THIS
TANZANIA HIV IMPACT SURVEY
A DROP THAT COUNTS

SAMPLING AND WEIGHTING TECHNICAL REPORT
THIS 2016-2017

DECEMBER 2018

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

Section **Page**

Appendix **Page**

Acronyms

CDC	US Centers for Disease Control and Prevention
CHAID	Chi-square Automatic Interaction Detector
CI	Confidence Interval
CV	Coefficient of Variation
DEFF	Design Effect
DHS	Demographic and Health Survey
DU	Dwelling Unit
EA	Enumeration Area
FTP	File Transfer Protocol
HH	Household
HIV	Human Immunodeficiency Virus
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
THIS	Tanzania HIV Impact Survey
THMIS	Tanzania HIV/AIDS and Malaria Indicator Survey
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

Introduction 1

The 2016 Tanzania HIV Impact Survey (THIS) is a Population-based HIV Impact Assessment (PHIA) designed to assess the prevalence of key human immunodeficiency virus (HIV)-related health indicators. Data collection for the THIS was conducted between November 2016 and August 2017, and included over 43,000 individuals in approximately 15,000 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 Tanzania HIV Impact Survey (THIS) is a stratified multistage probability sample design, with strata defined by the 31 regions of the country, first-stage sampling units defined by enumeration areas (EAs) within strata, second-stage sampling units defined by households within EAs, and finally eligible persons within households. Within each region, the first-stage sampling units (also referred to as "primary sampling units" or PSUs) were selected with probabilities proportionate to the number of households in the PSU based on the 2012 Population and Housing Census. The allocation of the sample PSUs to the 31 regions was made in a manner designed to achieve specified precision levels for (a) national estimates of HIV incidence rates, and (b) regional 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 self-weighting (i.e., equal probability) samples within each region to the extent feasible.

Within the sampled households, all eligible adults 15 years of age or older were included in the study sample for data collection. All eligible children 0-14 years of age in one third of the sampled households were included in the study for data collection.

Details of sample design employed for the THIS 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.
- Adjustment for PSU nonresponse to compensate for PSUs for which no household data were collected, if necessary.
- 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 for the THIS are provided in Sections 3 and 4.

Sample Design 2

2.1 Population of Inference

The population of inference for the THIS is comprised of individuals 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 THIS data collection; however, as discussed later in Section 2.5, only members of the *de facto* population are included in the THIS study population. Table 2-1 summarizes projections of the 2017 Tanzania *de facto* population by gender and age group.

Table 2-1 Summary of 2017 population projections for Tanzania by gender and age group

Age group	Gender		Total
	Male	Female	
14 years or younger	12,950,761	12,786,956	25,737,717
15 to 49 years	12,955,164	13,370,081	26,325,245
50 years or older	2,436,044	2,811,010	5,247,054
Total	28,341,969	28,968,047	57,310,016

Source: United Nations, Department of Economic and Social Affairs, Population Division (2015). World Population Prospects: The 2015 Revision, custom data acquired via website. https://esa.un.org/unpd/wpp/DataQuery/

2.2 Precision Specifications and Assumptions

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

Specifications

- The relative standard error (RSE) of the national estimate of annual HIV incidence among persons aged 15-49 should be 40% or less.

- 95% confidence interval bounds of ±0.10 or less for an estimated VLS rate among all HIV+ adults aged 15 to 49 years old in 10 high prevalence regions (i.e., regions with HIV prevalence of 5% or higher).
- For the five Zanzibar regions, 10 PSUs are to be selected for the Urban West region, and four PSUs for each of the remaining regions.
- A total overall sample size (including adults 15-49, adults 50+, and children 0-14 years of age) of approximately 40,000 analyzable blood draws.

Assumptions

- An overall HIV prevalence rate of 0.051 (5.1%) for adults 15-49 years of age that varies by region (see Table 2-2). Source: 2011-12 Tanzania HIV/AIDS and Malaria Indicator Survey (THMIS).
- An annual national HIV incidence rate for adults aged 15-49 of $P_a = 0.0037$ (0.37%). Source: 2012 UNAIDS estimate.
- A mean duration of recent infections (MDRI) of 130 days, yielding an annualization rate of 365/130= 2.8077. Hence, the estimated incidence rate for MDRI = 130 days is $P_m = 0.0037/2.8077 = 0.0013$ (0.13%).
- A viral load suppression (VLS) rate among HIV+ adults aged 15-49 in each region of $P_{vh} = 0.50$ (50%). This is a conservative assumption because it will overstate the actual variance of the VLS rate.
- An average of 30 occupied sampled households per sampled cluster (PSU).
- An intra-cluster correlation (ICC) of 0.05 for prevalence and VLS rates. The ICC provides an average measure of the homogeneity of responses within the first-stage sampling units.
- An occupancy rate of 97.4% for sampled dwellings. 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. Source: 2011-12 THMIS.
- An overall household response rate of 98.2% among occupied households. Source: 2011-12 THMIS.
- The average number of persons aged 15 to 49 per household is 2.07. Source: 2012 population and housing census.
- The percentage of persons in households who are 0-14 is 43.9%. Source: 2012 population and housing census.
- The percentage of persons in households who are 50+ years of age is 10.0%. Source: 2012 population and housing census.

- Among the eligible individuals 15+ years of age in households completing the household roster, a biomarker response rate of 80.0%. Source: Conservative assumption derived from the 2011-12 THMIS.
- Among the eligible children 0-14 years of age in the households designated for child data collection, a biomarker response rate of 75.0%. This value is the corresponding biomarker response rate for adults minus 5%.

Based on the specifications and assumptions listed above, a sample of 526 clusters (EAs) was determined to be the minimum needed to meet the specified precision goals. The allocation of the sample to the 31 regions (strata) of Tanzania is shown in Table 2-2. The expected numbers of households included in the study and the corresponding projected numbers of respondents by age group are also summarized in Table 2-2. The actual numbers of respondents achieved are presented in Sections 2.4 and 2.5 and differ from the counts in Table 2-2 because of differences between the response rates and other assumptions used to develop the sample design and those obtained during data collection. Further details about the sampling of households are given in Section 2.4.

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

Region code	Stratum (Region)	Est. HIV prevalence rate [1]	Sample clusters (EAs)	Target no. dwelling units to be sampled	Exp. no. households [2]	Projected number of respondents [3]		
						Adults 15-49	Adults 50+	Children 0-14 [4]
1	Dodoma	0.0284	11	339	330	537	127	520
2	Arusha	0.0318	9	277	270	439	104	426
3	Kilimanjaro	0.0379	10	308	300	488	115	473
4	Tanga	0.0237	10	308	300	488	115	473
5	Morogoro	0.0375	14	431	420	683	161	662
6	Pwani	0.0587	37	1,139	1,110	1,806	426	1,751
7	Dar es Salaam	0.0683	32	985	960	1,562	368	1,514
8	Lindi	0.0290	6	185	180	293	69	284
9	Mtwara	0.0409	8	246	240	391	92	379
10	Ruvuma	0.0693	32	985	960	1,562	368	1,514
11	Iringa	0.0905	25	770	750	1,220	288	1,183
12	Mbeya [5]	0.0893	26	801	780	1,269	299	1,230
13	Singida	0.0323	7	216	210	342	81	331
14	Tabora	0.0510	41	1,262	1,230	2,001	472	1,940
15	Rukwa	0.0616	35	1,078	1,050	1,709	403	1,656
16	Kigoma	0.0340	11	339	330	537	127	520
17	Shinyanga	0.0732	30	924	900	1,464	345	1,419
18	Kagera	0.0476	16	493	480	781	184	757
19	Mwanza	0.0418	19	585	570	927	219	899
20	Mara	0.0441	12	370	360	586	138	568
21	Manyara	0.0148	7	216	210	342	81	331
22	Njombe	0.1467	18	554	540	879	207	852
23	Katavi	0.0588	37	1,139	1,110	1,806	426	1,751
24	Simiyu	0.0356	9	277	270	439	104	426
25	Geita	0.0465	12	370	360	586	138	568
26	Songwe [5]	0.0893	26	801	780	1,269	299	1,230
51	Kaskazini Unguja	0.0030	4	123	120	195	46	189
52	Kusini Unguja	0.0030	4	123	120	195	46	189
53	Mjini Magharibi	0.0140	10	308	300	488	115	473
54	Kaskazini Pemba	0.0030	4	123	120	195	46	189
55	Kusini Pemba	0.0030	4	123	120	195	46	189
	TOTAL	0.0507	526	16,197	15,780	25,677	6,053	24,888

[1] Source: 2011-12 Tanzania HIV/AIDS and Malaria Indicator Survey (HMIS)

[2] Assumes occupancy rate of 96.0%.

[3] Entries are projected counts based on the assumptions used to develop the sample design.

[4] All children 0-14 years of age in a random one-third sample of households (see Section 2.4.5).

[5] Region was split off from original Mbeya region.

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 THIS are defined to be the Enumeration Areas (EAs) created for the 2012 Tanzania Population and Housing Census. The 2012 sampling frame consisted of approximately 106,000 EAs containing an estimated 9.2 million households and 44.9 million persons.

2.3.2 Selection of the PSU Sample

A stratified sample of 526 EAs was selected from the final EA sampling frame in accordance with the sample allocation given in Table 2-2. The 31 strata specified for sampling were the 31 regions of Tanzania, including the five regions in Zanzibar. The EA samples were selected systematically and with probabilities proportionate to a measure of size (MOS) equal to the number of households in the EA based on the 2012 Population and Housing Census. Prior to selection, the EAs were sorted by type of EA (i.e., urban vs. rural), district within type of EA, ward within district, village within ward, and finally by EA within village. The sorting of the EAs prior to sample selection induces an implicit geographic stratification. 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.[1]

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 THIS because they were known to contain occupied dwelling units, but were inaccessible for various reasons (e.g., flooding, land disputes with the government). 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

[1] The EA sample was actually selected in two phases. Initially, a sample of 714 EAs had been selected based on precision requirements that were more stringent than those described in Section 2.2. However, the size of the sample was subsequently deemed to be impracticable with the available resources, and a decision was made to reduce the sample size. The final sample of 526 EAs was designed to satisfy the precision requirements listed in Section 2.2.

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.

In addition to the three EAs mentioned above, there was one EA in which all of the dwelling units had been demolished. This PSU is "out of scope" of the study since there are no households located within it. In general, substitution is not appropriate for out-of-scope (ineligible) EAs. Thus, there were 525 eligible EAs in the final sample for THIS.

2.3.4 Segmentation

Of the 526 sampled EAs, 147 were considered to be too large to be listed in their entirety. Thus, these 147 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 are described in the listing manual developed for the THIS.

2.4 Selection of Households

The selection of households for the THIS 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 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 (EA).

2.4.2 Listing

In essence, the listing process involves compiling complete, up-to-date, and accurate lists of all dwelling units and households for each sampled EA through a field operation using trained staff referred to as "listers." Local leaders and knowledgeable community members were consulted to assist in the listing process. For each of the 525 eligible EAs selected for the study, listers were provided with maps from which to delineate the boundaries of the EA, and to record the locations of the dwelling units/households found by the listers in the field. Information about the listed dwelling units/households was entered into computer tablets. The information recorded in the tablets included the address or description of the listed dwelling unit/household, the name of the head of household, the type of structure (house, apartment, compound, etc.), occupancy status, and GPS coordinates. Vacant structures were listed along with households in occupied dwelling units. Approximately 56,000 eligible dwelling units/households were listed for the THIS.

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 2-3 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 2-3 Distribution of records in listing file by type of record, eligibility status, and stratum

Region code	Stratum (Region)	Number of dwelling units/ house-holds in listing file	Number of discard-ed listings [1]	Number of unoccupied dwelling units [2]	Number of unoccupied dwelling units eligible for sampling [3]	Number of occupied dwelling units/ house-holds [4]	Number of occupied dwelling units/ house-holds eligible for sampling	Total number of dwelling units/ households eligible for sampling
1	Dodoma	1,498	0	16	16	1,482	1,482	1,498
2	Arusha	1,192	0	20	20	1,172	1,172	1,192
3	Kilimanjaro	1,035	0	65	65	970	970	1,035
4	Tanga	1,271	0	32	32	1,239	1,239	1,271
5	Morogoro	2,049	230	33	31	1,786	1,786	1,817
6	Pwani	3,784	0	42	41	3,742	3,742	3,783
7	Dar es Salaam	2,917	1	60	60	2,856	2,855	2,915
8	Lindi	663	0	16	16	647	647	663
9	Mtwara	935	0	39	39	896	896	935
10	Ruvuma	3,520	0	97	97	3,423	3,423	3,520
11	Iringa	2,351	0	44	44	2,307	2,307	2,351
12	Mbeya	2,708	0	75	75	2,633	2,633	2,708
13	Singida	810	0	6	6	804	804	810
14	Tabora	3,630	0	64	64	3,566	3,566	3,630
15	Rukwa	3,862	0	160	150	3,702	3,702	3,852
16	Kigoma	1,066	0	34	34	1,032	1,032	1,066
17	Shinyanga	3,496	0	55	55	3,441	3,441	3,496
18	Kagera	1,148	0	47	47	1,101	1,101	1,148
19	Mwanza	2,112	0	43	42	2,069	2,069	2,111
20	Mara	1,431	0	66	66	1,365	1,365	1,431
21	Manyara	1,277	0	23	22	1,254	1,254	1,276
22	Njombe	2,139	0	96	96	2,043	2,043	2,139
23	Katavi	4,091	0	51	51	4,040	4,040	4,091
24	Simiyu	730	0	50	50	680	680	730
25	Geita	1,185	0	13	13	1,172	1,172	1,185
26	Songwe	2,467	47	59	59	2,361	2,361	2,420
51	Kaskazini Unguja	303	0	1	1	302	302	303
52	Kusini Unguja	490	57	6	6	427	427	433
53	Mjini Magharibi	1,194	0	4	4	1,190	1,190	1,194
54	Kaskazini Pemba	325	0	0	0	325	325	325
55	Kusini Pemba	369	0	1	1	368	368	369
	TOTAL	56,048	335	1,318	1,303	54,395	54,394	55,697

[1] One record for which GPS coordinates were not obtained was also discarded, resulting in a total of 336 discarded records.

[2] Records coded as vacant, under construction, or with no residents at time of listing (see Appendix A).

[3] Subset of the unoccupied dwelling units that could potentially serve as residential quarters (see Appendix A).

[4] 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 equal probability samples of dwelling units within each of the 31 regions of the country, 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 2012 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 or less than the desired 30 households in EAs where the sampling measure of size differs from the actual listing count.

The calculation of the required within-EA sampling rates proceeded as follows. First, the target overall sampling rate for region h = 1, 2, ..., 31, was computed as:

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

where

T_h = target sample size for region h given in Table 2-2 ;
m_h = number of sample EAs in region h ;
N_{hi} = number of eligible dwelling units in PSU i in region h based on listing counts;
P_{hi} = probability of selecting PSU i in region h .

The total *targeted* number of listings to be selected across all 31 regions is $\sum_{h=1}^{31} T_h$ = 16,197 (see Table 2-2).[2] To obtain an equal probability sample within region 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 region 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 and very small workloads in others. To reduce the variation in workload across the sampled EAs, the maximum number of dwelling units to be selected in any EA was capped at 60, and the minimum number of dwelling units to be selected in any PSU was set to 15. The difference between

[2] The actual final sample count differed trivially due to randomness in selection (see Table 2-4).

the number of dwelling units that would have been selected using the rates, f_{hi}^{within}, and the specified maximum and minimum numbers was then re-distributed to the other EAs in the same region so as to maintain the desired total sample size for the region. The within-EA sampling rates, f_{hi}^{within}, were therefore adjusted to reflect the redistribution of the sample within the region. 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 $15 \leq N_{hi} A_{hi} f_{hi}^{within} \leq 60$ 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, village/neighborhood name, 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.

2.4.5 Results of Second-Stage Sampling

Table 2-4 summarizes the numbers 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 (DEFF) to be expected for the selected sample. The UEW design effect provides a measure of the increase in the variance of a sample-based estimate resulting from variable overall sampling fractions within a stratum (e.g., see Kish, 1965, page 403). With an equal probability sample within a stratum, the design effects would ordinarily equal 1.0. However, with the capping and redistribution of the sample described previously, the overall sampling rates (and, hence, household weights) will vary within a stratum (region). Despite the variation in weights, the UEW design effects are all very close to 1.0 (indicating minimal increase in variance due to unequal weighting) for all strata.

Table 2-4 Number of sampled dwelling units/households and expected unequal weighting design effects by region

Region code	Stratum (Region)	Number of sample PSUs (EAs)	Number of sampled dwelling units/house-holds	Number of dwelling units/house-holds flagged for child data collection	Minimum PSU sample size	Maximum PSU sample size	UEW DEFF for household sample after capping
1	Dodoma	11	339	112	21	21	1.00
2	Arusha	9	277	91	15	15	1.00
3	Kilimanjaro	10	308	101	19	19	1.01
4	Tanga	10	308	102	18	18	1.00
5	Morogoro	14	431	142	15	15	1.00
6	Pwani	37	1,139	377	15	15	1.00
7	Dar es Salaam	31 [1]	985	325	15	15	1.06
8	Lindi	6	185	61	19	19	1.00
9	Mtwara	8	246	81	21	21	1.00
10	Ruvuma	32	985	325	19	19	1.00
11	Iringa	25	770	254	15	15	1.08
12	Mbeya	26	801	264	15	15	1.11
13	Singida	7	216	71	21	21	1.00
14	Tabora	41	1,262	416	15	15	1.01
15	Rukwa	35	1,078	356	15	15	1.00
16	Kigoma	11	339	112	15	15	1.00
17	Shinyanga	30	924	305	16	16	1.00
18	Kagera	16	493	163	19	19	1.02
19	Mwanza	19	585	193	15	15	1.01
20	Mara	12	370	122	20	20	1.04
21	Manyara	7	216	71	17	17	1.00
22	Njombe	18	554	183	15	15	1.08
23	Katavi	37	1,139	376	15	15	1.10
24	Simiyu	9	277	92	20	20	1.00
25	Geita	12	370	122	15	15	1.59
26	Songwe	26	801	265	19	19	1.07
51	Kaskazini Unguja	4	123	41	24	24	1.00
52	Kusini Unguja	4	123	40	27	27	1.00
53	Mjini Magharibi	10	308	102	20	20	1.00
54	Kaskazini Pemba	4	123	41	27	27	1.00
55	Kusini Pemba	4	123	40	29	29	1.00
	Total	525	16,198	5,346	15	61	1.63 [2]

[1] Thirty-two PSUs had originally been selected from this region. However, one was later determined to be out of scope (contained no households).

[2] This DEFF reflects total variation in weights within and across regions.

Table 2-5 summarizes the number of dwelling units selected for the THIS by final household response status. Of the 16,198 sampled dwelling units 557 (3.4%) were determined during data collection to be vacant/unoccupied, 137 (0.8%) for which eligibility for the survey (i.e., occupancy status) could not be established, 693 (4.3%) were determined to be eligible for the study (i.e., contained eligible household members) but did not complete the household roster, and 14,811

(91.4%) completed the household roster. The overall unweighted household response rate was 94.7%.

Table 2-5 Distribution of dwelling unit sample by region and response status

Region code	Stratum (Region)	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 nonresp-onding households	Unweighted response rate [3]
1	Dodoma	339	3	3	296	37	0.881
2	Arusha	277	11	0	244	22	0.917
3	Kilimanjaro	308	4	0	298	6	0.980
4	Tanga	308	5	0	301	2	0.993
5	Morogoro	431	8	1	417	5	0.986
6	Pwani	1,139	32	118	971	18	0.880
7	Dar es Salaam	985	12	0	946	27	0.972
8	Lindi	185	7	0	175	3	0.983
9	Mtwara	246	9	0	229	8	0.966
10	Ruvuma	985	21	0	925	39	0.960
11	Iringa	770	24	0	700	46	0.938
12	Mbeya	801	30	1	717	53	0.930
13	Singida	216	7	0	179	30	0.856
14	Tabora	1,262	76	3	1,127	56	0.950
15	Rukwa	1,078	37	1	1,011	29	0.971
16	Kigoma	339	5	0	321	13	0.961
17	Shinyanga	924	38	0	851	35	0.960
18	Kagera	493	21	0	433	39	0.917
19	Mwanza	585	20	3	549	13	0.972
20	Mara	370	12	0	345	13	0.964
21	Manyara	216	10	2	191	13	0.928
22	Njombe	554	21	3	475	55	0.891
23	Katavi	1,139	60	1	1,053	25	0.976
24	Simiyu	277	9	0	262	6	0.978
25	Geita	370	9	0	353	8	0.978
26	Songwe	801	36	0	715	50	0.935
51	Kaskazini Unguja	123	3	0	117	3	0.975
52	Kusini Unguja	123	11	0	108	4	0.964
53	Mjini Magharibi	308	8	1	281	18	0.937
54	Kaskazini Pemba	123	4	0	109	10	0.916
55	Kusini Pemba	123	4	0	112	7	0.941
	TOTAL	16,198	557	137	14,811	693	0.947

[1] Vacant or unoccupied dwelling units, households with no persons eligible for THIS.

[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 THIS involved the following steps: (1) compiling a list of all individuals known to reside in the household or who slept in the household during the night prior to data collection; (2) identifying those rostered individuals who are eligible for data collection; and (3) selecting for the study those individuals meeting the age and residency requirements of the study. However, as noted below, only those individuals who were present in the household the night before the interview (i.e., the *de facto* population) are 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 2-6 summarizes the number of households completing the roster and the corresponding number of rostered individuals by stratum and resident status.

Table 2-6 Number of households completing rosters and number of persons by resident status

Region code	Stratum (Region)	Number of households completing rosters	Usual resident but did not sleep here	Usual resident and slept here	Nonresident but slept here	Total
1	Dodoma	296	99	1,144	38	1,281
2	Arusha	244	27	998	19	1,044
3	Kilimanjaro	298	29	1,096	40	1,165
4	Tanga	301	33	1,221	32	1,286
5	Morogoro	417	60	1,520	57	1,637
6	Pwani	971	154	3,495	144	3,793
7	Dar es Salaam	946	128	3,263	101	3,492
8	Lindi	175	27	614	16	657
9	Mtwara	229	39	730	16	785
10	Ruvuma	925	221	3,686	98	4,005
11	Iringa	700	163	2,501	123	2,787
12	Mbeya	717	182	2,491	178	2,851
13	Singida	179	60	774	33	867
14	Tabora	1,127	248	7,034	179	7,461
15	Rukwa	1,011	239	4,765	100	5,104
16	Kigoma	321	51	1,774	33	1,858
17	Shinyanga	851	205	4,649	120	4,974
18	Kagera	433	91	1,785	34	1,910
19	Mwanza	549	131	2,515	106	2,752
20	Mara	345	79	1,978	56	2,113
21	Manyara	191	40	846	46	932
22	Njombe	475	86	1,711	77	1,874
23	Katavi	1,053	295	5,046	136	5,477
24	Simiyu	262	61	1,709	38	1,808
25	Geita	353	92	1,802	67	1,961
26	Songwe	715	199	2,858	94	3,151
51	Kaskazini Unguja	117	52	510	45	607
52	Kusini Unguja	108	70	407	32	509
53	Mjini Magharibi	281	133	1,374	70	1,577
54	Kaskazini Pemba	109	50	540	49	639
55	Kusini Pemba	112	53	593	61	707
	TOTAL	14,811	3,397	65,429	2,238	71,064

2.5.2 Selecting Individuals for Data Collection

All of the individuals listed in the household rosters who were 15+ years of age and were either usual residents of the household or who slept in the household were eligible for data collection. However, children 0-14 years of age were eligible for data collection only if the household in which they resided had been randomly designated for child data collection (see Section 2.4.5). Table 2-7

summarizes the number of individuals eligible for data collection by region, age group, and resident status.

Although data collection was attempted for all of the 38,734 adults and 10,642 children indicated in Table 2-7, 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 36,144 adults and 10,402 children who slept in the household during the night prior to the interview.

Table 2-7 Number of individuals eligible for data collection

Region code	Stratum (Region)	Adults 15-59 [1]				Children 0-14 [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
1	Dodoma	67	586	25	678	8	186	4	198
2	Arusha	23	556	17	596	2	128	0	130
3	Kilimanjaro	26	687	28	741	1	132	4	137
4	Tanga	27	698	18	743	0	167	5	172
5	Morogoro	44	882	34	960	4	207	7	218
6	Pwani	111	2089	80	2,280	16	461	12	489
7	Dar es Salaam	100	2155	74	2,329	14	354	9	377
8	Lindi	21	362	9	392	0	81	5	86
9	Mtwara	28	463	14	505	3	92	1	96
10	Ruvuma	164	2054	69	2,287	14	567	7	588
11	Iringa	136	1453	89	1,678	9	357	9	375
12	Mbeya	126	1465	96	1,687	15	356	21	392
13	Singida	43	416	23	482	5	114	4	123
14	Tabora	209	3434	121	3,764	13	1220	21	1,254
15	Rukwa	182	2279	54	2,515	24	814	11	849
16	Kigoma	45	829	17	891	1	289	6	296
17	Shinyanga	166	2369	86	2,621	13	783	20	816
18	Kagera	63	911	23	997	8	245	4	257
19	Mwanza	108	1343	67	1,518	5	408	15	428
20	Mara	69	917	34	1,020	1	361	8	370
21	Manyara	31	451	30	512	3	132	0	135
22	Njombe	74	948	56	1,078	3	249	6	258
23	Katavi	225	2470	85	2,780	15	819	20	854
24	Simiyu	48	777	27	852	3	313	3	319
25	Geita	69	853	42	964	9	313	10	332
26	Songwe	139	1482	57	1,678	19	442	6	467
51	Kaskazini Unguja	39	299	23	361	7	78	6	91
52	Kusini Unguja	46	241	14	301	2	63	3	68
53	Mjini Magharibi	90	756	45	891	14	208	7	229
54	Kaskazini Pemba	37	243	16	296	4	106	13	123
55	Kusini Pemba	34	274	29	337	5	103	7	115
	Total	2,590	34,742	1,402	38,734	240	10,148	254	10,642

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

2.5.3 Distribution of Person Samples

Tables 2-8A through 2-8C summarize the number of individuals selected for data collection and the corresponding numbers completing the interview and blood test, for adults 15+ years, 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 THIS study population are shown under the *de facto* heading in these tables. Note that for children 0-9 years in Table 2-8C, the counts of completed "interviews" refer to the number of children for whom a parent completed the child questionnaire module for that particular child.

Table 2-8A Distribution of completed interviews and blood tests for adults 15+ years

Region code	Stratum (Region)	*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]
1	Dodoma	611	550	507	67	34	31
2	Arusha	573	479	431	23	8	8
3	Kilimanjaro	715	643	585	26	9	8
4	Tanga	716	677	653	27	6	6
5	Morogoro	916	856	817	44	14	14
6	Pwani	2,169	1,977	1,834	111	48	46
7	Dar es Salaam	2,229	1,957	1,775	100	31	27
8	Lindi	371	349	338	21	12	12
9	Mtwara	477	446	415	28	6	5
10	Ruvuma	2,123	1,995	1,915	164	84	83
11	Iringa	1,542	1,411	1,344	136	69	65
12	Mbeya	1,561	1,428	1,324	126	51	47
13	Singida	439	366	333	43	21	20
14	Tabora	3,555	3,147	3,067	209	93	87
15	Rukwa	2,333	2,142	2,084	182	67	63
16	Kigoma	846	788	781	45	24	24
17	Shinyanga	2,455	2,254	2,208	166	64	63
18	Kagera	934	877	858	63	34	34
19	Mwanza	1,410	1,327	1,284	108	41	39
20	Mara	951	885	865	69	20	20
21	Manyara	481	442	421	31	13	13
22	Njombe	1,004	933	884	74	37	37
23	Katavi	2,555	2,378	2,328	225	94	89
24	Simiyu	804	736	726	48	17	17
25	Geita	895	834	820	69	27	27
26	Songwe	1,539	1,418	1,370	139	63	62
51	Kaskazini Unguja	322	292	285	39	25	23
52	Kusini Unguja	255	237	231	46	25	24
53	Mjini Magharibi	801	703	666	90	39	39
54	Kaskazini Pemba	259	238	212	37	13	10
55	Kusini Pemba	303	279	265	34	15	15
	Total	36,144	33,044	31,626	2,590	1,104	1,058

[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 more information about the response status categories defined for the individual interview.

[4] These are cases that provided an analyzable blood sample, regardless of whether the individual interview was completed. Of the 31,626 de facto cases completing the blood test, five did not complete the interview but are treated as interview respondents for weighting purposes. See Appendix B for more information about the response status categories defined for the blood tests.

Table 2-8B Distribution of completed interviews and blood tests for adolescents 10-14 years

Region code	Stratum (Region)	*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]
1	Dodoma	57	43	39	5	2	2
2	Arusha	27	21	19	1	1	1
3	Kilimanjaro	48	46	43	0	0	0
4	Tanga	58	57	54	0	0	0
5	Morogoro	73	70	68	1	0	0
6	Pwani	151	142	138	8	5	5
7	Dar es Salaam	110	98	94	0	0	0
8	Lindi	28	26	24	0	0	0
9	Mtwara	25	24	23	0	0	0
10	Ruvuma	173	169	160	3	2	2
11	Iringa	109	103	100	5	3	3
12	Mbeya	113	105	104	3	0	0
13	Singida	31	21	20	4	2	2
14	Tabora	323	286	283	7	4	4
15	Rukwa	213	192	188	5	1	1
16	Kigoma	87	85	83	0	0	0
17	Shinyanga	205	198	198	3	2	2
18	Kagera	81	77	77	1	1	1
19	Mwanza	121	117	113	1	0	0
20	Mara	103	97	97	1	1	1
21	Manyara	34	33	33	1	1	1
22	Njombe	82	80	78	0	0	0
23	Katavi	236	230	225	3	2	2
24	Simiyu	79	72	71	1	1	1
25	Geita	95	95	94	3	3	3
26	Songwe	134	128	128	7	2	2
51	Kaskazini Unguja	15	15	15	1	1	1
52	Kusini Unguja	26	25	24	1	1	1
53	Mjini Magharibi	61	60	58	6	2	2
54	Kaskazini Pemba	41	37	35	3	0	0
55	Kusini Pemba	30	27	27	2	0	0
	Total	2,969	2,779	2,713	76	37	37

[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 more information about the response status categories defined for the individual interview.

[4] These are cases that provided an analyzable blood sample, regardless of whether the individual interview was completed. Of the 2,713 de facto cases completing the blood test, one case did not complete the interview but is treated as an interview respondent for weighting purposes. See Appendix B for more information about the response status categories defined for the blood tests.

Table 2-8C Distribution of completed interviews and blood tests for children 0-9 years

Region code	Stratum (Region)	*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]
1	Dodoma	133	131	107	3	3	3
2	Arusha	101	99	87	1	1	1
3	Kilimanjaro	88	88	79	1	1	0
4	Tanga	114	113	111	0	0	0
5	Morogoro	141	140	130	3	3	2
6	Pwani	322	320	284	8	7	3
7	Dar es Salaam	253	247	222	14	13	2
8	Lindi	58	57	51	0	0	0
9	Mtwara	68	68	61	3	3	1
10	Ruvuma	401	395	381	11	10	4
11	Iringa	257	249	232	4	4	3
12	Mbeya	264	257	230	12	12	3
13	Singida	87	83	64	1	1	1
14	Tabora	918	887	822	6	6	2
15	Rukwa	612	603	570	19	18	12
16	Kigoma	208	208	208	1	1	1
17	Shinyanga	598	592	576	10	10	7
18	Kagera	168	168	162	7	7	3
19	Mwanza	302	299	289	4	4	3
20	Mara	266	266	259	0	0	0
21	Manyara	98	95	88	2	2	2
22	Njombe	173	173	159	3	3	2
23	Katavi	603	595	576	12	12	5
24	Simiyu	237	237	224	2	2	2
25	Geita	228	228	223	6	6	5
26	Songwe	314	310	299	12	12	5
51	Kaskazini Unguja	69	67	63	6	6	1
52	Kusini Unguja	40	40	38	1	1	0
53	Mjini Magharibi	154	152	139	8	3	2
54	Kaskazini Pemba	78	75	53	1	1	0
55	Kusini Pemba	80	80	69	3	3	2
	Total	7,433	7,322	6,856	164	155	77

[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 more information about the response status categories defined for the individual interview.

[4] These are cases that provided an analyzable blood sample, regardless of whether the individual interview was completed. Of the 6,856 de facto cases completing the blood test, 105 did not complete the interview but are treated as interview respondents for weighting purposes. See Appendix B for more information about the response status categories defined for the blood tests.

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 THIS, nonresponse can occur at different stages of data collection, for example, (1) before the enumeration of individuals in the household, (2) after household enumeration and selection of persons but before completion of the individual interview, and (3) after completion of the interview but before collection of a usable blood sample. The procedures used to compensate for nonresponse at each of the relevant stages of data collection are described in Section 3.4.

Noncoverage arises when some members of the survey population have no chance of being selected for the sample. For example, noncoverage can occur if the field operations fail to enumerate all dwelling units during the listing process, or if certain household members are omitted from the household rosters. To compensate for such omissions, the poststratification procedures described in Sections 3.4.3 and 3.4.4 are used to calibrate the weighted sample counts to available population projections.

3.1 Overview of the Weighting Process

The overall weighting approach for THIS includes several steps.

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

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

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

Calculation of Household Weights: The next step is to calculate household weights. The household base weights are calculated as the PSU weights times the reciprocal of the within-PSU 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 2-5 above, this only happened for 0.8 % 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 poststratifying (weighting up) to 2017 population projections.

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

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

3.2 Preparation for Weighting

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 THIS survey data that are used to construct the sampling weights are contained in the following data files.

- **ptz_ffcorr_hh_qx_STAT_20170731**: A household (HH) file that contains the majority of household data collected in the HH questionnaire.
- **ptz_ffcorr_death_STAT_20170731**: A household (HH) file that contains data collected in the HH questionnaire regarding any deaths that have occurred in the household since 2013.
- **ptz_ffcorr_roster_STAT_20170731**: A file that contains the roster of household members collected in the HH questionnaire with a record for each rostered person.
- **ptz_ffcorr_indiv_STAT_20170731**[3]: 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 and over, 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.
- **TanBiomarker20170918**: 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.

[3] A later version of this file, **ptz_ffcorr_indiv_STAT_20170911**, that included two additional variables (BIRTHMON and BIRTHDAY) was used to create the weight delivery files.

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 region.
- 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 THIS: 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 THIS 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

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

As described in Section 2.3, the PSUs were selected systematically from a list of PSUs that had been ordered by type of EA (i.e., urban vs. rural), district within type of EA, ward within district, village within ward, and finally by EA within village. To take account of the precision benefits of implicit stratification as fully as possible, the sampled PSUs within each region 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 region, one of the variance-estimation strata was defined to contain three sampled PSUs. To fully reflect the sample design, the formation of the variance-estimation strata was applied to all of the sampled PSUs, including those that may later become a "nonresponse" (e.g., a sampled PSU containing households that was found to be inaccessible at the time of data collection) or ineligible (e.g., the PSU was found to contain no households).

For the THIS, a total of 257 variance-estimation strata were created. A jackknife replicate was then formed by randomly deleting a PSU from a particular variance-estimation stratum k, say, and retaining all of the PSUs in the remaining variance-estimation strata. For a variance-estimation stratum consisting of a pair of PSUs, the weight of the retained PSU within the variance-estimation stratum k was doubled. For a variance-estimation stratum consisting of three PSUs, the weight of the two retained PSUs within the variance-estimation stratum were increased by 1.5 (see Section 3.4.1). This process was repeated for all $r = 1, 2, ..., 257$ variance-estimation strata, resulting in a total of 257 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

Region code	Stratum (Region)	No. PSUs	No. variance strata consisting of pairs	No. variance strata consisting of triplets	Number of jackknife replicates
1	Dodoma	11	4	1	5
2	Arusha	9	3	1	4
3	Kilimanjaro	10	5	0	5
4	Tanga	10	5	0	5
5	Morogoro	14	7	0	7
6	Pwani	37	17	1	18
7	Dar es Salaam	32	16	0	16
8	Lindi	6	3	0	3
9	Mtwara	8	4	0	4
10	Ruvuma	32	16	0	16
11	Iringa	25	11	1	12
12	Mbeya	26	13	0	13
13	Singida	7	2	1	3
14	Tabora	41	19	1	20
15	Rukwa	35	16	1	17
16	Kigoma	11	4	1	5
17	Shinyanga	30	15	0	15
18	Kagera	16	8	0	8
19	Mwanza	19	8	1	9
20	Mara	12	6	0	6
21	Manyara	7	2	1	3
22	Njombe	18	9	0	9
23	Katavi	37	17	1	18
24	Simiyu	9	3	1	4
25	Geita	12	6	0	6
26	Songwe	26	13	0	13
51	Kaskazini Unguja	4	2	0	2
52	Kusini Unguja	4	2	0	2
53	Mjini Magharibi	10	5	0	5
54	Kaskazini Pemba	4	2	0	2
55	Kusini Pemba	4	2	0	2
	Total	526	245	12	257

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 THIS data files.

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

3.4 Development of Weights

3.4.1 PSU Weights

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

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

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

where P_{hi}^{PSU} = probability of selecting PSU *i* from region *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. Using the PSU weights defined above, the sampled PSUs (i.e., whole EAs or segments) weight up to the numbers shown in the second column of Table 3-2.

As indicated in Table 3-1, 257 jackknife replicates were formed from the 526 sampled PSUs. For variance estimation, replicate-specific PSU weights, $W_{(r)hi}^{(1)}$, $r = 1, 2, ..., 257$ were created to provide the basis for calculating the required replicate weights in subsequent stages of the weighting process. Let *h* denote one of the variance-estimation strata created for jackknife replication (Section 3.3.1)

and let *i* denote the PSU within variance-estimation stratum *h*. For a given jackknife replicate, *r* = 1, 2, ..., 257, the corresponding replicate-specific PSU base weight was computed as

$W_{(r)hi}^{(1)} = a\, W_{hi}^{(1)}$ if *h* = *r* and PSU *i* in variance-estimation stratum *h* is included in replicate *r*

$= 0$ if *h* = *r* and PSU *i* in variance-estimation stratum *h* is not included in replicate *r*

$= W_{hi}^{(1)}$ if $h \neq r$

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

Table 3-2 Number of PSUs and weighted sums by region

Region code	Stratum (Region)	Number of sampled PSUs (EAs)	PSUs weighted by PSU base weights [1]	Number of PSUs with responding households	PSUs with resp. households weighted by PSU base weights [2]
1	Dodoma	11	5,227	11	5,227
2	Arusha	9	5,448	9	5,448
3	Kilimanjaro	10	5,327	10	5,327
4	Tanga	10	4,110	10	4,110
5	Morogoro	14	6,303	14	6,303
6	Pwani	37	4,235	37	4,235
7	Dar es Salaam	32	15,841	31 [2]	15,216
8	Lindi	6	2,610	6	2,610
9	Mtwara	8	4,359	8	4,359
10	Ruvuma	32	4,163	32	4,163
11	Iringa	25	3,934	25	3,934
12	Mbeya	26	6,339	26	6,339
13	Singida	7	3,686	7	3,686
14	Tabora	41	6,252	41	6,252
15	Rukwa	35	3,232	35	3,232
16	Kigoma	11	3,782	11	3,782
17	Shinyanga	30	3,229	30	3,229
18	Kagera	16	10,053	16	10,053
19	Mwanza	19	5,800	19	5,800
20	Mara	12	5,313	12	5,313
21	Manyara	7	2,883	7	2,883
22	Njombe	18	2,533	18	2,533
23	Katavi	37	1,538	37	1,538
24	Simiyu	9	3,552	9	3,552
25	Geita	12	5,943	12	5,943
26	Songwe	26	3,297	26	3,297
51	Kaskazini Unguja	4	407	4	407
52	Kusini Unguja	4	336	4	336
53	Mjini Magharibi	10	1,460	10	1,460
54	Kaskazini Pemba	4	564	4	564
55	Kusini Pemba	4	446	4	446
	Total	526	132,201	525	131,576

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

[2] One PSU in region 7 was determined to be ineligible. The sum of the weights without this PSU in region 7 drops to 15,216.

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 adjusted for PSU nonresponse (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 region *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 region *h*

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

The corresponding weights for jackknife replicate r = 1, 2, …, 257, were computed as:

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

where $W_{(r)hi}^{(1)}$ is the PSU base weight for PSU *I* in region *h* 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 3-3. In Table 3-4, we show the corresponding weighted sums by response status and region 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 region.

Table 3-3 Response-status groups specified for household weighting

Household response status code [1]	Description	Number of dwelling units/households
1	Eligible respondent	14,811
2	Eligible nonrespondent	693
3	Ineligible/out-of-scope	557
4	Unknown eligibility status	137

[1] See Appendix B for definitions.

Table 3-4 Weighted sums of household base weights by response status

Region code	Stratum (Region)	Household Response Status				
		Status code 1: Eligible Respondents	Status code 2: Eligible Nonresp-ondents	Status code 3: Not Eligible (vacant, destroyed, not a DU, etc.)	Status code 4: Could not determine eligibility	Weighted Count of Households [1]
1	Dodoma	575,276	71,910	5,831	5,831	658,847
2	Arusha	540,316	48,775	24,497	-	613,588
3	Kilimanjaro	507,856	10,164	6,560	-	524,581
4	Tanga	470,870	3,129	7,822	-	481,820
5	Morogoro	771,432	9,258	14,812	1,852	797,354
6	Pwani	336,385	6,127	10,893	39,447	392,853
7	Dar es Salaam	1,238,960	33,112	16,114	-	1,288,185
8	Lindi	279,265	4,787	11,171	-	295,223
9	Mtwara	350,626	12,249	13,780	-	376,655
10	Ruvuma	391,229	16,495	8,882	-	416,606
11	Iringa	343,389	23,313	11,760	-	378,463
12	Mbeya	563,064	46,878	22,054	719	632,715
13	Singida	351,122	58,847	13,731	-	423,700
14	Tabora	433,773	21,404	28,965	1,143	485,285
15	Rukwa	307,093	8,738	11,278	305	327,414
16	Kigoma	341,962	13,876	5,337	-	361,175
17	Shinyanga	334,397	13,764	14,932	-	363,093
18	Kagera	626,779	54,771	29,432	-	710,982
19	Mwanza	611,362	14,080	22,978	3,293	651,713
20	Mara	549,435	18,828	20,935	-	589,197
21	Manyara	385,593	26,245	20,188	4,038	436,063
22	Njombe	238,246	29,963	11,192	1,344	280,745
23	Katavi	159,030	3,753	8,363	139	171,284
24	Simiyu	265,463	6,079	9,119	-	280,662
25	Geita	627,363	16,451	14,376	-	658,190
26	Songwe	252,139	17,764	13,197	-	283,101
51	Kaskazini Unguja	28,092	720	720	-	29,532
52	Kusini Unguja	31,926	1,182	3,252	-	36,360
53	Mjini Magharibi	149,102	9,551	4,245	531	163,429
54	Kaskazini Pemba	37,123	3,406	1,362	-	41,891
55	Kusini Pemba	36,474	2,280	1,303	-	40,056
	Total	12,135,143	607,897	389,081	58,641	13,190,762

[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 THIS), 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 THIS, all PSUs were found to have response rates above 50%. Therefore, no PSUs were collapsed for this purpose. There were, however, four PSUs in Pwani that could not be visited due to security concerns. All of the sampled dwelling units in these PSUs were categorized as nonresponding households and the four PSUs were collapsed with other PSUs to form nonresponse adjustment cells for weighting purposes.

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 region *h*. Note that n_{hi}^{samp} is the sum of the sample sizes in each of the four response status groups defined in Table 3-3, 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 region *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 region *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 region *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 region *h*

The first-phase household nonresponse adjustment factor for PSU weighting cell *i* in region *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 region *h*, and where the sum in the numerator extends over the entire sample of dwelling units/households in PSU weighting cell *i* in region *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 region *h* was then computed as:

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

The corresponding replicate weights for replicate *r* = 1, 2, …, 257 were computed in similar fashion as:

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

where

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

Note that for the sampled dwelling units/households in response-status group 4, $W_{hij}^{HH1} = W_{(r)hij}^{HH1} =$ 0 for $r = 1, 2, \ldots, 257$.

The effect of this adjustment is to distribute the total weight of the undetermined-eligibility cases (i.e., the estimated 58,641 dwelling units shown in the next-to-last column of Table 3-4 to the combined weight of the remaining three groups of sampled dwelling units/households. The resulting weighted counts using W_{hij}^{HH1} as computed above are given in Table 3-5.

Table 3-5 Weighted sums of household weights adjusted for unknown eligibility

Stratum code	Stratum (Region)	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
1	Dodoma	579,487	73,529	5,831	658,847	653,016
2	Arusha	540,316	48,775	24,497	613,588	589,091
3	Kilimanjaro	507,856	10,164	6,560	524,581	518,021
4	Tanga	470,870	3,129	7,822	481,820	473,998
5	Morogoro	773,193	9,303	14,857	797,354	782,496
6	Pwani	374,418	6,571	11,864	392,853	380,989
7	Dar es Salaam	1,238,960	33,112	16,114	1,288,185	1,272,071
8	Lindi	279,265	4,787	11,171	295,223	284,052
9	Mtwara	350,626	12,249	13,780	376,655	362,875
10	Ruvuma	391,229	16,495	8,882	416,606	407,724
11	Iringa	343,389	23,313	11,760	378,463	366,703
12	Mbeya	563,664	46,963	22,088	632,715	610,627
13	Singida	351,122	58,847	13,731	423,700	409,969
14	Tabora	434,611	21,549	29,125	485,285	456,160
15	Rukwa	307,384	8,738	11,292	327,414	316,122
16	Kigoma	341,962	13,876	5,337	361,175	355,838
17	Shinyanga	334,397	13,764	14,932	363,093	348,161
18	Kagera	626,779	54,771	29,432	710,982	681,549
19	Mwanza	614,655	14,080	22,978	651,713	628,735
20	Mara	549,435	18,828	20,935	589,197	568,263
21	Manyara	389,150	26,437	20,476	436,063	415,586
22	Njombe	239,435	29,995	11,316	280,745	269,429
23	Katavi	159,144	3,761	8,379	171,284	162,905
24	Simiyu	265,463	6,079	9,119	280,662	271,543
25	Geita	627,363	16,451	14,376	658,190	643,814
26	Songwe	252,139	17,764	13,197	283,101	269,903
51	Kaskazini Unguja	28,092	720	720	29,532	28,812
52	Kusini Unguja	31,926	1,182	3,252	36,360	33,108
53	Mjini Magharibi	149,598	9,586	4,245	163,429	159,184
54	Kaskazini Pemba	37,123	3,406	1,362	41,891	40,528
55	Kusini Pemba	36,474	2,280	1,303	40,056	38,754
	Total	12,189,524	610,505	390,734	13,190,762	12,800,029

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 region *h* was computed as the ratio:

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

where W_{hij}^{HH1} is the first-phase adjusted weight for dwelling unit/household *j* in PSU weighting cell *i* in region *h*, and where the sum in the number extends over the sample of responding and nonresponding households in PSU weighting cell *i* in region *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 region *h* was then computed as:

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

The corresponding replicate weights for replicate $r = 1, 2, \ldots, 257$ 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 3-5.

Weight Trimming

The Geita region was one of the regions that saw a tremendous growth in population in many of its EAs. As a result, it was necessary to cap the number of sampled dwelling units for many of the EAs in this region at 60 (see Section 2.4.4). This resulted in very large household base weights in some EAs, which in turn led to highly variable nonresponse-adjusted weights across the region. 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 of the sampled dwelling units in this region. For this purpose, a weight outlier is defined to be a weight that is greater than 2 times the median *nonresponse-adjusted* weight[4] within the corresponding sampling stratum. Such weights were capped at 2 times the median weight. The resultant weights were then recalibrated to household control totals for Geita region through a poststratification adjustment. Details on the computation of household control totals is provided in Appendix D. The resulting weighted counts using the trimmed and recalibrated nonresponse-adjusted households weights are shown in Table 3-6.

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

Table 3-6 Weighted sums of nonresponse-adjusted household weights after trimming and recalibration

Stratum code	Stratum (Region)	Eligible responding households	Ineligible dwellings	Total dwelling units/households
1	Dodoma	653,016	5,831	658,847
2	Arusha	589,091	24,497	613,588
3	Kilimanjaro	518,021	6,560	524,581
4	Tanga	473,998	7,822	481,820
5	Morogoro	782,496	14,857	797,354
6	Pwani	380,989	11,864	392,853
7	Dar es Salaam	1,272,071	16,114	1,288,185
8	Lindi	284,052	11,171	295,223
9	Mtwara	362,875	13,780	376,655
10	Ruvuma	407,724	8,882	416,606
11	Iringa	366,703	11,760	378,463
12	Mbeya	610,627	22,088	632,715
13	Singida	409,969	13,731	423,700
14	Tabora	456,160	29,125	485,285
15	Rukwa	316,122	11,292	327,414
16	Kigoma	355,838	5,337	361,175
17	Shinyanga	348,161	14,932	363,093
18	Kagera	681,549	29,432	710,982
19	Mwanza	628,735	22,978	651,713
20	Mara	568,263	20,935	589,197
21	Manyara	415,586	20,476	436,063
22	Njombe	269,429	11,316	280,745
23	Katavi	162,905	8,379	171,284
24	Simiyu	271,543	9,119	280,662
25	Geita	403,785	11,931	415,716
26	Songwe	269,903	13,197	283,101
51	Kaskazini Unguja	28,812	720	29,532
52	Kusini Unguja	33,108	3,252	36,360
53	Mjini Magharibi	159,184	4,245	163,429
54	Kaskazini Pemba	40,528	1,362	41,891
55	Kusini Pemba	38,754	1,303	40,056
	Total	12,560,000	388,288	12,948,288

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 THIS 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 THIS are categorized into three age groups for which different data elements are collected: (1) adults aged 15 or older, 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 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-third 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 3-7 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 (trimmed and recalibrated) household weight, $W_{hij}^{(2A)}$. The sample person's base weight is the same as the nonresponse-adjusted household weight for adults (persons 15 or older), but it is three times 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 region *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 years or older, or $K_k = 3$ if the roster age of person *k* is 14 years or younger in households designated for child data collection.

The corresponding replicate base weights, $W_{(r)hijk}^{(3)}$, $r = 1, 2, \ldots, 257$, were computed in an analogous manner, with $W_{hij}^{(2A)}$ replaced by $W_{(r)hij}^{(2A)}$ in the above formula.

Table 3-7 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 3-7 include a small number of persons who did not complete the interview but did provide an analyzable blood test.

Table 3-7 Distribution of eligible sample persons by age group and interview response status

Group	Age[1]	Interview Status[2]	Count	Weighted count [3]
Adults	15+	Eligible Respondent	33,044	27,405,211
		Eligible Nonrespondent	3,094	2,694,456
Adolescents	10-14	Eligible Respondent	2,779	6,558,381
		Eligible Nonrespondent	190	546,311
Children	0-9	Eligible Respondent	7,322	17,168,365
		Eligible Nonrespondent	110	240,537

[1] Based on age reported in roster.

[2] Eligible respondents include cases that completed the individual interview or the blood test. See Appendix B for definitions of response status categories.

[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 region 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 50 variables from the household questionnaire and EA sampling frame that could potentially be used for nonresponse adjustment. The LASSO regression is used to reduce the number of variables to a manageable subset of the most important and relevant predictors that would subsequently be entered into the CHAID algorithm to define the final nonresponse adjustment weighting cells. The LASSO is a restrictive procedure similar to linear regression that

shrinks regression coefficient estimates to zero. In other words, predictors that are found to be nonsignificant have their regression coefficients set to 0 (Hastie, Tibshirani, and Friedman, 2009).

In the final model produced by the LASSO, only the most significant variables predictive of the response variable were identified and kept. The HPGENSELECT procedure (Johnston and Rodriguez, 2015) with selection method=lasso in SAS 9.4 was used to select the variables, with the weight set to the person base weight, $W_{hijk}^{(3)}$. Separate models were fitted for the three age groups defined in Table 3-7. 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 50, 49, and 49 variables used in the initial models for adults, adolescents, and children, respectively, the LASSO identified 43, 30, and 36 variables to be significant predictors of response for the three respective age groups, as shown in Table 3-8.

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

The next step was to apply the CHAID algorithm (Magidson, 2005) to the variables selected by the LASSO procedure. CHAID classifies the sampled individuals (i.e., the respondents and nonrespondents) into "cells" based on information available for all sample persons. The cells are formed in such a way that persons belonging to the same cell 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) was forced into the model to make the initial splits. The reason for doing this was because males and females received different questions; without forcing this variable into the model, the resulting tree would not have been created correctly. After forcing the SEX variable in the model, the tree was then allowed to grow freely. The CHAID algorithm selected 2, 8, and 11 variables for adults, adolescents, and children, respectively,

that were used to create the weighting classes for nonresponse adjustment. Table 3-9 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. This occurred in one instance for adolescents, where one cell had an adjustment factor of 2.17. After collapsing the cell with a neighboring cell, the final adjustment factor became 1.38. A total of 6 final weighting adjustment cells were created for adults, 21 cells for adolescents, and 24 cells for children. The final weighting cells created for nonresponse adjustment are documented in Appendix C.

Table 3-8 Variables in the original model, variables selected by LASSO, and variables selected by CHAID, and final adjustment cells

Group	Age	Variables in original model	Variables selected by LASSO	Variables selected by CHAID	Number of nonresponse adjustment cells
Adults	15+	50	43	2	6
Adolescents	10-14	49	30	8	21
Children	0-9	49	36	11	24

Table 3-9 Variables selected by CHAID to produce classes for interview nonresponse adjustment

Age group	Number	Variable name	Description
Adults	1	H_HH_SIZE_C	1-9, where 9 includes all HHs with 9 or more people
	2	SEX	Gender based on roster
Adolescents	1	H_ECON12	Received economic support in the past 12 months: 1 - Yes; 0 - No
	2	H_HH_SIZE_C	1-9, where 9 includes all HHs with 9 or more people
	3	H_OWNTRNSPRT	Household owns transportation: 1 - owns car and no other transport; 2 - owns bike and no other transport; 3 - owns moto and no other transport; 4 - owns bike and moto and no other transport; 5 - does not own any transport; 6 - owns some transport
	4	H_ROOMSLEEP	Number of rooms used for sleeping: 1, 2,..., 6+
	5	H_TLETTYP	Toilet type: see TOILETTYPE; 96 - otherwise
	6	HAVRAD	A radio in working condition?
	7	STRATA	Numeric code for EA sampling stratum
	8	WATERSOURCE	What is the main source of drinking water for members of your household?
Children	1	H_COOKFUEL	Cooking Fuel: 1 - Electricity, Gas, paraffin; 2 - Coal/charcoal; 3 - firewood, 9 - other
	2	H_HAVE_ELEC_DEVICE	Household has: 1 - electricity and both fridge * tele; 2 - electricity and either fridge or tele; 3 - electricity but no fridge nor tele; 4 - no electricity but either fridge or tele; 5 - no electricity and no fridge nor tele
	3	H_HH_SIZE_C	1-9, where 9 includes all HHs with 9 or more people
	4	H_HHLIGHTS	Main source of energy for lighting in HH: 1 - Electricity; 2 - Solar; 3 - Paraffin, 9 - other
	5	H_MATRF	Roof material: 1 - natural roof; 2 - rudimentary roofing; 3 - finished roofing; 9 - other
	6	H_MATWALL	Wall material: 1 - natural walls; 2 - rudimentary walls; 3 - cement/stone; 4 - bricks; 5 - cement blocks; 9 - other
	7	H_OWNANML	Household owns animals: 1 - cow(s) only; 2 - cow(s) and poultry; 3 - owns cow(s) and goats/sheep; 4 - cow(s) and (and some other animal); 5 - only poultry; 6 - no cow(s) but other animals, including poultry; 9 - doesn't own animals
	8	H_TLETTYP	Toilet type: see TOILETTYPE; 96 - otherwise
	9	SICKFLAGHH	flag household with sick adult
	10	STRATA	Numeric code for EA sampling stratum
	11	WATERSOURCE	What is the main source of drinking water for members of your household?

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, ..., 257 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, ..., 257.

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, ..., 257 were similarly computed as

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

Table 3-10 summarizes the number of weighting cells created for nonresponse adjustment, the overall weighted response rate, and the minimum and maximum adjustment for each of the three major age groups.

Table 3-10 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	Maxi	before adjustment [1]	after adjustment [2]
Adults	15+	33,044	6	91.05	1.00	1.15	27,405,211	30,099,667
Adolescents	10-14	2,779	21	92.31	1.00	1.68	6,558,381	7,104,693
Children	0-9	7,322	24	98.62	1.00	1.45	17,168,365	17,408,902

[1] Weight is person base weight, $W_{mk}^{(3)}$.

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

3.4.3.3 Poststratification Adjustment

The final step in computing the individual interview weights was to adjust the nonresponse-adjusted interview weights 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 2017 national population counts by gender and five-year age groups published by the United Nations (UN). The poststratified interview weights were computed as follows.

Let N_{ga}^{2017} denote the 2017 Tanzania 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:

$$T_{ga}^{2017} = N_{ga}^{2017} / \sum_{k=1}^{n_{ga}^{resp}} W_{gak}^{(int)},$$

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:

$$T_{(r)ga}^{2017} = N_{ga}^{2017} / \sum_{k=1}^{n_{(r)ga}^{resp}} W_{(r)gak}^{(int)}$$

for the *r* = 1, 2, …, 257 jackknife replicates.

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

$$W_{gak}^{(ps-int)} = T_{ga}^{2017} W_{gak}^{(int)},$$

and the corresponding poststratified replicate weights were computed as:

$$W_{(r)gak}^{(ps-int)} = T_{ga}^{2017} W_{(r)gak}^{(int)}$$

for *r* = 1, 2, …, 257.

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

Table 3-11 2017 Tanzania 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	4,974,120	4,648,842	1.0700	4,888,303	4,523,105	1.0807	9,862,423	9,171,946	1.0753
5-9	4,330,119	4,288,742	1.0096	4,233,202	4,030,009	1.0504	8,563,321	8,318,751	1.0294
10-14	3,646,522	3,493,060	1.0439	3,665,451	3,519,786	1.0414	7,311,973	7,012,846	1.0427
15-19	3,032,666	2,423,323	1.2514	3,032,037	2,643,553	1.1470	6,064,703	5,066,876	1.1969
20-24	2,511,475	1,844,393	1.3617	2,553,336	2,633,598	0.9695	5,064,811	4,477,990	1.1310
25-29	2,079,597	1,657,406	1.2547	2,203,407	2,318,440	0.9504	4,283,004	3,975,846	1.0773
30-34	1,747,429	1,428,452	1.2233	1,861,761	1,905,185	0.9772	3,609,190	3,333,637	1.0827
35-39	1,467,926	1,287,320	1.1403	1,544,828	1,689,410	0.9144	3,012,754	2,976,730	1.0121
40-44	1,191,963	1,149,008	1.0374	1,226,271	1,353,682	0.9059	2,418,234	2,502,690	0.9663
45-49	924,108	866,002	1.0671	948,441	1,007,167	0.9417	1,872,549	1,873,170	0.9997
50-54	697,713	749,534	0.9309	743,399	824,489	0.9016	1,441,112	1,574,022	0.9156
55-59	540,555	560,442	0.9645	599,059	550,802	1.0876	1,139,614	1,111,244	1.0255
60-64	402,499	480,773	0.8372	482,460	540,485	0.8926	884,959	1,021,258	0.8665
65+	795,277	1,038,202	0.7660	986,092	1,158,054	0.8515	1,781,369	2,196,256	0.8111
Total	28,341,969	25,915,498	1.0936	28,968,047	28,697,764	1.0094	57,310,016	54,613,262	1.0494

[1] Source: 2017 UN 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. Similarly to the construction of the interview weights described previously, development of the final blood test weights involves adjustments for nonresponse and poststratification to 2017 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 $r = 1, 2, \ldots, 257$), respectively, where *k* denotes the interview respondent, *h* denotes the region, *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 weighting purposes (e.g., see Tables 2-8A through 2-8C). 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+	Male	Respondent	13,750	12,723,841
			Nonrespondent	682	761,013
		Female	Respondent	17,829	15,832,688
			Nonrespondent	743	792,177
Adolescents	10-14	Male	Respondent	1,345	3,392,613
			Nonrespondent	30	100,447
		Female	Respondent	1,377	3,440,415
			Nonrespondent	27	79,371
Children	0-9	Male and female	Respondent	6,894	16,219,250
			Nonrespondent	468	1,271,448

[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 region 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 initial 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 respective age groups.

Due to high blood test response rates conditional on interview response, that ranged from 97 to 98 percent depending on the age group, the LASSO procedure did not select any variables predictive of response for adolescent females and selected only one variable for adolescent males. To adjust for nonresponse at the blood test level, four variables that were typically selected as significant predictors of blood test response in previous countries, namely, a categorized age based on interview (AT_BESTAGE_C), a stratification variable (STRATA), urban / rural definition

(URBAN_RURAL), and self-reported HIV status derived from interview items (KNOWN_HIV_STATUS) were added manually to the CHAID analyses for both adolescent males and females. For children, in addition to household and child-specific variables selected by the LASSO, LASSO variables from the responding parent or guardian were also added to the CHAID model as potential predictors.

To create the CHAID tree for children, gender of the responding parent or guardian (PROXY_GENDER) was forced into the model to make the initial splits. The reason for doing this was because males and females received different questions; without forcing this variable into the model, the resulting tree would not have been created correctly. After forcing the PROXY_GENDER variable in the model, the tree was then allowed to grow freely.

Table 3-14 summarizes variables that were included in the final CHAID models. The trees produced by CHAID 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 initial model	Variables selected by the LASSO	Variables selected by CHAID	Number of nonresponse adjustment cells
Adults	15+	Male	99	58	21	36
		Female	108	51	21	55
Adolescents	10-14	Male	92	1 [1]	2	4
		Female	94	0 [1]	1	3
Children	0-9	Male and female	64	30	25	50

[1] Four additional variables (STRATA, URBAN_RURAL, AT_BESTAGE_C, and KNOWN_HIV_STATUS) added 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_ALCNUMDAY	How many drinks containing alcohol do you have in a typical day?
	2	AT_BESTAGE_C	Categorical age based on interview age (BEST AGE)
	3	AWAYNIGHT12MMONTH	In the last 12 months, have you been away from home for more than one month at a time?
	4	BUYFOOD	Would you buy fresh vegetables from a shop keeper or vendor if you knew the person had HIV?
	5	CONDOMWHAT	Do you use condoms?
	6	H_COOKFUEL	Cooking Fuel: 1 - Electricity, Gas, paraffin; 2 - Coal/charcoal; 3 - firewood, 9 - other
	7	H_HH_SIZE_C	1-9, where 9 includes all HHs with 9 or more people
	8	H_HHDISWALK	Have difficulty walking or climbing steps?
	9	H_MATRF	Roof material: 1 - natural roof; 2 - rudimentary roofing; 3 - finished roofing; 9 - other
	10	H_MOMGUARD	Does kid's natural mother or female guardian usually live in this HH: 1 - Yes, 2 - No
	11	H_OWNANML	Household owns animals: 1 - cow(s) only; 2 - cow(s) and poultry; 3 - owns cow(s) and goats/sheep; 4 - cow(s) and (and some other animal); 5 - only poultry; 6 - no cow(s) but other animals, including poultry; 9 - doesn't own animals
	12	HIVTSTEVER	Have you ever tested for HIV?
	13	MCPLANS	Are you planning to get circumcised?
	14	MCSTATUS	Some men are uncomfortable talking about circumcision but it is important for us to have this information. Some men are circumcised. Are you circumcised?
	15	MOSNETS	Does your household have any mosquito nets that can be used while sleeping?
	16	OKHITSEX	Do you believe it is right for a man to hit or beat his wife if she refuses to have sex with him?
	17	PAINURIN	During the last 12 months, have you had pain on urination?
	18	PNSORE	During the last 12 months, have you had an ulcer or sore on or near your penis?
	19	SCHLHI	What is the highest level of school you attended?
	20	STRATA	Numeric code for EA sampling stratum
	21	WATERSOURCE	What is the main source of drinking water for members of your household?
Adult Female	1	AT_BESTAGE_C	Categorical age based on interview age (BEST AGE)
	2	AT_PART12MONUM	Recoded: People often have sex with different partners over their lifetime. In total, with how many different people have you had sex in the last 12 months?
	3	BUYFOOD	Would you buy fresh vegetables from a shop keeper or vendor if you knew the person had HIV?
	4	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?

Age group	Number	Variable name	Description
	5	H_COOKFUEL	Cooking Fuel: 1 - Electricity, Gas, paraffin; 2 - Coal/charcoal; 3 - firewood, 9 - other
	6	H_DEATHS	Has any usual resident of your household died since 2014?: 1 = Yes, 0 = Otherwise
	7	H_HH_SIZE_C	1-9, where 9 includes all HHs with 9 or more people
	8	H_MATRF	Roof material: 1 - natural roof; 2 - rudimentary roofing; 3 - finished roofing; 9 - other
	9	H_MATWALL	Wall material: 1 - natural walls; 2 - rudimentary walls; 3 - cement/stone; 4 - bricks; 5 - cement blocks; 9 - other
	10	H_MOMDADSICKDEAD	Parent sick/dead flag: 1 - both parents are sick/dead; 2 - both parents are not sick/dead; 3 - mother is sick/dead, father is alive; 3 - one parent is sick/dead; 4 - father is sick/dead, mother is alive
	11	H_OWNANML	Household owns animals: 1 - cow(s) only; 2 - cow(s) and poultry; 3 - owns cow(s) and goats/sheep; 4 - cow(s) and (and some other animal); 5 - only poultry; 6 - no cow(s) but other animals, including poultry; 9 - doesn't own animals
	12	H_TLETTYP	Toilet type: see TOILETTYPE; 96 - otherwise
	13	HEALTHC	Who usually makes decisions about health care for yourself: you, your (spouse/partner), you and your (spouse/partner) together, or someone else?
	14	HIVTSBP	Have you ever tested for HIV before your pregnancy with ${namedis}?
	15	HUSOTWIF	Does your husband or partner have other wives or does he live with other women as if married?
	16	KNOWN_HIV_STATUS_R	Categorical known HIV status
	17	RELATTOHH	What is the relationship of name to the head of the household?
	18	SELLSX12MO	In the last 12 months, have you received payment for sex?
	19	SICKFLAGHH	flag household with sick adult
	20	STRATA	Numeric code for EA sampling stratum
	21	VGSORE	During the last 12 months, have you had an ulcer or sore on or near your vagina?
Adolescent Male	1	CH_KIDHIVTESTEVR	Has ${curchnm}* ever been tested for HIV?
	2	STRATA	Numeric code for EA sampling stratum
Adolescent Female	1	STRATA	Numeric code for EA sampling stratum
Children	1	ALCFREQ	How often do you have a drink containing alcohol?
	2	AT_PART12MONUM	Recoded: People often have sex with different partners over their lifetime. In total, with how many different people have you had sex in the last 12 months?
	3	BUYFOOD	Would you buy fresh vegetables from a shop keeper or vendor if you knew the person had HIV?

Age group	Number	Variable name	Description
	4	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?
	5	H_HH_SIZE_C	1-9, where 9 includes all HHs with 9 or more people
	6	H_HHLIGHTS	Main source of energy for lighting in HH: 1 - Electricity; 2 - Solar; 3 - Paraffin, 9 - other
	7	H_HUNGRY4WKYN	In the past 4 weeks, did you or any household member go to sleep at night hungry because there was not enough food?
	8	H_MATWALL	Wall material: 1 - natural walls; 2 - rudimentary walls; 3 - cement/stone; 4 - bricks; 5 - cement blocks; 9 - other
	9	H_OWNANML	Household owns animals: 1 - cow(s) only; 2 - cow(s) and poultry; 3 - owns cow(s) and goats/sheep; 4 - cow(s) and (and some other animal); 5 - only poultry; 6 - no cow(s) but other animals, including poultry; 9 - doesn't own animals
	10	H_OWNTRNSPRT	Household owns transportation: 1 - owns car and no other transport; 2 - owns bike and no other transport; 3 - owns moto and no other transport; 4 - owns bike and moto and no other transport; 5 - does not own any transport; 6 - owns some transport
	11	H_ROOMSLEEP	Number of rooms used for sleeping: 1, 2,..., 6+
	12	H_TLETTYP	Toilet type: see TOILETTYPE; 96 - otherwise
	13	H_WTRSFR	Do you do anything to the water to make it safer to drink?
	14	HEALTHC	Who usually makes decisions about health care for yourself: you, your (spouse/partner), you and your (spouse/partner) together, or someone else?
	15	HIVTSBP	Have you ever tested for HIV before your pregnancy with ${namedis}?
	16	MCPLANS	Are you planning to get circumcised?
	17	ONEPARTNR	Can the risk of HIV transmission be reduced by having sex with only one uninfected partner who has no other partners?
	18	P_BESTAGE_C	Categorical age based on interview age (BEST AGE)
	19	PROXY_GENDER	Gender of responding parent/guardian
	20	PROXY_STATUS	Interview response status of responding parent/guardian
	21	RELATTOHH	What is the relationship of name to the head of the household?
	22	SCHLCUR	Are you enrolled in school?
	23	WATERSOURCE	What is the main source of drinking water for members of your household?
	24	WORK12MO	Have you done any work in the last 12 months for which you received cash or goods as payment?
	25	WORK7DAY	Have you done any work in the last seven days for which you received cash or goods as payment?

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, ..., 257 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, ..., 257.

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, ..., 257 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+	Male	13,750	36	94.36	1.0000	1.3704	12,723,841	13,484,853
		Female	17,829	55	95.23	1.0000	1.4345	15,832,688	16,624,865
Adolescents	10-14	Male	1,345	4	97.12	1.0000	1.2485	3,392,613	3,493,060
		Female	1,377	3	97.75	1.0000	1.1033	3,440,415	3,519,786
Children	0-9	Male and female	6,894	50	92.73	1.0000	1.7849	16,219,250	17,490,697

[1] Among the interview respondents.

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

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

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 (e.g., see Valliant, Dever, and 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.

As shown in Table 3-16, there were eight weight outliers at the blood test level. Outliers were present in weights for adult females (i.e., 1 outlier) and children (i.e., 7 outliers). The adult female with an outlier weight was an "age switcher," where she was reported and sampled as an adolescent age 10 to 14 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 and adolescents in households designated for child data collection were multiplied by a subsampling factor of $K_k = 3$; this factor was 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 adult females, children, and overall.

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

Region code	Stratum (Region)	Adults (15+)				Adolescents (10-14)				Children (0-9)	
		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
1	Dodoma	226	280	0	0	20	18	0	0	109	0
2	Arusha	175	255	0	0	10	9	0	0	88	0
3	Kilimanjaro	241	341	0	0	26	21	0	0	78	0
4	Tanga	283	367	0	0	31	28	0	0	109	0
5	Morogoro	324	491	0	0	21	49	0	0	130	0
6	Pwani	763	1,066	0	0	64	76	0	0	287	0
7	Dar es Salaam	724	1,049	0	0	47	49	0	0	222	0
8	Lindi	156	181	0	0	14	11	0	0	51	0
9	Mtwara	206	208	0	0	12	13	0	0	60	0
10	Ruvuma	852	1,062	0	0	73	85	0	0	384	0
11	Iringa	550	792	0	0	49	54	0	0	231	0
12	Mbeya	559	765	0	1	49	52	0	0	233	2
13	Singida	139	193	0	0	10	11	0	0	64	0
14	Tabora	1,454	1,607	0	0	135	149	0	0	827	0
15	Rukwa	896	1,183	0	0	88	103	0	0	572	0
16	Kigoma	329	448	0	0	43	42	0	0	210	0
17	Shinyanga	1,003	1,206	0	0	104	95	0	0	574	0
18	Kagera	410	447	0	0	41	36	0	0	163	0
19	Mwanza	584	701	0	0	63	48	0	0	290	0
20	Mara	353	511	0	0	58	39	0	0	260	0
21	Manyara	191	227	0	0	15	20	0	0	89	0
22	Njombe	349	534	0	0	37	36	0	0	165	0
23	Katavi	1,024	1,302	0	0	120	100	0	0	583	5
24	Simiyu	303	424	0	0	37	32	0	0	225	0
25	Geita	384	436	0	0	51	43	0	0	223	0
26	Songwe	600	767	0	0	50	77	0	0	303	0
51	Kaskazini Unguja	106	177	0	0	11	6	0	0	63	0
52	Kusini Unguja	108	126	0	0	8	13	0	0	38	0
53	Mjini Magharibi	274	390	0	0	28	32	0	0	139	0
54	Kaskazini Pemba	78	134	0	0	15	17	0	0	56	0
55	Kusini Pemba	106	159	0	0	15	13	0	0	68	0
Overall		13,750	17,829	0	1	1,345	1,377	0	0	6,894	7

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+	Male	13,750	0	13,484,853	980.72	1.5982	13,484,853	980.72	1.5982
		Female	17,829	1	16,624,865	932.46	1.5921	16,620,916	932.24	1.5903
Adolescents	10-14	Male	1,345	0	3,493,060	2597.07	1.6611	3,493,060	2597.07	1.6611
		Female	1,377	0	3,519,786	2556.13	1.6887	3,519,786	2556.13	1.6887
Children	0-9		6,894	7	17,490,697	2537.09	1.6196	17,489,093	2536.86	1.6193
Overall			41,195	8	54,613,262	1325.73	2.0433	54,607,708	1325.59	2.0428

[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 2017 population counts within classes defined by gender an five-year age group.

Let N_{ga}^{2017} denote the 2017 Tanzania 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:

$$T_{ga}^{2017} = N_{ga}^{2017} / \sum_{k=1}^{n_{ga}^{BT}} W_{gak}^{(BT)},$$

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:

$$T_{(r)ga}^{2017} = N_{ga}^{2017} / \sum_{k=1}^{n_{(r)ga}^{BT}} W_{(r)gak}^{(BT)}$$

for the $r = 1, 2, \ldots, 257$ jackknife replicates.

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

$$W_{gak}^{(ps-BT)} = T_{ga}^{2017} W_{gak}^{(BT)},$$

and the corresponding poststratified replicate weights were computed as:

$$W_{(r)gak}^{(ps-BT)} = T_{ga}^{2017} W_{(r)gak}^{(BT)}$$

for $r = 1, 2, \ldots, 257$.

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

Table 3-18. 2017 Tanzania 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	4,974,120	4,630,642	1.0742	4,888,303	4,494,236	1.0877	9,862,423	9,124,877	1.0808
5-9	4,330,119	4,297,461	1.0076	4,233,202	4,066,755	1.0409	8,563,321	8,364,215	1.0238
10-14	3,646,522	3,493,060	1.0439	3,665,451	3,519,786	1.0414	7,311,973	7,012,846	1.0427
15-19	3,032,666	2,437,905	1.2440	3,032,037	2,663,977	1.1382	6,064,703	5,101,883	1.1887
20-24	2,511,475	1,850,321	1.3573	2,553,336	2,638,413	0.9678	5,064,811	4,488,734	1.1283
25-29	2,079,597	1,634,529	1.2723	2,203,407	2,281,761	0.9657	4,283,004	3,916,290	1.0936
30-34	1,747,429	1,415,696	1.2343	1,861,761	1,913,581	0.9729	3,609,190	3,329,277	1.0841
35-39	1,467,926	1,287,981	1.1397	1,544,828	1,684,235	0.9172	3,012,754	2,972,216	1.0136
40-44	1,191,963	1,134,798	1.0504	1,226,271	1,348,076	0.9096	2,418,234	2,482,874	0.9740
45-49	924,108	875,216	1.0559	948,441	1,013,029	0.9362	1,872,549	1,888,245	0.9917
50-54	697,713	757,026	0.9216	743,399	821,985	0.9044	1,441,112	1,579,011	0.9127
55-59	540,555	556,712	0.9710	599,059	547,757	1.0937	1,139,614	1,104,470	1.0318
60-64	402,499	481,828	0.8354	482,460	549,282	0.8783	884,959	1,031,110	0.8583
65+	795,277	1,052,840	0.7554	986,092	1,158,819	0.8509	1,781,369	2,211,659	0.8054
Total	28,341,969	25,906,016	1.0940	28,968,047	28,701,692	1.0093	57,310,016	54,607,708	1.0495

[1] Source: 2017 UN 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, $W_{gak}^{(BT)}$.

Weights for Analysis of the Hepatitis B Blood Test Results

4

Adults 15 years of age or older who reached the Biomarker module of the questionnaire were randomly selected for Hepatitis B (HepB) testing. A separate set of weights was constructed for analysis of the HepB blood test results as described below.

4.1 Selection Criteria

A probability of 0.04 was applied independently to each eligible adult to select participants for HepB testing. The criteria used to identify persons eligible for Hepatitis B testing are given in Appendix E.

4.2 Definition of Response Status

The THIS data set contains a variable HEPBFLAG which is set to 1 if the case is to receive a HepB blood test. However, some cases that are not survey eligible have a value of 1 for this flag, as do non-adult cases. Thus, it was necessary to define a new variable to determine the set of cases eligible for the adult HepB weighting.

The cases eligible to be weighted for analysis of the HepB results are those meeting the following conditions:

- The respondent has a confirmed age in years of 15 or older;
- The person is survey eligible for THIS (i.e., has an INDIV_STATUS code of 1 or 2);
- The person is flagged for the Hepatitis B test (i.e., HepBFlag = 1)

Table 4-1 summarizes the distribution of the survey eligible cases (INDIV_STATUS = 1) by age group (ADULT 15+), whether or not the person was flagged for HepB testing (HEPBFLAG), whether or not the person provided a blood test, and whether the HepB test provided an analyzable result (reactive/non-reactive). It can be seen that there are 1,312 cases meeting the above conditions, where all but one of these provided a valid HepB test result.

Note that there are also 84 cases where the confirmed age is 15+, the INDIV_STATUS is code 1, and the HepBFlag = 0 but the case has a valid HepB test result (i.e. ResultHepB = reactive or non-reactive). Since these cases were not part of the probability sample chosen for HepB testing, they will not receive a HepB weight.

For the 1,312 eligible adults who were selected to receive a Hepatitis B blood test, their response status for weighting purposes was based on whether the result of the test (RESULTHEPB) had a valid value of R (reactive) or NR (non-reactive). The response status variable Adult_HepB_STATUS was defined as follows:

- 0 if the person was not an adult, not survey eligible or not selected for HepB testing;
- 1 if the person is a survey-eligible adult selected for HepB testing and has a valid result; or
- 2 if the person is a survey-eligible adult selected for HepB testing and does not have a valid result

Table 4-1 Distribution of responses to the adult HepB questions

ADULT 15+ (1 = 15+, 0 = under 15)	INDIV_STATUS (1 = survey eligible)	HEPBFLAG (1 = yes, 0 = no)	BT_STATUS (1 = yes, 2 = no)	RESULTHEPB	Frequency
0	1		2		508
0	1	0	1		9,196
0	1	0	1	Non-reactive	25
0	1	0	2		17
0	1	1	1		20
0	1	1	1	Non-reactive	374
0	1	1	1	Reactive	1
1	1		2		1,412
1	1	0	1		30,190
1	1	0	1	Non-reactive	79
1	1	0	1	Reactive	5
1	1	0	2		6
1	1	1	1		1
1	1	1	1	Non-reactive	1,273
1	1	1	1	Reactive	38

4.3 Weighting

The following steps were implemented to construct the HepB weights.

- Each eligible person *k* in household *j* who was selected to be tested for HepB was assigned a base weight, $W_{jk}^{HepB:bw} = 25\, W_{jk}^{BT-nr}$, where W_{jk}^{BT-nr} = the nonresponse-adjusted blood test weight from the regular weighting process (see Section 3.4.4.2). The factor 25 is the reciprocal of the sampling rate used to randomly select individuals for HepB testing.
- Since only one of the 1,312 eligible persons selected for testing did not provide a valid HepB test result (see table 4-1), no further nonresponse adjustment of the HepB base weights was made.
- The last step was to apply a ratio-raking algorithm to simultaneously poststratify the $W_{jk}^{HepB:bw}$s to appropriate population counts defined by broad age groups and HIV status.

Table 4-2 summarizes results of the HepB weighting process.

Table 4-2 Selected statistics on the creation of the HepB weights

Sex-age group	Number selected for HepB testing	Base -weighted count of persons selected for HepB testing	Number of HepB respondents	Base -weighted count of HepB respondents	Weighted count of respondents after post-stratification (raking)
Males 15-49	437	13,503,247	437	13,503,247	12,820,599
Females 15-49	615	14,713,406	615	14,713,406	13,392,734
Males 50+	119	2,731,894	119	2,731,894	2,570,609
Females 50+	140	2,749,110	139	2,733,192	2,788,357
Male 15+	556	16,235,141	556	16,235,141	15,391,208
Female 15+	755	17,462,516	754	17,446,598	16,181,091
Total	1,311	33,697,656	1,310	33,681,739	31,572,299

Weights for Analysis of the Hepatitis C Blood Test Results

5

Adults 15 years of age or older who reached the Biomarker module of the questionnaire were randomly selected for Hepatitis C (HepC) testing. A separate set of weights was constructed for analysis of the HepC blood test results as described below.

5.1 Selection Criteria

A region-stratified sample of adults who participated in biomarker testing was eligible for HepC testing, whereby 171 eligible adults in each of the 31 regions were randomly selected to receive HepC testing.

5.2 Definition of Response Status

The THIS data set contains a variable HEPCFLAG which is set to 1 if the case is to receive a HepC blood test.

The cases eligible to be weighted for analysis of the HepC results are those meeting the following conditions:

- The respondent has a confirmed age in years of 15 or older;
- The person participated in biomarker testing (i.e., has a BT_STATUS code of 1);

Table 5-1 summarizes the distribution of the eligible cases, as determined by biomarker test participation (BT_STATUS = 1), age group (ADULT 15+), and results of the HepC test (reactive/non-reactive/not tested). It can be seen that there are 5,301 cases meeting the above conditions, where all but one of these was in fact tested for HepC.

All 5,301 adults who were eligible and selected to receive a HepC blood test were included for weighting purposes. The response status variable HEPCRESULT was defined as follows:

- . (blank) if the person was not an adult

- 99 (missing) if the person was an adult but did not participate in biomarker testing or was not selected for HepC testing
- -8 if the person was eligible and selected but was not tested
- 1 if the person is an eligible adult selected for HepC testing and has a reactive result; or
- 2 if the person is an eligible adult selected for HepC testing and has a non-reactive result

Table 5-1 Distribution of HepC eligibility criteria and results

ADULT 15+ (1 = 15+, 0 = under 15)	BT_STATUS (1 = received testing, 2 = not tested)	HEPCFLAG (1 = yes, 0 = no)	HEPCRESULT	Frequency
0	2			525
0	1			9619
1	2	0	Missing	1425
1	1	0	Missing	26278
1	1	1	Reactive	52
1	1	1	Non-reactive	5248
1	1	1	Not tested	1

5.3 Weighting

The following steps were implemented to construct the HepC weights.

- Each eligible person *k* in household *j* who was selected to be tested for HepC was assigned a weight, $W_{hjk}^{HepC:bw} = F_h^{HepC} W_{hjk}^{BT-ps}$,

  where $F_h^{HepC} = N_h/171$ is a weighting factor for each region *h* = 1, 2, …, 31 with N_h indicating the number of eligible individuals in that region. F_h^{HepC} thus represents the reciprocal of the probability that an individual in region *h* is one of the 171 individuals randomly selected for HepC testing;

  and W_{jk}^{BT-ps} = the nonresponse- and post-stratification adjusted blood test weight from the regular weighting process (see Section 3.4.4.2).

- Since only one of the 5,301 eligible persons selected for testing did not provide a valid HepC test result (see table 5-1), no further nonresponse adjustment of the HepB base weights was made.
- Post-stratification of the HepC weights was not done. Although the HepC weights are derived from the final blood test weights which had been poststratifed to population

control totals, weighted counts of respondents in the HepC subsample by sex and age group will differ from the corresponding weighted counts of respondents in the full sample of blood test respondents because the HepC subsample was not stratified by sex and age.

Table 5-2 summarizes results of the HepC weighting process.

Table 5-2 Selected statistics on the creation of the HepC weights

Sex-age group	Number selected for HepC testing	Number of HepC respondents	Weighted count of persons tested for HepC
Males 15-49	1,797	1,797	12,772,813
Females 15-49	2,406	2,405	13,034,698
Males 50+	501	501	2,370,427
Females 50+	597	597	2,927,451
Male 15+	2,298	2,298	15,143,240
Female 15+	3,003	3,002	15,962,148
Total	5,301	5,300	31,105,388

Weights for Analysis of Violence Module Data 6

Females 13 years of age or older who responded to the individual interview were eligible to receive the violence module (VM) of the survey. The sampling algorithm was designed to select one eligible female from each responding household. Females age 13-14 were to be selected from a subset of one-third of responding households designated for child data collection (child-flagged households). However, due to a programming error, the intended sampling algorithm was not implemented correctly. As a result, the selection probabilities for sampled females could not be determined for approximately one-half of sample respondents. Therefore, violence module data and weights are not provided in PHIA datasets. Please contact PHIA via the help request page at <https://phia-data.icap.columbia.edu/help/create> for more information.

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 variable as either being eligible for the study (ELIG_FLAG = Y) or not (ELIG_FLAG = N). In addition, dwelling units for which GPS coordinates were missing were coded as not eligible for sampling purposes.

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

Of the 56,048 listing records, 335 were coded as "discard" for various reasons and one had missing data for all three of the variables required to determine eligibility. These 336 cases were not eligible for sample selection. Of the remaining cases, 55,697 listing records in the Master Listing File were eligible for sampling. Examples of the eligible listing records are:

The 53,918 listings coded as 1, 1, 1 (that is, a single house/compound of HH; not vacant and not under construction; Occupied as a residence) were eligible for sampling (ELIG_FLAG=Y).

The 298 listings with codes 1, 1, 2 (that is, a single house/compound of HH; not vacant and not under construction; no one living in structure) were also eligible for sampling since such dwelling units could potentially be occupied at the time of interview.

The 693 listings with codes 1, 2, 2 were dwelling units that were vacant with no one currently living there, but could potentially have residents at the time of interview, and so they were considered eligible for sampling.

The 80 cases with codes 1, 3, 1 were currently under construction, but appeared to have people living there and were also considered eligible for sampling.

The 183 listings with codes 1, 3, 2 were under construction and had no one currently living there. Although it was not possible to ascertain whether these dwelling units would be completed by the time of interview, they were considered to be eligible to avoid possible undercoverage.

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

Structure type (STOBS_D)	Vac/Constr. status (STVAC_D)	Resid. status (RESYN_D)	ELIG_FLAG	Total in master file	Eligible
Cases with DM_FLAG = DISCARD			N	335	0
Cases with no data for any of the three variables			N	1	0
1 = Single house/Compound of houses	1 = Not Vacant/not under construct.	1 = Yes	Y	53,918	53,918
1 = Single house/Compound of houses	1 = Not Vacant/not under construct.	2 = No	Y	298	298
1 = Single house/Compound of houses	2 = Vacant	1 = Yes	Y	37	37
1 = Single house/Compound of houses	2 = Vacant	2 = No	Y	693	693
1 = Single house/Compound of houses	3 = Under Construction	1 = Yes	Y	80	80
1 = Single house/Compound of houses	3 = Under Construction	2 = No	Y	183	183
2 = Apartment bldg./Gated comm	1 = Not Vacant/not under construct.	1 = Yes	Y	187	187
2 = Apartment bldg./Gated comm	1 = Not Vacant/not under construct.	2 = No	Y	7	7
2 = Apartment bldg./Gated comm	3 = Under Construction	Missing	Y [1]	2	2
2 = Apartment bldg./Gated comm	3 = Under Construction	1 = Yes	Y	1	1
3 = Church/Mosque/Temple	1 = Not Vacant/not under construct.	1 = Yes	Y	8	8
3 = Church/Mosque/Temple	1 = Not Vacant/not under construct.	2 = No	N	1	0
4 = Community center	1 = Not Vacant/not under construct.	1 = Yes	Y	23	23
5 = School/University	1 = Not Vacant/not under construct.	1 = Yes	Y	93	93
5 = School/University	2 = Vacant	2 = No	Y	2	2
5 = School/University	3 = Under Construction	2 = No	N	1	0
6 = Shop/Business/Commercial bldg	1 = Not Vacant/not under construct.	1 = Yes	Y	161	161
6 = Shop/Business/Commercial bldg	1 = Not Vacant/not under construct.	2 = No	N	8	0
6 = Shop/Business/Commercial bldg	2 = Vacant	2 = No	N	3	0
7 = Clinic/Hospital/Dr. office	1 = Not Vacant/not under construct.	1 = Yes	Y	4	4
7 = Clinic/Hospital/Dr. office	3 = Under Construction	2 = No	N	1	0
8 = Other	3 = Under Construction	2 = No	N	1	0
TOTAL				56,048	55,697

[1] Cases are eligible for sampling because of availability of GPS data.

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

Definition of Household, Interview, and Blood Test Response Status

Definition of Household, Interview, and Blood Test Response Status

B.1 Survey Status for Household: HH_STATUS

Table B-1 Household response status codes (HH_STATUS)

Value	Meaning	Comments
1	Responding household	All households with Roster records
2	Nonresponding in-scope household	Household with a record, no roster data, and judged in-scope for the survey based on the RESULTNDT or RESULTNDTOTHR variables
3	Household not in scope for the survey	Households with a record, no roster data, and judged not in-scope for the survey based on the RESULTNDT or RESULTNDTOTHR variables
4	Household with no roster data, but unable to determine whether the household was in scope for the survey	In the weighting process the base weights for these cases is distributed among the other household records.

SAS Code for HH_STATUS

```
 attrib HH_eligible length=3 label="Household Eligibility flag – will be used to create
HH_STATUS";
   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 length=3 label="HH disposition code";
   if     HH_eligible = 1     then HH_STATUS= 1;                    /* Eligible Respondent */
   else if HH_eligible in(2,3) then HH_STATUS= 2;                    /* Eligible
NonRespondent */
   else if HH_eligible = 4     then HH_STATUS= 3;                    /* Ineligible */
```

else if HH_eligible = 5 then HH_STATUS= 4; /* Unknown eligibility Status */

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

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

Notes regarding this code:

The statement "if HH_ELIGIBLE = 2 and ROSTERCOUNT > 0 then HH_STATUS= 1" resets HH_STATUS to 1 = Eligible Respondent for "partly complete" households that have roster records. (The variable ROSTERCOUNT is created earlier in the program; it counts the number of individual records on the file phiazim_cff_roster_20161201 for each value of EA_HHID_FIXED.)

The statement "if HH_ELIGIBLE = 5 and UPCODE_STAT_HH in (2,3,4) then HH_STATUS = UPCODE_STAT_HH;" moves cases from HH_Status 4 = Unknown Eligibility Status to one of the other status codes that apply to household records with no response. (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.)

B.2 Survey Status for Individual Interview: INDIV_STATUS

Table B-2 Individual response status codes (INDIV_STATUS)

Value	Meaning	Comments
1	Responding, in-scope individual	Individual from in-scope household; for children, must also be in household with ChildFlag turned on, has questionnaire data and/or biomarker data
2	Nonresponding in-scope individual	Individual from in-scope household; for children, must also be in household with ChildFlag turned on, no questionnaire data or biomarker data
7	Rostered in error	based on “reason for no data” (15 cases)
8	Not Sampled	Child in household with Child Flag not turned on
9	De Jure Ineligible	Slept here last night? = NO

SAS Code for INDIV_STATUS

```
label CHILD_SMPYN = "CHILD IS SAMPLED Y/N"

if 0 <= AGEYEARS <= 14 and
   CHILD_SMPFLG   = "1"  then CHILD_SMPYN = 1;
else
  CHILD_SMPYN = 0;

label INDIV_AGEGROUP = "INDIVIDUAL AGE GROUP BASED ON BEST_AGE"

    IF CONFAGEY_RECODE ^= . THEN DO;
        BEST_AGE = CONFAGEY_RECODE;
        BEST_AGE_FLAG = 1;
    END;
    ELSE DO;
        BEST_AGE = AGEYEARS;
        BEST_AGE_FLAG = 2;
    END;

    IF GENDR ^= . THEN DO;
       BEST_GENDER = GENDR;
       BEST_GENDER_FLAG = 1;
    END;
    ELSE DO;
       BEST_GENDER = SEX;
       BEST_GENDER_FLAG = 2;
    END;
```

```
    if 0 <= BEST_AGE <= 9 then INDIV_AGEGROUP = 1;
    else
        if 10 <= BEST_AGE <= 14 then INDIV_AGEGROUP = 2;
        else
            if BEST_AGE >= 15 then INDIV_AGEGROUP = 3;
```

NOTE: Section above creates INDIV_AGEGROUP based on CONFAGEY_RECODE when available, otherwise AGEYEARS.

```
label INDIV_QXSTATUS = "Completion of questionnaire";

  if (INDIV_AGEGROUP=1 and (CH_KIDAGEY => 0 and CH_KIDGENDER > "0") and
        (CH_KIDENROLL > "0" or CH_KIDHIVTESTEVR > "0" or CH_KIDVISTTBCLIN >
        "0")) or
    (INDIV_AGEGROUP=2 and INDFINRSLT in ("1","2") and ADOLTSEND ^= .) or
    (INDIV_AGEGROUP=3 and INDFINRSLT in ("1","2") and MILESTONEA1 ="1" and
        MILESTONEA2 ^= "2" and MILESTONEA3 ^= "2" and MILESTONEA4 ^= "2" and
        MILESTONEA5 ^= "2" and MILESTONEA7 ^="2" and MILESTONEA8 ^= "2" and
        MILESTONEA9 ^= "2" and MILESTONEA10 ^= "2" and MILESTONEA11 ^= "2" and
        MILESTONEA12 ^= "2") Then INDIV_QXSTATUS = 1;
  else
    INDIV_QXSTATUS = 0;
end;
```

NOTE: INDIV_QXSTATUS analyzes the relevant interview variables for each INDIV_AGEGROUP. Value INDIV_QXSTATUS = 1 indicates enough interview data to consider the interview completed. For ages 0 -9 the determination is based on the Module 3A variables from the linked adult.

```
label INDIV_STATUS = "Individual Response Status"

IF SLEEPHERE=2 then INDIV_STATUS =9;
ELSE
  if 0<= AGEYEARS <=14 and CHILD_SMPYN = 0 then INDIV_STATUS = 8;
  ELSE
    if upcase(HIV1STATUSFINALSURVEY) in ("NEGATIVE", "POSITIVE") OR
(INDIV_QXSTATUS =1) then INDIV_STATUS = 1;
    ELSE
      INDIV_STATUS = 2;

IF UPCODE_STAT ^= . THEN DO;
       If UPCODE_STAT in (6, 7) then INDIV_STATUS = UPCODE_STAT;
       Else
```

```
        if INDIV_AGEGROUP =1  and
          INDIV_STATUS  ^= 9 and
          UPCODE_STAT    = 9 then INDIV_STATUS = UPCODE_STAT;
        Else
         if INDIV_AGEGROUP in (2, 3) and
           INDIV_STATUS not in (5, 8, 9) then INDIV_STATUS = UPCODE_STAT;
end;

IF BEST_AGE = . and
  BEST_GENDER = " " THEN INDIV_STATUS = 7;
```

B.3 BTEST Survey Status for Individual blood test data

Table B-3 Blood test response status codes (BTEST)

Value	Meaning	Comments
1	Has blood test	Responding individuals with hiv1statusFinalSurvey with values 'Positive' or 'Negative'
2	Does not have blood test	All other responding individuals

SAS Code for BTEST

```
 ATTRIB BTEST LABEL="Was blood test done: 1=YES, 2=NO";
   IF HIV1statusfinalsurvey In (1,3) THEN BTEST=1;
   ELSE                                BTEST=2;

NOTE: HIV1statusfinalsurvey is changed to numeric when read in:
VALUE   1 = '1 - Negative'
        2 = '2 - Unknown'
        3 = '3 - Positive'
```

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+ years)

TN_INDIV_STATUS.pdf (Adolescents 10-14 years)

CH_INDIV_STATUS.pdf (Children 0-9 years)

Blood Test

AM_BTEST.pdf (Males 15+ years)

AF_BTEST.pdf (Females 15+ 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

Tan_AD_INDIV.xlsx (Persons 15+ years)

Tan_TN_INDIV.xlsx (Adolescents 10-14 years)

Tan_CH_INDIV.xlsx (Children 0-9 years)

Blood Test

Tan_AM_BT.xlsx (Males 15+ years)

Tan_AF_BT.xlsx (Females 15+ years)

Tan_TM_BT.xlsx (Males 10-14 years)

Tan_TF_BT.xlsx (Females 10-14 years)

Tan_CH_BT.xlsx (Children 0-9 years)

Appendix D

Derivation of Household Control Totals Used in Poststratification

Derivation of Household Control Totals Used in Poststratification

The nonresponse-adjusted household weights for Geita region were trimmed and recalibrated to household control totals (i.e., estimated number of households in the region). The household control total for Geita was derived by dividing the 2017 population projections by the estimated average number of persons per household within the region.

The population projections used for person and blood test level poststratification was the 2017 national population counts published by the United Nations (UN). However, the UN population counts are not provided at the region level. Consequently, a regional population projection for Geita was derived using 2017 UN population projections, 2012 census counts and 2016 population projections published by the Tanzania NBS, and estimates of average household size obtained rom the THIS survey results.

An estimate of the 2017 Tanzania regional population was derived using data from the 2012 Tanzania Population Census and 2016 national population projections published by the Tanzania NBS as follows::

Let N_h^{2012} denote the 2012 Tanzania population count (based on the 2012 census) for region *h* and let N_h^{2016} denote the corresponding 2016 Tanzania population projection for region *h*. The average annual growth rate for region *h* was then computed as:

$$G_h = [(N_h^{2016} - N_h^{2012})/N_h^{2012}]/[2016 - 2012]$$

The 2017 population estimate for Tanzania based on data published by Tanzania NBS for region *h* was then computed as:

$$N_h^{2017} = G_h \, N_h^{2016}$$

The corresponding 2017 population projection for region *h* based on data from the 2017 UN projections was computed as:

$$U_h^{2017} = \left(U^{2017} / \sum_{h=1}^{h} N_h^{2017} \right) N_h^{2017}$$

where U^{2017} is the 2017 UN national population projection for Tanzania.

Next, the average number of persons per household in region *h* was derived from the THIS as:

$$A_h = \sum_h \sum_i \sum_{k=1}^{n_{hij}^{(1)}} W_{hijk}^{(int)} / \sum_h \sum_i \sum_{j=1}^{n_{hi}^{(1)}} W_{hij}^{(2A)}$$

where $W_{hij}^{(2A)}$ is the final nonresponse-adjusted household weight prior to trimming and recalibration for household *j* in PSU *i* in region *h*, and $W_{hijk}^{(int)}$ is the nonresponse-adjusted interview weight for person *k* in household *j* in PSU *i* in region *h*.

The estimated 2017 household control total for region *h* was then computed as:

$$D_h^{2017} = U_h^{2017} / A_h$$

The steps below provide specific details on the calculation of the household control total for Geita region, where the subscript *h* now refers specifically to Geita region:

G_h = [(N_h^{2016}-N_h^{2012})/ N_h^{2012}]/[2016-2012]
= [(1,932,230-1,739,530)/1,739,530]/[2016-2012]
= 2.77%

N_h^{2017} = $G_h\, N_h^{2016}$
= 2.77% x 1,932,230
= 1,985,742

U_h^{2017} = (U^{2017} / $\sum_{h=1}^{h} N_h^{2017}$) N_h^{2017}
= (57,310,016 / 51,630,039) x 1,985,742
= 2,204,199

A_h = $\sum_h \sum_i \sum_{k=1}^{n_{hij}^{(1)}} W_{hijk}^{(int)} / \sum_h \sum_i \sum_{j=1}^{n_{hi}^{(1)}} W_{hij}^{(2A)}$
= 3,424,675 / 627,363
= 5.46

D_h^{2017} = U_h^{2017} / A_h
= 2,204,199/5.46
= 403,785

Appendix E

Hepatitis B Eligibility Criteria, and Program Code

E.1 Eligibility Criteria for Hepatitis B Testing

The response status variable Adult_Hep_STATUS was created to identify individuals selected for HepB testing and their HepB test response status.

Adult_HepB_STATUS	Description
0	Not an adult (Under 15), Not survey eligible, or Not selected for HepB testing
1	Survey-eligible adult with valid HIV blood test selected for HepB testing, and has a valid HepB test result
2	Survey-eligible adult with valid HIV blood test, selected for HepB testing, and does not have a valid HepB test result

E.2 Code to Define HepB Response Status

```
DATA HEPB_prep;
   SET W100.BLOOD_DELIVERY(keep=EA_HHID_LN_FIXED CONFAGEY_RECODE
BT_STATUS HEPBFLAG RESULTHEPB HIV1STATUSFINALSURVEY);

    LABEL ADULT_HEPB_STATUS='ADULT HepB STATUS disposition status';
   If 15 <= CONFAGEY_RECODE & BT_STATUS =1 & HEPBFLAG= 1 then do;
    if RESULTHEPB in ("R","NR") then ADULT_HepB_STATUS = 1;
    Else
      if RESULTHEPB NOT in ("R","NR") then ADULT_HepB_STATUS = 2;
   end;
   Else ADULT_HepB_STATUS = 0;
   RUN;
```