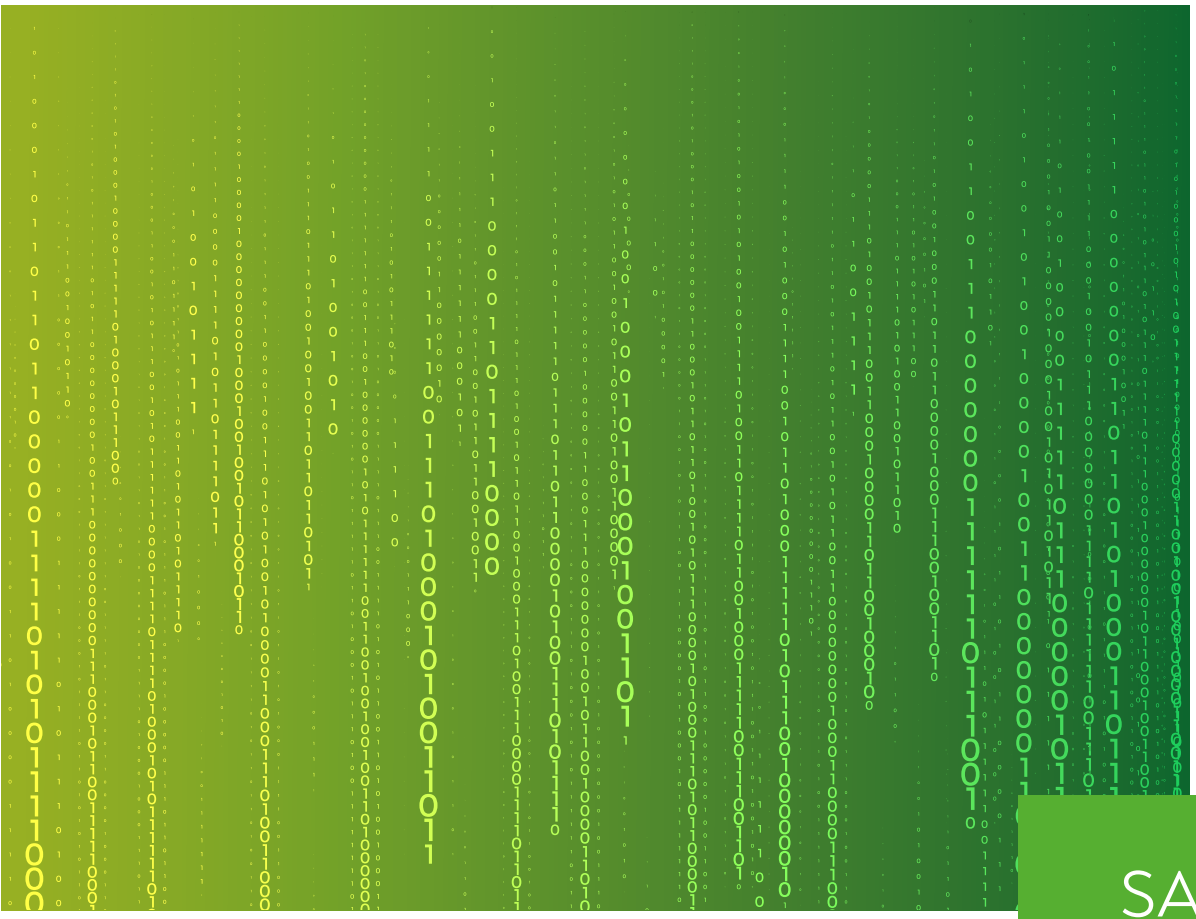
THIS 2022 2023

TANZANIA HIV IMPACT SURVEY



SAMPING AND WEIGHTING TECHNICAL REPORT

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

TANZANIA HIV IMPACT SURVEY 2022-2023
THIS 2022-2023

This project is supported by the US President's Emergency Plan for AIDS Relief (PEPFAR) through CDC under the terms of cooperative agreement U2GGH002173. The findings and conclusions are those of the authors and do not necessarily represent the official position of the funding agencies.

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

Table of Contents

Contents (continued)

Acronyms

CHAID	Chi-square Automatic Interaction Detector
CI	Confidence Interval
DEFF	Design Effect
DU	Dwelling Unit
EA	Enumeration Area
HH	Household
HIV	Human Immunodeficiency Virus
ICC	Intra Cluster Correlation
LASSO	Least Absolute Shrinkage and Selection Operator
MDRI	Mean Duration of Recent Infection
MOS	Measure of Size
NBS	Tanzania National Bureau of Statistics
PHIA	Population-based HIV Impact Assessment
PSU	Primary Sampling Unit
SAS	Statistical Analysis System
THIS	Tanzania HIV Impact Survey
UEW	Unequal Weighting
VLS	Viral Load Suppression

1. Introduction

The 2022-2023 Tanzania HIV Impact Survey (THIS 2022-2023) is a cross-sectional sample survey designed to assess the prevalence of key human immunodeficiency virus (HIV)-related health indicators among individuals 15 years or older. Data collection for THIS 2022-2023 was conducted between November 2022 and March 2023. The survey included over 35,000 interviewed individuals and over 33,000 individuals with valid blood tests in over 17,000 randomly-selected households. The purpose of this report is to document the procedures used to select the households and individuals for the study and the subsequent weighting of the respondent sample.

1.1 Overview of Sample Design

The sample design for THIS 2022-2023 is a stratified multistage probability sample design, with first-stage sampling units defined by enumeration areas (EAs) within twenty-eight[1] strata, second-stage sampling units defined by households within EAs, and finally age-eligible persons within households. Within each sampling stratum, the first-stage sampling units (also referred to as "primary sampling units" or PSUs) were selected with probabilities proportionate to number of households in the PSU based on the 2022 Population and Housing Census obtained from Tanzania National Bureau of Statistics (NBS). The allocation of the sample PSUs to the twenty-eight strata was based on HIV prevalence rate and the highest number of persons living with HIV that are not virally suppressed in a manner designed to achieve specified precision levels for: (a) national estimate of HIV incidence among persons 15 to 49 years old; (b) regional estimates of viral load suppression (VLS) rates among HIV-positive persons 15 to 49 years old for three priority tiers based on HIV prevalence rate[2]; and (c) regional estimates of viral load suppression (VLS) rates among HIV-positive 15 to 49 year olds for 3 regions with the highest number of persons living with HIV that are not virally suppressed[3].

[1] Out of 31 regions in Tanzania, the 5 Zanzibar regions were aggregated into two strata Pemba (which include 2 regions Kaskazini Pemba and Kusini Pemba) and Unguja (which include 3 regions Kusini Unguja, Mjini Magharibi and Kaskazini Unguja). The other 26 regions were treated as separate strata.

[2] Tier 1 low priority (<3% HIV prevalence rate), tier 2 medium priority (3-5.9% HIV prevalence rate) and tier 3 high priority (≥6% HIV prevalence rate)

[3] The 3 regions include Morogoro, Dodoma, and Dar es Salaam.

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

Within the responding households, all eligible persons 15 years of age and older who were present in the household on the night prior to the interview were included in the study sample for THIS 2022-2023 data collection.

Details of the sample design employed for THIS 2022-2023 are provided in Section 2.

1.2 Overview of Weighting Process

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

The main steps of the weighting process include

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

- Calculation of person-level blood test weights to reflect the differential probabilities of selecting individuals within households, compensate for nonresponse to the blood test; and

- Post-stratification of the person-level blood test weights to calibrate the weighed counts of persons so that they match external population counts to adjust for potential undercoverage.

Technical details of the weighting procedures employed for THIS 2022-2023 are provided in Section 3.

2. Sample Design

2.1 Population of Inference

The population of inference for THIS 2022-2023 is comprised of the *de facto* population of individuals 15 years of age and older. The *de facto* population is comprised of all individuals who were present in households (i.e., "slept in the household") on the night prior to the date of interview. In contrast, those individuals who are usual residents of the household regardless of whether they were present in the household during the previous night comprise the *de jure* population. Individuals belonging to either the *de facto* or *de jure* populations were included on the rosters compiled for sampling purposes; however, only members of the *de facto* population were eligible for data collection. Table 2-1 summarizes estimates (projections) of the 2022 Tanzania population by gender and age group.

Table 2-1 2022 population estimates for Tanzania by gender and age group

Age group	Gender		Total
	Male	Female	
15 to 49 years	13,750,057	14,992,264	28,742,321
50 years or older	3,109,923	3,488,887	6,598,810
Total	16,859,980	18,481,151	35,341,131

Source: 2022 Population and Housing Census from National Bureau of Statistics (NBS)

2.2 Precision Specifications and Assumptions

The following specifications and assumptions were used to develop the sample design for THIS 2022-2023. Three priority tiers based on HIV prevalence rates (among persons 15-49) were established to set precision requirements. The high priority tier consisted of 5 regions with HIV prevalence of 6% or higher plus 3 additional regions with the highest number of persons living with HIV that are not virally suppressed (Morogoro, Dodoma and Dar es Salaam); medium priority tier consisted of 13 regions with HIV prevalence between 3 and 6 percent; and the low priority tier consisted of 7 regions with less than 3 percent HIV prevalence rate.

2.2.1 Specifications

- Relative standard error of the national estimate of HIV incidence among adults 15 to 49 years old should be 30% or less
- The initial VLS precision goal for the high priority tier (Mwanza, Kagera, Mbeya, Iringa and Njombe, Morogoro, Dodoma and Dar es Salaam) was to achieve a 95% CI of +/- 10 percent in each region, resulting in 288 clusters assigned to this tier. These 288 clusters were reallocated to select 36 clusters in each of the 8 regions, resulting in a range of expected 95% CI's between +/-7.7 and +/-12 percent.
- The initial VLS precision goal for the medium priority tier was to achieve a 95% CI of +/- 20 percent in each region, resulting in 195 clusters assigned to this tier. These 195 clusters were reallocated to select 15 clusters in each of the 13 regions, resulting in a range of expected 95% CI's between +/-17 and +/-22 percent.
- The low priority tier did not have a VLS precision goal specified, but each region was assigned 12 clusters.

2.2.2 Statistical Assumptions

- National HIV prevalence rate of 4.7% for adults 15-49 years old that varies by region (see Table 2-2), (Source: 2016 Tanzania HIV Impact Survey (THIS 2016-2017));
- Annual national incidence rate for adults aged 15-49 of $p_a = 0.24\%$. (Source: THIS 2016-2017);
- Stratum-level incidence rates of p_{ah}, $h = 1, 2, \ldots, 28$, which are obtained by adjusting the national incidence rate using the regional prevalence rates as follows:

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

  where p_h and p are the HIV prevalence rates for stratum h and the country, respectively, and p_a is the annual national incidence rate (Source: THIS 2016-2017);
- Mean duration of recent infection (MDRI) of 130 days, yielding an annualization rate of 365/130= 2.8077;
- Estimated incidence rate for MDRI = 130 days of $p_m = 0.0024/2.8077 = 0.0855\%$. and the corresponding regional estimates are obtained as $p_{mh} = p_{ah}/2.8077$;
- Viral load suppression rate among HIV positive adults aged 15-49 of $p_{VLS} = 50\%$ in each stratum, which results in a conservative estimate of the underlying population variance associated with VLS rate;

- Intracluster correlation (ICC) of 0.0545 for VLS and 0.003 for prevalence (Source: tabulations of THIS 2016-2017 data);
- ICC of 0.000 for incidence (Source: analyses of prior PHIA surveys);
- Overall age distributions (Source: tabulations of THIS 2016-2017 data); and
- Regional population distribution obtained from NBS.

2.2.3 Operational Assumptions

- Varying numbers of households to be sampled per PSU, resulting in an average of 35 sampled households per PSU;
- Occupancy rate of 96.6% for sampled dwelling units (Source: THIS 2016-2017);
- Household response rate of 94.7% among occupied households (Source: THIS 2016-2017);
- Average household size of 4.396 (*de facto)* persons per household (Source: THIS 2016-2017);
- Overall percentage of *de facto* persons 15-49 years of age per household of 43.54%; and an overall percentage of *de facto* persons 50+ years of age of 10.91% (Source: THIS 2016-2017);
- Within the responding households, a person-level interview response rate of 91.5% (Source: THIS 2016-2017); and
- Among persons completing the interview, a blood test response rate of 94.9% (Source: THIS 2016-2017). Thus, among the persons selected for THIS 2022-2023, the assumed overall response rate for the blood tests is 86.9% (91.5% * 94.9%).

Based on the specifications and assumptions listed above, a sample of 567 EAs (clusters) was determined to be the minimum needed to meet the specified precision goals. The allocation of the sample to the twenty-eight strata is shown in Table 2-2. The expected numbers of households included in the study and the corresponding projected numbers of respondents by age group are also summarized in this table. The actual numbers of respondents achieved are presented in Sections 2.4 and 2.5 and differ from the counts in Table 2-2 because of differences between the response rates and other assumptions used to develop the sample design and those achieved during data collection. Further details about the sampling of households are given in Section 2.4.

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

Stratum code	Stratum name	Tier	HIV prevalence rate[1]	Total Number of sample clusters	Target Number of DUs/HHs to be sampled	Target Number of participating HHs[2]	Projected Number of respondents[3]	
			Adults 15-49				Adults 15-49	Adults 50+
1	Lindi	Low	0.003	12	420	384	638	160
2	Pemba	Low	0.002	12	420	384	638	160
3	Unguja	Low	0.005	12	420	384	638	160
4	Manyara	Low	0.018	12	420	384	638	160
5	Arusha	Low	0.019	12	420	384	638	160
6	Mtwara	Low	0.021	12	420	384	638	160
7	Kilimanjaro	Low	0.022	12	420	384	638	160
8	Kigoma	Medium	0.030	15	525	480	798	200
9	Mara	Medium	0.033	15	525	480	798	200
10	Simiyu	Medium	0.035	15	525	480	798	200
11	Singida	Medium	0.035	15	525	480	798	200
12	Rukwa	Medium	0.041	15	525	480	798	200
13	Tabora	Medium	0.048	15	525	480	798	200
14	Geita	Medium	0.052	15	525	480	798	200
15	Pwani	Medium	0.053	15	525	480	798	200
16	Katavi	Medium	0.054	15	525	480	798	200
17	Tanga	Medium	0.055	15	525	480	798	200
18	Ruvuma	Medium	0.055	15	525	480	798	200
19	Shinyanga	Medium	0.055	15	525	480	798	200
20	Songwe	Medium	0.056	15	525	480	798	200
21	Morogoro	High	0.039	36	1,260	1,152	1,915	480
22	Dodoma	High	0.042	36	1,260	1,152	1,915	480
23	Dar es Salaam	High	0.043	36	1,260	1,152	1,915	480
24	Mwanza	High	0.065	36	1,260	1,152	1,915	480
25	Kagera	High	0.068	36	1,260	1,152	1,915	480
26	Mbeya	High	0.092	36	1,260	1,152	1,915	480
27	Iringa	High	0.112	36	1,260	1,152	1,915	480
28	Njombe	High	0.116	36	1,260	1,152	1,915	480
All	Tanzania		0.047	567	19,845	18,144	30,160	7,560

DU = dwelling unit; HH= household

[1] Source: THIS 2016-2017

[2] Assumes occupancy rate of 96.6% and household response rate of 94.7%.

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

2.3 Selection of the Primary Sampling Units (PSUs)

2.3.1 Definition of PSUs

In THIS 2022-2023, the first-stage sampling units, PSUs, were EAs defined by NBS. The term PSU is the more general statistical term. The first-stage THIS 2022-2023 sample was selected from a sampling frame of EAs created and updated by NBS for the 2022 Tanzania Population and Housing Census. The updated sampling frame consisted of 103,385 EAs containing an estimated 12,120,213 households as of 2022.

2.3.2 Selection of the PSU Sample

A stratified sample of 567 EAs was selected from the updated EA sampling frame in accordance with the sample allocation given in Table 2-2. The following procedure was used to select the EAs for THIS 2022-2023. Within each stratum, the EAs in the updated sampling frame were sorted in the same way they had been sorted in the THIS 2016-2017 frame to the extent feasible; i.e., by region code, district code, council code, constituency code, division code, ward code, village code and EA number[4] . The sorting of EAs prior to sample selection induces an implicit geographic substratification within each stratum.

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

Following recommendations by NBS there was no need to replace any EAs found to have been selected previously for THIS 2016-2017.

[4] A 16-digit EA code defined by NBS with region code, district code, council code, constituency code, division code, ward code, village code and EA number was used.

2.3.3 Out-of-Scope PSUs

Out-of-scope PSUs are defined to be those EAs with no households (e.g., EAs that are no longer occupied due to flooding or other natural disasters, or where all residents have been permanently relocated). These are also sometimes referred to as "empty" PSUs. One PSU was out-of-scope because it was designated as a wildlife management area.

2.3.4 Non-responding PSUs and Substitution

A sampled PSU that contains eligible households is considered nonresponding if it cannot be entered. There was one PSU sampled for THIS 2022-2023 that was inaccessible because it included a prison. The data collection team was denied entry to the entire PSU by prison officers and families for security reasons. This PSU was replaced by a PSU in the same general area following guidelines developed for PHIA.

2.3.5 Summary of the PSU Sample

As indicated in the previous sections, 567 PSUs (EAs) were selected for THIS 2022-2023. There was one out-of-scope (ineligible) PSU and no nonresponding PSUs. Table 2-3 summarizes the distribution of the sampled PSUs by stratum and sampling status of the PSU.

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

Stratum code	Stratum name	Sampled PSUs	Nonresponding PSUs excluded from 2nd stage DU/HH selection	Ineligible PSUs excluded from 2nd stage DU/HH selection	Number of in-scope PSUs included in study
1	Lindi	12	0	0	12
2	Pemba	12	0	0	12
3	Unguja	12	0	0	12
4	Manyara	12	0	1	11
5	Arusha	12	0	0	12
6	Mtwara	12	0	0	12
7	Kilimanjaro	12	0	0	12
8	Kigoma	15	0	0	15
9	Mara	15	0	0	15
10	Simiyu	15	0	0	15
11	Singida	15	0	0	15
12	Rukwa	15	0	0	15
13	Tabora	15	0	0	15
14	Geita	15	0	0	15
15	Pwani	15	0	0	15
16	Katavi	15	0	0	15
17	Tanga	15	0	0	15
18	Ruvuma	15	0	0	15
19	Shinyanga	15	0	0	15
20	Songwe	15	0	0	15
21	Morogoro	36	0	0	36
22	Dodoma	36	0	0	36
23	Dar es Salaam	36	0	0	36
24	Mwanza	36	0	0	36
25	Kagera	36	0	0	36
26	Mbeya	36	0	0	36
27	Iringa	36	0	0	36
28	Njombe	36	0	0	36
All	Tanzania	567	0	1	566

DU = dwelling unit; HH= household

2.4 Selection of Households

The selection of dwelling units/households for THIS 2022-2023 involved the following steps: (1) listing all potentially eligible dwelling units/households within the sampled EAs, (2) assigning eligibility codes to the listed dwelling unit/household records based on characteristics of the listed units, and (3) selecting the sample of dwelling units/households from those records determined to be eligible for selection.

2.4.1 Definition of Second-Stage Sampling Units

For both sampling and analysis purposes, a household is defined to be a group of individuals who reside in a physical structure such as a house, apartment, compound, or homestead, and share in housekeeping arrangements. The physical structure in which people reside is referred to as the "dwelling unit" which may contain more than one household meeting the above definition. Households are eligible for participation in the study if they are located within the sampled EA. For PHIA, the sampling units are households. When vacancy of a "dwelling unit" cannot be determined, the "dwelling unit" is included on the household sampling frame. Therefore, the sampling frame and the sample of second-stage sampling units is a mixture of households and dwelling units.

2.4.2 Listing

A total of 243 sampled EAs were found to be too large for in-field listing and divided into segments, with one segment chosen for listing. There were three segmentation processes for Tanzania: (1) desk-segmentation, (2) field-segmentation, and (3) post-segmentation. Desk-segmentation was conducted by NBS prior to field listing to account for segments known to be too large to field based on the number of dwelling units in the census data. Field-segmentation was conducted during listing based on observations by the field data collectors. A post-segmentation process that was a correction applied to the listing file after field listing was completed for segments that did not appear to be fully listed. Estimates of the size of the segments included in listings and those segments excluded from listing were used to update the probability of selecting PSU i in stratum h prior to selecting dwelling units/households such that the PSU weights accounted for the population of the entire PSU. Table 2.4 summarizes the number of PSUs segmented by type of segmentation.

Table 2.4 Sampled and listed PSUs by type of segmentation

Type of segmentation	Number of PSUs
Not segmented	323
Desk-segmentation only	56
Field-segmentation only	93
Post-segmentation only	45
Desk- and field-segmentation	37
Desk- and post-segmentation	12
Total	566

In essence, the listing process involves compiling complete, up-to-date, and accurate lists of all dwelling units and households for each sampled EA or EA segment through a field operation using trained staff referred to as "listers." Local leaders and knowledgeable community members were consulted to assist in the listing process. Listers were provided with maps from which to delineate the boundaries of the EA, and to record the locations of the dwelling units/households found by the listers in the field. Information about the listed dwelling units/households was entered into computer tablets. The information recorded in the tablets included the address or description of the listed dwelling unit/household, the name of the head of household (where available), the type of structure (house, apartment, compound, etc.), occupancy status, and GPS coordinates. Vacant structures were listed along with occupied households. Slightly over 60,000 eligible dwelling units and households for sampling were listed for THIS 2022-2023.

2.4.3 Determination of Eligibility for Sampling

As indicated above, all known households at the time of listing, plus vacant dwelling units that could potentially be occupied at the time of interview, were initially entered into the tablets as separate records. However, not all of these records were eligible for subsequent sampling purposes. Those records marked with the notation "discard" were data entry errors and were eliminated from the listing file. To establish eligibility for the remaining records, three key variables collected during listing were used: (1) the structure type, (2) whether the listed structure was vacant or under construction, and (3) whether anyone was living in the structure at the time of listing. Based on the values of these three variables, those records meeting the criteria specified in Appendix A were eligible for second-stage sampling. Table 2-5 summarizes the total number of records entered into the tablets, the numbers of unoccupied dwelling units, households eligible for sampling, and the total number of dwelling units/households (records) eligible for sampling.

Table 2-5 Distribution of records in listing file by type of record, eligibility status, and stratum

Stratum code	Stratum name	Number of records (DUs/HHs) in listing file[1]	Number of unoccupied DUs/HHs[2]	Number of unoccupied DUs/HHs eligible for sampling[3]	Number of occupied DUs/HHs eligible for sampling[4]	Total number of DUs/HHs eligible for sampling
1	Lindi	1,138	22	22	1,116	1,138
2	Pemba	1,344	111	111	1,233	1,344
3	Unguja	1,111	76	75	1,034	1,109
4	Manyara	1,153	12	12	1,141	1,153

5	Arusha	1,156	18	18	1,138	1,156
6	Mtwara	952	4	4	948	952
7	Kilimanjaro	1,331	50	50	1,280	1,330
8	Kigoma	1,378	38	36	1,340	1,376
9	Mara	1,837	87	87	1,750	1,837
10	Simiyu	1,393	55	55	1,338	1,393
11	Singida	1,385	20	20	1,365	1,385
12	Rukwa	1,726	76	76	1,650	1,726
13	Tabora	1,357	45	45	1,312	1,357
14	Geita	1,947	116	116	1,831	1,947
15	Pwani	1,417	29	29	1,388	1,417
16	Katavi	1,725	63	62	1,660	1,722
17	Tanga	1,590	106	106	1,484	1,590
18	Ruvuma	1,521	39	39	1,482	1,521
19	Shinyanga	1,498	63	63	1,435	1,498
20	Songwe	1,421	74	72	1,346	1,418
21	Morogoro	3,699	172	172	3,527	3,699
22	Dodoma	3,763	112	112	3,651	3,763
23	Dar es Salaam	2,880	78	78	2,802	2,880
24	Mwanza	4,342	51	50	4,291	4,341
25	Kagera	4,561	219	216	4,341	4,557
26	Mbeya	3,851	184	184	3,667	3,851
27	Iringa	4,334	328	325	4,006	4,331
28	Njombe	4,249	316	314	3,933	4,247
All	Tanzania	60,059	2,564	2,549	57,489	60,038

DU = dwelling unit; HH= household

[1] See Appendix A for additional details.

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

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

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

2.4.4 Selection of Dwelling Units/Households

In order to achieve equal-probability samples of dwelling units/households within each of the twenty-eight sampling strata, the sampling rates required to select dwelling units/households within a PSU depend on the difference between the size measure used in sampling (i.e., the estimated number of households in the PSU based on the most recent preliminary census results) and the actual number of dwelling units/households found at the time of listing which took place between July and August 2022. Given that the EA counts were based on preliminary census results which were not yet finalized, there may be differences between the census counts and what is listed in the field. Thus, application of the within-PSU sampling rates based on the size measure used in sampling can yield more than the desired number of dwelling units/households in PSUs that list more dwelling units/households than is provided in the preliminary census counts, and fewer than the

desired number of dwelling units/households in PSUs that list fewer dwelling units/households than is provided in the preliminary census counts.

The calculation of the required within-PSU sampling rates proceeded as follows. First, the target overall sampling rate for stratum *h* = 1, 2, ..., 28, was computed as:

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

where

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

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

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

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

The total *expected* number of listings to be selected across all eight strata is $\sum_{h=1}^{28} T_h = 19{,}845$ (see Table 2-2). To obtain an equal probability sample within stratum *h*, the required within-PSU sampling rate for PSU *i* in stratum *h* was then computed as:

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

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

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

To reduce the variation in workload across the sampled PSUs, the maximum number of dwelling units/households to be selected in any PSU was capped at 70 and the minimum number was set to 15. Inspection of the values of $\mathrm{E}(n_{hi})$ indicated that the expected sample sizes for 43 PSUs would fall below 15, and 26 would exceed 70. The difference between the number of dwelling units/households that would have been selected using the rates, f_{hi}^{within}, and the specified maximum and minimum number was then re-distributed to the other PSUs in the same stratum so as to maintain as closely as possible the desired total sample size for the stratum. The within-PSU sampling rates, f_{hi}^{within}, were therefore adjusted to account for the redistribution of the sample

within the stratum. The adjusted within-PSU sampling rate used to select the sample of dwelling units/households, $f_{hi}^{adj(w)}$, was calculated as:

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

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

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

L = 15 = the minimum PSU sample size,

U = 70 = the maximum PSU sample size, and

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

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

2.4.5 Results of Second-Stage Sampling

Table 2-5 summarizes the numbers of dwelling units/households selected for the study and the minimum and maximum PSU sample size by stratum. The last column shows the unequal weighting (UEW) design effects (DEFF) to be expected for the selected sample. The UEW DEFF provides a measure of the increase in the variance of a sample-based estimate resulting from the use of variable overall sampling rates within a stratum (e.g., see Kish, 1965, page 403). With an equal-probability sample within each stratum, the DEFFs would ordinarily equal 1.0. Variable sampling rates within stratum will increase the DEFF, which would arise, for example, from the capping of sample sizes that is done to control workload across EAs. However, since the extent of the capping and redistribution of the sample described previously was moderate, the corresponding increase in the variation of the overall sampling rates was small, resulting in stratum-level UEW DEFFs that range from 1.00 to 1.20 (Table 2-5). The overall UEW DEFF is higher at 0 1.29 because it reflects total variation in weights within and across strata.

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

Stratum code	Stratum name	Number of PSUs	Number of sampled DUs/HHs	Minimum number of DUs/HHs selected per PSU	Maximum number of DUs/HHs selected per PSU	Unequal weighting DEFF
1	Lindi	12	420	7	70	1.02
2	Pemba	12	420	25	46	1.00
3	Unguja	12	420	20	54	1.00
4	Manyara	11	420	7	59	1.01
5	Arusha	12	420	10	70	1.05
6	Mtwara	12	420	15	57	1.00
7	Kilimanjaro	12	420	15	70	1.05
8	Kigoma	15	525	15	64	1.01
9	Mara	15	525	15	70	1.00
10	Simiyu	15	525	19	47	1.00
11	Singida	15	525	16	52	1.00
12	Rukwa	15	525	23	49	1.00
13	Tabora	15	525	15	70	1.02
14	Geita	15	525	18	52	1.00
15	Pwani	15	524	15	70	1.20
16	Katavi	15	525	15	70	1.00
17	Tanga	15	525	15	52	1.00
18	Ruvuma	15	525	15	70	1.01
19	Shinyanga	15	525	13	70	1.00
20	Songwe	15	525	15	70	1.00
21	Morogoro	36	1,260	1	70	1.01
22	Dodoma	36	1,260	15	70	1.01
23	Dar es Salaam	36	1,260	15	70	1.01
24	Mwanza	36	1,258	15	70	1.02
25	Kagera	36	1,260	16	70	1.00
26	Mbeya	36	1,260	15	70	1.05
27	Iringa	36	1,238	15	60	1.02
28	Njombe	36	1,260	15	70	1.01
All	Tanzania	566	19,820	15	70	1.29[1]

DU = dwelling unit; HH= household

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

Table 2-6 summarizes the distribution of the sampled dwelling units/households by final dwelling unit/household response status. Of the 19,820 sampled dwelling units, 1,171 (6.0%) were determined during data collection to be vacant/unoccupied, 62 (0.3%) were those for which eligibility for the survey (i.e., occupancy status) could not be established, 1,295 (6.5%) were determined to be eligible for the study (i.e., contained eligible household members) but did not complete the household interview, and 17,302 (87.3%) completed the household interview. Excluding the ineligible cases, the overall unweighted household response rate was 92.8%.

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

Stratum code	Stratum name	Number of sampled DUs/HHs	Number of ineligible DUs[1]	Number of DUs/HHs with unknown eligibility[2]	Number of responding households[3]	Number of eligible non-responding households[4]	Unweighted response rate[5]
1	Lindi	420	31	0	371	18	0.954
2	Pemba	420	24	1	386	9	0.975
3	Unguja	420	17	2	391	10	0.970
4	Manyara	420	38	2	338	42	0.885
5	Arusha	420	35	2	338	45	0.878
6	Mtwara	420	18	0	384	18	0.955
7	Kilimanjaro	420	22	0	372	26	0.935
8	Kigoma	525	24	1	472	28	0.942
9	Mara	525	27	4	459	35	0.922
10	Simiyu	525	27	0	475	23	0.954
11	Singida	525	22	0	462	41	0.918
12	Rukwa	525	34	2	436	53	0.888
13	Tabora	525	30	1	460	34	0.929
14	Geita	525	25	3	457	40	0.914
15	Pwani	524	22	0	471	31	0.938
16	Katavi	525	25	2	480	18	0.960
17	Tanga	525	34	1	463	27	0.943
18	Ruvuma	525	19	3	497	6	0.982
19	Shinyanga	525	41	6	437	41	0.904
20	Songwe	525	32	4	453	36	0.919
21	Morogoro	1,260	83	4	1,072	101	0.911
22	Dodoma	1,260	73	4	1,082	101	0.912
23	Dar es Salaam	1,260	57	11	1,112	80	0.925
24	Mwanza	1,258	45	0	1,098	115	0.905
25	Kagera	1,260	61	0	1,142	57	0.952
26	Mbeya	1,260	96	5	1,057	102	0.908
27	Iringa	1,238	97	2	1,075	64	0.942
28	Njombe	1,260	112	2	1,062	84	0.925
All	Tanzania	19,820	1,171	62	17,302	1,285	0.928

DU = dwelling unit; HH= household

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

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

[3] Households completing the household interview.

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

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

[6] A non-sampled household in Dodoma was inadvertently interviewed, included in the final dataset, and weighted. The error was discovered after data lock and weighting. The household was removed from the dataset with no change in the weighting. The subsequent tables in this report include the incorrectly sampled household and so, the number of people rostered, interviewed, and that had blood drawn will not match the counts in the final dataset.

2.5 Selection of Individuals

The selection of individuals for THIS 2022-2023 involved the following steps: (1) compiling a list of all individuals known to reside in the household or who slept in the household during the night prior to data collection; (2) identifying those rostered individuals who are eligible for data collection; and (3) selecting for the study those individuals meeting the age and residency requirements of the study. As noted below, only those individuals who were present (i.e., slept) in the household on the night prior to the time the household roster was compiled (i.e., the *de facto* population) were eligible for data collection and retained for subsequent weighting and analysis.

2.5.1 Household Rosters

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

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

Stratum code	Stratum name	Number of households completing interview	Rostered persons by resident status[1]				
			Usual resident/did not sleep here[2]	Usual resident/ slept here	Nonresident/ slept here	Nonresident/ did not sleep here[2]	Total rostered persons
1	Lindi	371	146	1,153	9	16	1,324
2	Pemba	386	203	2,016	79	3	2,301
3	Unguja	391	226	1,730	107	9	2,072
4	Manyara	338	118	1,393	27	32	1,570
5	Arusha	338	124	1,183	16	16	1,339
6	Mtwara	384	142	1,143	22	28	1,335
7	Kilimanjaro	372	158	1,072	36	57	1,323
8	Kigoma	472	223	2,468	55	80	2,826
9	Mara	459	191	2,183	54	64	2,492
10	Simiyu	475	241	3,129	93	149	3,612
11	Singida	462	198	2,618	71	38	2,925
12	Rukwa	436	206	1,877	27	25	2,135
13	Tabora	460	193	2,328	62	65	2,648
14	Geita	457	209	2,662	76	100	3,047
15	Pwani	471	208	1,624	54	55	1,941
16	Katavi	480	253	2,426	50	65	2,794
17	Tanga	463	301	1,913	59	43	2,316
18	Ruvuma	497	184	1,827	40	35	2,086
19	Shinyanga	437	143	2,053	42	48	2,286
20	Songwe	453	148	1,548	30	43	1,769
21	Morogoro	1,072	566	4,339	115	73	5,093
22	Dodoma	1,082	419	4,029	104	71	4,623
23	Dar es Salaam	1,112	421	3,421	94	122	4,058
24	Mwanza	1,098	451	5,527	150	274	6,402
25	Kagera	1,142	393	4,572	104	121	5,190
26	Mbeya	1,057	449	3,340	109	71	3,969
27	Iringa	1,075	383	3,625	131	186	4,325
28	Njombe	1,062	383	3,344	77	183	3,987
All	Tanzania	17,302	7,280	70,543	1,893	2,072	81,788

[1] Counts include persons of all ages.

[2] Not eligible to be surveyed for THIS 2022-2023.

2.5.2 Selecting Individuals for Data Collection

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

Table 2-8 Number of individuals eligible for data collection

Stratum code	Stratum name	Persons 15-49 years[1]			Persons 50 years or older[1]		
		Usual resident/ slept here	Nonresident/ slept here	Total sampled persons[2]	Usual resident/ slept here	Nonresident/ slept here	Total sampled persons[2]
1	Lindi	495	2	497	208	1	209
2	Pemba	780	44	824	256	1	257
3	Unguja	861	56	917	165	9	174
4	Manyara	567	13	580	136	5	141
5	Arusha	553	8	561	150	3	153
6	Mtwara	508	9	517	222	5	227
7	Kilimanjaro	438	20	458	262	10	272
8	Kigoma	951	29	980	268	8	276
9	Mara	857	33	890	225	8	233
10	Simiyu	1,200	43	1,243	273	19	292
11	Singida	970	42	1,012	330	10	340
12	Rukwa	727	17	744	188	6	194
13	Tabora	956	41	997	230	6	236
14	Geita	1,057	45	1,102	271	11	282
15	Pwani	761	24	785	228	3	231
16	Katavi	971	31	1,002	188	5	193
17	Tanga	735	35	770	291	6	297
18	Ruvuma	810	19	829	261	8	269
19	Shinyanga	900	19	919	191	9	200
20	Songwe	687	21	708	209	2	211
21	Morogoro	1,822	60	1,882	492	21	513
22	Dodoma	1,582	54	1,636	589	10	599
23	Dar es Salaam	1,910	50	1,960	332	7	339
24	Mwanza	2,286	91	2,377	539	14	553
25	Kagera	1,838	56	1,894	540	11	551
26	Mbeya	1,472	55	1,527	474	14	488
27	Iringa	1,444	64	1,508	607	21	628
28	Njombe	1,406	47	1,453	564	12	576
All	Tanzania	29,544	1,028	30,572	8,689	245	8,934

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

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

2.5.3 Distribution of Sampled Persons

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

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

Stratum code	Stratum name	Persons 15-49 years[1]			Persons 50 years or older[1]		
		Selected for data collection	Interview respondents[2]	Blood test respondent[3]	Selected for data collection	Interview respondents[2]	Blood test respondent[3]
1	Lindi	497	461	406	209	195	181
2	Pemba	824	760	751	257	243	238
3	Unguja	917	858	831	174	162	158
4	Manyara	580	485	424	141	113	105
5	Arusha	561	472	412	153	125	112
6	Mtwara	517	487	449	227	211	205
7	Kilimanjaro	458	425	404	272	254	238
8	Kigoma	980	880	853	276	251	241
9	Mara	890	801	766	233	221	216
10	Simiyu	1,243	1,105	1,021	292	234	225
11	Singida	1,012	905	839	340	316	306
12	Rukwa	744	672	638	194	178	172
13	Tabora	997	894	823	236	204	195
14	Geita	1,102	974	952	282	250	245
15	Pwani	785	720	629	231	215	188
16	Katavi	1,002	932	881	193	179	169
17	Tanga	770	699	658	297	279	262
18	Ruvuma	829	794	773	269	264	261
19	Shinyanga	919	826	800	200	177	174
20	Songwe	708	656	579	211	202	183
21	Morogoro	1,882	1,729	1,574	513	476	447
22	Dodoma	1,636	1,493	1,386	599	549	515
23	Dar es Salaam	1,960	1,714	1,452	339	294	249
24	Mwanza	2,377	2,156	2,016	553	490	469
25	Kagera	1,894	1,769	1,727	551	501	491
26	Mbeya	1,527	1,397	1,284	488	447	421
27	Iringa	1,508	1,388	1,293	628	584	557
28	Njombe	1,453	1,357	1,301	576	537	521
All	Tanzania	30,572	27,809	25,922	8,934	8,151	7,744

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

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

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

3. Weighting and Estimation

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

Nonresponse is unavoidable in virtually all surveys of human populations. For THIS 2022-2023, nonresponse can occur at different stages of data collection. For example: (1) after the selection of the sample of PSUs at the first stage; (2) after the enumeration of dwelling unit/household in each PSU at the second stage; (3) after household enumeration and selection of persons but before completion of the individual interview; and (4) after completion of the interview but before collection of a usable blood sample. The procedures used to compensate for nonresponse at each of the relevant stages of data collection are described in Section 3.4.

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

3.1 Overview of the Weighting Process

The overall weighting approach for THIS 2022-2023 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. All of the subsequent weighting steps described below are applied to the full sample, and to each of the jackknife replicates.

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

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

Calculation of Person-Level Interview Weights: Once the household weights are determined, they become the individual base weights for individuals found from the household roster to be eligible for the survey. Similar to the household weights, the first phase of individual weight adjustment is for any individuals whose eligibility is unknown. Eligibility is unknown when age was not confirmed at the interview stage. These adjusted individual weights are then adjusted for nonresponse among the eligible individuals, with a final poststratification adjustment for the individual weights to compensate for undercoverage in the sampling process by adjusting the weighted frequencies to correspond to 2022 population projections.

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

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

3.2 Preparation for Weighting

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

3.2.1 Data Files for Weighting

The THIS 2022-2023 survey data used to construct the sampling weights are contained in the following data files.

- **tz_CFF_hh_int_STAT_20230419**: A household (HH) file that contains the household data collected in the HH questionnaire.
- **tz_CFF_roster_STAT_20230419**: A file that contains the roster of household members collected in the HH questionnaire with a record for each rostered person.
- **tz_CFF_ind_int_STAT_20230419**: 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.
- **TZ2Biomarker20230426**: A biomarker file containing identifying information and results for lab analyses of blood samples for individuals whose blood was drawn and analyzed in the lab.

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

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

3.2.2 Checks of Data Files

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

- Check identification variables, merging household survey files with sampling files, and accounting for records found in one file and not the other. (This type of check for the EAs occurs as part of the HH selection process.)
- Check counts of sampled and responding HHs against what was expected, overall and by sampling stratum.
- Adjust for substitution of EAs, if applicable. Check that guidelines have been followed and selection probabilities are consistent with guidelines.
- Set disposition codes (respondent, eligible nonrespondent, ineligible, unknown eligibility) to be used for weighting purposes based on data elements received for (a) sampled households, (b) sampled individuals, and (c) individuals selected for blood draws.

3.3 Creation of Variables for Variance Estimation

Two general methods can be used for estimating the sampling errors of survey-based estimates derived from THIS 2022-2023: the jackknife replication and Taylor Series methods. The jackknife replication variance estimation method is a widely used method for producing variance estimates using data from a complex survey. This method can correctly account for the stratification, clustering, and sample weighting, including nonresponse and poststratification weighting adjustments, from the THIS 2022-2023 complex sample design. Taylor 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 Series

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

3.3.1 Jackknife Replication

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

As described in Section 2.3.2, the PSUs were selected systematically from a list of PSUs that had been ordered geographically within sampling stratum. To take account of the precision benefits of implicit stratification as fully as possible, the sampled PSUs within each sampling stratum 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 sampling stratum, 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 567 of the sampled PSUs.

For THIS 2022-2023, 277 variance-estimation strata were created. A jackknife replicate was then formed by randomly deleting a PSU from a particular variance-estimation stratum k, say, and retaining all of the PSUs in the remaining variance-estimation strata. For a variance-estimation stratum consisting of a pair of PSUs, the weight of the retained PSU within the variance-estimation stratum k was doubled. For a variance-estimation stratum consisting of three PSUs, the weight of the two retained PSUs within the variance-estimation stratum were increased by 1.5 (see Section 3.4.1). The process was repeated for all $k = 1, 2, 3, \ldots 277$ variance-estimation strata, resulting in a total of 277 jackknife replicates. Table 3-1 summarizes the number of jackknife replicates that were created for variance estimation.

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

Sampling stratum code	Sampling stratum name	Sampled PSUs[1]	Variance strata consisting of pairs	Variance strata consisting of triplets	Number of jackknife replicates
1	Lindi	12	6	0	6
2	Pemba	12	6	0	6
3	Unguja	12	6	0	6
4	Manyara	12	6	0	6
5	Arusha	12	6	0	6
6	Mtwara	12	6	0	6
7	Kilimanjaro	12	6	0	6
8	Kigoma	15	6	1	7
9	Mara	15	6	1	7
10	Simiyu	15	6	1	7
11	Singida	15	6	1	7
12	Rukwa	15	6	1	7
13	Tabora	15	6	1	7
14	Geita	15	6	1	7
15	Pwani	15	6	1	7
16	Katavi	15	6	1	7
17	Tanga	15	6	1	7
18	Ruvuma	15	6	1	7
19	Shinyanga	15	6	1	7
20	Songwe	15	6	1	7
21	Morogoro	36	18	0	18
22	Dodoma	36	18	0	18
23	Dar es Salaam	36	18	0	18
24	Mwanza	36	18	0	18
25	Kagera	36	18	0	18
26	Mbeya	36	18	0	18
27	Iringa	36	18	0	18
28	Njombe	36	18	0	18
All	Tanzania	567	264	13	277

[1] Includes nonresponding and ineligible PSUs if applicable.

3.3.2 Taylor Series

Even though jackknife replication is the recommended method for variance estimation, not all software packages have a replication option to produce variance estimates. Therefore, information for producing Taylor Series estimates of variance is included in the THIS 2022-2023 data files.

The full-sample weight (see Section 3.4) is used as the weight to compute Taylor Series variance estimates. The variable **VarStrat** indicates the variance-estimation stratum and the variable **VarUnit** indicates the PSU within the variance-estimation stratum. This pair of variables allows the analyst to

produce variance estimates if their software does not easily accommodate replication methods but does have a Taylor Series capability.

3.4 Development of Weights

3.4.1 PSU Weights

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

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

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

where P_{hi}^{PSU} = segmentation and substitution-adjusted probability of selecting PSU *i* from sampling stratum *h*. Using the PSU weights defined above, the sampled PSUs weight up to the numbers shown in the fourth column of Table 3-2.

As described in Section 3.3.1, 277 jackknife replicates were formed from the 567 sampled PSUs. For variance estimation, replicate-specific PSU weights, $W_{(r)bi}^{(1)}$, $r = 1, 2, ..., 277$ were created to provide the basis for calculating the required replicate weights in subsequent stages of the weighting process. Let *b* denote one of the variance-estimation strata created for jackknife replication within sampling stratum *h* (Section 3.3.1) and let *i* denote the PSU within variance-estimation stratum *b*. For a given jackknife replicate, $r = 1, 2, ..., 277$, the corresponding replicate-specific PSU base was computed as

$W_{(r)bi}^{(1)}$ =	$a\, W_{hi}^{(1)}$	if $b = r$ and PSU i in variance-estimation stratum b is included in replicate r
=	0	if $b = r$ and PSU i in variance-estimation stratum b is not included in replicate r
=	$W_{hi}^{(1)}$	if $b \neq r$

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

Since there are no nonresponding PSUs in THIS 2022-2023, no nonresponse adjustment is needed. The variables in the file that carry the nonresponse adjusted PSU weights are set equal to the PSU base weights.

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

Sampling stratum Code	Sampling stratum Name	Number of sample EAs (PSUs)	Weighted number of EAs (PSUs) [1]	Weighted measure of size (MOS) [2]
1	Lindi	12	3,179.49	302,826
2	Pemba	12	1,055.19	95,107
3	Unguja	12	3,055.07	246,614
4	Manyara	12	2,818.61	312,194
5	Arusha	12	6,634.30	523,436
6	Mtwara	12	5,641.29	478,027
7	Kilimanjaro	12	6,991.73	371,860
8	Kigoma	15	3,876.46	417,243
9	Mara	15	4,658.02	420,854
10	Simiyu	15	3,336.34	265,759
11	Singida	15	3,125.50	307,179
12	Rukwa	15	2,411.06	256,723
13	Tabora	15	7,056.48	493,434
14	Geita	15	4,859.43	404,554
15	Pwani	15	8,848.67	491,803
16	Katavi	15	2,464.90	180,800
17	Tanga	15	6,259.59	530,979
18	Ruvuma	15	4,441.54	402,821
19	Shinyanga	15	4,655.31	376,689
20	Songwe	15	3,433.75	285,857
21	Morogoro	36	7,750.75	672,968
22	Dodoma	36	7,971.03	589,614
23	Dar es Salaam	36	16,176.27	1,465,263
24	Mwanza	36	7,160.51	652,945
25	Kagera	36	6,067.55	605,543
26	Mbeya	36	5,955.15	552,601
27	Iringa	36	2,275.74	210,182
28	Njombe	36	2,510.02	206,338
All	Tanzania	567	144,670	12,120,213

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

[2] The measure of size used to select the sample of PSUs; the PSU Measure of Size (MOS) equals the number of households in the frame. Weights are the PSU base weights, $W_{hi}^{(1)}$.

3.4.2 Dwelling Unit/Household Weights

3.4.2.1 Dwelling Unit/Household Base Weights

The household weighting process starts by calculating the dwelling unit/household level base weights. These are the product of the PSU weight adjusted for nonresponse (described in Section 3.4.1 as equal to the PSU base weight for THIS 2022-2023) and the reciprocal of the within-PSU dwelling unit/household selection probability; i.e., the dwelling unit/household base weight for sampled dwelling unit/household *j* in PSU *i* in sampling stratum *h* was computed as:

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

where

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

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

The corresponding weights for jackknife replicate $r = 1, 2, \ldots, 277$ were computed as:

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

where $W_{(r)bi}^{(1A)}$ is the PSU nonresponse-adjusted weight for PSU *i* in variance estimation stratum *b* 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. The specific rules used to classify dwelling units into the response status groups are given in Appendix B. In Table 3-4, we show the weighted counts of dwelling units/households by response status and sampling stratum using the dwelling unit/household base weights described above. The characteristics of the dwelling unit/household base weights were checked by examining statistical summaries of the weights such as the mean weight, coefficient of variation of the weights, sum of the weights, and the minimum and maximum values of the weights, both overall and by sampling stratum.

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

Response status group[1]	Description	Number of sampled dwelling units/households
1	Respondent (household with completed household interview)	17,302
2	Nonrespondent (household without a completed household interview)	1,285
3	Ineligible (dwelling units with no households)	1,171
4	Unknown eligibility (not known if dwelling unit contains household)	62
All	—	19,820

[1] See Appendix B for definitions.

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

Sampling stratum code	Sampling stratum name	Response status[1]				Total groups 1-4
		Group 1: responding household	Group 2: nonresponding household	Group 3: ineligible dwelling unit	Group 4: unknown eligibility	
1	Lindi	256,597	12,083	21,319	0	289,999
2	Pemba	107,526	2,507	6,686	279	116,997
3	Unguja	260,132	6,653	11,310	1,331	279,426
4	Manyara	238,520	30,282	26,252	1,442	296,496
5	Arusha	532,907	63,470	51,900	2,878	651,155
6	Mtwara	412,808	18,930	18,930	0	450,668
7	Kilimanjaro	595,939	40,664	37,486	0	674,089
8	Kigoma	307,061	18,236	15,126	667	341,090
9	Mara	452,234	34,589	26,651	4,003	517,478
10	Simiyu	253,678	12,283	14,420	0	280,381
11	Singida	249,971	22,184	11,903	0	284,058
12	Rukwa	213,829	25,993	16,675	981	257,477
13	Tabora	517,526	39,402	33,777	1,083	591,787
14	Geita	471,361	41,257	25,786	3,094	541,498
15	Pwani	703,403	53,380	31,581	0	788,364
16	Katavi	248,752	9,452	13,290	1,091	272,586
17	Tanga	490,449	28,599	36,027	1,060	556,134
18	Ruvuma	395,158	4,843	15,084	2,337	417,421
19	Shinyanga	304,252	28,884	28,790	4,448	366,373
20	Songwe	260,386	20,911	18,624	2,346	302,267
21	Morogoro	585,707	55,575	45,695	2,218	689,196
22	Dodoma	699,888	64,645	47,040	2,572	814,145
23	Dar es Salaam	1,095,929	77,357	55,360	6,633	1,235,279
24	Mwanza	692,264	71,615	27,482	0	791,362
25	Kagera	646,587	32,159	34,677	0	713,423
26	Mbeya	509,772	50,761	48,373	2,798	611,704
27	Iringa	217,784	12,992	20,016	399	251,192
28	Njombe	214,053	17,044	22,494	406	253,998
All	Tanzania	11,934,472	896,750	762,754	42,065	13,636,041

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

3.4.2.2 Adjustment for Dwelling Unit/Household Nonresponse

The general approach for handling dwelling unit/household nonresponse was to increase the weights of responding households so that they represent the nonresponding dwelling units/households in the same PSU. Because such nonresponse could occur before establishing whether or not a sampled dwelling unit/household is eligible for the study (i.e., whether or not the associated dwelling unit/household contains persons eligible for THIS 2022-2023), the nonresponse adjustment was implemented in two phases. In the first phase of adjustment, the base weights were adjusted to compensate for sampled dwelling units/households for which eligibility for the survey (e.g., occupancy status) was not ascertained. In the second phase of adjustment, the first-phase adjusted weights were further adjusted to compensate for the nonresponding dwelling units/households among those dwelling units/households known to be eligible for the study.

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

Phase 1 Adjustment

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

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

where

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

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

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

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

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

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

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

The corresponding replicate-specific first-phase dwelling units/households nonresponse adjustment factor for cell *f* were similarly computed for jackknife replicate *r* = 1, 2, ..., 277.

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 *i* in PSU weighting cell *f* in sampling stratum *h* was then computed as:

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

The corresponding replicate weights for replicate r = 1, 2, …, 277 were computed in similar fashion as:

$$W_{(r)bij}^{DU} = A_{(r)hi}^{(DU)} W_{(r)bij}^{(2)},$$

where

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

Note that for the dwelling units in response-status group 4 (dwelling units/households of unknown eligibility), $W_{hij}^{DU} = W_{(r)bij}^{DU} = 0$ for r = 1, 2, …, 277.

The effect of this adjustment is to distribute the total weight of the unknown-eligibility cases (i.e., the estimated 42,065 dwelling units/households 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}^{DU} as computed above are summarized in Table 3-5.

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

Sampling stratum code	Sampling stratum name	Response status				Total households: groups 1-2
		Group 1: responding household	Group 2: nonresponding household	Group 3: ineligible dwelling unit	Total status 1-3	
1	Lindi	256,597	12,083	21,319	289,999	268,680
2	Pemba	107,787	2,524	6,686	116,997	110,311
3	Unguja	261,355	6,679	11,392	279,426	268,034
4	Manyara	239,721	30,415	26,359	296,496	270,137
5	Arusha	534,007	64,317	52,831	651,155	598,324
6	Mtwara	412,808	18,930	18,930	450,668	431,738
7	Kilimanjaro	595,939	40,664	37,486	674,089	636,603
8	Kigoma	307,709	18,236	15,144	341,090	325,945
9	Mara	455,464	35,116	26,898	517,478	490,580
10	Simiyu	253,678	12,283	14,420	280,381	265,961
11	Singida	249,971	22,184	11,903	284,058	272,154
12	Rukwa	214,634	26,019	16,824	257,477	240,653
13	Tabora	518,347	39,551	33,889	591,787	557,898
14	Geita	473,847	41,622	26,028	541,498	515,469
15	Pwani	703,403	53,380	31,581	788,364	756,783
16	Katavi	249,734	9,487	13,365	272,586	259,220
17	Tanga	491,379	28,642	36,113	556,134	520,021
18	Ruvuma	397,454	4,843	15,125	417,421	402,296
19	Shinyanga	307,449	29,648	29,276	366,373	337,097
20	Songwe	262,467	20,999	18,801	302,267	283,466
21	Morogoro	587,564	55,681	45,951	689,196	643,245
22	Dodoma	701,956	64,941	47,248	814,145	766,897
23	Dar es Salaam	1,101,701	77,874	55,704	1,235,279	1,179,575
24	Mwanza	692,264	71,615	27,482	791,362	763,880
25	Kagera	646,587	32,159	34,677	713,423	678,746
26	Mbeya	511,979	51,050	48,676	611,704	563,029
27	Iringa	218,150	12,997	20,044	251,192	231,147
28	Njombe	214,369	17,101	22,527	253,998	231,470
All	Tanzania	11,968,321	901,039	766,681	13,636,041	12,869,359

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

Phase 2 Adjustment

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

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

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

The weighted household interview response rate for cell *f* is $R_{hf}^{(HH)} = 1/A_{hf}^{(HH)}$.

The corresponding replicate-specific interview nonresponse adjustment factor for cell *f* were similarly computed for jackknife replicate *r* = 1, 2, ..., 277.

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

$$W_{hij}^{(2A)} = A_{hf}^{(HH)} W_{hij}^{DU}.$$

The corresponding replicate weights for replicate *r* = 1, 2, …, 277 were computed in similar fashion as:

$$W_{(r)bij}^{(2A)} = A_{(r)hf}^{(HH)} W_{(r)bij}^{DU},$$

where

$$A_{(r)hf}^{(HH)} = \sum_{j=1}^{n_{(r)hf}^{(1)}+n_{(r)hf}^{(2)}} W_{(r)bij}^{DU} \; / \; \sum_{j=1}^{n_{(r)hf}^{(1)}} W_{(r)bij}^{DU}.$$

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

3.4.3 Person-Level Interview Weights

In this section, we detail the calculation of person-level sampling weights to be used to analyze the individual interview responses in theTHIS 2022-2023 data files. First, we define the initial person-level (interview) base weights in Section 3.4.3.1. Next, to compensate for interview nonresponse, the person base weights are adjusted within cells defined by variables available for both the responding

and nonresponding individuals. Like the dwelling unit/household nonresponse adjustments described previously, this person-level nonresponse adjustment was implemented in two phases.

3.4.3.1 Person Base Weights

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

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

The corresponding replicate base weights, $W_{(r)bijk}^{(base)}$, for $r = 1, 2, \ldots, 277$ were computed in an analogous manner, with $W_{hij}^{(2A)}$ replaced by $W_{(r)bij}^{(2A)}$ in the above formula.

3.4.3.2 Adjustment of Person Weights for Interview Nonresponse

Since the final eligibility of a rostered person cannot be determined until after the actual age is confirmed during the interview, the person-level base weights were adjusted in two phases. Table 3-6 summarizes the distribution of the rostered persons by the five response-status groups specified for the first-phase adjustment. Response status groups 4 and 5 are the cases determined to be ineligible for the study because they were either under 15 years old, or because they were neither present in the household nor a usual resident of the household at the time the household roster was compiled. All of these cases are treated as "known ineligible" cases and are excluded from the first-phase adjustment. The cases in response-status group 3 are cases for which final eligibility for the study is not known because actual age was not obtained. The combined weight of these individuals was distributed to the cases in response-status groups 1 and 2 within weighting classes defined by sex and age group as described below.

Table 3-6 Distribution of rostered persons by age group and de facto status and eligibility

First-phase eligibility group[1]	Resident status and age based on roster	Confirmed age based on interview	Eligibility Status	Number of rostered persons	Weighted number of rostered persons[2]
1	De facto person 15 years or older	15+	Eligible	39,445	29,405,928.45
2	De facto person 15 years or older	Under 15	Ineligible	55	43,922.04
3	De facto person 15 years or older	Unknown	Eligibility Unknown	6	4,724.95
4	Non de facto persons 15 years or older	NA	Ineligible	7,600	5,609,893.76
5	Persons under 15 years	NA	Ineligible	34,682	25,247,832.08
All	—	—		81,788[3]	60,312,301[2]

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

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

[3] Of the 81,788 rostered persons, 2,072 were those that neither slept in the household nor were usual residents (see Table 2-7). On a weighted basis, these 2,072 persons account for 1,454,915 of the total weighted count of 160,312,301 rostered persons.

Phase 1 Adjustment

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

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

where c denotes the first-phase adjustment cell, $W_{ck}^{(base)}$ is the base weight for person k in cell c, and $n_c^{(a)}$ = the number of cases in eligibility-status group a = 1, 2, 3 (Table 3-6) in weighting class c.

The corresponding replicate-specific first-phase interview nonresponse adjustment factors for cell c were similarly computed for jackknife replicate r = 1, 2, ..., 277.

The first-phase weighted interview response rate for cell c is $R_c^{(1)} = 1/A_c^{(1)}$ for the full sample, and $R_{(r)c}^{(1)} = 1/A_{(r)c}^{(1)}$ for jackknife replicate r = 1, 2, ..., 277.

The full-sample first-phase nonresponse-adjusted weight for person k in cell c was then computed as

$$W_{ck}^{(3)} = A_c^{(1)} W_{ck}^{(base)},$$

and the corresponding jackknife replicate weights for replicate $r = 1, 2, ..., 277$ were similarly computed as

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

Phase 2 Adjustment

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

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

Sex/Age group[1]	Interview response status[2]	Unweighted sample size	Weighted count[3]
Male 15 or older	Eligible respondent	15,405	11,313,272
	Eligible nonrespondent	2,016	1,651,683
	All response statuses	**17,421**	**12,964,955**
Female 15 or older	Eligible respondent	20,555	15,290,943
	Eligible nonrespondent	1,469	1,154,725
	All response statuses	**22,024**	**16,445,668**
Total 15 years or older	Eligible respondent	35,960	26,604,215
	Eligible nonrespondent	3,485	2,806,408
	All response statuses	**39,445**	**29,410,623**

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

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

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

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

There are 44 variables from the household questionnaire and EA sampling frame that could potentially be used for nonresponse adjustment. The LASSO regression was used to reduce the number of variables to a manageable subset that would subsequently be entered into the CHAID algorithm to define the final nonresponse adjustment weighting cells. The LASSO is a restrictive procedure similar to linear regression that shrinks regression coefficient estimates to zero. In other words, predictors that are found to be not significant have their regression coefficients set to zero (Hastie, Tibshirani, and Friedman, 2009).

In the final model produced by the LASSO, only the most significant variables predictive of the response variable were identified and kept. The HPGENSELECT procedure (Johnston and Rodriguez, 2015) with selection method=lasso in SAS 9.4 was used to select the variables, with the weight set to the base weight adjusted for unknown eligibility, $W_{ck}^{(3)}$. The final model was selected on the basis of cross validation with observations in the input data set partitioned into disjoint subsets, reserving 25% for training, 50% for validation, and 25% for testing. As there is some randomness in how the LASSO selects the variables, we set the seed to a known constant value so that if the program had to be re-run, the same results would be produced. Of the 44 variables used in the initial model, the LASSO identified 24 variables as significant predictors of response.

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

The next step was to apply the CHAID algorithm (Magidson, 2005) to the variables selected by the LASSO procedure. CHAID classifies the sampled individuals (i.e., the respondents and nonrespondents) into weighting cells based on information available for all sampled persons. The cells are formed in such a way that persons belonging to the same cell are expected to have similar propensities for responding to the study. Using the variables selected by the LASSO as input, CHAID uses a weighted log-linear modeling algorithm for the computation of chi-square statistics associated with each predictor, where the weight is the person first-phase nonresponse-adjusted weight, $W_{ck}^{(3)}$. An output of the CHAID procedure is a tree diagram that specifies the optimum number of final weighting cells, and their definitions based on the input predictor variables. The depth limit of the tree was set to 5, and the minimum subgroup size required to allow splitting and minimum terminal node size were set to 50 observations (both respondents and nonrespondents).

To create the CHAID tree, gender (variable SEX) and an indicator of whether or not the individual was under 18 years of age (H_AGETEENYEARS) were forced into the model to make the initial splits. The reason for doing this is that males and females in the specified age groups received different questions; without forcing this variable into the model, the resulting tree would not have been created correctly. After forcing these two variables into the model, the tree was then allowed to grow freely. The CHAID algorithm identified 19 variables to create the weighting classes for nonresponse adjustment. Table 3-8 lists the variables that were included in the final CHAID models. The final trees produced by the CHAID algorithm are documented in Appendix C.1. The corresponding nonresponse-adjustment classes used to adjust the person-level base weights are given in Appendix C.2.

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

Variable number	Variable name	Description
1	DEATHS	Has Any Usual Resident Of Your Household Died Since January 1, 2020?
2	H_AGETEENYEARS	Teen Indicator: 1 – 15-17 Years Old; 2 – Otherwise; Based On Ageyears (Roster)
3	H_AGEYEARS	Age (Categorical), Based On Roster Age. Matches Poststratification Cells
4	H_HHQITEMS	1:Electricity; 2:Working Radio; 3:Working Television; 4:Working Telephone/Mobile Telephone; 5:Working Refrigerator; 6:None Of The Above
5	H_HHQOWN	1-Bicycle; 2-Working Motorcycle Or Motor Scooter; 3-Working Car Or Truck; 4-A Working Boat With A Motor; 5-None Of The Above
6	H_HH_SIZE_C	1-9, Where 9 Includes All HHs With 9 Or More People
7	H_MATEXWALLS	Main Material Of Exterior Walls
8	H_OWNCHIKNNUM	Altogether, How Many Of The Below Listed Animals Do Members Of Your Household Own?
9	H_OWNCOWNUM	How Many Of The Below Listed Animals Do Members Of Your Household Own?
10	H_OWNDOGNUM	Altogether, How Many Of The Below Listed Animals Do Members Of Your Household Own?
11	H_RELATTOHH	1. Head 2. Wife/Husband/Partner 3. Son Or Daughter 4. Son-In-Law/Daughter-In-Law 5. Grandchild 6. Parent 7. Parent-In-Law 8. Brother/Sister 9. Co-Wife 10. Other
12	H_ROOMSLEEP	How Many Rooms Are Used For Sleeping?
13	H_TOILETTYPE	12. Flush; 22 Ventilated Improved Pit; 23. Pit Latrine With Slab; 24. Pit Latrine Without Slab; 96. Other
14	H_WATERSOURCE	What Is The Main Source Of Drinking Water For Members Of Your Household?
15	LIVEHERE	Usually Live Here?
16	MATFLO	Material of the floor?
17	SEX	Male Or Female?
18	STRATA	Numeric Code For EA Sampling Stratum
19	URBAN_RURAL	2=urban, 1=rural

Calculation of Phase 2 Nonresponse-Adjusted Person Weights

The general approach for computing the second-phase nonresponse-adjusted person-level interview weights was as follows. Within each of the final adjustment cells specified in Appendix C.2, the interview nonresponse adjustment factor for cell m $A_m^{(int)}$, was computed as

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

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

The corresponding replicate-specific interview nonresponse adjustment factor for cell *m* were similarly computed for jackknife replicate *r* = 1, 2, ..., 277 as

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

The weighted interview response rate for cell *m* is $R_m^{(int)} = 1/A_m^{(int)}$ for the full sample, and $R_{(r)m}^{(int)} = 1/A_{(r)m}^{(int)}$ for jackknife replicate *r* = 1, 2, ..., 277.

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, ..., 277 were similarly computed as

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

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

Table 3-9 Summary of the interview nonresponse adjustment process

Characteristic	Total sample
Number of variables in initial model	44
Number of variables selected by LASSO	24
Number of variables selected by CHAID	19
Number of final nonresponse-adjustment cells	86
Number of interview respondents	35,960
Minimum adjustment factor	1.00
Maximum adjustment	1.77
Weighted count of respondents before adjustment[1]	26,604,215
Weighted count of respondents after adjustment[2]	29,410,623

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

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

3.4.3.3 Poststratification Adjustment

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

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

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

To compensate for the types of coverage problems indicated above, the nonresponse-adjusted person weights were ratio-adjusted so that the resulting weighted sample counts match the population control totals indicated in Table 3-10. The population control totals given in this table are projected 2022 national population projections by gender and five-year age groups provided by the CSO. The poststratified interview weights were computed as follows.

Let N_{ga}^{2022} denote the 2022 Tanzania population control total for gender *g* and (five-year) age group *a* as given in Table 3-10. The poststratification ratio adjustment factor for gender *g* and age group *a* was then computed as:

$$T_{ga}^{2022} = N_{ga}^{2022} / \sum_{k=1}^{n_{ga}^{resp}} W_{gak}^{(int)},$$

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:

$$T_{(r)ga}^{2022} = N_{ga}^{2022} / \sum_{k=1}^{n_{(r)ga}^{resp}} W_{(r)gak}^{(int)}$$

for the $r = 1, 2, \ldots, 277$ jackknife replicates.

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

$$W_{gak}^{(ps-int)} = T_{ga}^{2022} W_{gak}^{(int)},$$

and the corresponding poststratified replicate weights were computed as:

$$W_{(r)gak}^{(ps-int)} = T_{(r)ga}^{2022} W_{(r)gak}^{(int)}$$

for $r = 1, 2, \ldots, 277$.

Table 3-10 provides the population control totals, weighted counts of the respondents before poststratification, and the ratio of the control totals to the nonresponse adjusted weights (poststratification adjustment factor) by age and gender.

.

Table 3-10 2022 Tanzania population projections and weighted counts before poststratification

Age group	Male			Female			Total		
	Population control total[1]	Weighted count before post-stratification[2]	Poststrat-ification ratio[3]	Population control total[1]	Weighted count before post-stratification[2]	Poststrat-ification ratio[3]	Population control total[1]	Weighted count before post-stratification[2]	Poststrat-ification ratio[3]
15-19	3,096,584	2,160,395	1.433	3,185,807	2,437,088	1.307	6,282,391	4,597,483	1.366
20-24	2,560,782	1,788,850	1.432	3,005,170	2,503,779	1.200	5,565,952	4,292,628	1.297
25-29	2,220,550	1,562,980	1.421	2,507,599	2,186,725	1.147	4,728,149	3,749,705	1.261
30-34	1,902,811	1,409,802	1.350	2,053,126	1,749,218	1.174	3,955,937	3,159,019	1.252
35-39	1,532,154	1,123,825	1.363	1,676,186	1,623,635	1.032	3,208,340	2,747,460	1.168
40-44	1,315,193	1,028,274	1.279	1,396,314	1,323,739	1.055	2,711,507	2,352,012	1.153
45-49	1,121,983	895,237	1.253	1,168,062	1,065,787	1.096	2,290,045	1,961,024	1.168
50-54	906,155	800,552	1.132	947,230	984,256	0.962	1,853,385	1,784,808	1.038
55-59	616,934	544,774	1.132	636,409	622,356	1.023	1,253,343	1,167,130	1.074
60-64	555,227	563,579	0.985	595,851	639,605	0.932	1,151,078	1,203,184	0.957
65+	1,031,607	1,077,274	0.958	1,309,397	1,318,896	0.993	2,341,004	2,396,170	0.977
Total 15+	16,859,980	12,955,541	1.301	18,481,151	16,455,082	1.123	35,341,131	29,410,623	1.202

[1] Source: Tanzania National Bureau of Statistics (NBS)

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

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

3.4.4 Person-Level Blood Test Weights

Not every interview respondent provided a useable blood sample. Thus, a separate set of weights is required for analysis of the blood test results. Similar to the construction of the interview weights described previously, development of the final blood test weights involves adjustments for nonresponse and poststratification to 2022 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 $W_{(r)bijk}^{(int)}$ (for replicate $r = 1, 2, \ldots, 277$), respectively, where *k* denotes the interview respondent, *h* denotes the sampling stratum (*b* denotes the variance stratum), *i* denotes the PSU, and *j* denotes the household. These weights have been adjusted for interview nonresponse, and thus act as the "base" weights for developing nonresponse adjustments for the blood test weights. Table 3-11 summarizes the counts of individuals by sex, age group and blood test response status, and the corresponding weighted counts using the nonresponse person-level interview weights, $W_{hijk}^{(int)}$.

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

Age group[1]	Sex	Blood test response status[2]	Unweighted sample size	Weighted count[3]
15 to 49 years	Male	Eligible respondent	10,833	9,186,920
		Eligible nonrespondent	833	782,442
	Female	Eligible respondent	15,067	11,945,664
		Eligible nonrespondent	1,050	944,305
50 years or older	Male	Eligible respondent	3,538	2,812,519
		Eligible nonrespondent	192	173,660
	Female	Eligible respondent	4,228	3,358,707
		Eligible nonrespondent	219	206,406
15 years or older	Male	Eligible respondent	14,371	11,999,439
		Eligible nonrespondent	1,025	956,102
	Female	Eligible respondent	19,295	15,304,371
		Eligible nonrespondent	1,269	1,150,711

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

[2] Status among the interview respondents. See Appendix B for definitions of the response status groups.

[3] Weighted count of interview respondents using final nonresponse-adjusted person-level interview weight, $W_{hijk}^{(int)}$.

3.4.4.2 Nonresponse Adjustment of Blood Test Weights

To compensate for blood test nonresponse, the nonresponse-adjusted person-level interview weights were further adjusted within cells defined by variables available for both the responding and nonresponding individuals (i.e., individuals completing the interview who may or may not have a final HIV status determination). These variables included data from the household roster and other information collected in the household questionnaire, selected PSU characteristics such as sampling stratum and urban/rural status, and the individual interview. The age and sex variables used to make the nonresponse adjustments are those reported in the interview.

For males, 79 potential predictor variables were available for initial selection. For females, 80 potential predictor variables were available for initial selection. The LASSO procedure was used to identify a reduced set of predictor variables to be used in the CHAID algorithm. From these initial sets of variables, the LASSO regression identified 53 significant variables for males and 45 significant variables for females. The selected variables were then input into the CHAID program to create the final weighting cells for nonresponse adjustment.

The CHAID algorithm identified 29 variables for males and 24 variables for females that were then used to create weighting classes for nonresponse adjustment. Table 3-12 lists the variables that were included in the final CHAID models. The final trees produced by the CHAID algorithm are documented in Appendix C.1. The corresponding nonresponse-adjustment classes used to adjust the person-level base weights are given in Appendix C.2.

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

Sex	Variable number	Variable name	Description
Male	1	ADDISHIV	Have You Ever Discussed HIV With Your Parents Or Guardian?
	2	ALCFREQ	How Often Do You Have A Drink Containing Alcohol?
	3	ATTTRANSMIT	Do You Agree a Person Living With HIV Who Is Taking HIV Medications Cannot Pass HIV To A Sexual Partner?
	4	ATTVIRALLOAD	Do you Agree HIV Medications Decrease The Amount of HIV In The Blood Of People Living With HIV?
	5	AT_BESTAGE_C	Categorical Age Based On Interview Age (Confagey)
	6	AT_FIRSTSXAGE	Age Of First Sexual Activity - Limited To Maximum Of 21
	7	AT_LIFETIMESEX	In Total, With How Many Different People Have You Had Sex In Your Lifetime? - Limited To Maximum Of 10
	8	AT_PART12MONUM	How Many Different People Have You Had Sex With In The Last 12 Months? - Limited To Maximum Of 8
	9	AVOIDPREG	Are You Or Your Partner Currently Doing Something Or Using Any Method To Delay Or Avoid Getting Pregnant?
	10	BUYFOOD	Would You Buy Fresh Vegetables From A Shopkeeper Or Vendor If You Knew That This Person Had HIV?
	11	CONDOMGET	If You Wanted A Condom, Would It Be Easy For You To Get One?
	12	COOKINGFUEL	What Type Of Fuel Does Your Household Mainly Use For Cooking?
	13	COVVACSAFE	How Safe Do You Think A Covid-19 Vaccine Is For You?
	14	CURMAR	What Is Your Marital Status Now: Are You Married, Living Together With Someone As If Married, Widowed, Divorced, Or Separated/Single?
	15	FAMSHAME	Do You Agree You Would Be Ashamed If Someone In Your Family Had HIV?
	16	HFHIVTSTOFFER	During Any Of Your Visits To The Health Facility In The Last 12 Months, Did A Doctor, Clinical Officer Or Nurse Offer You An HIV Test?
	17	H_AGETEENYEARS	Teen Indicator: 1 – 15-17 Years Old; 2 – Otherwise; Based On Ageyears (Roster)
	18	KIDSSCHOOL	Do You Think That Children Living With HIV Should Be Able To Attend School With Children Who Are HIV Negative?
	19	KNOWN_HIV_STATUS_R	Categorical Known HIV Status
	20	LITTLEINTEREST	Over The Past Two Weeks, How Often Have You Been Bothered By Having Little Interest In Doing Things?
	21	MALEGUARDHHM	Does Name Have A Male Guardian Who Usually Lives In This Household Or was A Guest Last Night?
	22	MATFLOOR	Main Material Of Floor
	23	OUTREGIONTYPE	Just Before You Moved Here, Did You Live In A City, In A Town, Or In A Rural area?
	24	PREPWDTK	Would You Take Prep To Help Prevent HIV?
	25	SCHLCUR	Are You Currently Enrolled In School?
	26	SCHLHI	What Is The Highest Level Of School You Attended?
	27	STRATA	Numeric code for EA sampling stratum
	28	TOILETSHARE	Do You Share This Toilet Facility With Other Households?
	29	WORKIND	What Is Your Occupation? That Is, What Kind Of Work Do You Mainly Do?

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

Sex	Variable number	Variable name	Description
Female	1	ALCFREQ	How Often Do You Have A Drink Containing Alcohol?
	2	ATTVIRALLOAD	Do you Agree HIV Medications Decrease The Amount of HIV In The Blood Of People Living With HIV?
	3	AT_BESTAGE_C	Categorical Age Based On Interview Age (CONFAGEY)
	4	AT_FIRSTSXAGE	Age Of First Sexual Activity - Limited To Maximum Of 21
	5	AT_LIFETIMESEX	In Total, With How Many Different People Have You Had Sex In Your Lifetime? - Limited To Maximum Of 10
	6	AT_LIVEB	How Many Times Have You Had A Pregnancy That Resulted In A Live Birth? - Limited To Maximum Of 14
	7	BUYFOOD	Would You Buy Fresh Vegetables From A Shopkeeper Or Vendor If You Knew That This Person Had HIV?
	8	CHENSCH	Is Name Currently Enrolled In School?
	9	COVVAC	Have You Received A Covid-19 Vaccine?
	10	COVVACSAFE	How Safe Do You Think A Covid-19 Vaccine Is For You?
	11	DEATHS	Has Any Usual Resident Of Your Household Died Since January 1, 2020?
	12	DEPRESSED	Over The Past Two Weeks, How Often Have You Felt Down, Depressed Or Hopeless?
	13	FAMSHAME	Do You Agree You Would Be Ashamed If Someone In Your Family Had HIV?
	14	HFHIVTSTOFFER	During Any Of Your Visits To The Health Facility In The Last 12 Months, Did A Doctor, Clinical Officer Or Nurse Offer You An HIV Test?
	15	KNOWN_HIV_STATUS_R	Categorical Known HIV Status
	16	MALEGUARDHHM	Does Name Have A Male Guardian Who Usually Lives In This Household Or was A Guest Last Night?
	17	MATFLOOR	Main Material Of Floor
	18	MONTHOUTEVER	Have You Ever Lived Away From Home For More Than 1 Month At A Time?
	19	OUTREGIONTYPE	Just Before You Moved Here, Did You Live In A City, In A Town, Or In A Rural area?
	20	PREPWDTK	Would You Take Prep To Help Prevent HIV?
	21	SICK3MO	Has HHrname Been Very Sick For At Least 3 Months During The Past 12 Months, That Is Name Was Too Sick To Work Or Do Normal Activities?
	22	STRATA	Numeric code for EA sampling stratum
	23	URBAN_RURAL	2=Urban, 1=Rural
	24	WORK7DAYS	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 blood test weights was as follows. Within each of the final adjustment cells specified in Appendix C.2 for blood-test nonresponse adjustment factor for cell m, $A_m^{(BT)}$, was computed as

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

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

The corresponding replicate-specific nonresponse adjustment factor for cell *m* were similarly computed for jackknife replicate *r* = 1, 2, ..., 277.

The weighted blood test response rate for cell *m* is $R_m^{(BT)} = 1/A_m^{(BT)}$ for the full sample, and $R_{(r)m}^{(BT)} = 1/A_{(r)m}^{(BT)}$ for jackknife replicate *r* = 1, 2, ..., 277.

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

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

and the corresponding jackknife replicate weights for replicate *r* = 1, 2, ..., 277 were similarly computed as

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

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

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

Characteristic	Male	Female
Number of variables in initial model	79	80
Number of variables selected by LASSO	53	45
Number of variables selected by CHAID	29	24
Number of final nonresponse-adjustment cells	70	97
Number of interview respondents	14,371	19,295
Minimum adjustment factor	1.00	1.00
Maximum adjustment	1.60	1.50
Weighted count of respondents before adjustment[1]	11,999,439	15,304,371
Weighted count of respondents after adjustment[2]	12,955,541	16,455,082

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

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

3.4.4.3 Poststratification Adjustment

Like the nonresponse-adjusted interview weights described previously, the nonresponse-adjusted blood test weights were poststratified to projected 2022 Tanzania population counts within classes defined by gender and five-year age group.

Let N_{ga}^{2022} denote the 2022 Tanzania population control total for gender *g* and (five-year) age group *a* as given in Table 3-14. The poststratification ratio adjustment factor used to adjust the blood test weights for gender *g* and age group *a* was computed as:

$$T_{ga}^{2022} = N_{ga}^{2022} / \sum_{k=1}^{n_{ga}^{BT}} W_{gak}^{(BT)},$$

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:

$$T_{(r)ga}^{2022} = N_{ga}^{2022} / \sum_{k=1}^{n_{(r)ga}^{BT}} W_{(r)gak}^{(BT)}$$

for the *r* = 1, 2, …, 277 jackknife replicates.

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

$$W_{gak}^{(ps-BT)} = T_{ga}^{2022} W_{gak}^{(BT)},$$

and the corresponding poststratified replicate weights were computed as:

$$W_{(r)gak}^{(ps-BT)} = T_{(r)ga}^{2022} W_{(r)gak}^{(BT)}$$

for $r = 1, 2, \ldots, 277$.

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

Table 3-14 2022 Tanzania population projections and weighted counts of blood test respondents before and after poststratification

Age group	Male			Female			Total		
	Population control total[1]	Weighted count before post-stratification[2]	Poststrat-ification ratio[3]	Population control total[1]	Weighted count before post-stratification[2]	Poststrat-ification ratio[3]	Population control total[1]	Weighted count before post-stratification[2]	Poststrat-ification ratio[3]
15-19	3,096,584	2,178,462	1.421	3,185,807	2,445,471	1.303	6,282,391	4,623,933	1.359
20-24	2,560,782	1,794,722	1.427	3,005,170	2,517,708	1.194	5,565,952	4,312,430	1.291
25-29	2,220,550	1,531,170	1.450	2,507,599	2,185,561	1.147	4,728,149	3,716,731	1.272
30-34	1,902,811	1,395,027	1.364	2,053,126	1,736,757	1.182	3,955,937	3,131,783	1.263
35-39	1,532,154	1,090,926	1.404	1,676,186	1,592,635	1.052	3,208,340	2,683,561	1.196
40-44	1,315,193	1,023,774	1.285	1,396,314	1,313,788	1.063	2,711,507	2,337,562	1.160
45-49	1,121,983	895,236	1.253	1,168,062	1,084,845	1.077	2,290,045	1,980,081	1.157
50-54	906,155	807,886	1.122	947,230	977,998	0.969	1,853,385	1,785,884	1.038
55-59	616,934	555,630	1.110	636,409	620,115	1.026	1,253,343	1,175,745	1.066
60-64	555,227	572,301	0.970	595,851	645,773	0.923	1,151,078	1,218,074	0.945
65+	1,031,607	1,110,407	0.929	1,309,397	1,334,432	0.981	2,341,004	2,444,838	0.958
Total 15+	16,859,980	12,955,541	1.301	18,481,151	16,455,082	1.123	35,341,131	29,410,623	1.202

[1] Source: Tanzania National Bureau of Statistics (NBS).

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

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

References

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

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

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

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

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

2022 Tanzania Census: Provided by Tanzania National Bureau of Statistics (NBS), 2023

Appendix A

Definition of Eligibility for Dwelling Unit/Household Sampling

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

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

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

Of the 60,059 dwelling unit/household records in the listing file, 21 were classified as ineligible for sampling based on the structure type, vacancy status, and residential status. Thus, a total of 60,038 records in the Master Listing File were eligible for household sampling.

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

Structure type (STOBS_D)	Vac/Constr. Status (STVAC_D)	Resid. Status (RESYN_D)	ELIG_FLAG	Total in master file	Eligible
Cases with no GPS information			N	0	0
1 = Single House / compound of houses	1 = Not Vacant and not under const.	1 = Yes	Y	53,625	53,625
1 = Single House / compound of houses	1 = Not Vacant and not under const.	2 = No	Y	474	474
1 = Single House / compound of houses	2 = Vacant	1 = Yes	Y	95	95
1 = Single House / compound of houses	2 = Vacant	2 = No	Y	1,786	1,786
1 = Single House / compound of houses	3 = Under Construction	1 = Yes	Y	54	54
1 = Single House / compound of houses	3 = Under Construction	2 = No	Y	361	361
2 = Flat/Block/Apartment building	1 = Not Vacant and not under const.	1 = Yes	Y	2,846	2,846
2 = Flat/Block/Apartment building	1 = Not Vacant and not under const.	2 = No	Y	21	21
2 = Flat/Block/Apartment building	2 = Vacant	1 = Yes	Y	9	9
2 = Flat/Block/Apartment building	2 = Vacant	2 = No	Y	223	223
2 = Flat/Block/Apartment building	3 = Under Construction	1 = Yes	Y	2	2
2 = Flat/Block/Apartment building	3 = Under Construction	2 = No	Y	16	16
3 = Church/Mosque/Temple	1 = Not Vacant and not under const.	1 = Yes	Y	38	38
3 = Church/Mosque/Temple	1 = Not Vacant and not under const.	2 = No	N	2	0
3 = Church/Mosque/Temple	2 = Vacant	2 = No	N	3	0
3 = Church/Mosque/Temple	3 = Under Construction	2 = No	N	2	0

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

Structure type (STOBS_D)	Vac/Constr. Status (STVAC_D)	Resid. Status (RESYN_D)	ELIG_FLAG	Total in master file	Eligible
4 = Shop/office/bus. cntr/comm. bldg.	1 = Not Vacant and not under const.	1 = Yes	Y	388	388
4 = Shop/office/bus. cntr/comm. bldg.	1 = Not Vacant and not under const.	2 = No	N	4	0
4 = Shop/office/bus. cntr/comm. bldg.	2 = Vacant	1 = Yes	Y	2	2
4 = Shop/office/bus. cntr/comm. bldg.	2 = Vacant	2 = No	N	4	0
4 = Shop/office/bus. cntr/comm. bldg.	3 = Under Construction	2 = No	N	1	0
5 = School/University	1 = Not Vacant and not under const.	1 = Yes	Y	72	72
5 = School/University	1 = Not Vacant and not under const.	2 = No	Y	2	2
5 = School/University	2 = Vacant	1 = Yes	Y	1	1
5 = School/University	2 = Vacant	2 = No	Y	3	0
6 = Clinic/hospital/Doctors office	1 = Not Vacant and not under const.	1 = Yes	Y	16	16
7 = Community Center/CBO	1 = Not Vacant and not under const.	1 = Yes	Y	3	3
96 = Other	1 = Not Vacant and not under const.	1 = Yes	Y	4	4
96 = Other	2 = Vacant	2 = No	N	2	0
				60,059	60,038

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

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

Appendix B

Definition of Household, Interview, and Blood Test Response Status

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

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

B.1 Survey Status for Household: HH_STATUS

B.1.1 Summary

HH_STATUS is defined for all sampled dwelling units. First, the variable UPCODE_RESLTNDT is derived using RESULTNDTOTHR. Next, the questionnaire completion variable and the upcoded RESULTNDT are used to calculate UPCODE_STAT_HH. Lastly, HH_STATUS is set equal to UPCODE_STAT_HH when the Data Lock files are delivered.

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

B.1.2 SAS code defining HH_STATUS

HH_STATUS = UPCODE_STAT_HH;

Definition for responding household:

UPCODE_STAT_HH = 1 if:

- RESULTNDT is NULL and (STARTINT = 1 AND HHELIG = 1 AND HHCONSTAT = 1 AND HHQDTHSINS is NOT NULL AND ROSTER_START and HHRINS3 is NOT NULL AND HHQINSHH is NOT NULL AND HHQASSIGN_INST is NOT NULL) OR

- RESULTNDT is NULL and (STARTINT = 4 and HHRINS3 is NOT NULL and HHQINSHH is NOT NULL AND HHQASSIGN_INST is NOT NULL)

Definitions for household without completed questionnaire:

The table below shows the values for RESULTNDT on the data file:

CANNOT COLLECT CSPRO CODE (RESULTNDT)	Map to UPCODE_STAT_HH
1 = HH NOT AVAILABLE AT ALL VISIT ATTEMPTS	2 = NONRESPONDING HH
2 = REFUSED	2 = NONRESPONDING HH
3 = DWELLING VACANT OR ADDRESS NOT A DWELLING	3 = INELIGIBLE HH
4 = DWELLING DESTROYED	3 = INELIGIBLE HH
5 = DWELLING NOT FOUND	4 = UNKNOWN STATUS HH
6 = HOUSEHOLD ABSENT FOR EXTENDED PERIOD OF TIME	3 = INELIGIBLE HH
96 = OTHER	Will be upcoded to UPCODE_RSLTNDT

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

UPCODE_STAT_HH = 2 if

- RESULTNDT OR UPCODE_RESLTNDT = 1 or 2 or 7 or 8 or 9
- If RESULTNDT=NULL, then
  - If HHELIG = 2 OR
  - (HHCONSTAT = 2 or 3) or
  - HHELIG = 1 AND HHCONSTAT=NULL OR
  - STARTINT = 4 and HHQASIGN_INST is NULL or
  - (HHQINSHH is NOT NULL) and HHQASSIGN_INST is NULL

UPCODE_STAT_HH = 3 if

- RESULTNDT OR UPCODE_RESLTNDT = 3 or 4 or 6 or 91

UPCODE_STAT_HH = 4 if

- RESULTNDT or UPCODE_RESLTNDT = 5, 89, 90, 92, 93, 94, 95, 97 or 99 or
- STARTDWELL = 10 and the record does not meet the criteria for 1, 2, or 3

UPCODE_RESLTNDT categories for RESULTNDT = '96' cases, based on text in RESULTNDOTH

Text meaning	UPCODE_RESLTNDT	UPCODE_STAT_HH
HOUSEHOLD NOT AVAILABLE AT ALL VISIT ATTEMPTS	1	2
REFUSED	2	2
DWELLING VACANT OR ADDRESS NOT A DWELLING	3	3
DWELLING DESTROYED	4	3
DWELLING NOT FOUND	5	4
HOUSEHOLD ABSENT FOR EXTENDED PERIOD OF TIME	6	3
Bereavement related	7	2
No capable Head of Household available to do survey	8	2
Out of Scope	91	3
COVID Delay – Unknown Eligibility	94	4
Cannot Trace	95	4
Recorded in another HH or tablet (discrepant record)	99	4

Table of examples for RESULTNDOTH upcoding

RESULTNDOTH	UPCODE_ RESLTNDT	UPCODE_ STAT_HH
Not available at three occasions	1	2
HOUSEHOLD HEAD TOO BUSY TO ACCOMODATE SURVEY		
HOUSEHOLD HEAD NOT AVAILABLE FOR AN EXTENDED PERIOD OF TIME		
HOUSEHOLD HEAD IS AWAY IN SOUTH AFRICA AND WIFE IS NOT ABLE TO MAKE DECISIONS OR GIVE PERMISSION		
HHH IS AN ARTISAN MINOR HE COMES BACK AROUND 10 PM AND GOES VERY EARLY IN THE MORNING AROUND 4 AM		
KEPT GIVING APPOINTMENTS BUT WAS NOWHERE TO BE FOUND ON LAST DAY		
PARTICIPANT 'S WORK SHIFTS COULD NOT ACCOMMODATE SURVEY ACTIVITIES TO BE CONDUCTED.		
Refusing Behavior	2	2
COULD NOT ACCOMODATE SURVEY DUE TO RELIGIOUS AFFILIATION.THEY ARE FROM THE JOHANNE MARANGE CHURCH		
DATA CANNOT BE COLLECTED DUE TO STRONG RELIGOUS BELIEF		
HEAD OF HOUSE STATED THAT IF THERE ARE NO MONETARY BENEFITS HIS HOUSEHOLD SHOULD NOT BE INCLUDED		
PARTICIPANT REFUSED TO PARTICIPATE IN THE SURVEY AND THE REASON BEING DOMESTIC ISSUES.		
THE FAMILY WAS RECENTLY ATTACHED AND ROBBED BY ARMED ROBBERS AT GUN POINT. WRONG TIMING		
HH HEAD LISTED AGREED HOWEVER THE SON IS NOT ALLOWING THE PROCEDURES TO BE DONE		
Vacant or not a dwelling	3	3
STRUCTURE UNDER CONSTRUCTION STILL AT FOUNDATION LEVEL		
NO ONE SLEEPS AT THE HOUSE		
HOUSEHOLD HEAD DECEASED. DWELLING VACANT		
VACANT		
DWELLING IS A BOTTLESTORE		
Household absent for extended period of time	6	3
MEMBERS OF THE HOUSEHOLD HAVE TRAVELLED FOR A LONG PERIOD OF TIME		
THE INDIVIDUAL STAYS ALONE AND HE HAS TRAVELLED TO ARGENTINA AND THERE IS NOONE STAYING AT THE HOUSE		
Death/Funeral		
SHE LOST HER BOYFRIEND WHO WAS BURIED LAST SUNDAY. HE DIED OF LIVER PROBLEMS IN SOUTH AFRICA	7	2
FUNERAL AT THE HOUSEHOLD		
GRIEVING.SHE RECENTLY LOST A SON AND MOURNERS ARE STILL GATHERED		
NOT IN AN EMOTIONAL STATE TO PARTICIPATE, HH MISSING, DEATH OF A GRANDCHILD AND BIRTH OF CHILD		
CLOSE RELATIVE (DAUGHTER-IN-LAW) TO THE DECEASED BURIAL SCHEDULED		
Participant/Household Head unable to do survey (incapacitated, language barrier, under age)	8	2
HOUSEHOLD HEAD INCAPACITATED MENTALLY CHALLENGED		
THE PARTICIPANT IS INCAPACITATED -DEAF		
SINGLE HOUSEHOLD MEMBER WHO IS TOO OLD AND INCAPACITATED		
HH IS 14 YEARS OLD SO PARTICIPANT IS INELIGIBLE		

Table of examples for RESULTNDOTH upcoding (continued)

RESULTNDOTH	UPCODE_ RESLTNDT	UPCODE_ STAT_HH
HOUSEHOLD HEAD UNABLE TO SPEAK ANY OF THE SURVEY LANGUAGES		
THE HOUSEHOLD HEAD PASSED ON IN BULAWAYO ON THE 3RD DAY VISIT. NO ONE TO CONSENT FOR THE HOUSEHOLD		
HOUSEHOLD HEAD INVOLVED IN A CAR ACCIDENT THEREFORE CANNOT ACCOMODATE AN INTERVIEW		
MEMBERS OF THE HOUSEHOLD HAVE TRAVELLED FOR A LONG PERIOD OF TIME		
THE INDIVIDUAL STAYS ALONE AND HE HAS TRAVELLED TO ARGENTINA AND THERE IS NOONE STAYING AT THE HOUSE		
Out of Scope	91	3
COVID Delay – Unknown Eligibility	94	4
Cannot Trace	95	4
Recorded in another HH or tablet (discrepant record)	99	4

B.2 INDIV_STATUS

B.2.1 Summary

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

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

B.2.2 SAS Code for INDIV_STATUS

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

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

if confageysource = .  and CONFAGEY_INELIG = 0 then confagey = .;
else
if confageysource = .  and CONFAGEY_INELIG > 0 then confagey = confagey_inelig;
else
  confagey = confagey_orig;
```

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

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

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

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

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

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

B.3 BT_STATUS

B.3.1 Summary

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

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

B.3.2 SAS Code for BT_STATUS

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

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

Appendix C

CHAID Trees and Definition of Final Nonresponse-Adjustment Weighting Cells

Appendix C - CHAID Trees and Definition of Final Nonresponse-Adjustment Weighting Cells

C.1 Final CHAID Trees

The final CHAID trees used to construct the weighting cells for nonresponse adjustment are documented in PDF files in the zipped file APPENDIX_C.zip. There are three PDF files corresponding to the groups for which the CHAID analysis was conducted for adjustment of the interview weights (Section 3.4.3.2) and the blood test weights (Section 3.4.4.2). The names of the PDF files containing the CHAID trees are listed below. Each tree indicates diagrammatically how the final weighting cells were created by successively partitioning the sample into heterogeneous subsets with respect to response propensity. The final cells (prior to collapsing, if done to control variation in weights) are indicated by the number underneath the box defining the cell.

Individual Interview

AD_INDIV_STATUS.pdf (Persons 15+ years)

Blood Test

AM_BT_STATUS.pdf (Males 15+ years)

AF_BT_STATUS.pdf (Females 15+ years)

C.2 Final Nonresponse-Adjustment Weighting Cells

The final nonresponse-adjustment weighting cells are documented in Excel files in the zipped file APPENDIX_C.zip. There are three Excel files corresponding to the groups for which the nonresponse adjustments were made. The names of the Excel files are listed below. Each row of the Excel file corresponds to a weighting cell, and shows the variables and the corresponding values used to define the weighting cell, the numbers of responding and nonresponding cases in the cell, the weighted counts of the responding and nonresponding cases, the weighted response rate, and

the nonresponse weight adjustment factor (which is defined to be the reciprocal of the weighted response rate).

Individual Interview

TAN_AD_INDIV.xlsx (Persons 15+ years)

Blood Test

tz_AM_BT.xlsx (Males 15+ years)

tz_AF_BT.xlsx (Females 15+ years)