NAIIS 2018
NIGERIA HIV/AIDS INDICATOR AND IMPACT SURVEY
...getting to zero

NIGERIA HIV/AIDS INDICATOR AND IMPACT SURVEY

DATA USE MANUAL

MARCH 2020

PARTNERS

FEDERAL MINISTRY OF HEALTH
NATIONAL AGENCY FOR THE CONTROL OF AIDS NACA
NATIONAL POPULATION COMMISSION
nbs NATIONAL BUREAU OF STATISTICS
CDC CENTERS FOR DISEASE CONTROL AND PREVENTION
The Global Fund
UNAIDS
World Health Organization
UNIVERSITY of MARYLAND SCHOOL OF MEDICINE
INSTITUTE OF HUMAN VIROLOGY
UNIVERSITY of WASHINGTON
ICF
AFENET African Field Epidemiology Network A healthier Africa

This project is supported by the U.S. President's Emergency Plan for AIDS Relief (PEPFAR) through CDC under the terms of cooperative agreement #U2GGH002108.
The contents of this document do not necessarily represent the official position of the funding agencies.

NIGERIA HIV/AIDS INDICATOR AND IMPACT SURVEY (NAIIS) 2018
DATA USE MANUAL

NAIIS 2018 Collaborating Institutions

Federal Ministry of Health, Nigeria (FMoH)
National Agency for the Control of AIDS, Nigeria (NACA)
National Population Commission, Nigeria (NPopC)
National Bureau of Statistics, Nigeria (NBS)
The United States Centers for Disease Control and Prevention (CDC)
The Global Fund to Fight AIDS, Tuberculosis and Malaria (GF)
Center for International Health, Education, and Biosecurity (Ciheb) at the University of Maryland, Baltimore (UMB)
ICF International
African Field Epidemiology Network (AFENET)
University of Washington (UW)
The Joint United Nations Programme on HIV and AIDS (UNAIDS)
World Health Organization (WHO)
United Nations Children's Fund (UNICEF)

Donor Support and Disclaimer

This project is supported by the President's Emergency Plan for AIDS Relief (PEPFAR) through the Centers for Disease Control and Prevention (CDC) under the cooperative agreement #U2GGH002108 to the University of Maryland, Baltimore (UMB) and by the Global Fund to Fight AIDS, Tuberculosis and Malaria through the National Agency for the Control of AIDS (NACA), Nigeria, under the contract #NGA-H-NACA to UMB. The findings and conclusions of this report are those of the authors and do not necessarily represent the official position of the funding agencies.

Suggested Citation

Federal Ministry of Health, Nigeria. Nigeria HIV/AIDS Indicator and Impact Survey (NAIIS) 2018: Data Use Manual. Abuja, Nigeria. March 2020.

Access This Report Online

www.ciheb.org/PHIA

Contact Information

Federal Ministry of Health
New Federal Secretariat Complex
Phase 3
Ahmadu Bello Way
PMB 083 Garki, Abuja
Phone: +234 9 5238362
Email: info@nigeria.gov.ng
Website: www.fmh.gov.ng

National Agency for the Control of AIDS
No.3 Ziguinchor Street
Wuse Zone 4, Abuja
Phone: +234 9 4613726
Email: info@naca.gov.ng
Website: www.naca.gov.ng

Contents

List of Abbreviations

NAIIS: Nigeria HIV/AIDS Impact and Indicator Survey

PHIA: Population-based HIV Impact Assessment

VL: Viral Load

VLS: Viral Load Suppression

WHO: World Health Organization

µL: Microliter

1. Background

1.1 Overview of the NAIIS Indicator and Impact Assessment Survey

The Nigeria HIV/AIDS Indicator and Impact Survey (NAIIS) was a Population-based HIV Impact Assessment (PHIA) conducted to measure important national and regional HIV-related indicators, including progress toward the achievement of the UNAIDS 90-90-90 targets (UNAIDS, 2014) and to guide policy and funding priorities. PHIAs are part of a multi-country project funded by the United States President's Emergency Plan for AIDS Relief (PEPFAR) to conduct national HIV-focused surveys that describe the status of the HIV epidemic.

NAIIS was led by the Government of Nigeria (GoN) under the Federal Ministry of Health (FMoH) and National Agency for the Control of AIDS (NACA). The survey was conducted with funding from PEPFAR and the Global Fund to Fight AIDS, Tuberculosis, and Malaria (GF) with technical assistance from the U.S. Centers for Disease Control and Prevention (CDC). The survey was implemented by the NAIIS Consortium and led by the University of Maryland, Baltimore (UMB) under the supervision of the NAIIS Technical Committee.

1.2 Purpose of this Manual

The purpose of the NAIIS Data Use Manual, or "Manual", is to guide users through the process of accessing, exploring, and analyzing data obtained from NAIIS. The Manual includes four major components: (1) survey design, sampling, and measures, (2) dataset structure and variables, (3) data access and analysis, and (4) example code and other documentation.

Although NAIIS is one of many PHIAs that have been carried out in PEPFAR-supported countries, certain specifications of the NAIIS survey design are unique. Data users who require additional information regarding sampling methods, eligibility criteria, survey implementation, biomarker testing, should refer to the NAIIS Technical Report, and detailed descriptions of survey weighting procedures can be found in the NAIIS Sampling and Weighting Document.

1.3 Survey Sampling and Measures

1.3.1 Survey Sampling

NAIIS sampled the population using a two-stage cluster sampling technique, selecting enumeration areas (EAs) followed by households. The sampling frame consisted of 662,855 EAs, a total of 28,900,478 households, and 140,431,798 persons based on the 2006 Census, with an average number of households and persons per EA of 44 and 212, respectively. The EAs were mutually exclusive (non-overlapping). This ensured that all households and residents had an equal chance of being included in the survey. Given the variability in household size across Nigeria (range of 4.0 to 5.7 individuals per household), state differences in household size based on the 2006 Census were considered when calculating the number of EAs or primary sampling units (PSUs) to be selected in each state.

The sample size was calculated to provide a representative national estimate of HIV incidence and HIV prevalence among adults aged 15-64 years with a relative standard error less than or equal to 9% and 2%, respectively, as well as representing national and state estimates of VLS prevalence among PLHIV with 95% confidence intervals (CIs) between 10% and 15%. The sample size also was calculated to provide HIV prevalence estimates at the state level. One-quarter of the households were randomly

selected for inclusion of children, which was designed to provide a representative national estimate of pediatric HIV prevalence with a relative standard error less than or equal to 0.1205%. The target sample size was 140,974 adults and 31,629 children, for an overall total of 172,603 adults and children.

The first stage of sampling selected 4,035 EAs using a probability proportional to size method. The 4,035 EAs were stratified by Nigeria's 36 states and the FCT. An equal-size approach was proposed with an estimated sample size of 3,700 blood specimens from each state. This number of blood specimens was sufficiently large to obtain robust estimates of HIV prevalence for the population and VLS among HIV-infected individuals in most states. The second stage selected a random sample of households within each EA using an equal probability method. The average number of households selected per cluster was 28. Distributions of samples EAs and households by the state are detailed in Table 2.A of the NAIIS technical report, and additional information on sampling procedures is provided in the NAIIS Sampling and Weighting Document.

1.3.2 Survey Questionnaires

In selected households, the household questionnaire was administered to the head of the household after consent was obtained. Then, individual questionnaires were administered to eligible and consented individuals in the household. Adults (15+ years) completed the adult questionnaire, and adolescents (10-14 years) completed the adolescent questionnaire. Adults also provided data on their children (0-14 years) as part of the "children" module of the adult questionnaire. Modules included in each questionnaire and their general eligibility criteria are listed in the table below.

Questionnaire Module	**Eligibility Criteria**
Household Questionnaire	Sample of households within selected EAs
Household roster	
Support for orphans and vulnerable children (OVC)	
Household spouses/live-in partners	
Deaths	
Household characteristics	
Economic support	
Individual questionnaire – adults (15+ years)	All rostered[1] and consenting adults
Respondent background	
Marriage	
Reproductive history	All women
Children	Parents or guardians of children or adolescents (age 0-14) in the household provide education, health, and HIV-related data about each of their children
Male circumcision	All men
Sexual activity	
HIV knowledge	
HIV status, care, and treatment	All self-reporting HIV-positive adults
Tuberculosis	

Questionnaire Module (continued)	Eligibility Criteria (continued)
Individual questionnaire – adolescents (10-14 years)	All rostered[1] and consenting adolescents from selected households
Sociodemographic characteristics	
HIV knowledge	All adolescents
HIV prevention interventions	
Sexual behavior	
Violence	A sub-sample of adolescents, meeting varied sex and age criteria in specific PHIA surveys
HIV risk perceptions	
Social norms, intention to abstain, self-efficacy, and assertiveness	
HIV testing	
Alcohol and drugs	
HIV stigma	
[1] Household members are eligible to be rostered if they were confirmed to have slept in the household the night before the interview.	

1.3.3 Biomarker Testing

All adults and adolescents who completed an individual interview and consented/ assented to biomarker testing provided blood samples for testing. Administration of tests for specific biomarkers depended on age, HIV serostatus, and other eligibility criteria presented in the table below.

Biomarker Test	Eligibility Criteria
HIV serostatus[1]	All participants
Recency of HIV infection[2]	All HIV+ participants ≥ 18 months old
CD4+ cell count	All HIV+ and 2% of HIV- participants
Antiretroviral (ARV) drug presence	All HIV+ participants
ARV drug resistance	All HIV+ participants
Hepatitis B surface antigen	Adults aged 18-64 years and emancipated minors aged 15-17 years: all HIV+ and 5,303 HIV- participants
Hepatitis C antibody	Adults aged 18-64 years and emancipated minors aged 15-17 years: all HIV+ and 5,303 HIV- participants
[1] HIV serostatus was determined using the Nigerian National Serial HIV Rapid Testing Algorithm that combined results from household-based rapid and confirmatory tests. Appendix B, Figure B.1 and accompanying text in the NAIIS Technical Report provide a detailed description of the algorithm. [2] HIV-1 LAg avidity plus viral load and HIV-1 LAg avidity plus viral load and ARV detection were used to distinguish recent from long-term infection.	

1.3.4 Other Documentation and Resources

Additional NAIIS documentation can be accessed through Nigeria's National Data Archive. This documentation includes:

- ***NAIIS Tabulation Plan***: A complete collection of tables presented in official NAIIS publications, along with relevant datasets and variables.
- ***Survey Questionnaires***: Three questionnaires are provided for NAIIS: one for the household, adult, and adolescent questionnaires. These questionnaires demonstrate the questionnaire's structure, including the order of questions, asked, each question's wording, variable names and labels, value coding and labels, and skip patterns.
- ***Codebooks***: All datasets are accompanied by a codebook listing all contained variables with names, types (numeric or character), labels, coding values, and labels, and source (raw data or derived).
- ***Variable Frequencies***: Variable frequencies are provided, which contain frequencies of all categorical variables in each dataset.
- ***CONSORT Diagrams***: CONsolidated Standard of Reporting Trials (CONSORT) diagrams define the creation process for variables that were derived by NAIIS analysts.
- ***Sampling and Weighting Document***: Describes the sampling and weighting procedures used for NAIIS data.
- ***NAIIS Publications***: Includes summary sheets, infographics, and the comprehensive NAIIS technical report.

2. Guide to Analysis

2.1 Data Access

All NAIIS data and technical documentation can be accessed through the Nigeria Data Archive (NADA) portal, https://www.nigerianstat.gov.ng/nada/index.php/catalog. Users are required to specify the purpose of their research, indicate which datasets are required, and fill out the Data Use Agreement to obtain access credentials. Once a data request is approved, users are permitted to download datasets directly from the NADA portal.

This form must be filled and submitted by the Lead Researcher. Lead Researcher refers to the person who serves as the main point of contact for all communications involving this agreement. Access to licensed datasets will only be granted when the Lead Researcher is an employee of a legally registered receiving agency (university, company, research centre, national or international organization, etc.) on behalf of which access to the data is requested. The Lead Researcher assumes all responsibility for compliance with all terms of this Data Access Agreement by employees of the receiving organization.

This request will be reviewed by a data release committee, who may decide to approve the request, to deny access to the data, or to request additional information from the Lead Researcher. A signed copy of this request form may also be requested.

This request is submitted on behalf of:

* **Receiving organization name**	
* **Telephone (with country code)**	

* **Intended use of the data:**

Please provide a short description of your research project (project question, objectives, methods, expected outputs, partners)

* **List of expected output(s) and dissemination policy**

* **Expected completion date (DD-MM-YYYY) of the research project:**	

* **Research team members (other than the Lead Researcher)**

Provide names, titles, and affiliations of any other members of the research team who will have access to the restricted data.

DATA USE AGREEMENT

Data access agreement

The representative of the Receiving Organization agrees to comply with the following conditions:

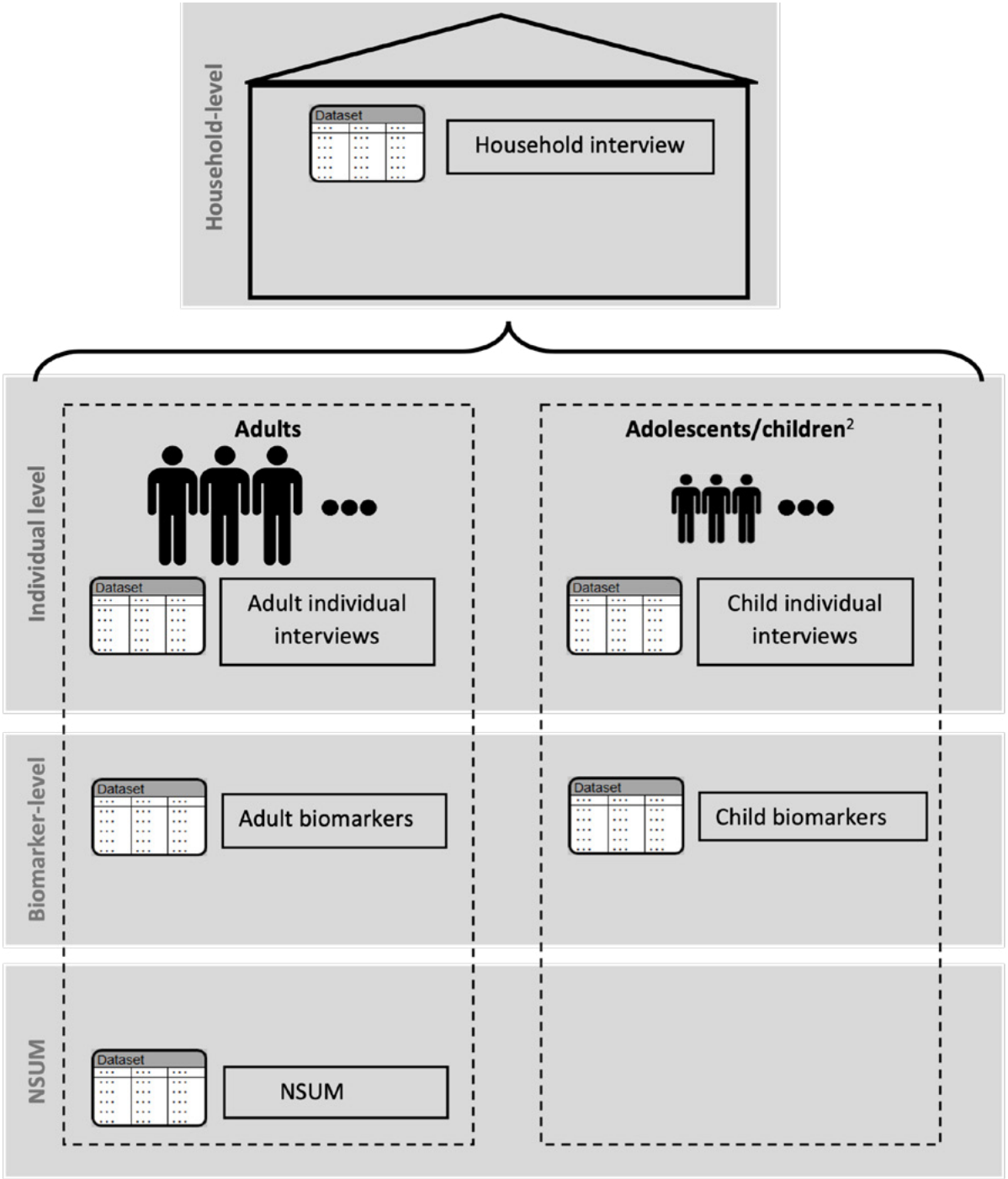
1. Access to the restricted data will be limited to the Lead Researcher and other members of the research team listed in this request.
2. Copies of the restricted data or any data created on the basis of the original data will not be copied or made available to anyone other than those mentioned in this Data Access Agreement, unless formally authorized by the Data Archive.
3. The data will only be processed for the stated statistical and research purpose. They will be used for solely for reporting of aggregated information, and not for investigation of specific individuals or organizations. Data will not in any way be used for any administrative, proprietary or law enforcement purposes.
4. The Lead Researcher must state if it is their intention to match the restricted microdata with any other micro-dataset. If any matching is to take place, details must be provided of the datasets to be matched and of the reasons for the matching. Any datasets created as a result of matching will be considered to be restricted and must comply with the terms of this Data Access Agreement.
5. The Lead Researcher undertakes that no attempt will be made to identify any individual person, family, business, enterprise or organization. If such a unique disclosure is made inadvertently, no use will be made of the identity of any person or establishment discovered and full details will be reported to the Data Archive. The identification will not be revealed to any other person not included in the Data Access Agreement.
6. The Lead Researcher will implement security measures to prevent unauthorized access to licensed microdata acquired from the Data Archive. The microdata must be destroyed upon the completion of this research, unless the Data Archive obtains satisfactory guarantee that the data can be secured and provides written authorization to the Receiving Organization to retain them. Destruction of the microdata will be confirmed in writing by the Lead Researcher to the Data Archive.
7. Any books, articles, conference papers, theses, dissertations, reports, or other publications that employ data obtained from the Data Archive will cite the source of data in accordance with the citation requirement provided with the dataset.
8. An electronic copy of all reports and publications based on the requested data will be sent to the Data Archive.
9. The original collector of the data, the Data Archive, and the relevant funding agencies bear no responsibility for use of the data or for interpretations or inferences based upon such uses.
10. This agreement will come into force on the date that approval is given for access to the restricted dataset and remain in force until the completion date of the project or an earlier date if the project is completed ahead of time.
11. If there are any changes to the project specification, security arrangements, personnel or organization detailed in this application form, it is the responsibility of the Lead Researcher to seek the agreement of the Data Archive to these changes. Where there is a change to the employer organization of the Lead Researcher this will involve a new application being made and termination of the original project.
12. Breaches of the agreement will be taken seriously and the Data Archive will take action against those responsible for the lapse if willful or accidental. Failure to comply with the directions of the Data Archive will be deemed to be a major breach of the agreement and may involve recourse to legal proceedings. The Data Archive will maintain and share with partner data archives a register of those individuals and organizations which are responsible for breaching the terms of the Data Access Agreement and will impose sanctions on release of future data to these parties.

☐ I have read and agree with the conditions Submit

Data are available for download in SAS (.sas7bdat), Stata (.dta), and csv formats. Users may also download comma-separated value type datasets for analysis outside of the aforementioned software packages.

2.2 Structure of NAIIS Datasets

2.2.1 NAIIS Data Structure[1]



[Order of variables]

Each row in the household dataset represents a single, sampled household, and includes those that were deemed ineligible or non-responding. Individual households are identified with the householdid variable, and those that participated in the survey are indicated as eligible when the hhstatus variable = 1.

The individual interview datasets contain survey questionnaire responses for all eligible and consenting participants. Each participant is identified by the PERSONID variable. Participants are categorized into three age groups: adults aged 15-64 years, adolescents aged 10-14 years, and children aged 0-9 years. The adult interview dataset contains questionnaire responses from those aged 15-64 years, and the child interview dataset contains questionnaire responses from those aged 0-14 years. Note that certain information for children aged 0-14 years is provided by an adult in the household. In the child interview dataset, all information for children aged 0-9 years is provided by their parent/ guardian and not the child him/ herself. Records in the child interview dataset for adolescents aged 10-14 years contain information provided by a parent /guardian, as well as information provided by the adolescent him/ herself as part of the adolescent questionnaire.

If a participant completed an individual interview and consented to undergo a blood draw for HIV testing, his/ her information will be contained in the biomarker dataset as a single row. Similar to the interview datasets, biomarker datasets are separated by age group into a child (aged 0-14 years) and adult (aged 15-64 years). The same identifier variable (PERSONID) is also used in the biomarker dataset, and individuals who underwent biomarker testing that resulted in valid laboratory test results are indicated by bt_status = 1.

2.2.2 Variable Labels and Formats

Information regarding variable labels, value codes, exact language used in questionnaire items can be obtained from the NAIIS codebook. Additionally, SAS users can pre-load formats for all NAIIS datasets by downloading format libraries (.sas7bcat) and applying formats to desired variables. Variable formats follow a consistent naming convention that begins with the name of the corresponding variable followed by the suffix “f”. For example, the corresponding format for the variable, “education”, is “educationf”.

2.2.3 Protecting Participant Confidentiality

To protect participant confidentiality, the EA, household and participant identifiers have been scrambled in ensure that participant identities cannot be ascertained. EAs and households were randomized, sorted and sequentially assigned new IDs. The householdid was then concatenated with the individual line number to create personid. As an additional precaution, all personally identifiable information including but not limited to names, phone numbers, addresses, and month/ day-specific birth dates have been excluded from NAIIS public release datasets.

2.3. Data Management and Cleaning

Field data collection for the household survey was conducted using CSPro. CSPro is an integrated data processing software developed by the U.S. Census Bureau. The survey used the Computer Assisted Personal Interview (CAPI) module and deployed an Android version used on tablets issued to the field teams. Data was transmitted to the survey headquarters through mobile phone networks.

Once the data was received at the headquarters it was subject to various consistency checks and field checks which were corrected through immediate communication with the field teams. The process for assuring the data files was as follows:

EA Level Checks:

- Completion checks: These checks assured that the expected households were interviewed in an EA and that the eligible persons were interviewed.
- Consistency and range checks: These were a series of checks that were done to assure that responses between various questions were reasonable for coded type questions and ranges for open questions reasonable.
- Frequency checks: This was a review of all various skip patterns to assure that the expected universes for filter questions were consistent. Because CAPI provides for in-line checks, such inconsistencies rarely occur. However, if there were such inconsistencies, these were addressed by calling the individual teams.

Concatenation of data files

Reporting was done at the state level. As the various states reported and were finalized, all data files for the states were joined and evaluated for completion. A similar redundant series of checks were undertaken. State data was then used to produce state level reports and eventually would be joined to produce a full national data set.

2.3.1 Missing Data and Other Exceptions

NAIIS data allows for various levels of coding to identify missing data. Please note the following:

- Not applicable: These data are blanks and are usually part of a skip. Blanks are not coded and should not be interpreted as missing. They are simply out of scope for the series of questions being administered. Don't Know: Responses, where the respondent did not know how to respond, are coded with the final character 8. This will depend on the length of the field. If the field is two characters long, the code is 98. The characters to the left of the 8 designating the field as "Don't Know" will always be filled with 9s.
- Refusal: Refusals are coded with a 9. All characters in a field are filled with 9s.
- Missing: Missing values that are expected to be in the path of the interview and not coded as refusals and unknown are left un-coded. During the analysis, these were coded by the analyst as missing values.

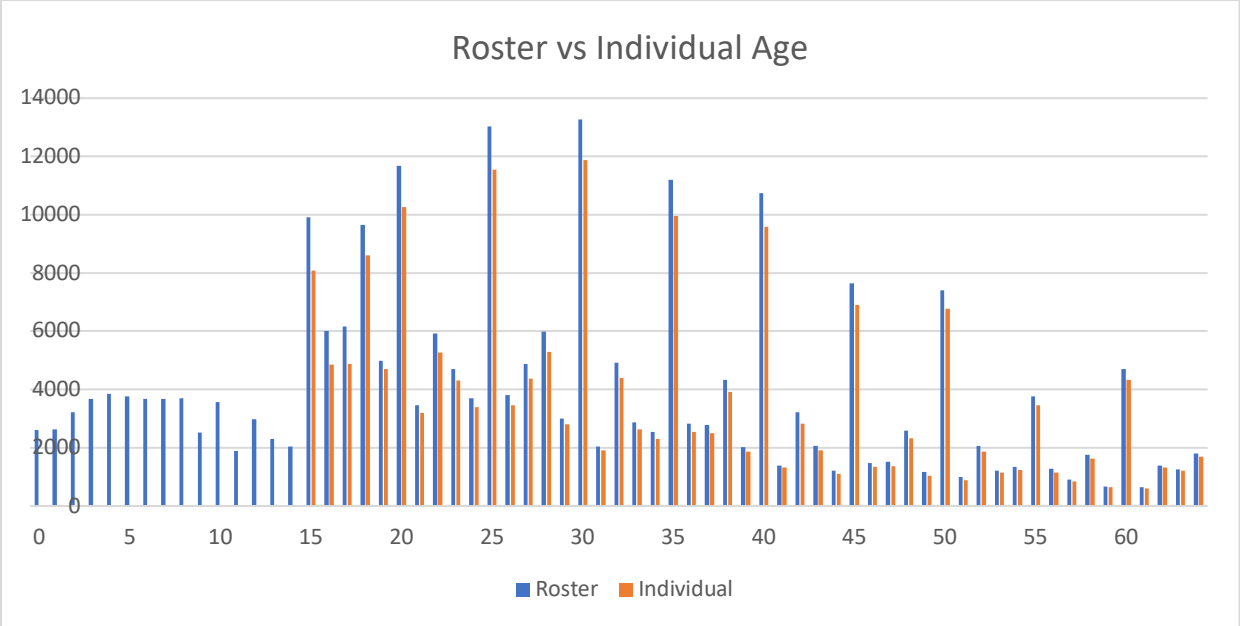
In many cases, NAIIS allowed for the selection of " Other " if none of the coded questions sufficiently responded. These were reviewed and, in most cases, coded back to existing categories if possible.

Multi-response questions: These are questions where the respondent code responds to multiple response possibilities. These are usually coded in the tablet as an alpha response depending on the number of responses, the letter A, B, C, and so forth are selected. These are later reformatted such that each response becomes binary and the response category incorporated into the variable name.

2.3.2 Age and Date Variables

The age of respondents was collected both in the household roster and from the individual respondent at the time of administering the household interview. The household roster was usually given by the head of the household who provided the household composition to the supervisor who was assigned to roster the household. The age variable preferred for the analysis was the age provided by the respondent. In the event of discrepancies, this age takes precedence.

Age heaping is notable in the survey. Age heaping is the preference for reporting age usually rounded to the 0- or 5-year increments. Age heaping was not smooth, and any user of the data will determine the value of applying age smoothing techniques to the age data. The graph provided below illustrates the extent of the age heaping with a preference on the 0 and 5 increments. It reviews both the reported age in the roster and the individual reported age.



Date Variables:

Date variables can be divided into process variables and respondent variables.

Process Dates: The process dates are usually a system generated dates that track the day, month, year, and time the interview starts and ends. This includes the time of individual visits should there be multiple visits to the household. The process dates are vital for monitoring field activities including the length of time that an individual interview takes. They may be of secondary interest to the analyst. The specific dates and times of individual visits are not provided in the data files.

Respondent dates: The survey asked the respondent to provide information on key dates of events such as birthdates and required day-month-year if possible. Dates for testing for HIV/AIDS for example were asked month and year.

2.3.3 Data Confidentiality Processes

The protection of participant privacy and confidentiality was maintained at each phase of NAIIS data collection and processing. To ensure the protection of participant privacy and confidentiality, NAIIS data processing encompasses various methods to reduce the risk of disclosure in the public-use data. The mitigation of potential risk disclosure occurs at the household-level and individual-level and addresses both direct and indirect identifiers in the public-use data.

In general, the following methods were used to minimize any privacy or confidentiality concerns in the NAIIS data:

- Redaction: removal of specific variables or removal of elements within the data variable (e.g. day from date).
- Top-coding: the process of re-coding continuous values above an upper bound to the value of the upper bound.
- Bottom-coding: the process of re-coding continuous values below a lower bound to the value of the lower bound.
- Small case count: defined as the lowest category containing at least 25 cases or 1 percent of households or individuals reporting the category; may be managed through top-coding, bottom-coding, or redaction.

The following risk mitigation methods are applied across all NAIIS public-use datasets:
- Removal of all direct identifiers (e.g. names, addresses, phone numbers)
- Household and participant IDs were randomly reassigned, as indicated in section 2.2.3.
- Days have been redacted from all date variables. Month and year were retained.
- All age variables have been top coded to 80.
- In certain circumstances, age variables were bottom-coded. See each NAIIS 2018 survey's ***Data Use Manual Supplement*** for specific details.
- For categorical variables, categories with counts of less than 25 were collapsed into "other", if "other" is an option. Response types "Don't know" and "Refused" were not collapsed into "other" because these response options are not identifying. Special circumstances may exist. See each NAIIS 2018 survey's ***Data Use Manual Supplement*** for additional details, including variables with this method applied.
- For dichotomous variables (i.e. variables with yes/no response options), the variables may have been redacted from the data if there were no risk remediation measures possible. See each NAIIS 2018 survey's ***Data Use Manual Supplement*** for additional details, including variables with this method applied.
- For continuous variables, top-coding or bottom-coding may have been used. See each NAIIS's Supplement for additional details, including variables with this method applied.

Recodes and redactions may introduce some data limitations. Some variables were redacted altogether and collapsed categories lost some detail. It may not be possible to reproduce all standard analytic variables from the variables available on the public-use datasets.

For more information about redactions to specific variables, see each NAIIS 2018 survey's ***Data Use Manual Supplement***.

2.4 Survey Weights

2.4.1 Weighting Approach

The main purpose of the survey weights calculated for NAIIS is to 1) account for unequal selection probabilities at different stages of sampling, 2) adjust for nonresponse at different stages of data collection, 3) reduce the variability of the weights using a weight trimming procedure, and 4) calibrate the weights to the 2018 population projections using data from the NPopC.

The process of calculating the weights started by calculating the design weighs that account for the selection probabilities of the different sampling units in different sampling stages. The design weights were adjusted to account for nonresponse in the primary sampling unit (PSU) and household levels. When weights were calculated for individuals, such as adults or adolescents, the weights were adjusted for individual-level nonresponse to the survey questionnaire. When weights were calculated for measurements, such as blood draws for HIV, the weights were adjusted for nonresponse to the test.

All weights were trimmed, where outliers were capped at a maximum value. Finally, all weights were calibrated based on the percentage of total distributions of the projected population.

2.4.2 Survey Weight Variables in NAIIS Datasets

The table below lists the names, eligible units, and completed units for the seven weight variables calculated for the NAIIS data. Additional details regarding the process of calculating each weight can be found in the NAIIS Sampling and Weighting Document.

Weight	Eligible units		Completed units
	Description	Variables (codes)	Variables (codes)
Household survey weight (hhwt0)	All selected households	hhstatus (1,2,3,4)	hhstatus (1)
Adult interview weight (intwt0)	All de-facto adults 15-64 in completed households	age (15:64) + sleephere (1) + hhstatus (1)	indstatus (1)
Adolescent interview weight (intwt0)	All de-facto adolescents 10-14 in completed households in the pediatric subsample	age (10:14) + pediatric (1) + sleephere (1) + hhstatus (1)	indstatus (1)
Blood draw weight (btwt)	All de-facto adults 15-64 who completed the adult questionnaire and all de-facto children 0-14 in completed households in the pediatric subsample	age (15:64) + sleephere (1) + indstatus (1) & age (0:14) + pediatric (1) + sleephere (1) + hhstatus (1)	hivstatusfinal (1,2)
Hepatitis weight (hepwgt)	All de-facto adults 15-64 who tested positive for HIV and all de-facto adults 15-64 who were selected for Hepatitis testing in the Hepatitis subsample (one adult 15-64 per household)	age (15:64) + sleephere (1) + hivstatusfinal (1) & age (15:64) + elghep (1) + sleephere (1) + hivstatusfinal (2) + hepbresult (1,2)	hepbresult (1,2)

2.4.3 Variance Estimation

The NAIIS dataset includes complex survey design variables, such as sampling strata, primary sampling unit or cluster, and survey weights. Users will need to specify these three variables for their analysis of interest at the national level. Several existing variance estimation methods can be used to estimate design-based standard errors for complex sample surveys. These methods require specifying appropriate survey weights, strata (state), and PSU (cluster).

The variance estimation method used in NAIIS was Taylor Series Linearization (TSL). TSL was utilized due to its simplicity of implementation and because it is not as computationally intensive as other replication methods, including Jackknife (JK) repeated replication that involves the calculation of estimates of subsamples obtained by resampling the full survey sample. Given the large NAIIS sample size, JK would be a computational challenge. Moreover, a linearization method and JK estimator may converge to the same value in large survey samples (Valliant, 2004).

For smooth functions of means, JRR and TSL perform equivalently. Thus, the choice of method is based on the feasibility of their implementation (Kovar, et al., 1988).

TSL uses the Taylor series expansion to approximate the sample estimator of interest by reducing the non-linear forms of variables to the linear function of sample totals or ratios (Woodruff, 1971). The Taylor linearization method treats any percentage or average as a ratio estimate, $r = y/x$, where y represents the total sample value for variable *y*, and *x* represents the total number of cases in the group or subgroup under consideration. The variance of *r* is computed using the formula given below, with the standard error being the square root of the variance:

$$SE^2(r) = \text{var}(r) = \frac{1-f}{x^2}\sum_{h=1}^{H}\left[\frac{m_h}{m_h-1}\left(\sum_{i=1}^{m_h} z_{hi}^2 - \frac{z_h^2}{m_h}\right)\right]$$

in which

$z_{hi} = y_{hi} - rx_{hi}$ and $z_h = y_h - rx_h$

where *h* represents the stratum which varies from 1 to *H*,

m_h is the total number of clusters selected in the h^{th} stratum,

y_{hi} is the sum of the weighted values of variable *y* in the i^{th} cluster in the h^{th} stratum,

x_{hi} is the sum of the weighted number of cases in the i^{th} cluster in the h^{th} stratum, and

f is the overall sampling fraction, which is so small that it is ignored.

2.4.4 Calculating Response Rates

Response rates are reported in ***PHIA Publications*** tables. In order to calculate household and individual response rates, the following procedure is used.

Household response rates. Sampled households were visited by field workers who determined household eligibility status, primarily based on the type of building and occupancy status. Household response status also depends on sufficient information being collected during the household interview. The variable hhstatus categorizes each household into one of four eligibility and response status categories:

hhstatus	Indicator of household eligibility and response status	1 - Eligible Responding Household 2 - Eligible Nonresponding Household 3 - Unknown Eligibility Status 4 - Ineligible (Vacant Household, Not a Dwelling, Dwelling Destroyed)

To calculate household response rates, PHIA uses the following procedure. Let I and P be the number of complete or partially complete interviews, R, NC and O represent Refusals, Non-contacts and other eligible, non-responding households respectively. UH represents Unknown if Household is occupied while OU represents other households with unknown eligibility. The estimated proportion of sampled households which are eligible is (I+P)+(R+NC+O) which is the eligibility rate among households with known eligibility. The estimated proportion of cases of unknown eligibility that are eligible denoted as e is used as an adjustment rate.

$$e = \frac{eligible}{eligible + ineligible}$$

Then, unweighted household response rates are calculated following AAPOR's Response Rate 4[2]:

$$\text{Response rate} = 100 \times \frac{I + P}{((I + P) + (R + NC + O) + e(UH + UO))}$$

Individual response rates. Individual response rates are based on individual eligibility and response status. The variables indstatus and bt_status categorize each individual for interview and blood draw eligibility and response status.

indstatus	Indicator of individual eligibility and response status	1 - Eligible Respondent 2 - Eligible Non-Respondent 8 - Not Sampled 9 - Non-de facto participants
bt_status	Did lab blood test have definite result?	1 - Lab blood test has a definite result 2 - Lab blood test does not have a definite result

Unweighted interview response rates are calculated by dividing the number of eligible respondents (indstatus = 1) divided by the total number of eligible respondents (indstatus = 1 or 2).

Unweighted blood draw response rates are calculated by dividing the number of individuals with definite lab blood test results (bt_status = 1) by the total number of interview respondents (indstatus = 1).

2.5 Linkages

For analyses that require information from multiple datasets, data users will need to link records from the household, individual interviews, and/ or biomarker datasets. The following sections provide instructions on how to perform these linkages. Instructions for special cases of mother-to-child (Section 2.5.3) and sexual partner (2.5.4) linkages are also provided.

2.5.1 Household, Individual Demographics, and HIV Status

For convenience, variables commonly used in the analysis have been made available in datasets outside of where they were initially collected. These variables include:

1. Collected in the household, also available in interview and biomarker
    a. Household ID (householdid)
    b. Geopolitical zone (zone_ng)
    c. Urban/ rural indicator (urban)
    d. Wealth quintiles (wealthquintile)
2. Collected in the interview, also available in biomarker
    a. Person ID (personid)
    b. Mother's ID (momid; child dataset only)
    c. Survey start year (surveystyear)
    d. Survey start month (surveystmonth)

e. Age (age)
f. Gender (gender)
g. Self-reported HIV status (hivselfreport; adult biomarker only)

3. Collected in biomarker, also available in interview
    a. Blood draw response status (bt_status)
    b. HIV status (hivstatusfinal)

2.5.2 General Procedures for Linking Datasets

Records from the household, individual interview, and biomarker datasets can be linked using a merge procedure available in statistical software packages and a common identifier variable, either Household ID (householdid) or Individual ID (personid). General guidelines are provided below, and specific examples can be found in Section 3. Example Code.

1. Household to an individual interview or household to biomarker
    a. Merging variable: Household ID (householdid)
    b. Type of merge: one-to-many
2. Biomarker to interview
    a. Merging variable: Individual ID (personid)
    b. Type of merge: one-to-one
    c. Note: biomarker records will only be available for the subset of participants who completed the blood draw
3. Interview to biomarker
    a. Merging variable: Individual ID (personid)
    b. Type of merge: one-to-one
    c. Note: interview records for participants without biomarker records should be dropped

2.5.3 Sexual Partner Linkages

Sexual and marital partnership data are collected as part of the Household Interview (Roster Information), and the Individual Interview (Marriage and Sexual Activity modules). For convenience in data users who wish to conduct analyses of partners, three types of partner linkage variables are provided in PHIA datasets. Note that survey weights are not provided for analyses with couples as the unit of analysis since sampling procedures do not identify couples during the household listing. For couples analyses, we suggest the use of the men's individual interview or blood weight (see Section 1.4.8, Survey weight variables in PHIA datasets).

HusID. The variable husid contains the personid of the husband reported by each female participant in the marriage module. If the husband is not a rostered household member, husid is blank. There is no analogous wifeid variable in the PHIA data sets. Husband-wife pair and polygamous relationships are identified only from husid.

PartID1-3. Three variables (*partid1, partid2, partid3*) contain the personid of up to 3 most recent sexual partners within the household as reported by the participant in the sexual activity module in the adult interview.

Lastpartner. The variable lastpartner contains the partid (1, 2, or 3) of the most recent sexual partner, if it is ascertainable from the data. Variables that contribute to lastpartner may differ by PHIA survey (refer to CONSORT diagram in each survey's Data Use Manual Supplement for details).

PartnerClusterID. Researchers may be interested in analyzing groups of 3+ individuals in the same household who are linked by sexual partnerships. Additionally, because HIV is a sexually transmitted disease, groupings of persons in a household who have had either direct sexual contact or indirect exposure via a mutual sexual partner or spouse are a potential unit of interest for study. These "partnership clusters" are relevant where an individual has multiple wives and/or sexual partners, thus pairs of individuals who are not themselves sexually partnered are connected indirectly through the common partner. The variable *partnerclusterid* captures these complex partnerships by assigning a unique ID to all individuals who are linked directly or indirectly by some marital or sexual relationship to any other individual in the same household. For partnership clusters formed by combinations of marital and sexual partnerships, linking individual records in a dataset is complex: the chains of partnership may require multiple links or joins. The inclusive definition of a partnership cluster and the addition of the unique number to the dataset enables analysts to easily examine both these complex linked groups and simple partnerships without having to do their own complex joining and sorting. This definition also avoids assigning persons to more than one cluster, which would require multiple partnership grouping variables. Note that only relationships within the same sampled households are included. Any relationships reported outside the household are not identified.

Construction. The variable *partnerclusterid* uniquely identifies each partnership cluster across the whole dataset. Partnership clusters are defined using the following rules:

1. All wives linked to their husbands using the *husid* variable are a part of the same cluster.
2. Any persons reported as sexual partners by a given person will be a part of that person's cluster.
3. A person can only be in one cluster: if a person is linked to two or more other people then all of them, and anyone linked to them as a sexual or marital partner, will be combined into a single larger cluster.
4. Self-reported information will be assumed to be correct, even if only one side of the partnership reports the partnership.

The table below lists expected types of partnership cluster structures. In all of these examples, other persons, such as children, grandparents, or other unrelated adults may be present in the household, but only the members related by spouse/sexual partner links are shown.

Case 1 is a household with two pairs of partners: a husband and wife who are recorded as spouses and who reported each other as their only recent sexual partners, and an unmarried couple who also recorded each other as their only recent sexual partners. In this case the married couple are assigned a cluster number of 1 and the second couple is assigned to cluster 2.

Case 2 shows another relatively simple situation. Each partner has reported another sexual partner who is outside the household. This example demonstrates the utility in distinguishing between null/none responses and 'individual outside household' responses to the sexual partner questions. The presence of the other partners does not change the cluster numbering. Note that the husband/wife does not need to be the primary or most recent partner and could be identified under *partid2* or *partid3*.

Case 3 shows a husband with multiple wives. All of the husband's wives are linked to the husband, and to each other, through the partnership cluster number.

Case 4 is similar to case 3, but there is an additional woman in the household who is linked to person 401 by sexual partnership reports. All three are linked in one partnership cluster.

Case 5 illustrates inconsistent reports of partnership within a household. Person 503 has reported a sexual partnership with person 501, but there was no reciprocal report by person 501. In this method, self-reports are treated as correct regardless of whether the relationship is reciprocated, so these people will be linked. In this example, person 501 is also married to and a reciprocal sexual partner with person 502. As a result, person 502 is linked together with person 503 in the same partnership cluster.

Case 6 demonstrates the (relatively rare) case of households with more complex connections. Here there are two married couples, but these are also connected by an additional non-marital sexual partnership. All four persons are part of the same partnership cluster. Note that in this case person 602 is linked to 603 and 604 by the partnership cluster number, but that neither of these numbers occur on her record at all, and 603 does not occur on either her or her husband's record. That is, persons 602 and 604 are indirectly linked through 603.

	personid	*gender*	*husid*	*partid1*	*partid2*	*partid3*	*partnerclusterid*
Case 1. Two simple couples in same household							
	101	M	.	102	.	.	1
	102	F	101	101	.	.	1
	103	M	.	104	.	.	2
	104	F	.	103	.	.	2
Case 2. Husband and wife with other partners outside household							
	201	M	.	202	(not in hh)	.	3
	202	F	201	201	(not in hh)	.	3
Case 3. Husband and two wives							
	301	M	.	302	303	.	4
	302	F	301	301	.	.	4
	303	F	301	301	.	.	4
Case 4. Husband and wife with another partner in the household							
	401	M	.	403	402	.	5
	402	F	401	401	.	.	5
	403	F	.	401	.	.	5
Case 5. Inconsistently reported partnership							
	501	M	.	502	.	.	6
	502	F	501	501	.	.	6
	503	F	.	501	.	.	6
Case 6. Complex/ chained partnership							
	601	M	.	602	604	.	7
	602	F	601	601	.	.	7
	603	M	.	604	.	.	7
	604	F	603	603	601	.	7

2.5.4 Mother-to-Child Linking

Rationale. PHIA surveys capture data regarding mother-child relationships, which may be of interest for analyses on children, including mother-to-child HIV transmission, among other topics. These data include information about children provided by their mother (e.g. breastfeeding, HIV testing and HIV status), as well as information about the mother (e.g., mother's age, HIV status, HIV testing, and care and treatment history of the mother).

Identification of mother-child pairs: In the reproductive module of the adult interview, women report if they had delivered a child in the last 3 years prior to the survey. If so, they provide pregnancy, childbirth and postpartum data on the last pregnancy and children born as a result. This data includes HIV testing, care and ARV use during pregnancy and children's breastfeeding history, and HIV status and testing. Thus, data to identify mother-child pairs is strongest for children born within the last 3 years of the survey, and reproductive module data is the primary source of information used to identify mother-child pairs for children under 3. Some data checks are performed to confirm the validity of mother-child relationships. Specifically, certain mother-child age combinations, such as when the mother's age is less than the child's age or less than 10 years older than the child, are considered impossible. In these cases, the identity of the mother is considered to be unknown.

For children aged 0-14 years, the identity of the mother is captured in the household roster by the line number of the mother and can be used to generate <personid> of the mother within the rostered household members. The personid is called momid in the child dataset and is blank for children whose mothers are unknown. The household head reports the information on the mother of the child,(eg mother alive, usual household resident, has been sick for at least 3 months in the last 12 months and mother's HIV status) as well as demographic information on the child(eg education, sex, age).

2.6 Analytic Variables

Analytic variables refer to those created by an analyst after the conclusion of the survey. These variables are derived by combining information from multiple original variables, and are included in PHIA datasets for convenience during the analysis process. A complete list of PHIA analytic variables can be found in NAIIS Data Use Supplement.

2.6.1 CONSORT Diagrams for Derived Variables

Derived variables refer to any variable that is dependent on results from multiple questionnaire responses and/or biomarker test results, i.e. those that were created by NAIIS analysts. For each of these variables, a CONSORT diagram is provided that details which source variables feed into the derived variable, and how exactly they are incorporated. NAIIS data users need to review these diagrams to ensure their interpretation of the derived variable is correct. A complete list of CONSORT diagrams for all derived variables can be found in the NAIIS Data Use Supplement.

2.6.2 Wealth Index

The use of a wealth index for measuring socioeconomic status, developed using survey data on household assets, materials, and durable goods, is a well-established method utilized in the DHS. Measurement of wealth using these indices is widely considered to be a more accurate construct than income when quantifying socioeconomic status in resource-limited settings and the indices are easily calculated using results from a household-level questionnaire. The DHS has provided commonly accepted guidelines for wealth index construction. Wealth index variables (continuous scores and quintiles) at both a national level and by urban/rural definitions have been constructed for further analysis using the DHS method, customized specifically for the NAIIS survey and the Nigerian context. Household dwelling characteristics and asset variables used to construct a wealth index vary by PHIA survey and country, and those used in the NAIIS wealth index are noted in Appendix 1. In the NAIIS

datasets, two national wealth index variables have been provided: a continuous score (wealthscorecont) and a categorical wealth quintile (wealthquintile).

Construction of wealth quintiles via DHS methods includes the following:

1. **Recode asset variables.** Household data include categorical variables about household characteristics, such as materials of walls, floors, and roof of the dwelling, source of water, type of cooking fuel and type of sanitation facilities used, as well as binary variables indicating ownership of durable goods such as beds, vehicles, radios, generators, etc. For a complete list of characteristics and assets included in the calculations for the NAIIS, please see the ***Data Use Manual Appendix*** and ***Survey Questionnaire***. Categorical variables are recoded as binary indicator variables (e.g. one variable was created for each floor type and a household receives 1 for the variable indicating their floor type and 0 for all others). Binary variables are coded as 1 (yes) or 0 (no, don't know, refused). Generally, missing data are treated as the absence of that asset, and households that do not have data on any assets are not assigned wealth index scores or wealth quintiles.
2. **Select the asset variables for inclusion.** Asset variables are analyzed using Principal Component Analysis (PCA), which is a statistical technique that transforms a number of (correlated) variables into uncorrelated components that capture variability (information) in decreasing order; thus, PCA is a useful dimension reduction technique. The DHS Program recommends using the first component of the model as a summary indicator of wealth. Since assets vary in relevance in urban and rural settings, PCAs are run separately for urban and rural households, and then for all households combined. Decisions to include or exclude asset variables from either setting are somewhat based on contextual knowledge, but largely are based on the findings in the survey data; for parsimony, all asset variables that have any variability are included in each analysis. Those assets or characteristics that do not have any variability in one or more of the urban, rural, or national calculations, are excluded from the final index construction (e.g. if no rural households had floors made of cement, then the cement floor variable would be excluded from the rural wealth index calculation).
3. **Run PCA and combine results.** Three PCAs were run in constructing the NAIIS wealth indices: a "common" model across all households, and models restricted to "urban" and "rural" households. As per convention, the first factor from each model is extracted to obtain three separate wealth indices. The common model wealth index is regressed separately on the urban or rural wealth index for households in those areas, and this regression model is then used to convert each household's (rural or urban) wealth index (wealthscorecontur) into a final "composite" wealth index (wealthscorecont).
4. **Generate wealth quintiles.** Households were then classified into quintiles (wealthquintileur, wealthquintilecont) using the composite wealth index. To account for the complex survey design, the weighted cumulative distribution of the wealth index is used to identify weighted quintile cut-points. Weights represent the household sampling weight, adjusted for the number of de jure members (that is, quintiles are determined from the cumulative distribution of the index for usual household members across all households). This ensures that at the national level when weighted, 20 percent of de jure members (usual household members) are included in each wealth quintile. The even quintile distribution of 20 percent does not hold at the household level or the individual interview level since the wealth index is a household-level measure, calculated for even distribution of all de jure members in the population. In a small number of cases in the NAIIS, if no household member is listed as de jure or de facto, a continuous score was calculated, but no categorical quintile was assigned since the quintiles are based on the total de jure population.

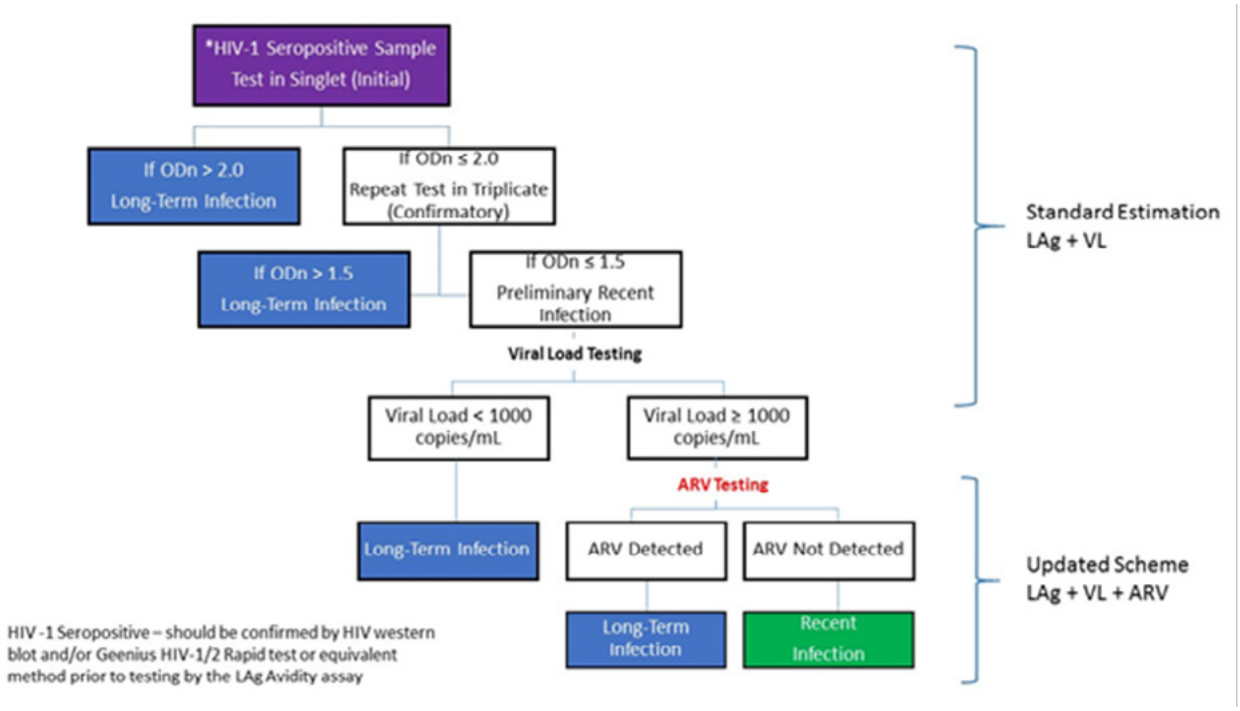
When using a survey constructed wealth index for analysis, there are important caveats and recommendations for interpretation to consider for the NAIIS wealth indices. Wealth indices and quintiles derived using this methodology are intended to represent relative measures of wealth as compared to other households in the same country, during the same survey period. It is important to note that the underlying PCA model simply finds the factors that best capture the variation in that specific survey data, and does not guarantee a straightforward interpretation. On average, households in higher wealth quintiles should be more wealthy, but there is considerable uncertainty as a result of the limitations of the available asset data and the model. Wealth is a complex concept that cannot be captured fully in this model, so wealth indices should be treated as approximate estimates rather than precise measures.

The value of the wealth index should not be thought of as directly proportional to household wealth, or as being measured along a standard baseline that can be compared between different countries or sub-populations, or across time. Relative measures should not be applied to subsets of the population unless one is willing to assume that the relative distribution of wealth is similar between the total and subsetted population.

For simplicity and to facilitate replication, variables were not selected differently in urban and rural models based on contextual or subjective knowledge. However, this may not be valid if assets are differentially related to wealth across contexts. Sensitivity analyses excluding variables considered to be context-specific (e.g. livestock) or which scored the most difference in the rural and urban models have typically shown that wealth indices are not sensitive to model specification. Alternative socioeconomic indicators are available and the merits of these alternatives are the subject of ongoing debate.

2.6.3 New HIV Infections and Annual HIV Incidence

In NAIIS, samples from participants who tested positive for HIV underwent additional testing to determine if the infection was recent or long-term. This test was carried out following the World Health Organization (WHO) consensus formula and began at the household with the rapid kit tests as recommended by the Nigerian national serial algorithm for HIV diagnosis. This involved tests with Determine, Unigold, and Statpak. All positive samples were tested with the Geenius assay.



Selected samples were:

- HIV-1 Positive as confirmed with Geenius and Western Blot
- From the 15-64 years old age group (adults)
- With Viral Load counts in excess of 1,000

The selected samples were tested using the Recent Infection Testing Algorithm (RITA). The samples were assayed with the Limiting Antigen Avidity Enzyme Immunoassay (LAg EIA). The tests took into consideration the probability of obtaining false positives that may be included with patients on ARVs and the elite controller group by excluding those with viral loads less than 1,000.

The samples were grouped into the plasma and DBS samples with the Plasma (2,723) tested with the Sedia Kit while the DBS samples (32) were tested using the Maxim kit. The assay consisted of 3 tests aimed at determining the normalized optical densities (ODn).

1. In the first test, all samples with ODn greater than 2 (ODn > 2) were classified as Long Term;
2. The second test was on samples with ODn less than 2. Those samples that had ODn between 0 and 1.5 were classified as Recent; and
3. The third test was on the set classed as recent to identify those with ODn less than 0.4 (ODn <0.4). These were classified as false positives.

For recency estimation, the samples that were initially classified as recent that also tested positive for ARVs were re-classified as Long Term. Therefore, an additional downward adjustment of incidence occurs when ARV data is made available which may affect specific gender/age groups.

HIV incidence estimates were calculated using a Mean Duration of Recent Infection (MDRI) of 130 days (118-142) and False Recent Ratio (FRR) of 0. The LAg ODn numbers from the assay were inputted to the CDC developed spreadsheet application, which is an Excel-based incidence calculator from which the preliminary incidence was determined.

3. Example Code

The complex design of the NAIIS 2018 survey sample dictate that in the analysis we account for multiple stages of sampling, stratification, and clustering. The NAIIS sample contains around 4,000 clusters, which makes the analysis using the jackknife (JK) replicate weights constructed with the method analogous to one used for other PHIA1 samples (with approximately 500 clusters) highly impractical. The following chapter contains examples of SAS, Stata, and R code illustrating how to load in and merge datasets, declare survey designs, and conduct analyses using the Taylor Linearization Method for NAIIS 2018 datasets.

The examples below depict how to obtain the point estimates and the confidence intervals for the following scenarios:

- Estimate adult HIV prevalence and 90-90-90 indicators (HIV care cascade steps: awareness, ARV use, and viral suppression).
- Estimate adult HIV status by TB diagnosis.
- Estimate HIV prevalence among adults and children, stratified by wealth quintile and 5-year age groups.
- Estimate arvstatus by sex and education
- Estimate HIV prevalence in a multi-country analysis using data from NAIIS2018, and Zamphia2016.
- Estimate HIV incidence among adults

3.1 SAS Code Examples

SAS code for examples 1-5 are shown below. To perform analysis of complex samples we can use the SAS survey procedures. These procedures analyze complex survey data by taking into account the sample design and the estimation procedures. The sample design and the estimation procedure (Taylor Linearization Method) are declared in the code of the SAS survey procedure. Two of the SAS survey procedures, the “Proc Surveymeans” and the “Proc Surveyfreq”, are both applicable to the examples that we illustrate below.

```
*************************************************************************;
***** Example 1. Estimate 90-90-90 indicators among adults          *****;
*************************************************************************;
*** Load in Adult BIO dataset;
LIBNAME filepath "C:\Desktop\";
DATA adultbio;
      SET filepath.NAIIS2018adultbio;
RUN;
*** Conduct analyses using the Taylor Linearization Method;
*   When using the Taylor Method strata and clusters have to be;
*   specified in the statements of survey procedures;
* HIV prevalence;
PROC SURVEYMEANS DATA = adultbio VARMETHOD = taylor mean clm nobs;
      WHERE hivstatusfinal ne 99;
      WEIGHT btwt0;
      STRATA varstrat;
      CLUSTER varunit;
      CLASS hivstatusfinal;
      VAR hivstatusfinal;
RUN;
* HIV awareness, conditional on HIV positive status;
PROC SURVEYMEANS DATA = adultbio VARMETHOD = taylor mean clm nobs;
      WHERE tri90 = 1 AND hivstatusfinal = 1 AND tri90aware ne 99;
      WEIGHT btwt0;
      STRATA varstrat;
      CLUSTER varunit;
      CLASS tri90aware;
      VAR tri90aware;
RUN;
* ARV use, conditional on awareness;
PROC SURVEYMEANS DATA = adultbio VARMETHOD = taylor mean clm nobs;
      WHERE tri90 = 1 AND tri90aware = 1 AND tri90art ne 99;
      WEIGHT btwt0;
      STRATA varstrat;
      CLUSTER varunit;
      CLASS tri90art;
      VAR tri90art;
RUN;
* Viral suppression, conditional on ARV use;
PROC SURVEYMEANS DATA = adultbio VARMETHOD = taylor mean clm nobs;
      WHERE tri90 = 1 AND tri90aware = 1 AND tri90art = 1 AND tri90vls ne 99;
      WEIGHT btwt0;
      STRATA varstrat;
      CLUSTER varunit;
      CLASS tri90vls;
      VAR tri90vls;
RUN;
```

```
*************************************************************************;
***** Example 2. Estimate adult HIV status by TB diagnosis          *****;
*************************************************************************;
*** Load in Adult IND dataset;
LIBNAME filepath "C:\Desktop\";
DATA adultind;
      SET filepath.NAIIS2018adultind;
RUN;
DATA adultbio;
      SET filepath.NAIIS2018adultbio;
RUN;

* Sort before merge;
PROC SORT DATA = adultbio;
      BY personid;
RUN;
PROC SORT DATA = adultind;
      BY personid;
RUN;

* Merge diagnosedtb_ng variable from Adult IND dataset;
DATA adultbio;
MERGE adultbio(in=a)
      adultind (keep=personid diagnosedtb_ng);
BY personid;
if a;
RUN;


*** Conduct analyses using the Taylor Linearization Method;
*   When using the Taylor Method strata and clusters must be;
*   specified in the statements of survey procedures;
* HIV status by TB diagnosis;
PROC SURVEYMEANS DATA = adultbio VARMETHOD = taylor df mean clm nobs;
      DOMAIN diagnosedtb_ng;
      WHERE (hivstatusfinal NE 99) AND (diagnosedtb_ng IN (1,2));
      WEIGHT btwt0;
      STRATA varstrat;
      CLUSTER varunit;
      CLASS hivstatusfinal;
      VAR hivstatusfinal;
RUN;
```

```
*************************************************************************;
***** Example 3. Estimate HIV prevalence among adults and children    *****;
*****    by 5-year age groups and by wealth quintiles                 *****;
*************************************************************************;

*** Load in Adult BIO datasets;
LIBNAME filepath "C:\Desktop\";
DATA bio;
      SET filepath.NAIIS2018adultbio
          filepath.NAIIS2018childbio;
RUN;
DATA ind;
      SET filepath.NAIIS2018adultind
          filepath.NAIIS2018childind;
RUN;
DATA hh;
      SET filepath.NAIIS2018hh;
RUN;

* Sort before merge;
PROC SORT DATA = bio;
      BY householdid;
RUN;
PROC SORT DATA = hh;
      BY householdid;
RUN;

* Merge wealth quintile variable from HH dataset;
DATA bio;
      MERGE bio (in=a)
            hh (keep=householdid wealthquintile);
      BY householdid;
        if a;
RUN;

* Sort before merge;
PROC SORT DATA = bio;
   BY personid;
RUN;
PROC SORT DATA = ind;
      BY personid;
RUN;

* Merge agegroup5population variable from Adult Interview dataset;
DATA bio;
      MERGE bio (in=a)
            ind (keep=personid agegroup5population);
      BY personid;
      IF a;
RUN;

*** Conduct analyses using the Taylor Linearization Method;
*   When using the Taylor Linearization method strata and clusters;
*   must be specified in the statements of survey procedures;
* HIV prevalence by wealth quintile;

PROC SURVEYMEANS DATA = bio VARMETHOD = taylor df mean clm nobs;
```

```
      DOMAIN wealthquintile;
      WHERE hivstatusfinal ne 99 AND wealthquintile ne 99;
      WEIGHT btwt0;
      STRATA varstrat;
      CLUSTER varunit;
      CLASS hivstatusfinal;
      VAR hivstatusfinal;
RUN;
* HIV prevalence by 5-year age groups;
PROC SURVEYMEANS DATA = bio VARMETHOD = taylor df mean clm nobs;
      DOMAIN agegroup5population;
      WHERE hivstatusfinal ne 99 AND agegroup5population ne 99;
      WEIGHT btwt0;
      STRATA varstrat;
      CLUSTER varunit;
      CLASS hivstatusfinal;
      VAR hivstatusfinal;
RUN;
```

```
*****************************************************************;
***** Example 4. Estimate arvstatus by sex and education    *****;
*****************************************************************;

*** Load in Adult BIO dataset;
LIBNAME filepath "C:\Desktop\";
DATA adultbio;
      SET filepath.NAIIS2018adultbio;
RUN;
DATA adultind;
      SET filepath.NAIIS2018adultind (keep=personid education);
RUN;

* Sort before merge;
PROC SORT DATA = adultbio;
      BY personid;
RUN;

PROC SORT DATA = adultind;
      BY personid;
RUN;

* Merge education variable from Adult Interview dataset;
DATA adultbio;
      MERGE adultbio (in=a)
            adultind;
      BY personid;
      IF a;
RUN;

*** Conduct analyses using the Taylor Linearization Method;
* arvstatus by gender and education for HIV positive adults;
PROC SURVEYMEANS DATA = adultbio VARMETHOD = taylor df mean clm nobs;
      DOMAIN gender*education;
      WHERE gender ne 99 AND education ne 99 AND (arvstatus IN (1,2));
      WEIGHT btwt0;
      STRATA varstrat;
      CLUSTER varunit;
      CLASS arvstatus;
      VAR arvstatus;
RUN;
```

```
*************************************************************************;
***** Example 5. Estimate Aware of HIV positive status in              *****;
***** multi-country analysis                                          *****;
*************************************************************************;
*** Load in Adult BIO datasets for 2 countries;
LIBNAME filepath "C:\Desktop\";

DATA naiisadultbio;
      SET filepath.NAIIS2018adultbio;
RUN;

* Renumber strata in the ZAMPHIA 2016 dataset;
* The highest numbered stratum in NAIIS survey is 66;
DATA zamadultbio;
      SET filepath.zamphia2016adultbio;
      varstrat = varstrat + 66;
RUN;

* Combine countries;
DATA combined;
      SET naiisadultbio
          zamadultbio;
      KEEP country aware varstrat varunit btwt0;
RUN;

*** Conduct analyses using the Taylor Linearization Method;
* Aware in 2 countries;
PROC SURVEYMEANS DATA = combined VARMETHOD = Taylor df mean clm nobs;
WHERE aware in (1,2);
WEIGHT btwt0;
STRATA varstrat;
CLUSTER varunit;
CLASS aware;
VAR aware;
RUN;
```

3.2 Example Code in STATA

Stata code for examples 1-5 are shown below. Note that Stata survey design and Taylor estimation procedures are declared prior to analysis, using the "svyset" command. Once this command has been issued, all you need to do for your analyses below is use the "svy:" prefix before each table command.

```
*************************************************************************
***** Example1. Estimate 90-90-90 indicators among adults       *****
*************************************************************************

*** Load in Adult BIO dataset
use "C:\Desktop\NAIIS2018adultbio.dta", clear

*** Declare survey design, for the Taylor Linearization Method
svyset varunit [pw=btwt0], strata(varstrat) vce(linearized) singleunit(scaled)

*** Conduct analyses
* HIV prevalence
svy: tab hivstatusfinal if hivstatusfinal!=99, se ci obs format(%8.3g)

* HIV awareness, conditional on HIV positive status
svy: tab aware if tri90==1 & hivstatusfinal==1 & aware!=99, se ci obs format(%8.3g)

* ARV use, conditional on awareness
svy: tab art if tri90==1 & aware==1 & art!=99, se ci obs format(%8.3g)

* Viral load suppression, conditional on ARV use
svy: tab vls if tri90==1 & aware==1 & art==1 & vls!=99, se ci obs format(%8.3g)
```

```
*************************************************************************************
***** Example 2. Estimate adult HIV status by TB diagnosis                    *****
*************************************************************************************

*** Load adult BIO dataset
use "C:\Desktop\NAIIS2018adultbio.dta", clear

*** Merge diagnosedtb_ng variable from Adult IND dataset
merge m:1 personid using "C:\Desktop\NAIIS2018adultind.dta", keep(match master)
keepusing(diagnosedtb_ng) nogen

*** Declare survey design for the Taylor Linearization Method
svyset varunit [pw=btwt0], strata(varstrat) vce(linearized) singleunit(scaled)

*** Conduct analyses using the survey weight, stratum, and psu.
* Note for bivariate analyses, outcome is specified first
* "col" option obtains column proportions

* HIV prevalence by TB diagnosis
svy: tab hivstatusfinal diagnosedtb_ng if hivstatusfinal!=99 &
(diagnosedtb_ng == 1 | diagnosedtb_ng == 2), se ci col obs format(%8.3g)
```

```
*******************************************************************************************
***** Example 3. Estimate HIV prevalence among adults & children                      *****
*****         by 5-year age groups and wealth quintiles                               *****
*******************************************************************************************

*** Prepare combined adult and child IND dataset for merging
use "C:\Desktop\NAIIS2018adultind.dta", clear
append using "C:\Desktop\NAIIS2018childind.dta"
save "C:\Desktop\NAIIS2018ind.dta", replace

*** Prepare combined adult and child BIO dataset for merging
use "C:\Desktop\NAIIS2018adultbio.dta", clear
append using "C:\Desktop\NAIIS2018childbio.dta"

*** Merge wealth quintile variable from HH dataset
merge m:1 householdid using "C:\Desktop\NAIIS2018hh.dta", keep(match master)
keepusing(wealthquintile) nogen

*** Merge agegroup5population variable from Adult and Child IND datasets
merge m:1 personid using "C:\Desktop\NAIIS2018ind.dta", keep(match master)
keepusing(agegroup5population) nogen

*** Declare survey design for the Taylor Linearization Method
svyset varunit [pw=btwt0], strata(varstrat) vce(linearized) singleunit(scaled)

*** Conduct analyses using the survey weight, stratum, and psu.
* Note for bivariate analyses, outcome is specified first
* "col" option obtains column proportions

* HIV prevalence by wealth quintile
svy: tab hivstatusfinal wealthquintile if hivstatusfinal!=99 & wealthquintile!=99,
se ci col obs format(%8.3g)


* HIV prevalence by 5-year age groups
svy: tab hivstatusfinal agegroup5population if hivstatusfinal!=99 &
agegroup5population!=99, se ci col obs format(%8.3g)
```

```
*******************************************************************************************
***** Example 4. Estimate arvstatus by sex and education                              *****
*******************************************************************************************

*** Load adult BIO dataset
use "C:\Desktop\NAIIS2018adultbio.dta", clear

*** Merge education variable from Adult IND dataset
merge m:1 personid using "C:\Desktop\NAIIS2018adultind.dta", keep(match master)
keepusing(education) nogen

*** Declare survey design for the Taylor Linearization Method
svyset varunit [pw=btwt0], strata(varstrat) vce(linearized) singleunit(scaled)

*** Conduct analyses using the survey weight, stratum, and psu.

* Arvstatus by sex and education
svy: prop arvstatus, over(gender education), if ((arvstatus ==1 | arvstatus == 2) &
    gender!= 99 & education!= 99)
```

```
*************************************************************************************
***** Example5. Estimate aware of HIV positive status in multi-country        *****
*****                          analysis                                       *****
*************************************************************************************

*** Combine the 2 countries
use "C:\Desktop\NAIIS2018adultbio.dta", clear
append using "C:\Desktop\zamphia2016adultbio.dta"

*** Recode variable varstrat into varstrat2 in Zamphia dataset;
*** Last stratum in Nigeria dataset is numbered 66;

generate varstrat2 = varstrat
replace varstrat2 = varstrat + 66 if country == "ZAMBIA"

*** Declare survey design for the Taylor Linearization Method
svyset varunit [pw=btwt0], strata(varstrat2) vce(linearized) singleunit(scaled)

*** Conduct analyses using survey weights, stratum, and psu
* HIV prevalence in 2 countries
svy: tab aware if (aware==1 | aware==2), se ci col obs format(%8.3g)
```

3.3 Example code in R

R codes for examples 1-5 are shown below (using .csv files). Note that R survey design and Taylor estimation procedures are declared prior to analysis by generating a survey object. Columns from the original dataset are separated into analytic variables (variables), base weights (weights), strata (strata), and PSUs (ids).

```
################################################################
### Example 1. Estimate 90-90-90 indicators among adults      ###
################################################################

### Install and load survey analysis and dplyr packages
memory.limit(size=56000)
install.packages("survey")
library(survey)
install.packages("dplyr")
library(dplyr)

### Load in Adult BIO dataset
adultbio <- read.csv('C:/Desktop/NAIIS2018adultbio.csv', na.strings = '.')
adultbio <- subset(adultbio, btwt0>0)

### Create survey object, Taylor series linearization
# Recode analytic outcomes to 0/1
vars <- c('hivstatusfinal', 'tri90', 'tri90aware', 'tri90art', 'tri90vls')
wtname <- 'btwt0'
strataname <- 'varstrat'
clustername <- 'varunit'
svydata <- adultbio
for(i in 1:length(vars)){
        varname <- vars[i]
        if(is.factor(svydata[,varname])) {
                svydata[,varname] <- as.numeric(levels(svydata[,varname]))[svydata[,varname]]
                # coerces factor variables to numeric before recoding
        }
        svydata[,varname][(svydata[,varname]==2)] <- 0
}

# Extract analytic variables, base weights and variance unit/strata variables
# and create survey object

svydata1 <- svydesign(
        variables=svydata[,vars],
        weights=svydata[,wtname],
        strata=svydata[,strataname],
        ids=~svydata[,clustername],
        nest=TRUE)

### Conduct analyses using base weights and variance unit/strata variables
# HIV prevalence
svydata1_hivstatusfinal <- svydata1[svydata1$variables$hivstatusfinal!=99]
res <- svyciprop(~I(hivstatusfinal==1),
        design=svydata1_hivstatusfinal,
        method="mean",level=0.95,df=25)
(c(res[[1]],attr(res,"ci"))) # display results
```

```
table(svydata1_hivstatusfinal$variables$hivstatusfinal)

# HIV awareness
svydata1_tri90aware <- svydata1[svydata1$variables$tri90==1 &
          svydata1$variables$hivstatusfinal==1 & svydata1$variables$tri90aware!=99]
res <- svyciprop(formula=~I(tri90aware==1),
          design=svydata1_tri90aware,
          method="mean",level=0.95,df=25)
(c(res[[1]],attr(res,"ci"))) # display results
table(svydata1_tri90aware$variables$tri90aware)

# ARV use
svydata1_art <- svydata1[svydata1$variables$tri90==1 &
svydata1$variables$tri90aware==1 & svydata1$variables$tri90art!=99]
res <- svyciprop(formula=~I(tri90art==1),
          design=svydata1_art,
          method="mean",level=0.95,df=25,na.rm = TRUE)
(c(res[[1]],attr(res,"ci"))) # display results
table(svydata1_art$variables$tri90art)

# Viral suppression
svydata1_vls <- svydata1[svydata1$variables$tri90==1 & svydata1$variables$tri90aware==1 &
svydata1$variables$tri90art==1 & svydata1$variables$tri90vls!=99]
res <- svyciprop(formula=~I(tri90vls==1),
          design=svydata1_vls,
          method="mean",level=0.95,df=25,na.rm = TRUE)
(c(res[[1]],attr(res,"ci"))) # display results
table(svydata1_vls$variables$tri90vls)
```

```
#################################################################
### Example 2. Estimate adult HIV status by TB diagnosis      ###
#################################################################

### Install and load survey analysis and dplyr packages
memory.limit(size=56000)
install.packages("survey")
library(survey)
install.packages("dplyr")
library(dplyr)

### Load in Adult BIO dataset
adultbio <- read.csv('C:/Desktop/NAIIS2018adultbio.csv', na.strings = '.')
adultbio <- subset(adultbio, btwt0>0)
```

```
### Merge TB diagnosis variable from Adult IND dataset
adultind <- read.csv('C:/Desktop/NAIIS2018adultind.csv', na.strings = ".")
adultind <- adultind[,c('personid','diagnosedtb_ng')]
adult <- merge(adultbio,adultind,by='personid')

### Create survey object
# Recode analytic variables to 0/1
vars <- c('hivstatusfinal', 'diagnosedtb_ng')
wtname <- 'btwt0'
strataname <- 'varstrat'
clustername <- 'varunit'
svydata <- adult
for(i in 1:length(vars)){
 varname <- vars[i]
 if(is.factor(svydata[,varname])) {
  svydata[,varname] <- as.numeric(levels(svydata[,varname]))[svydata[,varname]]
  # coerces factor variables to numeric before recoding
 }
 svydata[,varname][(svydata[,varname]==2)] <- 0
}
# Extract analytic variables,  weight, stratum, and psu
#   and create survey object
svydata1 <- svydesign(
 variables=svydata[,vars],
 weights=svydata[,wtname],
 strata=svydata[,strataname],
 ids=~svydata[,clustername],
 nest=TRUE)

### Conduct analyses
# HIV status by TB diagnosis
svydata1_diagnosedtb_ng <- svydata1[svydata1$variables$hivstatusfinal!=99 &
svydata1$variables$diagnosedtb_ng %in% c(1,2)]
svyby(formula=~I(hivstatusfinal==1),by=~diagnosedtb_ng,
    design=svydata1_diagnosedtb_ng,
    FUN=svyciprop,vartype="ci",method="beta",df=25)
table(svydata1_diagnosedtb_ng$variables$diagnosedtb_ng,
    svydata1_diagnosedtb_ng$variables$hivstatusfinal)
```

```
#########################################################################
### Example 3. Estimate HIV prevalence among adults & children          ###
###          by 5-year age groups and wealth quintiles                  ###
#########################################################################

### Install and load survey analysis and dplyr packages
memory.limit(size=56000)
install.packages("survey")
library(survey)
install.packages("dplyr")
library(dplyr)

### Load in and merge Adult and Child BIO datasets
adultbio <- read.csv('C:/Desktop/NAIIS2018adultbio.csv')
childbio <- read.csv('C:/Desktop/NAIIS2018childbio.csv')
bio <- bind_rows(adultbio,childbio)
bio <- subset(bio, btwt0>0)

### Merge wealth quintile variable from HH dataset
hh <- read.csv('C:/Desktop/NAIIS2018hh.csv')

# Select columns
hh2 <- select(hh, householdid, wealthquintile)
bio2 <- select(bio, householdid, personid, hivstatusfinal, btwt0,varstrat,varunit)

# Merge household and bio datasets
bio3 <- merge(bio2,hh2,by='householdid')

# Merge agegroup5population variable from Adult and Child IND datasets
adultind <- read.csv('C:/Desktop/NAIIS2018adultind.csv')
childind <- read.csv('C:/Desktop/NAIIS2018childind.csv')
ind <- bind_rows(adultind,childind)

# Select columns
ind2 <- select(ind, personid, agegroup5population, country)
data <- merge(bio3,ind2,by='personid',all.x=TRUE)

# Recode hivstatusfinal to 0/1
data$hivstatusfinal <- recode(data$hivstatusfinal,'1'=1,'2'=0)

# Extract analytic variables,  weight, stratum, and psu
vars <- c('hivstatusfinal','wealthquintile','agegroup5population')
wtname <- 'btwt0'
strataname <- 'varstrat'
```

```
clustername <- 'varunit'
svydata <- data

### Create survey object
svydata1 <- svydesign(
  variables=svydata[,vars],
  weights=svydata[,wtname],
  strata=svydata[,strataname],
  ids=~svydata[,clustername],
  nest=TRUE)

### Conduct analyses using survey weight, stratum, and psu

# HIV prevalence by wealth quintile
svydata1_hivbywealth <- svydata1[!(svydata1$variables$wealthquintile %in% c(99,NA))]
svyby(formula=~I(hivstatusfinal==1),by=~wealthquintile,
    design=svydata1_hivbywealth, FUN=svyciprop,vartype="ci",method="beta",df=25)
table(svydata1_hivbywealth$variables$wealthquintile,
    svydata1_hivbywealth$variables$hivstatusfinal)

# HIV prevalence by 5-year age groups
svyby(formula=~I(hivstatusfinal==1),by=~agegroup5population,
    design=svydata1_hivbywealth,
FUN=svyciprop,vartype="ci",method="beta",df=25)table(svydata1_hivbyagegroup$variables$agegroup,
    svydata1_hivbyagegroup$variables$hivstatusfinal)
```

```
#####################################################################
### Example 4. Estimate arvstatus by sex and education            ###
#####################################################################

### Install and load survey analysis and dplyr packages
memory.limit(size=56000)
install.packages("survey")
library(survey)
install.packages("dplyr")
library(dplyr)

### Load in Adult BIO
adultbio <- read.csv('C:/Desktop/NAIIS2018adultbio.csv', na.strings = ".")

### Merge education variable from Adult IND dataset
adultind <- read.csv('C:/Desktop/NAIIS2018adultind.csv', na.strings = ".")
adultind <- adultind[,c('personid','education')]
# Merge personid and education, all.x=TRUE option retains unmatched rows from master
```

```
adult <- merge(adultbio,adultind,by='personid',all.x=TRUE)
adult <- subset(adult, btwt0>0)

### Recode arvstatus to 0/1
adult$arvstatus <- recode(adult$arvstatus,'1'=1,'2'=0)

### Extract analytic variables,  weight, stratum, and psu
vars <- c('gender', 'education', 'arvstatus')
wtname <- 'btwt0'
strataname <- 'varstrat'
clustername <- 'varunit'
svydata <- adult

###   Create survey object
svydata1 <- svydesign(
  variables=svydata[,vars],
  weights=svydata[,wtname],
  strata=svydata[,strataname],
  ids=~svydata[,clustername],
  nest=TRUE)

### Conduct analyses using survey weights
# Viral load suppression by gender and education for HIV positive adults
svydata1_adult <- svydata1[svydata1$variables$gender!=99 & svydata1$variables$education!=99 &
              svydata1$variables$arvstatus!=99]
svyby(formula=~I(arvstatus==1),by=~gender+education,
design=svydata1_adult, FUN=svyciprop,vartype="ci",method="beta",df=25)
ftable(table(svydata1_adult$variables$gender, svydata1_adult$variables$education,
           svydata1_adult$variables$arvstatus))
```

```
#########################################################################
### Example 5. Estimate aware in multi-country analysis                 ###
#########################################################################

### Install and load survey analysis and dplyr packages
memory.limit(size=56000)
install.packages("survey")
library(survey)
install.packages("dplyr")
library(dplyr)

### Combine countries
naiisbio <- read.csv('C:/Desktop/NAIIS2018adultbio.csv', na.strings = ".")
zambio <- read.csv('C:/Desktop/zamphia2016adultbio.csv', na.strings = ".")
```

```
# Renumber strata in Zambia survey dataset
zambio$varstrat <- zambio$varstrat + 66

bio <- bind_rows(naiisbio,zambio)
bio <- subset(bio, btwt0>0)

### Recode aware to 0/1
bio$aware <- recode(bio$aware,'1'= 1,'2'= 0)

### Extract analytic variables,  weight, stratum, and psu
vars <- c('aware','art')
wtname <- 'btwt0'
strataname <- 'varstrat'
clustername <- 'varunit'
svydata <- bio

###   Create survey object
svydata1 <- svydesign(
  variables=svydata[,vars],
  weights=svydata[,wtname],
  strata=svydata[,strataname],
  ids=~svydata[,clustername],
  nest=TRUE)

# Estimate proportion of Aware of HIV positive status
svydata1_aware <- svydata1[svydata1$variables$aware!=99]
res <- svyciprop(~I(aware==1),
             design=svydata1_aware,
             method="mean",level=0.95,df=25,na.rm = TRUE)
(c(res[[1]],attr(res,"ci"))) # display results
table(svydata1_aware$variables$aware)
```

3.4 SAS Program for HIV Incidence Estimation

The program below calculates the weighted counts to be used as inputs to the CDC Incidence Calculator and the design effect to be used.

It also directly calculates the annual incidence, which can be checked against the incidence calculator result, and confidence intervals for this incidence.

Finally, it calculates the estimated number of new infections in the previous year.

```
/*

The required input dataset variables are:
  hivstatusfinal - hiv status (positive or negative) from blood testing
  recentlagvlarv - Recent infection indicator derived from LAg Avidity
                   testing and our recent infection algorithm
  btwt0 - blood test weight
  gender and age variables for table breakdowns
Input Parameters:
  diffvar - variable for which differentials will be calculated separately
            (i.e., gender = male, female)
  filter_var - variable preset to 1 for observations to be included in
               calculation (i.e., use age15_49 to select all ages 15-49)
  omega - the MDRI (Mean Duration of Recent Infection)
  sig_omega2 - the standard deviation (sigma**2) of omega, derived as
               =(12/(NORMSINV(0.975)))**2 using CDC figures

We use the Taylor Linearization Method in the proc surveyfreq to get the design
effect.


*/


Libname filepath "C:\Desktop";

Data NAIIS2018adultbio;
  Set filepath.NAIIS2018adultbio;


  If Incidence_classification in ("ART/EC (LT)","LT") then recentlagvl = 2;
  Else If Incidence_classification = "RECENT" then recentlagvl = 1;

  recentlagvlarv = recentlagvl;
  if recentlagvl=1 and arvstatus=1 then recentlagvlarv=2;

  if age >= 15 and age <= 24 then age15_24 = 1; else age15_24 = 2;
  if age >= 25 and age <= 34 then age25_34 = 1; else age25_34 = 2;
  if age >= 35 and age <= 49 then age35_49 = 1; else age35_49 = 2;
  if age >= 15 and age <= 64 then age15_64 = 1; else age15_64 = 2;
  if age >= 15 and age <= 49 then age15_49 = 1; else age15_49 = 2;

  all = 1; ** define by variable for getting totals when using the macro;
  where Hivstatusfinal in (1,2);
  ttl=1;
Run;


%macro incidence_diff(diffvar,filter_var,incidence_var = recentlagvlarv,
omega = 130, sig_omega2 = 37.48575911, pfr = 0, sig_pfr2 = 0);
data temp;
  set NAIIS2018adultbio;
  where &filter_var. = 1;
run;
Proc sort data=temp;
  by &diffvar.;
run;
```

```
proc summary data = temp;
  by &diffvar.;
  var bdwght;
output out = totals sum = sumbdwght;
run;
Proc sort data=totals;
  by &diffvar.;
run;

** Data step creates the normalized weights for observations and the variables
for getting aggregate counts ***;
data temp2;
  merge temp totals;
  by &diffvar.;
  norm_bdwght = bdwght*(_freq_ / sumbdwght); **normalized weight such that sum
of the weights = number of observation;
  if Hivstatusfinal = 1 then do;
    if &incidence_var. = 1 then when_infected = 1;
    else if &incidence_var. = 2 then when_infected = 2;
    else when_infected = 3; ** Positive but not classified by Lag test;
  end;
  else when_infected = 4; ** HIV negative;

PplusN = 1;
  n = (when_infected = 4);
  p = (when_infected in (1,2,3));
  q = (when_infected in (1,2));
  r = (when_infected = 1);
PplusN_wtd = PplusN * norm_bdwght;
  n_wtd = n * norm_bdwght;
  p_wtd = p * norm_bdwght;
  q_wtd = q * norm_bdwght;
  r_wtd = r * norm_bdwght;
PplusN_pop = PplusN * bdwght;
  p_pop = p * bdwght;
  q_pop = q * bdwght;
  r_pop = r * bdwght;

Recent_over_all = 100 * (&incidence_var. = 1);
run;

* Counts for the table;
proc summary data=temp2;
  by &diffvar.;
  var PplusN n_wtd p_wtd q_wtd r_wtd;
  output out=counts_aux sum=;
run;

data counts_aux;
set counts_aux;
rename n_wtd = n
       p_wtd = p
       q_wtd = q
       r_wtd = r;
run;

*** Summary gets aggregate weighted and unweighted counts ***;
```

```
proc summary data=temp2;
  by &diffvar.;
  var PplusN p q r PplusN_wtd p_wtd q_wtd r_wtd PplusN_pop p_pop q_pop r_pop;
  output out=counts sum=;
run;

data calc_incidence;
  set counts;
  neg_wtd = (PplusN_wtd - p_wtd) * (q_wtd / p_wtd);
  incid_instant = (r_wtd - &pfr. * q_wtd) / ((1 - &pfr. / &omega.) * (&omega.
                  /365) * neg_wtd);
  ** reduces to (r_wtd/neg_wtd)*(365/130) if false recent rate = 0;
  incid_annual = 100 *( 1 - exp(-incid_instant));
  ** expressed as annual percent;
  * Estimate number of new cases based on incidence and population at risk;
  neg_pop = PplusN_pop - p_pop;
  NewCases = neg_pop * incid_annual / 100;
run;

* Compute design effect for recent infection using blood test weights;
* Using the Taylor Linearization Method;

proc surveyfreq data=temp2 nosummary;
  by &diffvar.;
  ods output oneway=sfreq_results;
  weight btwt0;
  Stratum VarStrat;
  Cluster VarUnit;
  table Recent_over_all / DEFF;
run;

proc sort data=sfreq_results;
  by &diffvar.;
run;
data sfreq_results2;
  set sfreq_results;
  by &diffvar.;
  if first.&diffvar.; *select first line of results for each subpopulation;
run;

/* Add the design effect for the proportion recent/all to the counts needed for
Incidence Calculator */
data combine;
  merge calc_incidence sfreq_results2;
  by &diffvar.;
run;

/*
This formula comes from Kassanjee et al, Epidemiology, Vol 23, No 5, September
2012. A, B, and C are the 3 main terms in equation (e7) in the online appendix
*/

data variance;
  set combine;
  A = (1/q_wtd)*( (1/neg_wtd) + (r_wtd * ( p_wtd - r_wtd)) / (r_wtd - &pfr. *
      p_wtd * (&omega. - &pfr. * 365))**2);
  B = &sig_omega2. * 1 / (&omega. - &pfr. * 365 )**2;
```

```
  C = &sig_pfr2. * (( (p_wtd - r_wtd) * &omega. - r_wtd * (365 - &omega.) ) /
      ( (r_wtd - &pfr. * p_wtd) * (&omega. - &pfr. * 365) ))**2;

  * Confidence interval not including the design effect;
  * Note this formula is for instantaneous incidence;
  c_sq = (A + B + C);
  UCL_i = incid_instant * (1 + sqrt(c_sq) * probit(0.975));
  LCL_i = incid_instant * (1 - sqrt(c_sq) * probit(0.975));

  * Clopper-Pearson CI for cases where R = 0;
  if r_wtd = 0 then do;
    R_CPUCL = neg_wtd * (1 - (0.05/2)**(1/neg_wtd));
    UCL_i = (r_CPUCL - &pfr. * q_wtd) / ((1 - &pfr. / &omega.) * (&omega.
            /365) * neg_wtd);
    LCL_i = 0.0;
  end;

  /* Convert upper and lower limits of the interval into annual incidence
     limits */
  UCL_a = 100 *( 1 - exp(-UCL_i));
  LCL_a = 100 *( 1 - exp(-LCL_i));

  * Confidence interval adjusting for the design effect;
  * Design effect < 1 is treated as equal to 1;
  if (DesignEffect > 1) then
  adj_c_sq = DesignEffect * (A + B + C);
  else if (DesignEffect <= 1) then
  adj_c_sq = 1.0 * (A + B + C);
  adj_UCL_i = incid_instant * (1 + sqrt(adj_c_sq) * probit(0.975));
  adj_LCL_i = incid_instant * (1 - sqrt(adj_c_sq) * probit(0.975));

  * Use Clopper-Pearson CI for cases where R = 0;
  if r_wtd = 0 then do;
    R_CPUCL_adj = neg_wtd * (1 - (0.05/2)**( max(1.0, DesignEffect) /
                  neg_wtd));
    adj_UCL_i = (r_CPUCL_adj - &pfr. * q_wtd) / ((1 - &pfr. / &omega.) *
                (&omega. / 365) * neg_wtd);
    *(R_CPUCL_adj / neg_wtd) * (365 / &omega.);
    adj_LCL_i = 0.0;
  end;

  * Convert upper and lower limits of the interval into annual incidence;
  adj_UCL_a = 100 *( 1 - exp(-adj_UCL_i));
  adj_LCL_a = 100 *( 1 - exp(-adj_LCL_i));
run;

proc print data=variance;
  var &diffvar. PplusN p q r PplusN_wtd p_wtd q_wtd r_wtd DesignEffect
      neg_wtdn incid_instant incid_annual adj_LCL_a adj_UCL_a NewCases;
run;

* Create tables in appropriate format with the desired output variables;
data tab_output_incidence;
  length Row $ 8 gend $ 8;
  set variance;
  Row = translate( substr("&filter_var.", index("&filter_var.","e") + 1), '-
        ', '_');
```

```
  * User should customize to variable categories chosen;
  if gender = 1 then gend = 'Male';
  else if gender = 2 then gend = 'Female';
  else gend = 'Total';
run;

* Output weighted counts (normalized to the sample size);
data tab_output_x (keep = Row gend n p q r);
  length Row $ 8 gend $ 8;
  set Counts_aux;
  Row = translate( substr("&filter_var.", index("&filter_var.","e") + 1), '-
        ', '_');
  if gender = 1 then gend = 'Male';
  else if gender = 2 then gend = 'Female';
  else gend = 'Total';
run;

** Create columns for number of new cases;
proc surveyfreq data = naiis_adult_biomarker;
  where &filter_var. = 1 and bt_status = 1;
  Stratum VarStrat;
  Cluster VarUnit;
  weight btwt0;
  tables &diffvar. * Hivstatusfinal / row CL CLWT;
  ods output crosstabs = AdultPrev_&filter_var._out;
run;

data AdultPrev_&filter_var.;
  length gend $10;
  set AdultPrev_&filter_var._out;
  where Hivstatusfinal = 1;
  if gender = 1 then gend = 'Male';
  else if gender = 2 then gend = 'Female';
  else gend = 'Total';
  if gend ne 'Total' then do;
    Percent = RowPercent;
    StdErr = RowStdErr;
    LowerCL = RowLowerCL;
    UpperCL = RowUpperCL;
  end;
  drop Table F_: _SkipLine RowPercent RowStdErr RowLowerCL RowUpperCL
                 hivstatusfinal;
  rename wgtfreq = PLHIV
         lowerclwgtfreq = PLHIVLCL
         upperclwgtfreq = PLHIVUCL
         StdDev = PLHIVStdErr
         Percent = Prevalence
         LowerCL = PrevalenceLCL
         UpperCL = PrevalenceUCL
         StdErr = PrevalenceStdErr;
run;

proc sql;
  create table PLHIV_new as
    select i.*, p.PLHIV, p.PLHIVStdErr, p.PLHIVLCL, p.PLHIVUCL,
           p.Prevalence, p.PrevalenceStdErr, p.PrevalenceLCL, p.PrevalenceUCL
             from tab_output_incidence i
```

```
                left join adultprev_&filter_var. p
                  on i.gend = p.gend;
quit;
run;

data tab_output_newcases;
  set PLHIV_new;
  neg_pop = PLHIV*100.0/Prevalence - PLHIV;
  if incid_annual = 0 then do;
    NewCasesLCL = 0.0;
    NewCasesUCL = neg_pop * (adj_UCL_a/100) * (1 + (Prevalence -
                  PrevalenceLCL) / Prevalence);
    NewCasesRelErr = .;
  end;
  else do;
    NewCasesRelErr = SQRT( ((Prevalence - PrevalenceLCL) / Prevalence)**2 +
                     ((Incid_annual - adj_LCL_a) / Incid_annual)**2);
    NewCasesLCL = MAX(0.0, NewCases * (1 - NewCasesRelErr));
    NewCasesUCL = NewCases * (1 + NewCasesRelErr);
    NewCasesStdErr = (NewCases - NewCasesLCL) / 2.0635;
  end;
run;

* Filter and rearrange output variables in final tables;
data tab_output_incidence;
  retain Row gend Designeffect incid_annual LCL_a UCL_a adj_LCL_a adj_UCL_a;
  set tab_output_incidence;
  if LCL_a < 0 then LCL_a = 0;
  if adj_LCL_a < 0 then adj_LCL_a = 0;
  keep Row gend Designeffect incid_annual LCL_a UCL_a adj_LCL_a adj_UCL_a;
run;

data tab_output_newcases;
  retain Row gend PLHIV PLHIVStdErr PLHIVLCL PLHIVUCL
         NewCases NewCasesStdErr NewCasesLCL NewCasesUCL;
  set tab_output_newcases;
  keep Row gend PLHIV PLHIVStdErr PLHIVLCL PLHIVUCL
       NewCases NewCasesStdErr NewCasesLCL NewCasesUCL;
run;
** End new cases calculation;

%mend incidence_diff;


ODS Excel File="C:\Desktop\Incidence.xlsx";
Ods      Excel      Options(sheet_name="Incidence"      sheet_interval='none'
embedded_titles='yes' );

%incidence_diff(diffvar = ttl, filter_var = age15_24) ; * This will output
total;
Proc Print data=tab_output_incidence; Run;
%incidence_diff(diffvar = gender, filter_var = age15_24) ; * By gender;
Proc Print data=tab_output_incidence; Run;
%incidence_diff(diffvar = ttl, filter_var = age25_34) ;
Proc Print data=tab_output_incidence; Run;
%incidence_diff(diffvar = gender, filter_var = age25_34) ;
Proc Print data=tab_output_incidence; Run;
```

```
%incidence_diff(diffvar = ttl, filter_var = age35_49) ;
Proc Print data=tab_output_incidence; Run;
%incidence_diff(diffvar = gender, filter_var = age35_49) ;
Proc Print data=tab_output_incidence; Run;
%incidence_diff(diffvar = ttl, filter_var = age15_49) ;
Proc Print data=tab_output_incidence; Run;
%incidence_diff(diffvar = gender, filter_var = age15_49) ;
Proc Print data=tab_output_incidence; Run;
%incidence_diff(diffvar = ttl, filter_var = age15_64) ;
Proc Print data=tab_output_incidence; Run;
%incidence_diff(diffvar = gender, filter_var = age15_64) ;
Proc Print data=tab_output_incidence; Run;

ODS Excel Close;
```

4. References

[1] Population-based HIV Impact Assessment (PHIA) Data Use Manual. New York, NY. July 2019.

[2] Burgert, Clara R., Josh Colston, Thea Roy, and Blake Zachary. 2013. Geographic displacement procedure and georeferenced data release policy for the Demographic and Health Surveys. DHS Spatial Analysis Reports No. 7. Calverton, Maryland, USA: ICF International.

[3] The American Association for Public Opinion Research. 2015. Standard Definitions: Final Dispositions of Case Codes and Outcome Rates for Surveys. 8th edition. AAPOR.