match external population counts.
- Calculation of person-level blood test weights to reflect the differential probabilities of selecting individual within households, compensate for nonresponse to the blood test, and adjust for potential undercoverage through poststratification.

Technical details of the weighting procedures employed in ZAMPHIA are provided in Section 3.

Sample Design 2

2.1 Population of Inference

The population of inference for the ZAMPHIA is comprised of individuals 0-59 years of age who were present in households (i.e., "slept in the household") on the night prior to the date of interview. This population is referred to as the *de facto* population. In contrast, those individuals who are usual residents of the household regardless of whether they were present in the household during the previous night comprise the *de jure* population. All individuals belonging to either the *de facto* or *de jure* populations were included for PHIA data collection; however, as discussed later in Section 2.5, only members of the *de facto* population are included in the PHIA study population. Table 2-1 summarizes projections of the 2016 Zambia population by gender and age group.

Table 2-1 Summary of 2016 population projections for Zambia by gender and age group

Age group	Gender		Total
	Male	Female	
14 years or younger	3,683,036	3,643,560	7,326,596
15 to 49 years	3,614,536	3,769,247	7,383,783
50 to 59 years	304,848	316,903	621,751
Total	7,602,420	7,729,710	15,332,130

Source: 2016 *de jure* population projections produced by the Zambia Central Statistical Office (CSO). http://www.zamstats.gov.zm/report/Census/2010/National/Zambia%20Census%20Projection%202011%20-%202035.pdf.

2.2 Precision Specifications and Assumptions

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

2.2.1 Specifications

- The relative standard error (RSE) of the national estimate of HIV incidence among persons aged 15-49 should be 30% or less.
- 95% confidence bounds around the estimated viral load suppression (VLS) rate among HIV+ persons aged 15-49 for each of the 10 strata (provinces) should be ± 10% or less.

2.2.2 Assumptions

- An overall HIV prevalence rate of 0.103 (10.3%) that varies by province (e.g., see Table 2-2). Source: 2013-14 Zambia Demographic and Household Survey (Preliminary Report).
- An annual HIV incidence rate for adults aged 15-49 of $P_a = 0.0060$ (0.60%). Source: UNAIDS estimate for 2012 and 2013-14 Zambia Demographic and Household Survey (Preliminary Report).
- A mean duration of recent infection (MDRI) of 130 days, yielding an annualization rate of 365/130= 2.8077. Hence, the estimated incidence rate for MDRI = 130 days is $P_m = 0.0060/2.8077 = 0.0021$ (0.21%).
- A viral load suppression (VLS) rate among HIV+ adults aged 15-49 in each province *h* of $P_{vh} = 50\%$. This is a conservative assumption because it will overstate the actual variance of the VLS rate.
- An intra-cluster correlation (ICC) of $\rho = 0.05$ for both prevalence and incidence. The ICC provides an average measure of the homogeneity of responses within the first-stage sampling units.
- An occupancy rate of 91.9% for sampled dwelling units. Note that this is not included in the calculation of the overall survey response rate, but does determine the initial numbers of dwelling units to be sampled. A sample of 16,806 dwelling units will yield a sample of about 15,000 occupied dwelling units (households). Source: 2013-14 Zambia Demographic and Household Survey (DHS).
- An overall household response rate of 97.9% among the occupied dwelling units. Source: 2013-14 Zambia Demographic and Household Survey (DHS).
- The average number of persons aged 15 to 49 in a household is 1.89. Source: 2013-14 Zambia Demographic and Household Survey (DHS).
- The percentage of persons in households who are 0-14 is 49.6%. Source: 2013-14 Zambia Demographic and Household Survey (DHS).
- The percentage of persons in households who are 50-59 is 4.0%. Source: 2013-14 Zambia Demographic and Household Survey (DHS).
- Among the individuals 15-59 years of age in eligible responding households, a biomarker response rate of 68.9%. This corresponds to an overall biomarker response rate of 67.5%. This is a conservative estimate derived from response rates in the 2013-14 Zambia Demographic and Household Survey (DHS).

- Among the children 0-14 years of age in eligible responding households, a biomarker response rate of 62.1%.

Based on the assumptions listed above, a sample of 515 clusters (EAs) was determined to be the minimum needed to meet the precision requirements specified above. The distribution of the 515 sample clusters across the 10 provinces is shown in Table 2-2. Although the sample design originally called for the selection of an average of 30 households per cluster, this was later reduced to 24 households per cluster to ensure that the data collection activities could be completed as scheduled. As a result of the cutbacks in sample size, which were made uniformly across all EAs, the standard errors of the survey-based estimates were expected to increase by 4-10%. The expected numbers of households included in the study and the corresponding projected numbers of respondents by age group are summarized in Table 2-2. The actual numbers of respondents 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 the corresponding values achieved during data collection. Details about the additional subsampling of households are given in Section 2.4.

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

Stratum (Province)	Est. HIV prevalence rate [1]	Sample clusters (EAs)	Target no. dwelling units to be selected [2]	Dwelling units (DUs) selected [3]	Exp. no. house-holds [4]	Projected number of respondents [5]		
						15-49	50-59	0-14 [6]
Central	0.092	42	1,371	1,097	1,008	1,430	129	717
Copperbelt	0.128	74	2,415	1,932	1,776	2,519	228	1,262
Eastern	0.082	49	1,599	1,277	1,174	1,668	151	836
Luapula	0.084	32	1,044	835	767	1,089	98	546
Lusaka	0.136	86	2,806	2,241	2,059	2,928	265	1,467
Muchinga	0.047	50	1,632	1,308	1,202	1,702	154	853
Northern	0.054	45	1,468	1,174	1,079	1,532	138	768
North-Western	0.047	50	1,632	1,305	1,199	1,702	154	853
Southern	0.11	55	1,795	1,436	1,320	1,872	169	938
Western	0.125	32	1,044	836	768	1,089	98	546
TOTAL	0.103	515	16,806	13,441	12,352	17,533	1,585	8,786

[1] Source: Preliminary results from the 2013-14 Zambia Demographic and Health Survey (DHS).
[2] These are the original targets specified under the design. After the sample had been selected, the sample sizes were reduced as described in Section 2.4..6.
[3] These are the final sample sizes after sample reduction described in Section 2.4.6.
[4] Assumes occupancy rate of 0.919.
[5] Entries are projected counts based on the assumptions used to develop the sample design and reflect the additional subsampling of households within selected EAs.
[6] All responding children in 50% of the participating households.

2.3 Selection of the Primary Sampling Units (PSUs)

2.3.1 Definition of PSUs

The first-stage or primary sampling units (PSUs) for the ZAMPHIA were defined to be the Enumeration Areas (EAs) created for the 2010 Census of Population and Housing. The sampling frame consisted of 25,631 EAs containing 2.8 million households and 13.1 million persons as of the 2010 census. The EAs vary widely in size, with 97 EAs containing less than 30 households, and 149 containing more than 300 households. An attempt was made to combine the small EAs with an adjacent large EA for sampling purposes, but this was found to be impracticable, and a decision was made to exclude these EAs from the sampling frame. Thus, the final sampling frame contained 25,534 EAs. The total deletions accounted for approximately 0.08 percent of the 2010 population.

2.3.2 Selection of the PSU Sample

A stratified sample of 515 EAs was selected from the final EA sampling frame in accordance with the sample allocation given in Table 2-2. The strata specified for sampling were the 10 provinces of the country. The EA samples were selected systematically and with probabilities proportionate to a measure of size (MOS) from each stratum. The MOS used for sampling was equal to the number of households in the EA based on the 2010 Census of Population and Housing.

The first step of the sampling process was to divide the sampling frame of EAs into strata corresponding to the 10 provinces. Next, within each stratum, the EAs were sorted by the district code, urban/rural status within district, and then randomly within urban/rural status. The sorting of the EAs prior to sample selection induces an implicit stratification of the sampling frame designed to ensure that a representative mix of EAs with respect to geography and urban/rural status are included in the sample. To select the sample from a particular stratum, the cumulative MOS was determined for each EA in the ordered list of EAs, and the sample selections were designated using a sampling interval equal to the total MOS of the EAs in the stratum divided by the number of EAs to be selected and a random starting point. The resulting sample has the property that the probability of selecting an EA within a particular stratum is proportional to the MOS of the EA in the stratum.

2.3.3 Substitution

Three of the originally-sampled enumeration areas were replaced during listing. All three of these EAs are considered to be eligible for PHIA because they were known to contain occupied dwelling units, but were inaccessible for various reasons (e.g., due to flooding or surrounded by military barracks). The substitute EAs were identified by locating the position of the originally-sample EA in the ordered sampling frame, and then selecting the EA immediately preceding it on the list within the same substratum defined by the sorting variables used in sample selection. If there were no EAs preceding the original EA, the EA immediately following it was chosen. In this way, the substitute EA will have characteristics broadly similar to the originally-sampled EA. For subsequent sampling and weighting purposes, the probability of selecting the substitute EA was adjusted so that it reflected the probability of selection it would have had if it had originally been selected.

Data collection was not conducted in four of the sampled EAs. Two of these contained no households (and are therefore out of scope or "ineligible"). In general, substitution is not appropriate for out-of-scope EAs. The remaining two EAs were determined to have fewer than 30 households, and although they are inscope of the study, they were not released for fieldwork because of their small size. Thus, 511 eligible EAs are included in the final sample.

2.3.4 Segmentation

Of the 511 eligible EAs (including the three substitute EAs), 150 were considered to be too large to be listed in their entirety. Thus, these 150 EAs underwent another stage of sampling in which (a) the EA was subdivided into a specified number of segments of manageable size, (b) a rough measure of size was assigned to each defined segment, and (c) one segment was randomly selected with probability proportionate to the rough measure of size for listing. The segmentation procedures used in PHIA are described in **Zambia PHIA: Household Listing Manual**, July 2015, Central Statistical Office (CSO). Table 2-3 provides a summary of the 515 selected PSUs for each of the 10 provinces of Zambia, and the corresponding number of EAs that were replaced, segmented, out-of-scope, or nonresponding.

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

Stratum (Province)	Number of sample EAs	Number of replaced EAs	Number of out-of-scope PSUs [1]	Number of eligible non-responding EAs	Number of segmented EAs	Number of inscope EAs/segments included in study
Central	42	0	0	0	13	42
Copperbelt	74	0	0	0	11	74
Eastern	49	0	0	0	8	49
Luapula	32	0	0	0	11	32
Lusaka	86	2	1	0	24	85
Muchinga	50	0	0	0	18	50
Northern	45	0	0	2	14	43
North-Western	50	0	1	0	25	49
Southern	55	1	0	0	16	55
Western	32	0	0	0	10	32
Total	515	3	2	2	150	511

[1] EAs with no occupied (residential) dwelling units. Such EAs are out-of-scope (ineligible) for the survey.

2.4 Selection of Households

The selection of households for the ZAMPHIA involved the following steps: (1) listing the dwelling units/households within the sampled EAs, (2) assigning eligibility codes to the listed dwelling unit/household records, (3) selecting the samples of dwelling units/households, and (4) designating a subsample of households for child data collection.

2.4.1 Definition of Second-Stage Sampling Units

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

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. For each of the 515 EAs selected for the study, listers were provided with maps from which to delineate the boundaries of the EA, and to record the general locations of the dwelling units/households found by the listers in the field. Information about the listed dwelling units/households was entered into computer tablets. The information recorded in the tablets included the address or description of the listed dwelling unit/household, the name of the head of household, the type of structure (house, apartment, compound, etc.), occupancy status, and GPS coordinates. Vacant structures were listed along with households in occupied dwelling units. Over 60,000 dwelling units/households were listed for the ZAMPHIA. Additional details about the listing process are given in **Zambia PHIA: Household Listing Manual**, July 2015, Central Statistical Office (CSO).

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

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

Stratum (Province)	Number of dwelling units/ households in listing file	Number of discarded listings	Number of un-occupied dwelling units [1]	Number of un-occupied dwelling units eligible for sampling [2]	Number of occupied dwelling units/house-holds [3]	Number of occupied dwelling units/house-holds eligible for sampling	Total number of dwelling units/house-holds eligible for sampling
Central	4,182	0	513	436	3,669	3,669	4,105
Copperbelt	10,423	205	1,215	766	9,003	9,003	9,769
Eastern	5,126	5	659	447	4,462	4,462	4,909
Luapula	3,196	0	494	404	2,702	2,702	3,106
Lusaka	12,204	181	894	530	11,129	11,018	11,548
Muchinga	5,128	0	781	516	4,347	4,347	4,863
Northern	4,986	16	956	670	4,014	3,920	4,590
North-Western	5,421	0	852	598	4,569	4,514	5,112
Southern	6,420	0	690	401	5,730	5,730	6,131
Western	3,495	0	581	497	2,914	2,913	3,410
TOTAL	60,581	407	7,635	5,265	52,539	52,278	57,543

[1] Records coded as vacant, under construction, or with no residents at time of listing (see Appendix A).
[2] Subset of the unoccupied dwelling units that could potentially serve as residential quarters (see Appendix A).
[3] All records not coded as vacant, under construction, or with no residents at time of listing (see Appendix A).

2.4.4 Selection of Dwelling Units

In order to achieve an equal probability sample of dwelling units within a stratum, the sampling rates required to select dwelling units within an EA will depend on the difference between the size measure used in sampling (i.e., the number of households in the EA based on the 2010 census) and the actual number of dwelling units/households found at the time of listing. Thus, application of these within-EA sampling rates can yield more than the targeted number of households per EA in those EAs that have experienced growth in population since the 2010 census, and fewer than the targeted number in EAs that have declined in population.

The calculation of the required within-EA sampling rates proceeded as follows. First, the target overall sampling rate for stratum (province) $h = 1, 2, ..., 10$, was computed as:

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

where

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

m_h = number of inscope sample EAs in stratum *h* given in Table 2-2;

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

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

Note that the T_hs in the above formula for $F_h^{overall}$ refer to the target sample sizes under the original planned sample design. The total number of listings to be selected across all 10 strata as originally planned was $\sum_{h=1}^{10} T_h$ = 16,181 (see Table 2-2). The probabilities of selection, P_{hi}, for the three substitute EAs (see Section 2.3.3) were set to the probabilities they would have had if they had originally been selected for the sample. The probability of selection of a segmented EA was set to $P_{hs} = P_{hi}^{EA} P_{s|hi}^{seg}$, where P_{hi}^{EA} = the selection probability of EA *hi*, and $P_{s|hi}^{seg}$ = the conditional probability of selecting segment *s* in EA *hi*.

To obtain an equal probability sample within stratum *h*, the required within-EA sampling rate for EA *i* in stratum *h* was then computed as:

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

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

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

Inspection of the values of $\mathrm{E}(n_{hi})$ indicated that there would be unduly large workloads in some EAs. To maintain acceptable workloads in EAs that had experienced considerable growth, the maximum number of dwelling units to be selected in any EA was capped at 60, and the minimum number to be selected was set to 15. The difference between the number of dwelling units that would have been selected and the capped number was then re-distributed to the other EAs in the same stratum so as to maintain the desired total sample size. The within-EA sampling rates, f_{hi}^{within}, were adjusted to reflect the capping and the redistribution of the sample within the stratum. The adjusted within-EA sampling rate used to select the sample of dwelling units, $f_{hi}^{adj(w)}$, was calculated as:

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

where the adjustment factors, A_{hi}, were determined such that $15 \leq A_{hi}\, f_{hi}^{within} \leq 60$ and $\sum_{i=1}^{m_h} A_{hi}\, f_{hi}^{within} = T_h$. To preserve the geographical order in which they were listed, the eligible dwelling unit/household records in each EA were sorted by lister, segment, structure number, apartment number if applicable, and finally by household number assigned at the time of listing. 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. In addition, a random half sample of the selected dwelling units/households was designated (flagged) for child data collection.

2.4.5 Results of Second-Stage Sampling

Table 2-5 summarizes the number of dwelling units/households selected for the study, the number designated for child data collection, and the minimum and maximum EA sample size by stratum. The last column shows the unequal weighting (UEW) design effects to be expected for the selected sample. With an equal probability sample within each stratum, all of the design effects would ordinarily equal 1.0. Capping the samples at 60 per EA resulted in extra variation in weights within a stratum and hence increased design effects. Even with the capping, the stratum-level design effects are all less than 1.05 (indicating minimal increase in variance due to unequal weighting) for most strata.

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

Stratum (Province)	No. inscope sample PSUs (clusters)	Number of sampled dwelling units/house-holds	Number of dwelling units/households flagged for child data collection	Minimum PSU sample size	Maximum PSU sample size	UEW DEFF for PHIA sample after capping
Central	42	1,371	686	19	60	1.01
Copperbelt	74	2,415	1,207	15	60	1.02
Eastern	49	1,599	800	15	60	1.03
Luapula	32	1,044	522	15	60	1.01
Lusaka	85	2,800	1400	15	60	1.04
Muchinga	50	1,632	816	15	61	1.04
Northern	43	1,468	734	16	60	1.01
North-Western	49	1,632	816	15	60	1.01
Southern	55	1,795	897	15	60	1.00
Western	32	1,044	522	17	60	1.00
Total	511	16,800	8,400	15	61	1.07[1]

[1] Reflects variation in weights across and within EAs.

2.4.6 Reduction of the Dwelling Unit Sample

After the sample had been selected as described above, it was necessary to reduce the sample sizes by 20 percent to meet the very tight time schedule established for data collection. The reductions were made uniformly across all sample EAs by retaining a random subsample of 80 percent of the originally-selected dwelling units/households. In geographically large rural areas where households can be far apart, the subsamples were to be further clustered to reduce the amount of travel needed for data collection. Such clustering was done in those rural EAs where the listed households covered an area greater than 8 km^2 (about 3.1 square miles). In the remaining EAs, the subsamples were selected systematically from the lists of the originally-sampled dwelling units/households without additional clustering.

Among the 511 inscope sample EAs, 120 rural EAs met the criterion for clustering. However, four were so small in terms of the expected number of households to be sampled that they would have resulted in the creation of a single cluster, and hence were treated as a non-compact cluster EA for subsampling purposes. Within each of the remaining 116 EAs, it was possible to create either 2 or 3 compact clusters from which exactly one cluster per EA was random selected with probability proportional to the number of listed dwellings (households) in the cluster. Within the selected compact clusters, the next step was to select an equal probability systematic sample of households at rates designed to achieve the desired 80 percent subsample. Among the selected households, every other household was designated for child data collection.

For the 395 PSUs that did not undergo compact clustering, an equal probability systematic sample of 80% of the originally-selected households was selected from each PSU. Prior to selection, the eligible records were sorted by stratum (province), EA, and then by the child flag that had been assigned to the original sample.

A total of 13,441 dwelling units/households (listings) was selected across the 10 strata for the 80-percent subsample. Of these, 10,534 were selected from the 395 EAs that did not require clustering, and 2,907 households were selected from the remaining 116 clustered EAs. In addition, a systematic random subsample of 6,735 households among the 13,441 sampled households were flagged for child data collection; including 5,282 from the unclustered EAs, and 1,453 from the clustered EAs. Table 2-6 summarizes the number of dwelling units/households in the final reduced sample, the number designated for child data collection, and the minimum and maximum EA sample size by

stratum. The last column of the table shows the UEW design effects to be expected for the selected sample.

Table 2-6 Number of sampled dwelling units/households in final reduced sample, and expected unequal weighting design effects by stratum

Stratum (Province)	No. inscope sample EAs (clusters)	Number of sampled dwelling units/house-holds	Number of dwelling units/house-holds flagged for child data collection	Minimum EA sample size	Maximum EA sample size	UEW DEFF for the reduced sample
Central	42	1,097	550	15	48	1.01
Copperbelt	74	1,932	971	12	48	1.02
Eastern	49	1,277	641	12	48	1.03
Luapula	32	835	419	12	48	1.01
Lusaka	85	2,241	1,124	12	48	1.04
Muchinga	50	1,308	657	12	48	1.04
Northern	43	1,174	586	12	48	1.01
North-Western	49	1,305	648	12	48	1.01
Southern	55	1,436	719	12	48	1.00
Western	32	836	420	14	48	1.00
Total	511	13,441	6,735	12	48	1.07

Table 2-7 summarizes the number of dwelling units selected for PHIA by final household response status. Of the 13,411 sampled dwelling units, 1,133 (8.4%) were determined during data collection to be vacant/unoccupied, 115 (0.9%) for which eligibility for the survey (i.e., occupancy status) could not be established, 1,236 (9.2%) were determined to be eligible for the study (i.e., contained eligible household members) but did not complete the household roster, and 10,957 (81.7%) completed the household roster. The overall unweighted household response rate was 89.1%.

Table 2-7 Distribution of dwelling unit sample by province and response status

Stratum (Province)	Number of sampled dwelling units (DUs)	Number of ineligible DUs [1]	Number of DUs with unknown eligibility [2]	Number of households completing roster	Number of eligible non-responding households	Unweighted response rate [3]
Central	1,097	107	9	918	63	0.928
Copperbelt	1,932	97	9	1,661	165	0.905
Eastern	1,277	87	5	1,098	87	0.923
Luapula	835	101	5	665	64	0.907
Lusaka	2,241	140	18	1,841	242	0.877
Muchinga	1,308	119	8	1,067	114	0.898
Northern	1,174	128	15	924	107	0.885
North-Western	1,305	132	31	989	153	0.845
Southern	1,436	124	8	1,160	144	0.885
Western	836	98	7	634	97	0.860
Total	13,441	1,133	115	10,957	1,236	0.891

[1] Vacant or unoccupied dwelling units, households with no persons eligible for PHIA.
[2] Dwelling units for which occupancy status could not be determined.
[3] Computed as R/ [R + N + U*{ (R + N)/(R + N + I) }], where R = number of households completing roster; N = number of eligible nonresponding households; I = number of ineligible DUs, and U = number of DUs with unknown eligibility.

2.5 Selection of Individuals

The selection of individuals for the ZAMPHIA 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; (3) selecting for the study those individuals meeting the age and residency requirements of the study. In general, all household members 15-59 years of age in the sampled households were eligible for PHIA data collection, whereas children 0-14 years of age in a randomly selected one-half of the sampled households were eligible for data collection. However, as noted below, only those individuals who were present in the household the night before the interview (i.e., the *de facto* population) were 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. The rosters included 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 away 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 2-8 summarizes the number of households completing the roster and the corresponding number of rostered individuals by stratum and residency status. Note that the counts in Table 2-8 include children in households that were not flagged for child data collection.

Table 2-8 Number of households completing rosters and number of persons by residency status

Stratum (Province)	Number of households completing rosters	Usual resident but did not sleep here	Usual resident and slept here	Nonresident but slept here	Total
Central	918	156	4,290	80	4,526
Copperbelt	1,661	256	7,624	231	8,111
Eastern	1,098	146	4,948	34	5,128
Luapula	665	134	2,745	52	2,931
Lusaka	1,841	261	8,062	195	8,518
Muchinga	1,067	180	4,924	43	5,147
Northern	924	208	4,284	66	4,558
North-Western	989	328	4,861	78	5,267
Southern	1,160	338	5,439	130	5,907
Western	634	197	2,734	39	2,970
Total	10,957	2,204	49,911	948	53,063

2.5.2 Selecting Individuals for Data Collection

All of the individuals listed in the household rosters who were 15-59 years of age and were either usual residents of the household or who slept in the household were eligible for data collection. Basic information about all children was obtained from parents or guardians in the child module of the adult questionnaire, but children 0-14 years of age were eligible for additional data collection only if the household in which they resided had been randomly designated for child biomarker data collection (see Section 2.4.5). Table 2-9 summarizes the number of individuals eligible for data collection by stratum, age group, and residency status.

Although data collection was attempted for all of the 26,401 adults and 11,989 children indicated in Table 2-9, only those individuals in the *de facto* population will be weighted (see Section 3) and included in analysis. The *de facto* population is represented by the 24,945 adults and 11,685 children who slept in the household during the night prior to the interview.

Table 2-9 Number of individuals eligible for data collection

Stratum (Province)	Adults 15 to 59 [1]				Children 0-14 in households selected for child biomarker collection [1]			
	Usual resident but did not sleep here	Usual resident and slept here	Non-resident but slept here	Total	Usual resident but did not sleep here	Usual resident and slept here	Non-resident but slept here	Total
Central	100	1,985	51	2,136	15	1,005	15	1,035
Copperbelt	158	4,096	129	4,383	39	1,547	57	1,643
Eastern	95	2,302	20	2,417	19	1,155	5	1,179
Luapula	84	1,265	21	1,370	10	692	13	715
Lusaka	177	4,593	113	4,883	32	1,616	42	1,690
Muchinga	109	2,162	22	2,293	30	1,232	9	1,271
Northern	165	1,924	37	2,126	19	1,067	12	1,098
North-Western	231	2,328	49	2,608	47	1,199	9	1,255
Southern	213	2,567	61	2,850	54	1,282	37	1,373
Western	124	1,190	21	1,335	39	684	7	730
Total	1,456	24,421	524	26,401	304	11,479	206	11,989

[1] Age recorded in roster. In a small number of cases, the actual age at interview may be different. See Section 3.4.3.

2.5.3 Distribution of Person Samples

Tables 2-10A through 2-10C summarize the number of individuals selected for data collection and the corresponding numbers completing the interview and blood test, for adults 15-59 years old, adolescents 10-14 years, and children 0-9 years, respectively, where the age classification is based on the rostered age. The numbers of completed interviews and blood tests that can be weighted to represent the PHIA study population are shown under the *de facto* heading in these tables. Note that counts of children in these tables include only children in households selected for child blood draw, and that for children 0-9 years in Table 2-10C the counts of completed "interviews" refer to the number of children for whom a parent or guardian completed the child questionnaire module for that particular child.

Table 2-10A Distribution of completed interviews and blood tests for adults 15 to 59 years

Stratum (Province)	*De facto* [1]			*De jure but not de facto* [2]		
	Number selected for data collection	Number completing interview [3]	Number completing blood test [4]	Number selected for data collection	Number completing interview [3]	Number completing blood test [4]
Central	2,036	1,789	1,591	100	39	36
Copperbelt	4,225	3,643	3,298	158	86	76
Eastern	2,322	2,032	1,892	95	35	34
Luapula	1,286	1,145	1,022	84	28	27
Lusaka	4,706	3,649	3,247	177	61	55
Muchinga	2,184	1,932	1,742	109	40	37
Northern	1,961	1,735	1,481	165	61	54
North-Western	2,377	1,990	1,785	231	78	65
Southern	2,637	2,380	2,183	213	65	60
Western	1,211	1,046	931	124	37	32
Total	24,945	21,341	19,172	1,456	530	476

[1] Persons who were reported to have slept in the household last night.
[2] Usual residents of the household who did not sleep in the household last night.
[3] Persons who completed the blood test but not the interview are treated as interview respondents for weighting purposes. See Appendix B for programming details.
[4] These are cases that provided an analyzable blood sample, regardless of whether the individual interview was completed. Of the 19,172 de facto cases completing the blood test, 17 did not complete the interview but are treated as interview respondents for weighting purposes. See Appendix B for programming details.

Table 2-10B Distribution of completed interviews and blood tests for adolescents 10-14 years in households selected for child biomarker collection

Stratum (Province)	*De facto* [1]			*De jure but not de facto* [2]		
	Number selected for data collection	Number completing interview [3]	Number completing blood test [4]	Number selected for data collection	Number completing interview [3]	Number completing blood test [4]
Central	321	260	238	2	1	1
Copperbelt	522	425	376	7	4	4
Eastern	396	292	273	9	5	4
Luapula	206	157	146	3	1	1
Lusaka	507	379	354	10	3	1
Muchinga	369	277	257	12	4	4
Northern	322	223	183	15	3	3
North-Western	370	275	253	17	5	3
Southern	385	322	303	9	3	2
Western	185	133	118	20	8	8
Total	3,583	2,743	2,501	104	37	31

[1] Persons who were reported to have slept in the household last night.
[2] Usual residents of the household who did not sleep in the household last night.
[3] Persons who completed the blood test but not the interview are treated as interview respondents for weighting purposes. See Appendix B for programming details.
[4] These are cases that provided an analyzable blood sample, regardless of whether the individual interview was completed. Of the 2,501 de facto cases completing the blood test, all completed the interview. See Appendix B for programming details.

Table 2-10C Distribution of completed interviews and blood tests for children 0-9 years in households selected for child biomarker collection

Stratum (Province)	*De facto* [1]			*De jure but not de facto* [2]		
	Number selected for data collection	Number completing interview	Number completing blood test	Number selected for data collection	Number completing interview	Number completing blood test
Central	699	641	487	13	9	1
Copperbelt	1,082	1,003	793	32	28	18
Eastern	764	696	504	10	9	4
Luapula	499	462	338	7	6	5
Lusaka	1,151	1,020	727	22	14	2
Muchinga	872	834	636	18	11	3
Northern	757	666	398	10	9	3
North-Western	838	749	543	32	29	15
Southern	934	864	712	37	28	10
Western	506	469	319	19	11	3
Total	8,102	7,404	5,457	200	154	64

[1] Persons who were reported to have slept in the household last night.
[2] Usual residents of the household who did not sleep in the household last night.
[3] Persons who completed the blood test but not the interview are treated as interview respondents for weighting purposes. See Appendix B for programming details.
[4] These are cases that provided an analyzable blood sample, regardless of whether the individual interview was completed. Of the 5,457 de facto cases completing the blood test, 267 did not complete the interview but are treated as interview respondents for weighting purposes. See Appendix B for programming details.

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 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 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, page 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 PHIA, 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 3.4.3.4 and 3.4.4.4 are used to calibrate the weighted sample counts to available population projections.

3.1 Overview of the Weighting Process

The overall weighting approach for ZAMPHIA includes several steps.

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

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

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

Calculation of Household Weights: The next step is to calculate household weights. The household base weights are calculated as the EA base 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 (as shown in Table 2-7 above, this only happened for 0.9 % of the listings) and then the responding households have their weights adjusted to account for nonresponding eligible households. This adjustment is made based on the EA the households are in, and the resulting weight is the final household weight.

Calculation of Person-Level Interview Weights: Once the household weights are determined, they are used to calculate the individual base weights. The individual base weights are then adjusted for nonresponse among the eligible individuals, with a final adjustment for the individual weights to compensate for undercoverage in the sampling process by weighting up to 2016 population projections produced by the Zambia Central Statistical Office (CSO). For children in households not selected for child blood draws (see Section 2.4.5), data was collected from eligible parents or guardians, but the children were not assigned interview weights. For analysis of this full set of children, child module weights were generated after all other weighting was completed. See Appendix G for details.

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 any undercoverage in the sample.

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

Five 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 PHIA survey data that are used to construct the sampling weights are contained in the following data files. These are work files created and used during the weighting process and are not included in the public-use data.

- **Phiazam_ffcorr_hhqx_20170116**: A household (HH) file that contains the majority of household data collected in the HH questionnaire.
- **Phiazam_ffcorr_hhDeath_20170116**: A household (HH) file that contains data collected in the HH questionnaire regarding any deaths that have occurred in the household since 2013.
- **Phiazam_ffcorr_Roster_20170116**: A file that contains the roster of household members collected in the HH questionnaire with a record for each rostered person.
- **Phiazam_ffcor_individ_20170116**: 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. So variables for individual questionnaire data collected from persons aged 15 to 59, for individual questionnaire data collected from persons aged 10 to 14, for children under 10 for data collected from the child's parent or guardian are all included in every record, with values only for the applicable variables.

- **ZamBiomarker20170330**: 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.

For weighting purposes, each of these files except the biomarker file contains records for all sampled cases, irrespective of response and eligibility status.

3.2.2 Checks of Data Files

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

- Checking IDs, merging household survey files with sampling files, and accounting for records found in one file and not the other. (This type of check for the EAs occurs as part of the HH selection process.)
- Check counts of sampled and responding HHs against what was expected, overall and by province.
- Acknowledge/adjust for substitution, missed HH procedures, 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) all sampled households, (b) all sampled individuals, and (b) all sampled individuals for blood draws.
- Verify that the survey data, for all three components, have passed data cleaning.

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 PHIA: 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 PHIA 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 variables that indicate the "variance stratum" and the "variance unit" to which each sampled respondent belongs.

3.3.1 Jackknife Replication

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

As described in Section 2.3, the PSUs were selected systematically from a list of PSUs that had been ordered by province, district code within province, urban/rural status within district, and then randomly within each urban/rural status. To take account of the precision benefits of implicit stratification as fully as possible, the sampled PSUs within each province were paired off in the systematic order in which they were selected, treating each pair as a variance-estimation stratum. When there was an odd number of sampled PSUs in a province, one of the variance-estimation strata was defined to contain three sampled PSUs.

For the ZAMPHIA, a total of 253 variance-estimation strata were formed. A jackknife replicate was then formed by randomly deleting a PSU from a particular variance-estimation stratum k, say, and retaining all of the PSUs in the remaining variance-estimation strata. For a variance-estimation stratum consisting of a pair of PSUs, the weight of the retained PSU within the variance-estimation stratum k was doubled. For a variance-estimation stratum consisting of three PSUs, the weight of the two retained PSUs within the variance-estimation stratum were increased by 1.5 (see Section 3.4.1). This process was repeated for all $r = 1, 2, ..., 253$ variance-estimation strata, resulting in a total of 253 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

Sampling Stratum (Province)	No. PSUs	No. variance strata consisting of pairs	No. variance strata consisting of triplets	Number of jackknife replicates
Central	42	21	0	21
Copperbelt	74	37	0	37
Eastern	49	23	1	24
Luapula	32	16	0	16
Lusaka	85	41	1	42
Muchinga	50	25	0	25
Northern	43	20	1	21
North-Western	49	23	1	24
Southern	55	26	1	27
Western	32	16	0	16
Total	511	248	5	253

3.3.2 Taylor's Series

Even though jackknife replication is the recommended method for variance estimation, not all software packages have a replication option to produce variance estimates. For example, SPSS has built-in options for estimating variance using Taylor's Series methods, but the end user has to write a program within SPSS to produce replicate estimates of variance. Therefore, information for producing Taylor's Series estimates of variance is included in the PHIA 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 253 variance-estimation strata and the variable VarUnit indicates the primary sampling unit (PSU) or cluster within the variance-estimation stratum. This pair of variables allows the analyst to produce variance estimates if their software does not easily accommodate replication methods, but does have a Taylor's Series capability. Note that the variance-estimation strata and the sampling strata are not equivalent: as shown in Table 3-1, the sampling strata are defined by the province and urban/rural areas, while the variance-estimation strata are based on groupings of PSUs within each sampling stratum.

3.4 Development of Weights

3.4.1 PSU Weights

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

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

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

where P_{hi}^{PSU} = probability of selecting PSU *i* from province *h*. Note that if the PSU was segmented, then P_{hi}^{PSU} is the product of the probability of selecting the EA and the conditional probability of selecting the segment within the EA (e.g., see Section 2.4.4). If the PSU was a replacement PSU, then P_{hi}^{PSU} is the probability that the substitute PSU would have had if it had originally been selected for the sample.

As indicated in Table 3-1, 253 jackknife replicates were formed from the 511 PSUs. For variance estimation, replicate-specific PSU weights, $W_{(r)hi}^{(1)}$, r = 1, 2, ..., 253 were created to provide the basis for calculating the required replicate weights in subsequent stages of the weighting process. Let *h* denote one of the 253 variance-estimation strata created for jackknife replication (Section 3.3.1) and let *i* denote the PSU within variance-estimation stratum *h*. For a given jackknife replicate, r = 1, 2, ..., 253, the corresponding replicate-specific PSU base weight was computed as

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

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

3.4.2 Household Weights

3.4.2.1 Household Base Weights

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

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

where

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

$P_{j|hi}^{HH}$ = the conditional probability of selecting household *j* in PSU *i* in province *h*

The corresponding weights for jackknife replicate r = 1, 2, …, 253, were computed as:

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

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

Next, the sampled dwelling units/households were assigned to one of the four response status groups specified in Table 3-2. In Table 3-3, we show the corresponding weighted sums by response status and province using the household base weights calculated as just described. The characteristics of the household base weight 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, minimum and maximum values of the weights, both overall and by province.

Table 3-2 Response-status codes specified for household weighting

Household response status code [1]	Description	Number of dwelling units/households
1	Eligible respondent	10,957
2	Eligible nonrespondent	1,236
3	Ineligible/out-of-scope	1,133
4	Unknown eligibility status	115

[1] See Appendix B for programming details.

Table 3-3 Weighted sums of household base weights by response status

Stratum (Province)	Household Response Status				
	Status code 1: Eligible Respondents	Status code 2: Eligible Nonres-pondents	Status code 3: Not Eligible (Vacant, Destroyed, not a DU, etc.)	Status code 4: Could not determine eligibility	Weighted Count of Households [1]
Central	280,721	19,622	35,802	2,729	338,874
Copperbelt	416,528	40,917	23,607	2,196	483,247
Eastern	351,161	27,232	27,935	2,014	408,342
Luapula	220,880	21,932	34,511	1,840	279,163
Lusaka	493,862	65,661	37,881	5,262	602,667
Muchinga	175,342	18,716	20,463	1,540	216,061
Northern	248,289	28,760	35,729	3,934	316,712
North-Western	134,159	20,542	17,598	4,346	176,644
Southern	309,354	38,608	32,992	2,124	383,078
Western	177,023	26,903	27,526	1,934	233,386
Total	2,807,321	308,892	294,042	27,918	3,438,174

[1] Weights are the household base weights, $W_{hij}^{(2)}$ specified in Section 3.4.2.1.

3.4.2.2 Adjustment for Household Nonresponse

The general approach for handling household nonresponse was to increase the weights of responding households so that they represent the nonresponding households in the same 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 household contains persons eligible for PHIA), the household nonresponse adjustment was implemented in two phases. In the first phase of adjustment, the weights were adjusted to compensate for sampled dwelling units for which eligibility for the survey (e.g., occupancy status) was not ascertained. In the second phase of adjustment, the first-phase adjusted weights were further adjusted to compensate for the nonresponding households among those households known to be eligible for the study.

To account for variation in response rates across different types of PSUs, it is desirable to make the household nonresponse adjustments within weighting cells defined by the individual PSUs. However, if a PSU has a very low household response rate, such PSU-level adjustments can result in very large adjusted weights that would lead to increases in the variances of the survey estimates. To avoid this problem, such PSUs can be collapsed with a similar PSU to form a single non-response adjustment cell comprised of two or more PSUs. For the ZAMPHIA, a total of nine PSUs were found to have response rates at or below 50% which translates to an adjustment factor at or above

2.00. To dampen the effect of the adjustment for these PSUs, each was paired with the nearest PSU on the sorted list of sampled PSUs to form the final weighting cell for nonresponse adjustment. Without such collapsing, the adjustment factors would have ranged from 1.00 (for PSUs with 100% response rate) to 3.83 (for a PSU with a response rate of 26.1 %). After the grouping the highest adjustment factor was reduced to 1.875.

The procedures used to compute the nonresponse-adjusted household weights are described below.

Phase 1 Adjustment

As indicated above, the weighting cells for the household nonresponse adjustments are generally individual PSUs or a group of PSUs. We refer to these as "PSU weighting cells."

Let n_{hi}^{samp} denote the number of sampled dwelling units in PSU weighting cell *i* in province *h*. Note that n_{hi}^{samp} is the sum of the sample sizes in each of the four response status groups defined in Table 3-2, i.e.,

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

where

$n_{hi}^{(1)}$ = the number of responding households (i.e., households completing the roster) in PSU weighting cell *i* in province *h*

$n_{hi}^{(2)}$ = the number of eligible nonresponding households (i.e., households known to contain eligible persons but did not complete the roster) in PSU weighting cell *i* in province *h*

$n_{hi}^{(3)}$ = the number of known ineligible dwelling units (i.e., sampled dwelling units known to contain no persons eligible for the study) in PSU weighting cell *i* in province *h*

$n_{hi}^{(4)}$ = the number of sampled dwelling units for which eligibility for the study could not be ascertained in PSU weighting cell *i* in province *h*

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

$$A_{hi}^{(HH1)} = \sum_{j=1}^{n_{hi}^{samp}} W_{hij}^{(2)} \; / \; \sum_{j=1}^{n_{hi}^{(1)}+n_{hi}^{(2)}+n_{hi}^{(3)}} W_{hij}^{(2)}$$

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

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

$$W_{hij}^{HH1} = A_{hi}^{(HH1)} W_{hij}^{(2)}$$

The corresponding replicate weights for replicate $r = 1, 2, \ldots, 253$ were computed in similar fashion as:

$$W_{(r)hij}^{HH1} = A_{(r)hi}^{(HH1)} W_{(r)hij}^{(2)},$$

where

$$A_{(r)hi}^{(HH1)} = \sum_{j=1}^{n_{(r)hi}^{samp}} W_{(r)hij}^{(2)} \; / \; \sum_{j=1}^{n_{(r)hi}^{(1)}+n_{(r)hi}^{(2)}+n_{(r)hi}^{(3)}} W_{(r)hij}^{(2)} .$$

Note that for the sampled dwelling units/households in response-status group 4, $W_{hij}^{HH1} = W_{(r)hij}^{HH1} = 0$ for $r = 1, 2, \ldots, 253$.

The effect of this adjustment is to distribute the total weight of the undetermined-eligibility cases (i.e., the estimated 27,918 dwelling units shown in the next-to-last column of Table 3-3) to the combined weight of the remaining three groups of sampled dwelling units/households. The resulting weighted counts using W_{hij}^{HH1} as computed above are given in Table 3-4.

Table 3-4 Weighted sums of household weights adjusted for unknown eligibility

Stratum (Province)	Household Response Status				
	Status code 1: Eligible responding households	Status code 2: Eligible nonresponding households	Status code 3: Ineligible dwellings	Total dwelling units/households	Total eligible households
Central	282,871	19,896	36,108	338,874	302,767
Copperbelt	418,256	41,005	23,986	483,247	459,261
Eastern	352,883	27,323	28,135	408,342	380,206
Luapula	221,932	22,271	34,960	279,163	244,203
Lusaka	497,840	66,387	38,440	602,667	564,227
Muchinga	176,499	18,902	20,660	216,061	195,401
Northern	251,088	29,474	36,150	316,712	280,562
North-Western	137,232	21,086	18,326	176,644	158,318
SOUTHERN	311,018	38,916	33,145	383,078	349,934
WESTERN	178,382	27,275	27,730	233,386	205,657
Total	2,828,001	312,534	297,639	3,438,174	3,140,535

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

Phase 2 Adjustment

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

$$A_{hi}^{(HH2)} = \sum_{j=1}^{n_{hi}^{(1)}+n_{hi}^{(2)}} W_{hij}^{HH1} \Big/ \sum_{j=1}^{n_{hi}^{(1)}} W_{hij}^{HH1}$$

where W_{hij}^{HH1} is the first-phase adjusted weight for dwelling unit/household *j* in PSU weighting cell *i* in province *h*, and where the sum in the numerator extends over the sample of responding and nonresponding households in PSU weighting cell *i* in province *h*, while the sum in the denominator extends over the responding households.

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

$$W_{hij}^{(2A)} = A_{hi}^{(HH2)} W_{hij}^{HH1}.$$

The corresponding replicate weights for replicate $r = 1, 2, \ldots, 253$ were computed in similar fashion as:

$$W_{(r)hij}^{(2A)} = A_{(r)hi}^{(HH2)} W_{(r)hij}^{HH1},$$

where

$$A_{(r)hi}^{(HH2)} = \sum_{j=1}^{n_{(r)hi}^{(1)}+n_{(r)hi}^{(2)}} W_{(r)hij}^{HH1} \; / \; \sum_{j=1}^{n_{(r)hi}^{(1)}} W_{(r)hij}^{HH1} .$$

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 3-4.

3.4.3 Person-Level Interview Weights

Below, we detail the calculation of person-level base weights and nonresponse-adjusted person-level weights for analyzing the ZAMPHIA data files. Specifically, we first define the initial person-level (interview) base weights for adults, adolescents, and children in Section 3.4.3.1. Interview nonresponse adjustment using the LASSO and CHAID algorithms for variable selection is addressed in Section 3.4.3.2.

The samples for PHIA are categorized into three age groups for which different data elements are collected: (1) adults aged 15 to 59, with data collected using the adult questionnaire; (2) adolescents, aged 10-14, with survey responses collected from the adolescent using an adolescent questionnaire; and (3) children aged 0-9, with survey responses provided by a parent or guardian in the children's module of the adult questionnaire. Furthermore, some different questions are asked within the various age groups depending on the sex of the individual. All of the persons in sampled households are enumerated and placed into one of the three age categories based on the data collected in the household roster. Although all rostered adults are asked to participate in the study, only those individuals who are considered part of the *de facto* population are included in the weighting process. Adolescents and children are included in the study if they belong to the one-half subsample of households designated for child data collection.

3.4.3.1 Person Base Weights

The sampled individuals were classified into three groups as indicated in Table 3-6 based on the age reported in the household roster. As discussed in Section 3.4.2.2, the starting point for developing the interview nonresponse adjustments is the final nonresponse-adjusted household weight, $W_{hij}^{(2A)}$. The sample person's base weight is the same as the nonresponse-adjusted household weight for adults (persons 15-59), but it is twice the nonresponse-adjusted household weight for eligible adolescents (10-14) and children (0-9) in households designated for child data collection. That is, the base weight for sample person *k* in household *j* in PSU *i* in province *h* was computed from the formula

$$W_{hijk}^{(3)} = K_k \, W_{hij}^{(2A)} ,$$

where $K_k = 1$ if the roster age of person *k* is 15 to 59 years, or $K_k = 2$ if the roster age of person *k* is 14 years or younger in households designated for child data collection.

The corresponding replicate base weights, $W_{(r)hijk}^{(3)}$, $r = 1, 2, \ldots, 253$, were computed in an analogous manner, with $W_{hij}^{(2A)}$ replaced by $W_{(r)hij}^{(2A)}$ in the above formula.

Table 3-5 summarizes the counts of eligible individuals by age group and interview response status, and the corresponding weighted counts using the person-level base weights, $W_{hijk}^{(3)}$. As indicated earlier in Section 2.5.3, the counts of eligible interview respondents shown in Table 3-6 includes a small number of persons who did not complete the interview but did provide an analyzable blood test.

Table 3-5 Distribution of eligible sample persons by age group and interview response status

Group	Age [1]	Interview Status [2]	Count	Weighted count [3]
Adults	15-59	Eligible Respondent	21,341	6,069,845
		Eligible Nonrespondent	3,402	989,113
Adolescents	10-14	Eligible Respondent	2,743	1,557,186
		Eligible Nonrespondent	802	467,140
Children	0-9	Eligible Respondent	7,404	4,209,097
		Eligible Nonrespondent	633	372,203

[1] Based on age reported in interview.
[2] Eligible respondents include cases that completed the individual interview or the blood test. See Appendix B for programming details.
[3] Weighted by the person-level base weight, $W_{hijk}^{(3)}$.

3.4.3.2 Adjustment of Person Weights for Interview Nonresponse

To compensate for interview nonresponse, the person base weights were adjusted within cells defined by variables available for both the responding and nonresponding individuals. These variables included data from the household roster and other information collected in the household questionnaire, and selected PSU characteristics such as province and urban/rural status. The age and sex variables used to make the nonresponse adjustments are those reported in the household roster and not the interview-reported age and sex, because the latter values are not known for the nonrespondents.

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

There are approximately 80 variables from the household questionnaire and EA sampling frame that could potentially be used for nonresponse adjustment. The LASSO procedure was used for initial variable selection to reduce the number of variables to a manageable subset of the most important and relevant predictors. The LASSO is a restrictive procedure similar to linear regression that shrinks regression coefficient estimates to zero. In other words, predictors that are found to be nonsignificant have their regression coefficients set to 0 (Hastie, Tibshirani, and Friedman, 2009). The role of the LASSO is used to reduce the number of variables that would subsequently be entered into the CHAID algorithm to define the final nonresponse adjustment weighting cells.

In the final model produced by the LASSO, only the most significant variables predictive of the response variable were identified and kept. The HPGENSELECT procedure (Johnston and Rodriguez, 2015) with selection method=lasso in SAS 9.4 was used to select the variables, with the weight set to the person base weight, $W_{hijk}^{(3)}$. Separate models were fitted for the three age groups indicated in Table 3-6. The models were selected on the basis of cross validation with observations in the input data set partitioned into disjoint subsets for model, reserving 25% for training, 50% for validation, and 25% for testing. As there is some randomness in how the LASSO selects the variables, we set the seed to a known constant value to remove the randomness so that if the program had to be re-run, the same results would be produced. Out of 79, 78, and 78 variables used in the original models for adults, adolescents, and children, respectively, the LASSO identified 52, 30, and 21 variables to be significant predictors of response for the three age groups, respectively, as shown in Table 3-6. A complete list of these variables is given in Appendix C.

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

The next step was to apply the CHAID algorithm (Magidson, J., 2005) to the variables selected by the LASSO procedure. CHAID classifies the sampled individuals (i.e., the respondents and nonrespondents) into "cells" based on information available for all sample persons. The cells are formed in such a way that persons belonging to the same cell have similar propensities for being respondents. 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, $W_{hijk}^{(3)}$. An output of the CHAID procedure is a tree diagram that specifies the optimum number of final weighting cells, and their definitions based on the input predictor variables. The depth limit of the tree was set to 5, and the minimum subgroup size required to allow splitting and minimum terminal node size were set to 50 observations (both respondents and nonrespondents).

To create the CHAID tree for adults, gender (variable SEX) and an age-derived variable (specifically, whether the person was between the ages of 15-17 or 17+ (the derived variable H_AGETEENYEARS_C defined in Table 3-7), were forced into the model to make the initial splits. The reason for doing this was because males and females and adults 15-17 and adults 17+ received different questions; without forcing these variables into the model, the resulting tree would not have been created correctly. After forcing the two variables in the model, the tree was then allowed to grow freely. The CHAID algorithm selected 19, 11, and 13 variables for adults, adolescents, and children, respectively, that were used to create the weighting classes for nonresponse adjustment. Table 3-7 summarizes the variables that were included in the final CHAID models. The trees produced by CHAID are provided in Appendix C.

The final cells produced by CHAID were used to specify the nonresponse adjustment classes. However, cells that either had fewer than 30 respondents or had a weighted response rate of 50 percent or less, were collapsed with neighboring cells after reviewing the detailed CHAID trees. A total of 44 final weighting adjustment cells were created for adults, 17 cells for adolescents, and 31 cells for children. The final weighting cells created for nonresponse adjustment are also documented in Appendix C.

Table 3-6 Variables in the original model, variables selected by LASSO, and variables selected by CHAID, and final adjustment cells

Age Group	Variables in original model	Variables selected by the LASSO	Variables selected by CHAID	Number of nonresponse adjustment cells
Adults	79	52	19	44
Adolescents	78	30	11	17
Children	78	21	13	31

Table 3-7 Variables selected by CHAID to produce classes for interview nonresponse adjustment

Age group	Number	Variable name	Description
Adult	1	H_AGETEENYEARS_C	1: 15-17; 2: Other; based on AGEYEARS (roster)
	2	H_COOKFUEL_C	Cooking Fuel: Elect., Gas, Parfin/Kerosene/coal/charcoal/wood, Other
	3	H_DADGUARD	Father or male guardian in HH: 1: Yes, 2: No
	4	H_HAVERADTVREF_C	Household has radio, television, refrigerator
	5	H_HH_SIZE_C	1-9, where 9 includes all HHs with 9 or more rostered eligible people
	6	H_MOMGUARD	Mother or female guardian in HH: 1: Yes, 2: No
	7	H_OWNSHEEP_C	Household owns sheep: 0 - does not own cattle; 1 - owns 1 or more sheep
	8	H_TOILETSHARENUM_C	Number of households that use this toilet facility: 1 = 1, 2 = 2, 3= (3, 4), 4= 5+
	9	H_WTRSRC	Water Source: Pipe, Tube, Well, Spring/Rain, truck/bottled, other
	10	HAVEMICRO	calc - Does your household have a microwave? (hidden)
	11	HHR091	calc - What is name's current marital status?
	12	M_SPOUSEYN	calc - Does mname have a spouse or co-habitating partner who usually lives in the household or stayed here last night?
	13	NOEAT4WKFREQ	calc - How often did this happen in the past 4 weeks? (hidden)
	14	OWNLAND	calc - Does any member of your household own any agricultural land? (hidden)
	15	SCHLHLEVEL	calc - What is the highest level of school name has attended? (hidden)
	16	SEX	calc - Is name Male or Female? (hidden)
	17	SICKFLAGHH	calc - flag household with sick adult (hidden)
	18	STRATA	Design strata
	19	URBAN_RURAL	Urban_Rural indicator: 1=Urban; 2=Rural
Adolescent	1	H_DADGUARD	Father or male guardian in HH: 1: Yes, 2: No
	2	H_HAVECEPHCOWA_C	Household has Phone, Computer, watch
	3	H_HAVERADTVREF_C	Household has radio, television, refrigerator
	4	H_HH_SIZE_C	1-9, where 9 includes all HHs with 9 or more rostered eligible people
	5	H_MATFLR	RECODED MATFLOOR
	6	H_MOMGUARD	Mother or female guardian in HH: 1: Yes, 2: No
	7	H_OWNCHIKN_C	Household owns chickens: 0 - does not own chickens; 1 - owns (1-9) chickens; 2 - owns 10 or more chickens
	8	H_PARENTSICK_C	Categorical Parent Alive
	9	H_TLETTYP	RECODED TOILETTYPE
	10	HAVECASS	calc - Does your household have a cassette player? (hidden)
	11	HAVEMOSQNT	calc - Does your household have any mosquito nets that can be used while sleeping? (hidden)
Children	1	H_DADGUARD	Father or male guardian in HH: 1: Yes, 2: No
	2	H_HAVECEPHCOWA_C	Household has Phone, Computer, watch

Age group	Number	Variable name	Description
	3	H_HAVERADTVREF_C	Household has radio, television, refrigerator
	4	H_HH_SIZE_C	1-9, where 9 includes all HHs with 9 or more rostered eligible people
	5	H_MATFLR	RECODED MATFLOOR
	6	H_MOMGUARD	Mother or female guardian in HH: 1: Yes, 2: No
	7	H_OWNCHIKN_C	Household owns chickens: 0 - does not own chickens; 1 - owns (1-9) chickens; 2 - owns 10 or more chickens
	8	H_PARENTSICK_C	Categorical Parent Alive
	9	H_TLETTYP	RECODED TOILETTYPE
	10	HAVECASS	calc - Does your household have a cassette player? (hidden)
	11	HAVEMOSQNT	calc - Does your household have any mosquito nets that can be used while sleeping? (hidden)
	12	OWNLAND	calc - Does any member of your household own any agricultural land? (hidden)
	13	STRATA	Design strata

Calculation of Nonresponse-Adjusted Person Weights

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

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

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

The corresponding replicate-specific weighted response rates were similarly computed for jackknife replicate $r = 1, 2, ..., 253$ as

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

The interview nonresponse adjustment factor for cell m is $A_m^{(int)} = 1/R_m^{(int)}$ for the full sample, and $A_{(r)m}^{(int)} = 1/R_{(r)m}^{(int)}$ for jackknife replicate $r = 1, 2, ..., 253$.

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 $r = 1, 2, ..., 253$ were similarly computed as

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

Table 3-8 summarizes the number of weighting cells created for nonresponse adjustment, the overall weighted response rate, and the minimum and maximum adjustment factors for each of the three major age groups.

Table 3-8 Characteristics of the weighting cells developed for interview nonresponse adjustment and weighted counts before and after adjustment

Group	Age	Number of interview respondents	Number of adjustment cells	Overall weighted response rate	Adjustment factor		Weighted count of respondents	
					Min.	Max.	before adjustment [1]	after adjustment [2]
Adults	15-59	21,341	44	85.99	1.00	2.80	6,069,845	7,058,958
Adolescents	10-14	2,743	17	76.92	1.02	2.80	1,557,186	2,024,327
Children	0-9	7,404	31	91.88	1.00	2.35	4,209,097	4,581,300

[1] Weight is person base weight, $W_{mk}^{(3)}$.
[2] Weight is nonresponse-adjusted person weight, $W_{(r)mk}^{(int)}$.

3.4.3.3 Weight Trimming

To reduce the variability of the weights which can lead to inflated sampling variances, an adjustment known as "weight trimming" was applied to the nonresponse-adjusted weights. For this purpose, a weight outlier is defined to be a weight that is greater than *3.5 times* the median *nonresponse-adjusted* weight (Valliant, Dever, Kreuter, 2013) within the corresponding sampling stratum and age group. Such weights were capped at 3.5 times the median weight. The resultant weights are then recalibrated to population control totals through the poststratification adjustment described in the following section. This procedure was performed for both interview and blood test weights.

As shown in Table 3-9, there were 23 weight outliers at the interview level that fit the definition of an outlier above. There were twelve weight outliers in adult weights, three in adolescent weights, and

eight in child weights. Table 3-10 shows the impact of trimming on the sum of weights. It can be seen that after trimming, the sum of weights decreases and the design effect is reduced slightly, as expected, for all age groups and overall.

Table 3-9 Number of weight outliers (3.5 * median weight) within stratum and age group for interview weights

Stratum (Zone)	Adults (15-59)		Adolescents (10-14)		Children (0-9)	
	Number of interview respondents	Number of outliers	Number of interview respondents	Number of outliers	Number of interview respondents	Number of outliers
Central	1,789	0	260	0	641	0
Copperbelt	3,643	0	425	0	1,003	0
Eastern	2,032	1	292	1	696	7
Luapula	1,145	0	157	0	462	0
Lusaka	3,649	1	379	1	1,020	1
Muchinga	1,932	8	277	1	834	0
North-Western	1,990	0	275	0	749	0
Northern	1,735	1	223	0	666	0
Southern	2,380	0	322	0	864	0
Western	1,046	1	133	0	469	0
Overall	21,341	12	2,743	3	7,404	8

Table 3-10 Weighted counts, mean, and design effect (DEFF) before and after trimming for interview weights

Group	Age	Number of Interview Respondents	Number of records trimmed	Before trimming			After trimming		
				Wtd. count of respondents	Mean	DEFF [1]	Wtd. count of respondents	Mean	DEFF [1]
Adult	15-59	21,341	12	7,058,957.99	330.77	1.104	7,057,320.05	330.69	1.103
Adolescents	10-14	2,743	3	2,024,326.60	738.00	1.158	2,023,553.31	737.72	1.156
Children	0-9	7,404	8	4,581,300.15	618.76	1.122	4,579,858.80	618.57	1.120
Overall	0-59	31,488	23	13,664,584.74	433.96	1.263	13,660,732.16	433.84	1.261

[1] DEFF is calculated as $1+CV^2$, where CV = the coefficient of variation of the weights.

3.4.3.4 Poststratification Adjustment

The final step in computing the individual interview weights was to adjust the nonresponse-adjusted interview weights to national population totals 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. Undercoverage can occur:

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

To compensate for the types of coverage problems indicated above, the nonresponse-adjusted person weights were ratio-adjusted so that the resulting weighted sample counts match the population control totals indicated in Table 3-11. The population control totals given in this table are projected 2016 national population counts by gender and five-year age groups published by the Zambia Central Statistical Office (CSO). The post-stratified interview weights were computed as follows.

Let N_{ga}^{2016} denote the 2016 Zambia population control total for gender *g* and (five-year) age group *a* as given in Table 3-11. The poststratification ratio adjustment factor for gender *g* and age group *a* was then computed as:

$$T_{ga}^{2016} = N_{ga}^{2016} / \sum_{k=1}^{n_{ga}^{resp}} W_{gak}^{(int)}$$

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:

$$T_{(r)ga}^{2016} = N_{ga}^{2016} / \sum_{k=1}^{n_{(r)ga}^{resp}} W_{(r)gak}^{(int)}$$

for the $r = 1, 2, \ldots, 253$ jackknife replicates.

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

$$W_{gak}^{(ps-int)} = T_{ga}^{2016} W_{gak}^{(int)}$$

and the corresponding poststratified replicate weights were computed as:

$$W_{(r)gak}^{(ps-int)} = T_{ga}^{2016} W_{(r)gak}^{(int)}$$

for $r = 1, 2, \ldots, 253$.

Weighted counts of the interview respondents before and after poststratification are summarized in Table 3-11.

Table 3-11 2016 Zambia population projections (overall and by age and gender) and weighted counts before and after poststratification

Age group	Male			Female			Total		
	Population control total [1]	Wtd. count before post-stratification [2]	Post-stratification adjustment factor [3]	Population control total [1]	Wtd. count before post-stratification [2]	Post-stratification adjustment factor [3]	Population control total [1]	Wtd. count before post-stratification [2]	Post-stratification adjustment factor [3]
0-4	1,455,749	1,131,946	1.2861	1,432,186	1,152,559	1.2426	2,887,935	2,284,504	1.2641
5-9	1,222,453	1,174,111	1.0412	1,213,180	1,133,375	1.0704	2,435,633	2,307,486	1.0555
10-14	1,004,834	997,305	1.0075	998,194	1,010,139	0.9882	2,003,028	2,007,444	0.9978
15-19	874,130	744,184	1.1746	887,398	792,638	1.1195	1,761,528	1,536,822	1.1462
20-24	750,271	542,422	1.3832	776,990	704,600	1.1027	1,527,261	1,247,022	1.2247
25-29	550,497	423,088	1.3011	629,157	559,774	1.1239	1,179,654	982,863	1.2002
30-34	467,068	410,132	1.1388	537,564	501,551	1.0718	1,004,632	911,683	1.1020
35-39	398,434	330,584	1.2052	405,066	399,478	1.0140	803,500	730,062	1.1006
40-44	334,869	292,926	1.1432	314,401	325,498	0.9659	649,270	618,425	1.0499
45-49	239,267	215,575	1.1099	218,671	202,392	1.0804	457,938	417,967	1.0956
50-54	175,303	161,742	1.0838	175,892	194,098	0.9062	351,195	355,840	0.9869
55-59	129,545	126,716	1.0223	141,011	133,897	1.0531	270,556	260,613	1.0382
Total	7,602,420	6,550,733	1.1605	7,729,710	7,109,999	1.0872	15,332,130	13,660,732	1.1224

[1] Source: 2016 Zambia population projections.

[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 also provided a useable blood sample. Thus, a separate set of weights is required for analysis of the blood test results. Like the construction of the interview weights described previously, development of the final blood test weights involves adjustments for nonresponse and poststratification to 2016 population control totals.

3.4.4.1 Initial Weights

The starting point for the construction of the blood test weights is the set of final full-sample nonresponse-adjusted interview weights and corresponding replicate weights described in Section 3.4.3.2. These weights are given by $W_{hijk}^{(int)}$ and $W_{(r)hijk}^{(int)}$ (for $r = 1, 2, \ldots, 253$), respectively, where *k* denotes the interview respondent, *h* denotes the province, *i* denotes the PSU, and *j* denotes the household. These weights have already been adjusted for interview nonresponse, and thus act as the "base" weights for developing nonresponse adjustments for the blood tests. Note that persons who provided a valid blood sample are considered to be interview respondents for the weighting purposes (e.g., see Tables 2-9A through 2-9C). Table 3-12 summarizes the counts of individuals by gender, age group and blood test response status, and the corresponding weighted counts using the person-level interview weights, $W_{hijk}^{(int)}$.

Table 3-12 Distribution of sample persons completing the blood test by age group and response status

Group	Age [1]	Gender	Blood Test Status [2]	Count	Weighted count [3]
Adults	15-59	Male	Respondent	8,142	2,875,151
			Nonrespondent	1,029	372,220
		Female	Respondent	10,973	3,445,804
			Nonrespondent	1,136	368,122
Adolescents	10-14	Male	Respondent	1,263	907,146
			Nonrespondent	117	90,159
		Female	Respondent	1,283	914,623
			Nonrespondent	126	95,515
Children	0-9	Male and female	Respondent	5,469	3,335,570
			Nonrespondent	1,950	1,256,420

[1] Age reported in the interview, which may differ from the age reported on the roster.
[2] Status among the interview respondents. Persons completing the blood test are considered to be interview respondents regardless of whether a completed interview was obtained.
[3]Weighted by the person-level interview weight, $W_{hijk}^{(int)}$.

3.4.4.2 Nonresponse Adjustment of Blood Test Weights

To compensate for blood test nonresponse, the person-level interview weights were adjusted within cells defined by variables available for both the responding and nonresponding individuals. These variables included data from the household roster and other information collected in the household questionnaire, and selected PSU characteristics such as province and urban/rural status, and the individual interview. The age and sex variables used to make the nonresponse adjustments are those reported in the interview.

The LASSO procedure was used to identify a reduced set of predictor variables to be used in the CHAID algorithm. Table 3-13 shows the number of variables used in the models for adults, adolescents, and children, respectively, and the number of variables identified by the LASSO to be significant predictors of response for the three age groups, respectively. Because LASSO selected only two variables for adolescent males, three variables that were found to be significant predictors of response in other countries, namely, a categorized age based on interview (BEST_AGE), the PSU stratification variable (STRATA), and urban/rural status (URBAN_RURAL) were added manually for both adolescent males and females. During cell formation in CHAID, no variables were identified as predictive of the response variable for adolescent males, so STRATA and URBAN_RURAL were manually forced in the tree for this group. Table 3-14 summarizes the

variables that were included in the final CHAID models. The trees produced by CHAID and the resulting nonresponse-adjustment classes are provided in Appendix C.

Table 3-13 Variables in the original model, variables selected by LASSO, variables selected by CHAID, and final adjustment cells for blood test weights

Group	Age	Gender	Variables in original model	Variables selected by the LASSO	Variables selected by CHAID	Number of nonresponse adjustment cells
Adults	15-59	Male	158	64	18	23
		Female	208	94	19	32
Adolescents	10-14	Male	124	2 [1]	2 [2]	4
		Female	118	8 [1]	1	2
Children	0-9	Male and female	99	51	41	53

[1] Three additional added manually
[2] Forced into decision tree manually

Table 3-14 Variables selected by CHAID to produce classes for blood test nonresponse adjustment

Age group	Number	Variable name	Description
Adult Male	1	AT_WIFLIVEEW	How many wives/partners do you have who live elsewhere?
	2	AWY12MOMS	In the last 12 months, have you been away from your home for more than one month at a time?
	3	FAMSHAME	Do you agree or disagree with the following statement: I would be ashamed if someone in my family had HIV.
	4	FIRSTEXCNDM	The first time you had vaginal or anal sex, was a condom used?
	5	H_OWNPIG_C	Household owns pigs: 0 - does not own pigs 1 - owns (1-2) pigs; 2 - owns 3 or more pigs
	6	H_WTRSRC	Water Source: Pipe, Tube, Well, Spring/Rain, truck/bottled, other
	7	HFHIVTSTOFFER	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?
	8	HIVTSTRSLT	What was the result of that HIV test?
	9	HIVTSTSELFKIT	If an HIV self-test kit were available in this country, would you use it?
	10	MCPLANS	Are you planning to get circumcised?
	11	NOEAT4WKYN	calc - In the past 4 weeks, did you or any household member go a whole day and night without eating anything because there was not enough food? (hidden)
	12	PNDSCHRG	During the last 12 months, have you had an abnormal discharge from your penis?
	13	SCHLATT2015	calc - Did name attend school at any time during the 2015 school year? (hidden)
	14	SCHLHI	What is the highest level of school you attended: primary, secondary, or higher?
	15	STDTRT	Did you get treatment for these problems?
	16	STRATA	Design strata
	17	TBDIAGN	Have you ever been told by a doctor, clinical officer or nurse that you had TB?
	18	WORK7DAY	Have you done any work in the last 12 months for which you received a salary, cash, or in kind as payment?
Adult Female	1	AT_BESTAGE_C	Categorical age based on interview age (BEST AGE)
	2	AT_PREGNUM	Categorical: How many times have you been pregnant including a current pregnancy?
	3	AVOIDPREG	Are you or your partner currently doing something or using any method to delay or avoid getting pregnant?
	4	CERVCNRSLT	What was the result of your last test
	5	CNDMSEX	Do you believe women who carry condoms have sex with a lot of men?
	6	FIRSTSXFRC	The first time you had vaginal or anal sex, was it because you wanted to or because you were forced to?
	7	H_HH_SIZE_C	1-9, where 9 includes all HHs with 9 or more rostered eligible people
	8	H_OWNCHIKN_C	Household owns chickens: 0 - does not own chickens; 1 - owns (1-9) chickens; 2 - owns 10 or more chickens

Age group	Number	Variable name	Description
	9	HAVEMOSQNT	calc - Does your household have any mosquito nets that can be used while sleeping? (hidden)
	10	HAVETABLE	calc - Does your household have a table? (hidden)
	11	HIVTSBP	Have you ever tested for HIV before your pregnancy with ${namedis}?
	12	HIVTSTRSLT	What was the result of that HIV test?
	13	HIVTSTSELFKIT	If an HIV self-test kit were available in this country, would you use it?
	14	KNOWN_HIV_STATUS_R	Categorical known HIV status
	15	STRATA	Design strata
	16	SYPHTOF	When you were pregnant with ${namedis}, were you offered a test for syphilis?
	17	TBDIAGN	Have you ever been told by a doctor, clinical officer or nurse that you had TB?
	18	VGSORE	During the last 12 months, have you had an ulcer or sore on or near your vagina?
	19	VLNC	Has anyone ever done any of these things to you: - Punched, kicked, whipped, or beat you with an object - Slapped you, threw something at you that could hurt you, pushed you or shoved you - Choked, smothered, tried to drown you, or burned
Adolescent Male	1	STRATA	Design strata
	2	URBAN_RURAL	Urban/Rural indicator: 1=Urban, 2=Rural
Adolescent Female	1	DEATHS	calc - Now I would like to ask you more questions about your household. Has any usual resident of your household died since 2013? (hidden)
Children	1	ALLSHDTEST	To what extent do you agree with the following statement: All HIV-negative people should test for HIV every year. Do you strongly agree, agree, disagree, or strongly disagree?
	2	AT_LIVEBNUM	Categorical number of live births
	3	AT_WIFLIVEEW	Categorical: How many wives/partners do you have who live elsewhere?
	4	AVOIDPREG	Are you or your partner currently doing something or using any method to delay or avoid getting pregnant?
	5	AWY12MOMS	In the last 12 months, have you been away from your home for more than one month at a time?
	6	BIRTHCERT	calc - Does name have a birth certificate? Has name's birth ever been registered with the civil authority? (hidden)
	7	CERVCNTST	Have you ever been tested for cervical cancer?
	8	CH_KIDCRCMFUTR	Are you planning to have ${curchnm} circumcised in the future?
	9	CH_KIDENROLL	Is ${curchnm} currently enrolled in school?
	10	CH_KIDHEPB	And now my last question about ${curchnm} is on Hepatitis B: Has your child ever been vaccinated for Hepatitis B? Please show us his/her under 5 vaccination card.

Age group	Number	Variable name	Description
	11	CH_KIDHIVLASTRESULT	What was ${curchnm}'s last HIV test result?
	12	ETHGRPTR	What is your ethnic group/tribe?
	13	FAMSHAME	Do you agree or disagree with the following statement: I would be ashamed if someone in my family had HIV.
	14	H_MATFLR	Recoded material of floor
	15	H_MATRF	Recoded material of roof
	16	H_OWNCATTLE_C	Household owns cattle: 0 - does not own cattle; 1 - owns (1-4) cattle; 2 - owns 5 or more cattle
	17	H_OWNCHIKN_C	Household owns chickens: 0 - does not own chickens; 1 - owns (1-9) chickens; 2 - owns 10 or more chickens
	18	H_OWNDCOW_C	Household owns dairy cattle: 0 - does not own dairy cattle; 1 - owns (1-4) dairy cattle; 2 - owns 5 or more dairy cattle
	19	H_TLETTYP	RECODED TOILETTYPE
	20	H_WTRSRC	Water Source: Pipe, Tube, Well, Spring/Rain, truck/bottled, other
	21	HAVEBEDMAT	calc - Does your household have a bed with mattress? (hidden)
	22	HAVECUPB	calc - Does your household have a cupboard? (hidden)
	23	HEALTHC	Who usually makes decisions about health care for yourself: you, your (spouse/partner), you and your (spouse/partner) together, or someone else?
	24	HEALTHYINF	Can a healthy-looking person have HIV?
	25	HFHIVTSTOFFER	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?
	26	HHR091	calc - What is name's current marital status? (hidden)
	27	HIVTOBR	Were you offered an HIV test during labor?
	28	HIVTSBP	Have you ever tested for HIV before your pregnancy with ${namedis}?
	29	HIVTSTSELFKIT	If an HIV self-test kit were available in this country, would you use it?
	30	KIDSSCHOOL	Do you think children living with HIV should be allowed to attend school with children who do not have HIV?
	31	MORECHILDWAIT	How long would you like to wait before the birth of a/another child?
	32	P_BESTAGE_C	Categorical age based on interview age (BEST AGE)
	33	PREGPLAN	When you were pregnant with ${namedis}, did you plan to get pregnant at that time?
	34	PROXY_GENDER	Gender of responding parent / guardian
	35	RESPECT	Do people living with HIV, or thought to be living with HIV, lose the respect of other people?
	36	SCHLHI	What is the highest level of school you attended: primary, secondary, or higher?
	37	SDFLAGMOM	calc - flag for whether mom is dead or sick (hidden)
	38	STRATA	Design strata

Age group	Number	Variable name	Description
	39	SYPHTOF	When you were pregnant with ${namedis}, were you offered a test for syphilis?
	40	TBCUREHIV	Can TB be cured in people living with HIV?
	41	VLNC	Has anyone ever done any of these things to you: - Punched, kicked, whipped, or beat you with an object - Slapped you, threw something at you that could hurt you, pushed you or shoved you - Choked, smothered, tried to drown you, or burned

Calculation of Nonresponse-Adjusted Blood Test Weights

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

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

where m denotes the adjustment cell, $W_{mk}^{(int)}$ is the final interview weight for interview respondent k in cell m, n_m^{BT} = the number of interview respondents in cell m who provided a useable blood sample, and n_m^{NBT} = the number of interview respondents in cell m who did not provide a useable blood sample.

The corresponding replicate-specific weighted response rates were similarly computed for jackknife replicate r = 1, 2, ..., 253 as

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

The blood test nonresponse adjustment factor for cell m is $A_m^{(BT)} = 1/R_m^{(BT)}$ for the full sample, and $A_{(r)m}^{(BT)} = 1/R_{(r)m}^{(BT)}$ for jackknife replicate r = 1, 2, ..., 253.

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

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

and the corresponding jackknife replicate weights for replicate $r = 1, 2, ..., 253$ were similarly computed as

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

Table 3-15 summarizes the number of weighting cells created for nonresponse adjustment of the blood test weights, the overall weighted response rate, and the minimum and maximum adjustment for each of the three major age groups.

Table 3-15 Characteristics of the weighting cells developed for blood test nonresponse adjustment and weighted counts before and after adjustment

Group	Age	Gender	Number of blood test respondents	Number of adjustment cells	Overall weighted response rate [1]	Adjustment factor		Weighted count of respondents	
						Min.	Max.	Before adjustment [2]	After adjustment [3]
Adults	15-59	Male	8,142	23	88.54	1.0000	1.8194	2,875,151	3,247,371
		Female	10,973	32	90.35	1.0000	1.7940	3,445,804	3,813,927
Adolescents	10-14	Male	1,263	4	90.96	1.0555	1.3149	907,146	997,305
		Female	1,283	2	90.54	1.0910	1.2263	914,623	1,010,139
Children	0-9	Male and female	5,469	53	72.64	1.0000	2.4974	3,335,570	4,591,991

[1] Among the interview respondents.
[2] Weight is person interview weight, $W_{mk}^{(int)}$.
[3] Weight is nonresponse-adjusted blood test weight, $W_{(r)mk}^{(BT)}$.

3.4.4.3 Weight Trimming

To reduce the variability of the weights which can lead to inflated sampling variances, an adjustment known as "weight trimming" was applied to the nonresponse-adjusted weights. For this purpose, a weight outlier is defined to be a weight that is greater than 3.5 times the median *nonresponse-adjusted* weight (Valliant, Dever, Kreuter, 2013) within the corresponding sampling stratum and age group. Such weights were capped at 3.5 times the median weight. The resultant weights are then recalibrated to population control totals through the poststratification adjustment described in the following section. This procedure had also been performed on the interview weights (see Section 3.4.3.3).

As shown in Table 3-16, there were 28 weight outliers at the blood test level. Outliers were present in weights for adults, adolescents, and children. Three out of the 12 adult females and six out of the nine adult males with outlier weights were "age switchers," where the individual was reported (and sampled) as age 14 or younger in the roster, but later was confirmed to be 15 or older

at interview. As described previously in Section 3.4.3.1 (Person Base Weights), the weights of children in households designated for child data collection were multiplied by a subsampling factor of $K_k = 2$; this factor was appropriately retained at the blood test level for "age switchers," which contributed to the increased weights. Table 3-17 shows the impact of trimming on the sum of weights. It can be seen that after trimming, the sum of weights decreases and the design effect is reduced slightly, as expected, for all age groups and overall.

Table 3-16 Number of weight outliers (3.5 * median weight) within stratum and age group for blood test weights

Stratum (Zone)	Adults (15-59)				Adolescents (10-14)				Children (0-14)	
	Male	Female	Male	Female	Male	Female	Male	Female		
	Number of blood test respondents		Number of weight outliers		Number of blood test respondents		Number of weight outliers		Number of blood test respondents	Number of weight outliers
Central	702	889	2	0	109	128	0	0	488	0
Copperbelt	1386	1898	1	0	184	207	0	0	792	0
Eastern	819	1070	1	2	152	121	1	0	507	1
Luapula	426	590	0	0	68	84	0	0	338	0
Lusaka	1,242	2,003	0	2	175	181	1	0	727	2
Muchinga	758	971	3	3	144	123	0	1	639	0
North-Western	765	1012	1	1	125	136	0	0	543	0
Northern	659	820	1	1	91	94	0	0	398	1
Southern	990	1186	0	1	160	145	0	0	717	0
Western	395	534	0	2	55	64	0	0	320	0
Overall	8,142	10,973	9	12	1,263	1,283	2	1	5,469	4

Table 3-17 Weighted counts, mean, and design effect (DEFF) before and after trimming for blood test weights

Age Group	Age	Gender	Number of blood test respondents	Number of records trimmed	Before trimming			After trimming		
					Wtd. Count of respondents	Mean	DEFF [1]	Wtd. Count of respondents	Mean	DEFF [1]
Adults	15-59	Male	8,142	9	3,247,371	398.84	1.12	3,244,779	398.52	1.12
		Female	10,973	12	3,813,927	347.57	1.11	3,812,053	347.4	1.10
Adolescents	10-14	Male	1,263	2	997,305	789.63	1.17	996,999	789.39	1.17
		Female	1,283	1	1,010,139	787.33	1.16	1,009,897	787.14	1.16
Children	0-14		5,469	4	4,591,991	839.64	1.17	4,590,598	839.39	1.17
Overall			27,130	28	13,660,732	503.53	1.35	13,654,325	503.29	1.35

[1] DEFF is calculated as $1+CV^2$, where CV = the coefficient of variation of the weights.

3.4.4.4 Poststratification Adjustment

Like the nonresponse-adjusted interview weights described previously, the nonresponse-adjusted blood test weights were poststratified to projected 2016 population counts within classes defined by gender and five-year age group.

Let N_{ga}^{2016} denote the 2016 Zambia population control total for gender *g* and (five-year) age group *a* as given in Table 3-18. The poststratification ratio adjustment factor used to adjust the blood test weights for gender *g* and age group *a* was computed as:

$$T_{ga}^{2016} = N_{ga}^{2016} / \sum_{k=1}^{n_{ga}^{BT}} W_{gak}^{(BT)},$$

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:

$$T_{(r)ga}^{2016} = N_{ga}^{2016} / \sum_{k=1}^{n_{(r)ga}^{BT}} W_{(r)gak}^{(BT)}$$

for the $r = 1, 2, \ldots, 253$ jackknife replicates.

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

$$W_{gak}^{(ps-BT)} = T_{ga}^{2016} W_{gak}^{(BT)},$$

and the corresponding poststratified replicate weights were computed as:

$$W_{(r)gak}^{(ps-BT)} = T_{ga}^{2016} W_{(r)gak}^{(BT)}$$

for $r = 1, 2, \ldots, 253$.

Weighted counts of the blood test respondents before and after poststratification are summarized in Table 3-18.

Table 3-18 2016 Zambia population projections (overall and by age and gender) and weighted counts of blood test respondents before and after poststratification

Age group	Male			Female			Total		
	Population control total [1]	Wtd. count before post-stratification [2]	Post-stratification adjustment factor [3]	Population control total [1]	Wtd. count before post-stratification [2]	Post-stratification adjustment factor [3]	Population control total [1]	Wtd. count before post-stratification [2]	Post-stratification adjustment factor [3]
0-4	1,455,749	1,088,292	1.3376	1,432,186	1,115,213	1.2842	2,887,935	2,203,504	1.3106
5-9	1,222,453	1,205,230	1.0143	1,213,180	1,181,863	1.0265	2,435,633	2,387,094	1.0203
10-14	1,004,834	996,999	1.0079	998,194	1,009,897	0.9884	2,003,028	2,006,896	0.9981
15-19	874,130	751,501	1.1632	887,398	796,984	1.1134	1,761,528	1,548,485	1.1376
20-24	750,271	548,884	1.3669	776,990	707,868	1.0976	1,527,261	1,256,753	1.2152
25-29	550,497	417,868	1.3174	629,157	553,745	1.1362	1,179,654	971,613	1.2141
30-34	467,068	406,805	1.1481	537,564	498,731	1.0779	1,004,632	905,536	1.1094
35-39	398,434	321,703	1.2385	405,066	393,013	1.0307	803,500	714,717	1.1242
40-44	334,869	289,572	1.1564	314,401	328,913	0.9559	649,270	618,485	1.0498
45-49	239,267	213,451	1.1209	218,671	204,561	1.0690	457,938	418,013	1.0955
50-54	175,303	165,182	1.0613	175,892	197,237	0.8918	351,195	362,419	0.9690
55-59	129,545	129,812	0.9979	141,011	130,999	1.0764	270,556	260,812	1.0374
Total	7,602,420	6,535,300	1.1633	7,729,710	7,119,025	1.0858	15,332,130	13,654,325	1.1229

[1] Source: 2016 Zambia population projections.

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

Special Purpose Weights 4

In addition to the analytic weights described in Section 3, three sets of special purpose weights were created for analysis of specific sections of the individual questionnaire. The three sections of interest are (a) the violence module (VM), (b) questions on HIV knowledge (HIVK), and (c) weight and height measurements for children 0-60 months of age. Special weights are required for analyses of these sections because the relevant questionaire items were administered to random subsamples of the interview respondents.

4.1 Weights for Analysis of the Violence Module

The violence module was administered to a random sample of women 15-59 years of age. The module does not apply to men 15-59 years of age nor children 0-14 years of age.

4.1.1 Selection Criteria for the Violence Module

One eligible adult female aged 15-59 was randomly selected per household to respond to the questions in the violence module. The criteria used to identify persons eligible for the violence module are given in Appendix D.

4.1.2 Definition of Response Status for the Violence Module

For adult females who were designated to receive the violence module, their violence respondent status is based on whether they answered key questions within the violence module. For weighting purposes, respondents are defined to be those women who (a) provided a VALID response to all four "how many times" questions, or (b) provided a VALID response to the VLNC question (see Appendix D for detailed descriptions of the VM variables). This definition results in an unweighted response rate of 91.9% (8,213/8,940). Table 4-1 summarizes the number of responses to the five key adult violence variables.

Table 4-1 Distribution of responses to key variables in the violence module

TOUCHTIMES	CMPLSXTIMES	FRCSXTIMES	PRSSXTIMES	VLNC	Frequency
Missing	Missing	Missing	Missing	Missing	711
Missing	Missing	Missing	Missing	Invalid	12
Missing	Missing	Missing	Missing	Valid	23
Missing	Missing	Missing	Valid	Invalid	2
Missing	Missing	Missing	Valid	Valid	7
Missing	Missing	Valid	Missing	Valid	1
Missing	Missing	Valid	Valid	Invalid	1
Missing	Missing	Valid	Valid	Valid	7
Missing	Valid	Missing	Missing	Valid	2
Missing	Valid	Valid	Valid	Invalid	1
Missing	Valid	Valid	Valid	Valid	14
Valid	Missing	Missing	Missing	Valid	1
Valid	Missing	Missing	Valid	Valid	2
Valid	Missing	Valid	Missing	Valid	1
Valid	Missing	Valid	Valid	Valid	6
Valid	Valid	Missing	Missing	Valid	3
Valid	Valid	Missing	Valid	Valid	7
Valid	Valid	Valid	Missing	Valid	10
Valid	Valid	Valid	Valid	Invalid	36
Valid	Valid	Valid	Valid	Valid	8,093

Notes: Missing = NULL, or blank
Invalid = -8 (don't know) or -9 (refused question). The "how many times?" questions did not show either of these answer codes, probably because they had separate "why no answer?" questions that show ONLY those codes.
Valid for "how many times?" is 0 or a positive integer; for VLNC is Yes or No.

4.1.3 Construction of Weights for the Violence Module

The following steps were implemented to construct the violence weights.

- Each eligible woman 15-59 years of age who was selected for the violence module was assigned an appropriate base weight, $W_{jk}^{viol-bw}$, reflecting the probability of selection for the violence module, as follows:

$$W_{jk}^{viol-bw} = W_{jk}^{bw} N_j^F,$$

where N_j^F = the number of eligible women 15-59 in household *j* (based on roster) if there were four or less eligible women in the household or $N_j^F = 4$ it there were five or more eligible women in the household, and where W_{jk}^{bw} is the corresponding base weight from the regular weighting process (see Section 3.4.3.1). The number of eligible women in the household used to compute the violence module initial weight was top-coded to a value of 4 as a way to prevent the creation of large person weights in households with a large number of eligible respondents. The small bias introduced by top coding is mitigated by the poststratification adjustment described below. The top-coded value was determined by examining the design effects and the bias and variance

trade-offs of estimates of the total population using nonresponse-adjusted weights based on different top-coded values.

- Next, the response-status for persons selected for the violence module was assigned as described in Section 4.2. Note that respondents to the violence module also completed the regular interview.
- A CHAID analysis was then applied to the sample of persons selected for the violence module, separately by sex, using the same predictors identified for the regular interview weights (see Table 3-8).
- The final cells identified from the CHAID analysis were used to compute the nonresponse-adjusted weights for the violence module, $W_{jk}^{viol-nr} = A_{jk}^{nradj} W_{jk}^{viol-bw}$.
- The last step was to poststratify the $W_{jk}^{viol-nr}$s to appropriate population counts by detailed age groups within the population of 15-59 year old females.

Table 4-2 lists the variables that were used to create the nonresponse-adjustment cells for creating the violence weights. Table 4-3 summarizes selected unweighted and weighted counts associated with the VM weighting process.

Table 4-2 List of variables identified by CHAID

Variable	Description
H_AGETEENYEARS_C	1: 15-17; 2: Other; based on AGEYEARS (roster)
H_AGEYEARS_C	Best AGEYEARS categorical
H_ECON3	Received some economic support on the past 3 months
H_HAVERADTVREF_C	Household has radio, television, refrigerator
H_HH_SIZE_C	1-9, where 9 includes all HHs with 9 or more rostered eligible people
H_MATWALL	RECODED MATEXWALLS
H_OWNBIGANIMAL_C	Household owns big animals
H_POWER_C	Power: Electricity, Solar energy, Battery, No Power
H_ROOFWALFLR_C	Roof/Wall/Floor materials: Natural, metal/cement, asbestos, etc
H_ROOMSLEEP_C	No. Rooms to sleep: 1, 2, 3, 4+
H_TOILET_C	Toilet Shared, Not shared: Flush, Latrine, Bucket/Other
H_WTRSRC	Water Source: Pipe, Tube, Well, Spring/Rain, truck/bottled, other
STRATA	Design strata

Table 4-3 Selected statistics on the creation of the weights for the violence module

Age group	Number selected for violence module	Base - weighted count of persons selected for violence module	Number of respondents	Base - weighted count of respondents to violence module	Weighted count of respondents after nonresponse adjustment	Weighted count of respondents after post-stratification
Females 15-49	8,224	3,463,655	7,534	3,108,625	3,436,520	3,769,247
Females 50-59	716	303,454	679	286,499	308,705	316,903
Total	8,940	3,767,109	8,213	3,395,124	3,745,225	4,086,150

4.2 Weights for Analysis of the HIV Knowledge Module

The HIV Knowledge (HIVK) module was administered to a random sample of adults 15-59 years of age. The adolescent version of the HIV Knowledge module was administered to children 10-14 years of age. Since all adolescents were selected to respond to this module, no separate HIVK weights were produced for adolescents (i.e., the regular adolescent interview weights described in Section 3.4.3 can be used for this purpose.) The module does not apply to children 0-9 years of age.

4.2.1 Selection Criteria for the HIV Knowledge Module

Each adult 15-59 years of age had an independent probability of selection of 50% for the HIVK module, regardless of the number of other adults in the household. The criteria used to identify persons eligible for the HIVK module are given in Appendix E.

4.2.2 Definition of Response Status for the HIV Knowledge Module

For weighting purposes, respondents are those individuals selected for HIVK with a valid answer to the HIVK question, MOSQUITO ("Can a person get HIV from mosquito bites?"). The valid answers are "Yes = 1", "No = 2", and "Don't Know = 3". The answer "Refused = -9" is considered invalid, i.e., nonresponse. Of the 10,647 adults (15 - 59) who were respondents to the individual interview and were selected for the HIVK module, 10,636 (99.90%) are HIVK "respondents" under the above definition. Table 4-4 summarizes the number of responses to key HIVK variables (see Appendix E for descriptions of the variables).

Table 4-4 Distribution of responses to key variables in the HIVK module

Variable Name	Total (# cases = 10,647) [1]		Male (# cases = 4,650)		Female (# cases = 5,997)	
	# with valid answer	Unwtd RR	# with valid answer	Unwtd RR	# with valid answer	Unwtd RR
ONEPARTNR	10,633	100%	4,640	100%	5,993	100%
MOSQUITO	10,636	100%	4,642	100%	5,994	100%
CONDOMS	10,635	100%	4,642	100%	5,993	100%
SHAREFOOD	10,635	100%	4,642	100%	5,993	100%
HEALTHYINF	10,635	100%	4,642	100%	5,993	100%
BUYFOOD	10,633	100%	4,642	100%	5,991	100%
KIDSSCHOOL	10,630	100%	4,641	100%	5,989	100%
FEARTEST	10,630	100%	4,640	100%	5,990	100%
TALKBAD	10,633	100%	4,642	100%	5,991	100%
RESPECT	10,632	100%	4,642	100%	5,990	100%
SALIVA	10,630	100%	4,639	100%	5,991	100%
FAMSHAME	10,625	100%	4,641	100%	5,984	100%

[1] Counts are of individuals 15-59 years of age who were selected for the HIVK module.

4.2.3 Construction of Weights for the HIV Knowledge Module

The following steps were implemented to construct the HIVK weights.

- Each eligible person 15-59 years of age who was selected for the HIVK module was assigned a base weight, $W_{jk}^{HIVK(bw)}$, reflecting the probability of selection for the HIVK module, as follows:

$$W_{jk}^{HIVK(bw)} = 2\, W_{jk}^{(int)},$$

  where $W_{jk}^{(int)}$ is the corresponding nonresponse-adjusted interview weight from the regular weighting process (see Section 3.4.3.1).

- To reduce the variability of the weights which can lead to inflated sampling variances, an adjustment known as "weight trimming" was applied to the $W_{jk}^{HIVK(bw)}$s. The same trimming rules described in Sections 3.4.3.3 and 3.4.4.3 were applied. As shown in Table 4-5, the weights of one female respondent 15-49 years of age and two male respondents 15-49 years of age were trimmed.

- Because nonresponse to the HIVK module among those individuals completing the regular interview was trivial (0.01%), the final step was to poststratify the trimmed weights $W_{jk}^{HIVK(trim)}$s to appropriate population counts using procedures similar to those described in Section 3.4.3.4.

Table 4-5 summarizes selected unweighted and weighted counts associated with the HIVK weighting process.

Table 4-5 Selected statistics on the creation of the weights for the HIV knowledge module

Sex/age group [1]	Number selected for HIVK module	Base - weighted count of persons selected for HIVK module	Number of HIVK respondents	Base - weighted count of HIVK respondents	Number of HIVK respondents trimmed	Weighted count of HIVK respondents after trimming	Weighted count of HIVK respondents after post-stratification
Females 15-49	5,504	3,471,192	5,501	3,469,614	1	3,469,595	3,769,247
Females 50-59	493	306,688	493	306,688	.	306,688	316,903
Males 15-49	4,233	2,999,645	4,227	2,995,964	2	2,994,143	3,614,536
Males 50-59	417	280,509	415	279,426	.	279,426	304,848
Total	10,647	7,058,035	10,636	7,051,692	3	7,049,852	8,005,534

[1] Sex and age are based on household roster data except for the post-stratified weighted counts in the last column of table. For the latter, sex and age are based on interview responses.

4.3 Weights for Analysis of Children's Weight and Height Measurements

A subsample of children 0-60 months of age was selected to obtain weight and height measurements.

4.3.1 Selection Criteria for the Weight and Height Measurements

All children 0-60 months of age who tested HIV positive and a random sample of approximately 5 percent of children 0-60 months of age who tested HIV negative were selected for the weight and height measurements.

4.3.2 Definition of Response Status for the Weight and Height Measurements

Table 4-6 summarizes the distribution of children 0-60 months old for whom a blood test weight had been computed by the standard PHIA weighting procedures described in Section 3.4.4 by (a) HIV testing status (HIVSTATUS, HIVSTATUSC), (b) weight/height measurement selection status (CWH_FLAG), and (c) the presence or absence of reported height (CWHHEIGHT) and weight (CWHWEIGHT). The number of cases to be weighted are shown in the last column of the table,

and are those for which CWH_FLAG = 1 and for which the weight and height measurements are not both missing. Additional details about the creation of the response status variable is given in Appendix F.

Table 4-6 Distribution of children 0-60 months old with a blood test weight by HIV test result and selection status

HIVSTATUS[1] (1 = pos.; 2 = neg.)	HIVSTATUSC[2] (1 = pos.; 2 = neg.)	CWH_FLAG 1=selected; 0 = not. sel.	CWHHEIGHT	CWHWEIGHT	Cases [3] with a blood test weight	Cases to weighted for weight and height analysis
	1	1	NON-MISS	NON-MISS	34	34
	2	0	MISS	MISS	620	0
	2	1	NON-MISS	NON-MISS	28	28
1		1	NON-MISS	NON-MISS	17	17
2		0	MISS	MISS	1,855	0
2		1	NON-MISS	NON-MISS	109	109
TOTAL	—	—	—	—	2,663	188

[1] HIVSTATUS is the HIV result variable for children who are older than 18 months.

[2] HIVSTATUSC is the HIV result variable for children 18 months or younger.

[3] Children with a confirmed age of 0-60 months for whom a blood test was previously computed (see Section 3.4.4)

4.3.3 Construction of Weights for the Weight and Height Measurements

The basic steps for creating the analytic weights required for analysis of the weight and height measurements were as follows:

- A "base" weight, $W_i^{WH:base}$, was assigned to those cases with CWH_FLAG = 1 as follows:

$$W_i^{WH:base} = K W_i^{BT}$$

where W_i^{BT} is the final blood test weight for child *i* (see Section 3.4.4) and

$K = 1$ if the child tested HIV positive;

$K = 20$ if the child tested HIV negative, was selected for weight and height measurements, and the reported weight and height measurements were not both missing.

From Table 4-6, it can be seen that 188 cases in Zambia were included in the weighting process. Note that since all the sampled children provided weight and height measurements, a separate nonresponse adjustment was not done.

- Next, the base weights, W_i^{BT}, were recalibrated so that the final weighted counts match the corresponding full-sample weighted counts by gender.

  Specifically, let $W_{gi}^{WH:PS}$ denote the final weight for child *i* of gender *g*. Then $W_{gi}^{WH:PS}$ was computed as:

$$W_{gi}^{WH:PS} = W_{gi}^{WH:base} \ (A_g \ / \ B_g \ )$$

  where

  $W_{gi}^{WH:base}$ = the base weight for child *i* of gender *g* as computed above,

  A_g = $\sum_{j=1}^{n_g} W_{gj}^{BT}$

  W_{gj}^{BT} = the previously-computed full-sample blood test weight for child *j* of gender *g*

  n_g = the number of children of gender *g* in the *full* sample for which $W_{gj}^{BT} > 0$

  B_g = $\sum_{j=1}^{n_g^{WH}} W_{gj}^{WH:base}$

  n_g^{WH} = the number of children of gender *g* who were selected for and provided weight/height measurements

- The above steps were repeated for each of the jackknife replicates to provide the corresponding jackknife weights for variance estimation.

Table 4-7 summarizes selected unweighted and weighted counts associated with the weighting process.

Table 4-7 Selected statistics on the creation of the weights for children's weight and height measurements

Sex/age group [1]	Number providing weight and height measurements (respondents)	Base-weighted count of respondents	Final (post-stratified) weighted count of respondents
Females 0-60 mos.	95	1,553,150	1,461,368
Males 0-60 mos.	93	1,424,010	1,484,821
Total	188 [2]	2,977,160	2,946,189

[1] Sex and age are based on household roster data except for the post-stratified weighted counts in the last column of table. For the latter, sex and age are based on interview responses.

[2] Represents an unweighted response rate of 188/188 = 1.000 (see Table 4-6).

References R

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning*. Springer Series in Statistics. http://www.springer.com/us/book/9780387848570

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

Valliant, R., Dever, J., & Kreuter, F. (2013). Practical Tools for Designing and Weighting Survey Samples. New York, NY: Springer.

APPENDIX A

Definition of Eligibility for Dwelling Unit/Household Sampling

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 variables 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 final listing file for Zambia. Table A-2 contains a description of the possible values taken on by the three variables.

Of the 60,581 records in the master listing file, 57,543 were eligible for sampling. Among the eligible listings, 44,017, were those coded as 1,1,1 (that is, a single house/compound of households; not vacant and not under construction; persons living in the structure) are eligible for sampling (ELIG_FLAG=Y).

The 202 listings with codes 1, 1, 2 (that is, a single house/compound of households; not vacant and not under construction; no persons living in the structure) are also eligible for sampling since they appear to be dwelling units (DUs) with no one living there at the time of listing, but could potentially have occupants at the time of interview.

The 2,691 listings with codes 1, 2, 1 (a single house/compound of households; vacant; persons living in the structure), though apparently contradictory appear to be potential households, and were made eligible for sampling.

The 1,279 listings with codes 1, 2, 2 are "vacants" with no one currently living there, but could have residents at the time of interview, and so they were considered eligible for sampling.

The 247 cases with codes 1, 3, 1 are currently under construction, but appear to have people living there and were also considered eligible for sampling.

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

Eligibility (ELIG_FLAG)	Structure Type (STOBS_D)	Structure vacant or under construction? STVAC_D	Anyone living in the structure? RESYN_D	Number of Records
2 erroneously listed replacement EAs for which the originally-sampled EA was listed and retained in the sample				**363**
N	1	1	1	178
N	1	1	2	2
N	1	2	1	4
N	1	3	2	2
N	2	1	1	175
N	2	1	2	1
N	5	1	2	1
2 Out of Scope EAs				**174**
N	1	1	1	150
N	1	2	2	2
N	2	1	1	10
N	3	2	2	2
N	5	2	2	1
N	6	1	1	6
N	6	2	1	1
N	6	2	2	2
2 "Small" EAs which were erroneously replaced				**121**
N	1	1	1	94
N	1	1	2	1
N	1	2	1	18
N	1	2	2	4
N	3	1	2	1
N	3	2	2	1
N	6	2	2	2
All other records coded as "Discarded" in the master listing file				**44**
N	1	1	1	15
N	1	2	1	2
N	2	1	1	21
N	2	1	2	1
N	3	1	1	2
N	4	3	2	1
N	6	1	1	2
Other ineligible listings in eligible EAs				**2,336**
N				1
N	1	3	2	378
N	2	3	2	28
N	3	1	2	71
N	3	2	2	455
N	3	3	2	26
N	4	1	2	4
N	4	2	1	2
N	4	2	2	8
N	5	3	1	1
N	5	3	2	8
N	6	2	2	1,016
N	6	3	1	7
N	6	3	2	70
N	7	2	2	78
N	7	3	2	2
N	8	2	1	1
N	8	2	2	135
N	8	3	1	5
N	8	3	2	40

Eligibility (ELIG_FLAG)	Structure Type (STOBS_D)	Structure vacant or under construction? STVAC_D	Anyone living in the structure? RESYN_D	Number of Records
TOTAL ELIGIBLE LISTINGS				**57,543**
Y	1	1	1	44,017
Y	1	1	2	202
Y	1	2	1	2,691
Y	1	2	2	1,279
Y	1	3	1	247
Y	2	1		10
Y	2	1	1	7,957
Y	2	1	2	160
Y	2	2	1	21
Y	2	2	2	17
Y	2	3		7
Y	2	3	1	15
Y	3	1	1	32
Y	3	2	1	1
Y	4	1	1	5
Y	5	1	1	28
Y	5	1	2	61
Y	5	2	1	3
Y	5	2	2	227
Y	6	1	1	195
Y	6	1	2	264
Y	6	2	1	29
Y	7	1	1	8
Y	7	1	2	23
Y	8	1	1	26
Y	8	1	2	18
TOTAL NUMBER OF RECORDS IN THE MASTER LISTING FILE				**60,581**

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

Structure Type (STOBS_D)

1 – single House/compound of hh
2 – apartment bldg./gated comm.
3 – church/mosque/temple
4 – Community center
5 – School/University
6 – Shop/business ctr/commerce bldg.
7 – Clinic/hospital/dr.office
8 - 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

Program Code Used to Create Household, Interview, and Blood Test Response Status

Program Code Used to Create 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 SAS Code for HH_STATUS

```
attrib HH_eligible length=3 label="Household Eligibility flag – will be used to create HH_STATUS_0";

if STARTINT='1'  and TAPGOOD='1'  and RESULTNDT=" " then HH_eligible = 1; /* Complete */
else if STARTINT='1'  then HH_eligible = 2;  /* Partial complete */
else if STARTINT='2' and RESULTNDT in ('3','5')    then HH_eligible = 3;   /* Eligible NR */
else if STARTINT='2' and RESULTNDT in ('6','7')    then HH_eligible = 4;   /* Known Ineligible */    else if STARTINT='2' and RESULTNDT in ('8', '10')  then HH_eligible = 5; /* Unknown Ineligible */

 attrib HH_STATUS_0 length=3 label="Intermediate HH disposition code";

if HH_eligible = 1 then HH_STATUS_0=1; /*Eligible Respondent*/
else if HH_eligible in (2,3) then HH_STATUS_0=2;/*Eligible NonRespondent*/
else if HH_eligible = 4     then HH_STATUS_0=3; /* Ineligible */
else if HH_eligible = 5     then HH_STATUS_0=4; /*Unknown eligibility Status*/

if HH_ELIGIBLE = 2 and ROSTERCOUNT > 0 then HH_STATUS_0 = 1;    /* Eligible Respondent */
```

if HH_ELIGIBLE = **5** and UPCODE_STAT_HH in (**2**,**3**,**4**) then HH_STATUS_0 = UPCODE_STAT_HH;

```
if EA_HHID_FIXED in ('080807110081012034', '050504079212011036') then do;
    HH_STATUS_0 = 4;
     HH_eligible = 5;
end;
```

Notes:

The variable ROSTERCOUNT is created earlier in the program; it counts the number of non-empty individual records on the roster file for each value of EA_HHID_FIXED. Households with no questionnaire record but with at least one valid roster record are eligible respondent households.

The variable UPCODE_STAT_HH is created based on the text in RESULTNDTOTHR. The DM team, the ICAP team and the statistical team all contributed to evaluating the text comments and assigning codes based on the text. It is used to assign nonresponding households where RESULTNDT = 10 "other, specify" into the three categories of households with no response.

B.2 SAS Code for INDIV_STATUS

```
label indiv_status   = "Individual Response Status"
indiv_qxstatus = "Completion of questionaire";

    indiv_status = 0;
    indiv_qxstatus = 0;

if (INDIV_AGEGROUP = 1 and
    (CH_KIDAGEY       => 0 or
    CH_KIDGENDER     => "0" or
    CH_KIDENROLL     => "0" or
    CH_KIDHIVTESTEVR => "0" or
    CH_KIDSLAST12UN  => "0" or
    ch_KIDVISTTBCLIN => "0" or
    CH_KIDHEPB       => "0")) then indiv_qxstatus = 1;
else
```

```
            if indiv_agegroup=2 and icnsnt="1" and indfinrslt in ("1","2") and ADOLTSEND ^= .
then indiv_qxstatus = 1;
            else
            if (indiv_agegroup =3 and icnsnt = "1" and indfinrslt in ("1","2")) then do;
 two_flag = 0;
            do i = 1 to 12;
                if miles(i) = "2" then two_flag = 1;
            end;
            if two_flag = 0 then indiv_qxstatus = 1;
      end;

if sleephere = "2" then indiv_status = 9;
    else
     if indiv_nonelig_reason in (5,10) then indiv_status = 8;
     else
      if indiv_nonelig_reason in (6,9) then indiv_status = 2;
      else
       if in_indiv = . and indiv_elig_check = 1 then indiv_status = 2;
       else
        if in_indiv = . and AGEYEARS = . and Sleephere = " " then indiv_status = 7;
        else
         if upcase(hiv1statusfinalsurvey) in ("NEGATIVE", "POSITIVE") then indiv_status = 1;
         else
          if indiv_qxstatus = 1 then indiv_status = 1;
          else
           indiv_status = 2;
```

B.3 SAS Code for BT_STATUS

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

ELSE BT_STATUS=**2**;

APPENDIX C

CHAID Trees and Definition of Final Nonresponse-Adjustment Weighting Cells

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 a total of eight PDF files corresponding to the three groups for which the CHAID analysis was conducted for adjustment of the interview weights (Section 3.4.3.2) and the five groups for which the CHAID analysis was conducted for adjustment of the blood test weights (Section 3.4.4.2). The names of the eight 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 subsets that varied 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 15-59 years)

TN_INDIV_STATUS.pdf (Adolescents 10-14 years)

CH_INDIV_STATUS.pdf (Children 0-9 years)

Blood Test

AM_BTEST.pdf (Males 15-59 years)

AF_BTEST.pdf (Females 15-59 years)

TM_BTEST.pdf (Males 10-14 years)

TF_BTEST.pdf (Females 10-14 years)

C_BTEST.pdf (Children 0-9 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 eight 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). Cells that were collapsed to control the variation in weights are highlighted.

Individual Interview

Zam_AD_INDIV.xlsx (Persons 15-59 years)

Zam_TN_INDIV.xlsx (Adolescents 10-14 years)

Zam_CH_INDIV.xlsx (Children 0-9 years)

Blood Test

Zam_AM_BT.xlsx (Males 15-59 years)

Zam_AF_BT.xlsx (Females 15-59 years)

Zam_TM_BT.xlsx (Males 10-14 years)

Zam_TF_BT.xlsx (Females 10-14 years)

Zam_CH_BT.xlsx (Children 0-9 years)

APPENDIX D

Adult Violence Module Variables, Eligibility Criteria, and Program Code

Adult Violence Module Variables, Eligibility Criteria, and Program Code

D.1 Variables Included in the Adult Violence Module

Variable	Question Text
touchtimes	How many times has anyone ever touched you in a sexual way without your permission, but did not try and force you to have sex?
touchtimesdk	Please provide the reason this previous question was left blank: How many times has anyone ever touched you in a sexual way without your permission, but did not try and force you to have sex?
touchage	How old were you the first time this happened?
touchagedk	Please provide the reason this previous question was left blank: How old were you the first time this happened?
cmplsxtimes	How many times in your life has anyone tried to make you have sex against your will but did not succeed? This includes someone using harassment, threats, tricks, or physical force.
cmplsxtimesdk	Please provide the reason this previous question was left blank: How many times in your life has anyone tried to make you have sex against your will but did not succeed? This includes someone using harassment, threats, tricks, or physical force.
cmplsxage	How old were you the first time someone tried to make you have sex against your will but did not succeed?
cmplsxagedk	Please provide the reason this previous question was left blank: How old were you the first time someone tried to make you have sex against your will but did not succeed?
frcsxtimes	How many times in your life have you been physically forced to have sex?
frcsxtimesdk	Please provide the reason this previous question was left blank: How many times in your life have you been physically forced to have sex?
frcsxage	How old were you the first time someone physically forced you to have sex?
frcsxagedk	Please provide the reason this previous question was left blank: How old were you the first time someone physically forced you to have sex?
frcsx12mo	In the past 12 months, did someone physically force you to have sex?
frcsx12mopt	In the past 12 months, did a partner physically force you to have sex?
prssxtimes	How many times in your life has someone pressured you to have sex through harassment, threats and tricks and did succeed?
prssxtimesdk	Please provide the reason this previous question was left blank: How many times in your life has someone pressured you to have sex through harassment, threats and tricks and did succeed?
prssxage	How old were you the first time someone pressured you to have sex and did succeed?
prssxagedk	Please provide the reason this previous question was left blank: How old were you the first time someone pressured you to have sex and did succeed?
prssx12mo	In the past 12 months, did someone pressure you to have sex and did succeed?
prssx12mopt	In the past 12 months, did a partner pressure you to have sex and did succeed?
uwntsxhelp	After any of these unwanted sexual experiences, did you try to seek professional help or services from any of the following?
unwntsxnohlp	What was the main reason that you did not try to seek professional help or services?
vlnc	Has anyone ever done any of these things to you: - Punched, kicked, whipped, or beat you with an object - Slapped you, threw something at you that could hurt you, pushed

Variable	Question Text
	you or shoved you - Choked, smothered, tried to drown you, or burned you intentionally - Used or threatened you with a knife, gun or other weapon?
vlncfrstage	How old were you the first time one of these things happened to you?
vlncfrstagedk	Please provide the reason this previous question was left blank: How old were you the first time one of these things happened to you?
vlnc12motimes	In the past 12 months, how many times did someone: - Punched, kicked, whipped, or beat you with an object - Slapped you, threw something at you that could hurt you, pushed you or shoved you - Choked, smothered, tried to drown you, or burned you intentionally - Used or threatened you with a knife, gun or other weapon?
vlnc12moptnr	In the past 12 months, did a partner do any of these things to you?
seekhelp	Thinking about all these experiences that we just discussed, whether someone has done the following: - Punched, kicked, whipped, or beat you with an object - Slapped you, threw something at you that could hurt you, pushed you or shoved you - Choked, smothered, tried to drown you, or burned you intentionally - Used or threatened you with a knife, gun or other weapon Did you try to seek professional help or services for any of these incidents from any of the following?
seekhelpwhynot	What was the main reason that you did not try to seek professional help or services?

D.2 Eligibility Criteria for the Violence Module

One female 15-59 years of age was randomly selected to answer questions in the violence module from each household that contained one or more such females. The variable VM_STATUS was created to identify individuals eligible to receive the violence module and was assigned to every rostered record, with values as shown in the table below. Codes 1 through 9 were assigned only to cases flagged to receive the violence module.

VM_STATUS	Description
0	Not selected for Violence Module
1	Violence Module Respondent
2	In-scope for Violence Module, Non-Respondent
3	Out of scope for Violence Module, changed to male in Interview
4	Out of scope for Violence Module, changed age out of age range for Violence Module in Interview
5	No data, unknown whether eligible for survey
6	Collected in Another Tablet
7	Rostered in Error
8	Not Sampled (adults over the age limit of participation for the country and children in households with child flag = NO)
9	Extraneous Cases – De Jure Ineligible

D.3 Code to Define Violence Module Status (VM_STATUS)

```
DATA HH_QX;
        LENGTH EA_HHID_VIOL $20;
        LENGTH VIOLFLAG_X $2;
        SET w11.HH_QX(KEEP=EA_HHID_FIXED CHILDFLAG VIOLFLAG);

        VIOLFLAG_X = PUT(INPUT(VIOLFLAG, 3.), Z2.);

        IF VIOLFLAG ^= ' ' THEN DO;
                EA_HHID_VIOL = EA_HHID_FIXED||VIOLFLAG_X;
        END;
RUN;

DATA ROSTER;
```

```
        SET w11.ROSTER;

        IF AGEYEARS < 15  THEN ROSTER_VIOL_AGECAT = 1;  /* Roster age less than 15
        */
        ELSE IF 14 < AGEYEARS <= 59 THEN ROSTER_VIOL_AGECAT = 2;  /* Roster age
        BETWEEN 15 - 59 */

        LABEL ROSTER_VIOL_AGECAT = "Violence weighting age categories from Roster
        Age";
RUN;

PROC SORT DATA=HH_QX; BY EA_HHID_FIXED; RUN;
PROC SORT DATA=ROSTER; BY EA_HHID_FIXED; RUN;

DATA NEW_ROSTER;
        MERGE ROSTER (IN=AA) HH_QX (IN=BB);
        BY EA_HHID_FIXED;

        LABEL VM_FLAG = "Adult Female age 15 to 59 Selected for Violence Module"

        VM_FLAG = 0;

        IF AA AND BB then do;
                IF ROSTER_VIOL_AGECAT = 2 THEN DO;
                        IF EA_HHID_LN_FIXED=EA_HHID_VIOL THEN VM_FLAG = 1;
                END;
        END;
        ELSE IF AA THEN OUTPUT;
RUN;

DATA INDIV;
        SET w30.W30_indiv_qx_reduced;

        IF (TOUCHTIMES >= 0 AND CMPLSXTIMES >= 0 AND FRCSXTIMES >= 0 AND
        PRSSXTIMES >= 0) OR compress(VLNC) in ('1','2') THEN VM_QXSTATUS = 1;
        ELSEVM_QXSTATUS = 0;
RUN;

PROC SORT DATA=NEW_ROSTER; BY EA_HHID_LN_FIXED; RUN;
PROC SORT DATA=INDIV; BY EA_HHID_LN_FIXED; RUN;

DATA INDIV w31.W31_viol;
        MERGE INDIV(IN=A) NEW_ROSTER(KEEP=EA_HHID_LN_FIXED VM_FLAG
        ROSTER_VIOL_AGECAT);
        BY EA_HHID_LN_FIXED;
        IF A;
```

```
        Label INDIV_VIOL_AGEGROUP = "Violence age group from Best Age";

        INDIV_VIOL_AGEGROUP = 0;
        IF INDIV_AGEGROUP = 3THEN INDIV_VIOL_AGEGROUP = 2; /* Adult (15 - 59)
        */
        ELSE IF INDIV_AGEGROUP in (1,2) THEN INDIV_VIOL_AGEGROUP = 1;/*
        Child/Adolescent (0-14) */
        ELSE IF INDIV_AGEGROUP = 4THEN INDIV_VIOL_AGEGROUP = 3; /* Seniors
        60 and older */

        IF VM_FLAG = 0 THEN VM_STATUS = 0; /* Not selected for Violence Module */
        ELSE IF INDIV_STATUS NOT IN (1, 2) THEN VM_STATUS = INDIV_STATUS;
        /* others */
        ELSE IF BEST_GENDER ^= '2' THEN VM_STATUS = 3; /* Out of scope for Violence
        Module, changed to male in Interview */
        ELSE IF INDIV_VIOL_AGEGROUP ^= 2THEN VM_STATUS = 5;/* Out of scope for
        Violence Module, changed age out of 15 - 59 in Interview */
        ELSE IF VM_QXSTATUS = 1    THEN VM_STATUS = 1; /* Violence Module
        Respondent */
        ELSE   VM_STATUS = 2; /* In-scope for Violence Module, Non-Respondent */
RUN;
```

APPENDIX E

HIV Knowledge Module Variables, Eligibility Criteria, and Program Code

HIV Knowledge Module Variables, Eligibility Criteria, and Program Code

E.1 Variables Included in the HIV Knowledge Module

NAME	LABEL
ONEPARTNR	Can the risk of HIV transmission be reduced by having sex with only one uninfected partner who has no other partners?
MOSQUITO	Can a person get HIV from mosquito bites?
CONDOMS	Can a person reduce their risk of getting HIV by using a condom every time they have sex?
SHAREFOOD	Can a person get HIV by sharing food with someone who has HIV?
HEALTHYINF	Can a healthy-looking person have HIV?
BUYFOOD	Would you buy fresh vegetables from a shop keeper or vendor if you knew the person had HIV?
KIDSSCHOOL	Do you think children living with HIV should be allowed to attend school with children who do not have HIV?
FEARTEST	Do you think people hesitate to take an HIV test because they are afraid of how other people will react if the test result is positive for HIV?
TALKBAD	Do people talk badly about people living with HIV or who are thought to be living with HIV?
RESPECT	Do people living with HIV, or thought to be living with HIV, lose the respect of other people?
SALIVA	Do you fear that you could get HIV if you come into contact with the saliva of a person living with HIV?
FAMSHAME	Do you agree or disagree with the following statement: I would be ashamed if someone in my family had HIV.

E.2 Eligibility Criteria for HIV Knowledge Module

Each interviewed adult 15-59 years of age had an independent probability of selection of 50% for the HIVK module, regardless of the number of other adults in the household. HIV Knowledge respondents are those interviewed individuals selected for the HIVK module with a valid answer to the HIVK question, MOSQUITO ("Can a person get HIV from mosquito bites?"). The valid answers are "Yes = 1", "No = 2", and "Don't Know = 3". The variable HIVK_STATUS was created to identify individuals eligible to receive the HIVK module and was assigned to every rostered record, with values as shown in the table below. Codes 1 through 9 were assigned only to cases flagged to receive the HIVK module.

HIVK_STATUS	Description
0	Not selected for HIVK Module
1	HIVK Module Respondent
2	HIVK Module Eligible Non-Respondent
4	Unknown Survey Eligibility
6	Collected in Another Tablet
7	Rostered in Error
8	Not Sampled (adults over the age limit of participation for the country, 59, and children in households with child flag = NO)
9	Extraneous Cases – De Jure Ineligible

E.3 Program Code for HIVK Response Status

```
data eligibles (keep = ea_hhid_ln_fixed hivk_status mosquito);
set “data set containing all individual records”;
  where 15 <= confagey_RECODE <= 59 and
        indiv_hivkflag= "1" and
          indiv_status = 1;

if MOSQUITO   in ("1","2","3") then HIVK_STATUS = 1;
else
  if MOSQUITO   in ("-9"," ") then HIVK_STATUS = 2;
run;

proc sort data = eligibles (drop = mosquito);
  by ea_hhid_ln_fixed;
run;

proc sort data = “data set containing all individual records”;
  by ea_hhid_ln_fixed;
run;

data W32_HIVK;
merge eligibles(in=a“data set containing all individual records”( (in=b);
  by ea_hhid_ln_fixed;
  if b;
  if b and not a then HIVK_STATUS=0;
run;
```

```
data w32.w32_hivk;
set w32_hivk;
if indiv_status => 3 then hivk_status = indiv_status;
run;
```

APPENDIX E

Eligibility Criteria and Program Code for Weight and Height Measurements

Eligibility Criteria and Program Code for Weight and Height Measurements

F.1 Eligibility Criteria for Weight and Height Measurements

The variable CWH_STATUS was created to identify children eligible to receive weight and height measurements and was assigned to children 0-60 months old who had a blood test weight, with values as shown in the table below.

CWH_STATUS	Description
1	Provided W/H measurements
2	Did not provide W/H measurements
.	Not selected for W/H measurements

F.2 Program Code for Response Status for Weight and Height Measurements

```
DATA CWH;
   SET W100.Blood_delivery;

   IF CONFAGEM not in (' ' 'AGE NOT RECORDED')
              THEN CONFAGEM_r = CONFAGEM+0;

            IF HIVNEGWH not in (' ')
              THEN HIVNEGWH_r = HIVNEGWH + 0;

            CWH_FLAG = 0;
            IF HIVNEGWH_r => 95 OR HIVSTATUS = 1 OR HIVSTATUSC = 1
              THEN CWH_FLAG = 1;

   IF 0<= CONFAGEM_r <=60 AND BTWT0 > 0;

   RUN;
```

```
DATA FRM;
   SET CWH (RENAME=(CWHHEIGHT=CWHHEIGHT_A
CWHWEIGHT=CWHWEIGHT_A ));

   CWHHEIGHT=INPUT(CWHHEIGHT_A,8.2);
   CWHWEIGHT=INPUT(CWHWEIGHT_A,8.2);

   IF     CWH_FLAG=1 and CWHHEIGHT ^= . AND CWHWEIGHT ^= . THEN
CWH_RESP = 1;
   ELSE IF CWH_FLAG=1 and (CWHHEIGHT = . OR CWHWEIGHT = .)  THEN
CWH_RESP = 2;
   ELSE       CWH_RESP = .;

   RUN;
```

Appendix G

Child module weight creation and eligibility criteria

G.1 Purpose of the child module weights

As described in Section 2.4.5, a subset of all sampled households was randomly selected for additional child data collection. In these selected households, children were eligible for blood testing, and additional interview questions were asked either of the child (for adolescents) or the parent/guardian. In other households this additional data was not collected. The blood test and interview weights (btwt and intwt, respectively) on the child biomarker and individual datasets allow for analysis of the variables only collected in the households selected for additional child data collection.

Although the information available for children in the selected households is more detailed, questions included in the child module of the adult interview were administered to parents and guardians of all children in the household. The household roster also contains information about all children in the household. If an analysis aims to use these data, the sample population is different: specifically, this sample includes all rostered children who would have been eligible to participate, irrespective of whether their household was flagged for child data collection. In these situations, a separate set of weights is needed. These are referred to hereafter as child module weights.

G.2 Child module weight creation process

Three main steps were carried out to create the child module weights:

1. Create a list of all children aged 0-14 rostered in any responding household who were de facto eligible (i.e., slept in the household the night before) and had a responding parent or guardian, and link each child to their parent or guardian using the line number of the responding adult in the household (parentguardqx variable in the child interview dataset).
2. Assign each child an initial weight equal to the linked adult's non-response adjusted (but not post-stratified) interview weight (trmpnr1w0 from the intermediary weights file). We refer to this weight as the child module base weight, chmodbw0.
3. Post-stratify the resulting set of weights to ensure that the total populations by five-year age group and gender sum to the control totals used for the blood test and interview weights. We refer to the resulting weight as the child module final weight, chmodfw0.

In step one, individuals in the child dataset were included as possible guardians, because there can be cases where someone under 15 years of age responded as the parent or guardian of another child in the household. Records for children who would not have been eligible for the survey were excluded.

Potentially eligible children have indstatus = 1 or 8 (see section E.5 below for full details on the eligibility criteria).

In step two, if the adult did not respond or was deemed ineligible for some reason (for example, if they did not stay in the household the previous night), their interview weight was set to zero, so their associated children will also have a child module weight of zero.

The post-stratification in step three used an adjustment factor that was computed for each cell defined by gender and five-year age group of the rostered children. This adjustment factor is equal to the control total in each cell divided by the sum of the chmodbw0 weights of the children in the cell. Each child's chmodbw0 weight in the given cell was multiplied by the corresponding adjustment factor to obtain the final weight, chmodfw0.

Steps two and three were repeated for each replicate weight set (trmpnr1w001-trmpnr1wXXX) to create the associated jackknife replicate weights for the child module. First, the child module replicate base weights were computed as chmodbw001 = trmpnr1w001, chmodbw002 = trmpnr1w002, ..., chmodbwXXX = trmpnr1wXXX. Each set of jackknife replicate weights was then used to compute the corresponding replicate-specific post-stratification adjustment factors and final post-stratified replicate weights, chmodfw001, chmodfw002, ..., chmodfwXXX.

G.3 Variables available for all children and when to use these weights

The child module weights should only be used when the analysis variables are collected for all rostered children (i.e., eligibility for data collection is not restricted to whether the household was flagged for child data collection). In general, this includes variables from the roster, such as age and gender, as well as questions from the adult questionnaire's children module that have been attached to the child records. These variables can be identified by filtering the variable category in the child interview dataset codebook to "Adult questionnaire - Module 3A: children" (note that the module number may vary by country). Most of these variables have the prefix "ch_" in their variable names to assist with identification. Additional information about the mother is available for linked children in the variables prefixed "mom". Questions from the "Household questionnaire – Child" category are also available for all children because these are completed by the head of household.

Variables which are asked in the adolescent interview or related to blood testing are not available for children in non-selected households, so the child module weights should not be used for these.

G.4 Further non-response adjustments

The child module weights are general-purpose weights which are a reasonable approximation of the weights that would be obtained through a more complex non-response adjustment procedure like that used for the main child interview weights. A major assumption is that the non-response pattern for children is captured fully by the non-response adjustments carried out for the linked adults. It is possible that these non-response adjustments do not fully account for some specific characteristics of the child. For example, older children may tend to have more missing data than younger ones, and missing parent/guardian links may occur at different rates for different ages or other groups of children. To more fully compensate for these patterns a precise definition of response status for children would have to be developed based on the questions answered, and non-response adjustments applied to relevant response cells based on child-level characteristics. For highly detailed or specialized analysis we recommend that the non-response patterns be checked for the particular groups of interest for the analysis to determine whether any further adjustments may be needed.

G.5 Child module weight eligibility criteria

The following table shows all combinations of values for variables defining eligibility for child module weights. Children who were unable to be linked to an adult (linked adult indstatus = .) or whose linked adult was not an eligible respondent (linked adult indstatus = 2, 7, 8, or 9) are ineligible. Among children who had an eligible, responding, linked adult, those with indstatus = 2, 6, 7, 9 were also ineligible (2 = non-responding sampled child, 6 & 7 = were duplicated or erroneous child records, 9 = de jure ineligible).

Only those children in rows 1 and 5 below in the table, with indstatus = 1 or 8, linked adult indstatus = 1, and sleephere = 1, are assigned child module weights.

Table G-1 Variables determining child module weight eligibility criteria

Linked adult's indstatus	Child's indstatus	Child's sleephere	Explanation of child eligibility status
1	1	1	Eligible: Sampled child with responding adult. These children have valid individual weights (intwt)
1	2	1	Ineligible: Sampled child with linked adult, but considered non-respondent (e.g., parent refused consent or did not provide sufficient data)
1	6	1	Ineligible: the child record was collected in another tablet
1	7	1	Ineligible: the child was rostered in error
1	8	1	Eligible: Child with linked, responding adult, in a household not sampled for child blood testing. These children do not have individual weights (intwt) but are eligible for child module weights (chmodfwt)
1	9	1	Ineligible: Non-de facto child. The adult was an eligible respondent, but the child had ind0040 = 3 (not available), 6 (incapacitated), or did not sleep in HH the night before.
2,7,8,9	1	1	Ineligible: Ineligible or non-responding linked adult
2,7,8,9	2	1	Ineligible: Ineligible or non-responding linked adult
2,7,8,9	6	1	Ineligible: Ineligible or non-responding linked adult
2,7,8,9	7	1	Ineligible: Ineligible or non-responding linked adult
2,7,8,9	8	1	Ineligible: Ineligible or non-responding linked adult
2,7,8,9	9	1	Ineligible: Ineligible or non-responding linked adult
.	2	1	Ineligible: Not able to be linked to an adult
.	8	1	Ineligible: Not able to be linked to an adult
.	9	2	Ineligible: Not able to be linked to an adult