Published in final edited form as:

J Acquir Immune Defic Syndr. 2021 August 01; 87(Suppl 1): S28–S35. doi:10.1097/QAI.0000000000002703.

Data Architecture to Support Real-Time Data Analytics for the Population-Based HIV Impact Assessments

Melissa Metz, MS[a], Rebecca Smith, BBA[a], Rick Mitchell, MS[a,b], Yen T. Duong, PhD[a], Kristin Brown, MPH[c], Steve Kinchen, BS[c], Kiwon Lee, MPH[a], Francis M. Ogollah, MSc[a], Tafadzwa Dzinamarira, PhD[a], Vusumuzi Maliwa, BS[a], Carole Moore, BS[c], Hetal Patel, MSc[c], Hannah Chung, MPH[a], Helecks Mtengo[a], Suzue Saito, PhD, MIA, MA[a,d]

[a]ICAP at Columbia University, New York, NY

[b]Clinical Trials Unit, Westat, Rockville, MD

[c]Division of Global HIV and TB, Center for Global Health, U.S. Centers for Disease Control and Prevention, Atlanta, GA

[d]Department of Epidemiology, Mailman School of Public Health at Columbia University, New York, NY

Abstract

Background and Setting: Electronic data capture facilitates timely use of data. Population-based HIV impact assessments (PHIAs) were led by host governments, with funding from the President's Emergency Plan for AIDS Relief, technical assistance from the Centers for Disease Control, and implementation support from ICAP at Columbia University. We described data architectures, code-based processes, and resulting data volume and quality for 14 national PHIA surveys with concurrent timelines and varied country-level data governance (2015–2020).

Methods: PHIA project data were collected through tablets, point-of-care and laboratory testing instruments, and inventory management systems, using open-source software, vendor solutions, and custom-built software. Data were securely uploaded to the PHIA data warehouse daily or weekly and then used to populate survey-monitoring dashboards and return timely laboratory-based test results on an ongoing basis. Automated data processing allowed timely reporting of survey results.

Results: Fourteen data architectures were successfully established, and data from more than 450,000 participants in 30,000 files across 13 countries with completed PHIAs, and blood draws producing approximately 6000 aliquots each week per country, were securely collected, transmitted, and processed by 17 full-time equivalent staff. More than 25,600 viral load results

Correspondence to: Melissa Metz, MS, Data Management and Informatics Lead, ICAP at Columbia University, 722 W 168th St, 13th Floor, New York, NY 10032 (mm33@columbia.edu).

A portion of this topic was previously presented in poster form as Achieving real-time quality and efficiency monitoring through electronic data sharing architecture for national population-based biomarker surveys in developing countries: the population-based HIV impact assessments (PHIA) project 2015–2017, 21st International AIDS Conference (AIDS 2016); Durban, South Africa.

The authors have no funding or conflicts of interest to disclose.

Supplemental digital content is available for this article. Direct URL citations appear in the printed text and are provided in the HTML and PDF versions of this article on the journal's Web site (www.jaids.com).

were returned to clinics of participants' choice. Data cleaning was not needed for 98.5% of household and 99.2% of individual questionnaires.

Conclusion: The PHIA data architecture permitted secure, simultaneous collection and transmission of high-quality interview and biomarker data across multiple countries, quick turnaround time of laboratory-based biomarker results, and rapid dissemination of survey outcomes to guide President's Emergency Plan for AIDS Relief epidemic control.

Keywords

HIV; population-based surveys; electronic data collection; data management; data architecture; PHIA

INTRODUCTION

The population-based HIV impact assessment (PHIA) project, led by host governments with technical assistance from the Centers for Disease Control and implementation support from ICAP at Columbia University, is assessing the status of HIV epidemic and the impact of response efforts in select countries. Each cross-sectional, household-based survey used a 2-stage cluster design. Interviewers collected demographic, behavioral, and clinical information from eligible household participants. Blood draws were conducted for home-based HIV testing and counseling and point-of-care $CD4^+$ T-cell enumeration, with results immediately returned to the participants. Blood samples that were positive for HIV underwent laboratory-based confirmatory testing, HIV incidence testing, RNA polymerase chain reaction (PCR) [viral load (VL)], DNA PCR (early infant diagnosis), and serum antiretroviral drug detection. Data were weighted for survey design, and the χ^2 automatic interaction detection–based methods were used to adjust for nonresponse. Additional details are available elsewhere.[1] As of this writing, 13 surveys have been completed, and one is being weighted (2015–2020; Table 1).

Many studies[2–7] have discussed the advantages of electronic data collection, which was a guiding principle in designing the PHIA data architecture. The main challenge for the PHIA project was to establish a data architecture that could ensure the timely return of laboratory-based test results to survey participants as per the new Joint United Nations Programme on HIV/AIDS/World Health Organization guidelines on population-based surveys.[8] This required a high level of precision and accuracy in the collection and management of both interview and laboratory-based data. Additional challenges included developing a secure and cost-conscious core architecture that could be adapted to multiple countries simultaneously based on variability across countries (Table 2) in data governance rules and practices, existing information infrastructure, and human resource capacity. PHIA prioritized open-source technology heavily used in Africa, particularly in service delivery and research settings,[9] automated processes for data transactions, and a modular build to increase adaptability to different country contexts. This study details the core PHIA data architecture (including software and processes) and lessons learned as adaptations were made in each country.

METHODS

The core data architecture (Fig. 1) enabled secure, near real-time data sharing between the field and laboratory teams and the PHIA data warehouse (DWH). The core data architecture comprises 3 access zones. Access zone 1 represents data and blood collection at the participating households, satellite laboratories (SLs) for specimen processing, and central laboratories (CLs) for VL and PCR testing. Access zone 2 represents where zone 1 data are warehoused, merged, and cleaned. In addition to data staff in zone 1, 17 full-time equivalent staff (FTE) were allocated to zone 2 to develop, test, and implement data capture tools and provide ongoing data management and analytic support. Access zone 3 is a copy of data in zone 2 but is local to the survey country.

Questionnaire Programming and Household Data Collection (Access Zones 1 and 2)

PHIA questionnaires were programmed in Open Data Kit (ODK; Seattle, WA) open-source software on encrypted Google Nexus 9 Android tablets (HTC Corporation; New Taipei City, Taiwan). The PHIA programmers coded various features to facilitate high-quality data capture. This included a routing structure capturing eligibility of rostered individuals and their data nested within a household on a master tablet used by team leads; the roster was transferred from the master tablet to interviewer tablets assigned to each eligible individual, allowing for multiple interviewers to work in parallel while preserving the roster. In addition, multiple data links were programmed, including linking the individual interview data to the roster data by household identifier so that all eligible individuals were associated with a selected household; linking by roster line number for quality control purposes to compare responses provided in the household questionnaire with the individual questionnaire; and linking by roster line number between sexual partners, parent/guardian and child, and mother and child to facilitate paired analyses.

During the interview process, several tablet features were leveraged to ensure integrity of the selected household sample and household-level and individual-level data. First, Global Positioning System data were recorded to confirm selected households; the consent forms signed on the tablet's screen were stored as images to keep an audit trail of appropriate consenting practices; and randomly generated participant identifiers (PTIDs) were assigned using scannable preprinted quick response barcodes, which were then scanned using the camera on the tablet and affixed on specimen tubes to minimize data entry mistakes and ensure linkage between questionnaire and specimen data. High accuracy in the linkage between questionnaire and specimen data was essential for timely return of laboratory-based biomarker results, such as VL, to participants. Completed questionnaires were uploaded daily to the ODK server in the DWH (Fig. 1, arrow 1).

Specimen Collection, Testing, and Inventory Management (Mostly Access Zone 1)

Blood specimens were collected from consented participants for HIV and other rapid tests (detailed elsewhere[10]), with results recorded in the tablets. CD4 counts for all specimens showing positive results for HIV and a random sample showing negative results for HIV were measured using the Alere Pima Analyzer (Abbott, Abbott Park, IL) and uploaded to the secure Alere server using a vendor-provided modem (Fig. 1, arrow 2).

Blood specimens from the field were sent daily to PHIA SLs set up within on average 2 hours of driving distance from households, along with specimen tracking forms containing PTID labels and specimen information. PTIDs affixed to the specimen tubes were scanned into the LDMS laboratory information management software (Frontier Science, Boston, MA; Fig. 1, arrow 3), and the participant's household HIV test result and blood draw time also were entered.

LDMS assigned a specimen ID to each expected aliquot (subspecimen), producing scannable labels. At all data entry points, PTID and specimen barcodes were scanned for accuracy. If any aliquots were not produced, because of a short draw or unusable specimens, reasons were recorded into LDMS to help track specimen quality. Freezer locations and freezer entry times for each aliquot also were saved in LDMS.

The Test Results Entry Module was developed specifically for the PHIA surveys to track which specimens required repeat rapid tests for quality assurance (QA) or confirmatory Geenius testing (Geenius HIV 1/2 Supplemental Assay; Bio-Rad, Hercules, CA). The module selected specimens for QA and Geenius testing based on survey-specific algorithms and generated worksheets to record the results. The results were entered into the test results entry module using validation and drop-downs to simplify data entry.

Daily, SLs securely uploaded data to the DWH in access zone 2. The data included the LDMS data through its export function (Fig. 1, arrow 4), transferring inventory and testing records to Frontier, and Geenius results extracted as comma-separated values (CSV) files (Fig. 1, arrow 5).

Weekly, specimens were shipped from the SL to the CL, with a flash drive containing encrypted LDMS inventory and freezer box location data (Fig. 1, arrow 6). When specimens were shipped, shipping dates and destination were recorded, allowing for full accountability of specimen quantities, quality, and location. Importing data into the CL's instance of LDMS provided a local inventory of received specimens, without handling and recording each specimen. CL LDMS exports (Fig. 1, arrow 7) informed the survey team that the specimens had arrived. The CL uploaded HIV VL results (Fig. 1, arrow 8) weekly to the secure File Transfer Protocol (sFTP) server. Additional laboratory-based biomarker test results (Table 2) also were uploaded from reference laboratories (Fig. 1, arrow 9).

Repository and Analytic Data Sets (Access Zones 2 and 3)

The PHIA DWH was a secure repository for all survey data, including data pulled from the Alere server (Fig. 1, arrow 10), the ODK server (Fig. 1, arrow 11), and LDMS/Frontier (Fig. 1, arrow 12), along with tools for merging, analyzing, and sharing data (described further).

Automated processes extracted data collected in the field (excluding laboratory-based data) from the ODK PostgreSQL[11] (Postgres) database daily to approximately 50 raw SAS (SAS Institute, Cary, NC) data files, stripping personally identifiable information (PII), to make the data shareable with the survey management team. The 50 raw files were merged based on links described earlier, producing easy-to-use analytic data sets. These data sets were uploaded and were updated daily, reflecting corrections made to the master change log containing all QA data corrections submitted from the field on an ongoing basis, including but not limited to missing household and individual survey outcomes, broken links between households and individuals, and missing biomarker data (see the “Dashboard and Monitoring” section further). These data sets were used by the survey management team to conduct ad hoc analyses to investigate survey implementation issues, such as low response rates, performance of certain questions, and any errors in consenting processes. The latter was performed at least twice during each survey by an external monitoring group to ensure compliance with all screening and consenting procedures as specified by the protocol. PII was managed by extracting data separately from the Postgres database and storing it in a restricted area with limited access by data management staff, who signed confidentiality agreements, to resolve data discrepancies.

The DWH transferred raw and cleaned extracts of the ODK data to an in-country Ministry of Health server daily (Fig. 1, arrow 13), providing the most recent version of the PHIA data. Final survey data, including all laboratory-based data, were also transferred to the Ministry of Health (Fig. 1, arrow 14). The sFTP server in the DWH shared these raw and corrected data sets, VL and Geenius results, and process files, such as orders for laboratory testing.

Variations to Core Architecture (Access Zones 1 and 2)

As mentioned earlier, this core architecture was customized to address country-level variations in data governance rules and practices, existing information infrastructure, human resource capacity, and national testing guidelines (Table 2). For example, existing information infrastructure in Kenya, Namibia, and Rwanda allowed for primary servers to be hosted in-country. Elsewhere, infrastructure was not easily available and ready for use within the survey budget.

VL testing and transfer of results also differed (Table 2). In most of the countries, VL results were shared in CSV format exported directly from the testing instrument. In Cameroon and Haiti, the instrument produced a CSV format with different column headers, 2 lines for each result, and certain columns presented in an initial stanza of common information. Furthermore, in Zimbabwe and Lesotho, the testing instrument was integrated with each CL’s laboratory information system, and the results were shared as Extensible Markup Language (XML) and CSV extracts, respectively. Third, although VL results were always returned, some countries had an additional laboratory-based testing that needed to be returned. For example, in Côte d’Ivoire where HIV-2 is prevalent, the differentiation of HIV-1 and HIV-2 infection was only made at the laboratory; the results were required to be returned to participants. In Kenya and Namibia,[12] GeneXpert (Cepheid, Sunnyvale, CA) was used for preliminary early infant diagnosis with CL confirmation because regional laboratories where SLs were hosted had existing capacity to run the tests. Each of these tests types was an additional data source for data collection and processing.

Gathering and Cleaning Laboratory Data

Twice a day, DWH files were automatically curated, a process that standardized file names, variable names, and formats and organized files in specific folders. Files of the same type were merged in a multicountry laboratory database using Postgres and programs written in

C# (Microsoft; Redmond, WA) and then output as CSV analytic data sets for use in SAS, Stata (StataCorp, College Station, TX), or Excel (Microsoft). For example, for different VL result formats described earlier, the laboratory database merge code mapped key fields into a single format that was standardized across all surveys.

Merging Household, LDMS, and Laboratory Data

A staging database excluding all PII was set up in Google Drive within a G Suite for Education[12] instance to facilitate cleaning, merging, and preparing analytic data sets containing field-based and laboratory-based data, including data from household, SL, and CL, which were linked through unique PTIDs. The staging database was used to merge household and LDMS data twice a week in the DWH to order laboratory-based tests in the CL. The results were linked back to participant data, creating portable document format reports, downloaded in-country, and sent through courier to health facilities for return of results to participants.[13] Return and turnaround time were tracked in spreadsheets shared between headquarters and in-country coordinators.

The staging database was also used to identify missing data and mismatched data, including laboratory data without household data or vice versa, missing VL or CD4 data for an HIV-positive participant, or discrepant HIV status data between the household, SL, and CL data sets. This merged data set was also used to select specimens for recency, drug resistance, and antiretroviral testing after fieldwork completion.

Dashboards and Monitoring

The PHIA dashboard was a secure website providing 20 daily reports to facilitate monitoring of the quality, efficiency, and progress of each survey. The dashboard was accessible to all partners to facilitate evidence-based decisions to improve survey performance and data quality in real time. Area closeout and response rate reports tracked teams in the field, and SL reports tracked the collection of blood specimens, specimen processing and storage in SL freezers, and SL testing. These dashboard reports were improved in granularity over time, particularly after the first 3 surveys in Zimbabwe, Malawi, and Zambia, to incorporate lessons learned. Incompleteness and inconsistencies in the data highlighted in these reports were tracked and reconciled centrally. As described earlier, all data corrections resulting from this process were tracked in a master data change log that was then applied to the analytic data sets on an ongoing basis by the project team. As batches of data corrections were made, dashboard reports were updated to show the corrected information.

HIV Discrepancy Resolution

When SL or CL testing contradicted the original household HIV result, the data team highlighted relevant data so that the laboratory team could follow a retesting and investigation process to resolve these discrepancies. The staging database flagged mismatched results, which were presented for review in a system implemented using the Tracker Capture module of the open source District Health Information System (DHIS2).[14] DHIS2 simplified data entry, minimized transcription errors, and allowed a structured review of the relevant facts for each case, allowing for effective and timely resolution.

Workflow involved a list of cases filtered by reviewer: the in-country laboratory advisor, the central data team, or an overall senior laboratory advisor (Fig. 2). Flagged cases were examined, and additional tests or documentation review was performed and routed to the senior laboratory advisor for final review and determination for case resolution (see Fig. 1, Supplemental Digital Content, http://links.lww.com/QAI/B662). An extract from this system was integrated back into the staging database, replacing discrepant results with a final HIV status and triggering any additional needed laboratory-based testing.

Security and Access Control

PHIA data management prioritized information security by preventing sensitive data from being exposed and by ensuring that all study data were available for analysis. Tablet data were uploaded as soon as feasible after completion of questionnaires to prevent loss. To prevent exposure, tablets were encrypted; data could not be viewed without entering a randomly assigned PIN. If a tablet was stolen despite extensive security efforts (such as extensive training for staff, provision of lock boxes to survey teams, and documentation of the chain of custody of tablets at all times), the random PIN would prevent unauthorized access.[15–17] In 2018, Find My Device[18] was added to wipe data still stored in stolen tablets remotely. PII was shared only on an as-needed basis.

The staff limited data exposure by entering only the PTID, no PII, and minimal additional information into LDMS and Geenius laptops. Laptops were password protected and encrypted. Data were protected from loss by being backed up and uploaded each day.

The DWH spanned several platforms; all were professionally managed, redundant, housed in secure data centers, and backed up for recovery purposes. Servers for sFTP, ODK, and data cleaning complied with the National Institute of Standards and Technology 800-53[19] security standard. Servers for Google Drive "regularly undergo independent verification of security, privacy, and compliance controls, achieving certifications against global standards,"[20] such as the International Organization for Standardization's ISO 27001.[21] Data transfers used the Secure Sockets Layer, which confirmed the identity of the server and encrypted data over common carriers so that the data cannot be intercepted.

Hundreds of PHIA staff and partners had access to files on the sFTP server, controlled by assigning unique usernames and passwords to each staff member, grouping staff by roles, and assigning read or write access to each folder, further limiting staff by country. All requests for access were centrally channeled through the PHIA account administrator for approval. Similar access restrictions were in place for all servers.

RESULTS

The PHIA project began data collection in late 2015 (Table 1), completed 450,000 interviews and more than 370,000 blood draws across 13 countries (Table 3), and was finalizing data sets for Haiti in December 2020. The discrepancy resolution process facilitated review of more than 2000 cases.

More than 30,000 files were shared through the sFTP server. The laboratory database curated the majority of these, which included test results and inventory information. Each country had 1–10 LDMS licenses for different SL teams, for a total of up to 18 laptops in use at any one time, paralleled by 18 Geenius laptops and readers. In total, 420 PIMA devices submitted more than 280,000 CD4 results, including quality control testing. Using almost 2000 tablets distributed across up to 4 simultaneous countries, staff surveyed more than 170,000 households. Electronic data collection enabled high-quality data entry: 98.5% of household forms and 99.2% of individual forms were submitted with complete individual and household outcomes and intact links between households and individuals. Preliminary analytic data sets and documentation with final HIV status resolved were available on average at 6 weeks after fieldwork completion.

The staging database allowed the return of more than 25,000 VL results, 91.5% within the defined turnaround time. The return of results was monitored on paper for the first 3 surveys (Zimbabwe, Malawi, and Zambia) but moved to a shared spreadsheet (in Google Drive) so that headquarters and in-country staff could collaborate and monitor on a real-time basis. Electronic tracking improved returns: the number of VL results not returned decreased from dozens in the first 3 surveys to fewer than 5 VL results in the next 8 surveys. In addition, the median turnaround time improved from 52 days in the first survey to 17 and 24 days in the last 2 surveys completed.

Real-time monitoring through dashboards allowed the PHIA team to address issues quickly and collaboratively throughout the survey period. For example, response rate dashboards highlighted not only low response rates, which were addressed by increasing community mobilization efforts, but also higher than expected response rates, which required the procurement team to increase the supplies needed to complete the survey. Finally, in addition to real-time reports and participant results, these systems facilitated timely generation of standardized summary sheets, public-use data sets, and a dynamic, online data visualization tool[22] for 10 countries and 12 in-depth reports[23] using standardized variables and measures.

DISCUSSION

The PHIA data architecture was modularly designed to adapt to diverse data sources and data governance environments rapidly and at scale. Depending on the data governance rules and regulations, human resource capacity, and type of laboratory instruments used in a country, the architecture was flexibly adapted to fit the varied needs in each country (Table 2). The design was further shaped by the project's unique real-time need; to our knowledge, these surveys are the first large-scale household surveys to return actionable laboratory results (as per recent WHO guidelines[8]) within a predefined 8- to 12-week timeframe, based on real-time merging of survey and laboratory data. Other surveys were at a smaller scale or did not return laboratory-based results and generally merged data only after the survey was completed.[2–4,6,7,24–27]

Electronic data collection made PHIA's real-time data merging and cleaning possible, enabling timely turnaround of both individual results and survey-level results. These processes were resource intensive, involving 17 FTEs across the survey partners dedicated to central data management and cleaning. This model was needed because surveys were often being implemented in several countries concurrently, and FTEs were distributed across, on average, 3 concurrent surveys. As a comparison point, the New York City Health and Nutrition Examination Survey (NYC HANES),[25] with similar result return processes, required an average of one FTE for every 1000 survey participants [conversation between Sharon E. Perlman (NYC DOHMH, HANES project), Melissa Metz, and Suzue Saito (ICAP, Mailman School for Public Health, Columbia University); January 2019]. PHIA's automation allowed similar work to be performed with a ratio closer to 4000 participants per FTE.

Other limitations included finite design time and incomplete knowledge of how the survey would differ in each country, leading to less-than-optimal solutions. Insufficient attention to return of results in the first countries led to dozens of results not being returned. Improved systems led to better turnaround time and completeness, but dedicated staff time was necessary to achieve these goals.

Real-time data management with less resource-intensive support could improve future surveys. The PHIA survey architecture took advantage of prebuilt solutions. However, because of extremely tight start-up timelines, the designs could have been more streamlined. In preparation for the second round of PHIA surveys, some efficiency gains were made. For example, listing and questionnaire data collection now use CSPro (US Census Bureau, Suitland, MD); this pivot from ODK enabled 2-way data synchronization, which cut down on manual processes for data corrections after initial form submission. Teams can monitor enumeration area and household completeness as soon as interviewers synchronize the data and make data corrections while still in the same enumeration area. In addition, DHIS2 instead of Excel spreadsheets has been used to streamline result return. With its built-in security controls and user roles, DHIS2 also allowed expanded real-time access to monitor discrepant field-laboratory serological results. These changes reduced manual transcriptions and substantially reduced data cleaning efforts. Large-scale national and multinational surveys requiring daily data review and return of actionable laboratory results should pursue electronic, preferably open-source software, vs. paper-based data collection, and data architecture design with maximum automated processing so that large numbers of records can be processed with speed, and design adaptations can be made efficiently when required. The PHIA data architecture permitted secure, simultaneous collection and transmission of high-quality interview and biomarker data across multiple countries with varied data governance landscape, quick turnaround time of laboratory-based biomarker results, and rapid dissemination of survey outcomes.

ACKNOWLEDGMENTS

The authors thank the many people who have worked toward the success of the PHIA surveys, including foremost the participants, interviewers, laboratory technicians, support staff, analysts, and partners.

Supported by the President's Emergency Plan for AIDS Relief (PEPFAR) through the US Centers for Disease Control and Prevention under the terms of Cooperative Agreements 1U2GGH000994 and 5NU2GGH001226. The findings and conclusions in this study are those of the authors and do not necessarily represent the official position of the funding agencies.

REFERENCES

1. Justman JE, Mugurungi O, El-Sadr WM. HIV population surveys—bringing precision to the global response. N Engl J Med. 2018;378:1859–1861. [PubMed: 29768142]
2. Ojwang' JK, Lee VC, Waruru A, et al. Using information and communications technology in a national population-based survey: the Kenya AIDS indicator survey 2012. J Acquir Immune Defic Syndr. 2014;66(Suppl 1):S123–S129. [PubMed: 24732816]
3. Leisher C. A comparison of tablet-based and paper-based survey data collection in conservation projects. Social Sci. 2014;3:264–271.
4. Walther B, Hossin S, Townend J, et al. Comparison of electronic data capture (EDC) with the standard data capture method for clinical trial data. PLoS One. 2011;6:e25348. [PubMed: 21966505]
5. Martin J PAPI to CAPI: The OPCS Experience; Second International Blaise Users Conference. Groβbritannien, ed. London: Stationery Off; 1993.
6. Thriemer K, Ley B, Ame SM, et al. Replacing paper data collection forms with electronic data entry in the field: findings from a study of community-acquired bloodstream infections in Pemba, Zanzibar. BMC Res Notes. 2012;5:113. [PubMed: 22353420]
7. Paudel D, Ahmed M, Pradhan A, et al. Successful use of tablet personal computers and wireless technologies for the 2011 Nepal Demographic and Health Survey. Glob Health Sci Pract. 2013;1:277–284. [PubMed: 25276539]
8. UNAIDS/WHO Working Group on Global HIV/AIDS and STI Surveillance. Monitoring HIV impact using population-based surveys. 2015. Available at: http://www.unaids.org/sites/default/files/media_asset/JC2763_PopulationBasedSurveys_en.pdf. Accessed March 19, 2019.
9. DHIS2 in action|DHIS2. Available at: https://www.dhis2.org/in-action. Accessed May 31, 2020.
10. Patel H, Pottinger Y, Birhanu S, et al. A comprehensive approach to assuring quality of laboratory testing in HIV surveys: lessons learned from population-based HIV impact project. J Acquir Immune Defic Syndr.
11. PostgreSQL. The world's most advanced open source database. Available at: https://www.postgresql.org/. Accessed March 2, 2019.
12. Domaoal RA, Sleeman K, Sawadogo S, et al. Successful use of near point-of-care early infant diagnosis in NAMPHIA to improve accuracy and turnaround times in a national household survey. J Acquir Immune Defic Syndr.
13. G Suite for Education. Google for education. Available at: https://edu.google.com/products/gsuite-for-education/. Accessed March 2, 2019.
14. Saito S, Duong YT, Metz M, et al. Returning HIV-1 viral load results to participant-selected health facilities in national Population-based HIV Impact Assessment (PHIA) household surveys in three sub-Saharan African Countries, 2015 to 2016. J Int AIDS Soc. 2017;20:e25004. [PubMed: 29171193]
15. About DHIS2|DHIS2. Available at: https://www.dhis2.org/about. Accessed March 3, 2019.
16. Security. Android open source project. Available at: https://source.android.com/security. Accessed March 3, 2019.
17. Carpene C How secure is your smartphone's lock screen? The conversation. Available at: http://theconversation.com/how-secure-is-your-smartphones-lock-screen-56987. Accessed March 3, 2019.
18. What if the FBI tried to crack an android phone? We attacked one to find out—motherboard. Available at: https://motherboard.vice.com/en_us/article/nz7ejb/what-if-the-fbi-tried-to-crack-an-android-phone-we-attacked-one-to-find-out. Accessed March 3, 2019.
19. Find, lock, or erase a lost Android device—Nexus Help. Available at: https://support.google.com/nexus/answer/6160491?hl=en. Accessed March 3, 2019.
20. NVD—800-53. Available at: https://nvd.nist.gov/800-53. Accessed March 3, 2019.
21. Cloud compliance—regulations & certifications. Google cloud. Available at: https://cloud.google.com/security/compliance/. Accessed March 3, 2019.

22. 14:00-17:00. ISO/IEC 27001:2013. ISO. Available at: http://www.iso.org/cms/render/live/en/sites/isoorg/contents/data/standard/05/45/54534.html. Accessed June 22, 2019.

23. PHIA project document manager—about. Available at: https://phia-data.icap.columbia.edu/. Accessed June 30, 2019.

24. Resources—PHIA. Available at: https://phia.icap.columbia.edu/resources/. Accessed June 30, 2019.

25. Justman J, Reed JB, Bicego G, et al. Swaziland HIV Incidence Measurement Survey (SHIMS): a prospective national cohort study. Lancet HIV. 2017;4:e83–e92. [PubMed: 27863998]

26. Thorpe LE, Greene C, Freeman A, et al. Rationale, design and respondent characteristics of the 2013–2014 New York city health and nutrition examination survey (NYC HANES 2013–2014). Prev Med Rep. 2015;2:580–585. [PubMed: 26844121]

27. King JD, Buolamwini J, Cromwell EA, et al. A novel electronic data collection system for large-scale surveys of neglected tropical diseases. PLoS One. 2013;8:e74570. [PubMed: 24066147]

28. Rainey E, Morton J, Litavecz S, et al. Tobacco surveillance through electronic data collection on the Android operating system—evidence from the Global Adult Tobacco Survey. Tob Induced Dis. 2018;16. Available at: https://doaj.org. Accessed December 30, 2018.

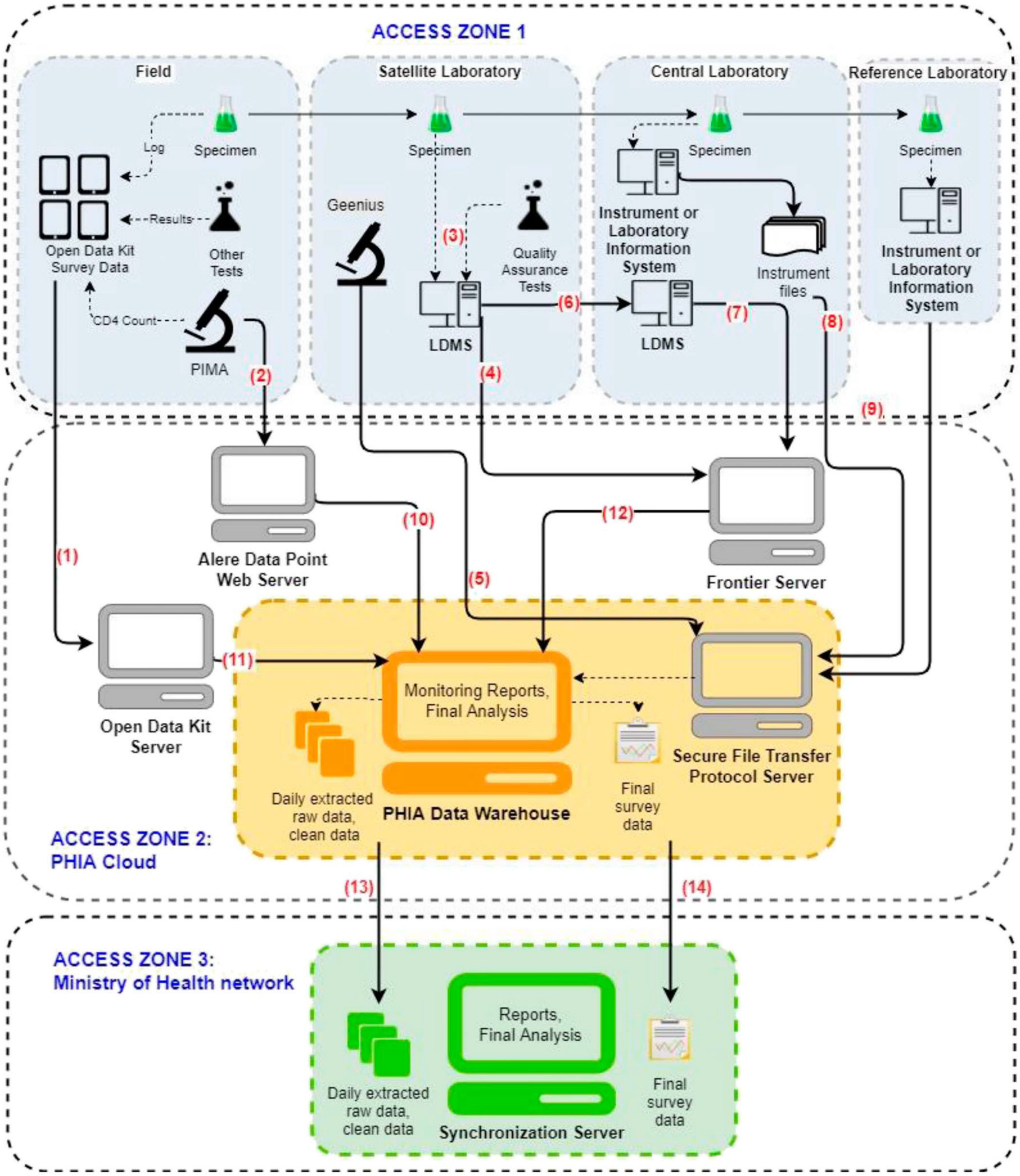

FIGURE 1.

Population-Based HIV Impact Assessment (PHIA) data architecture (2015–2020).

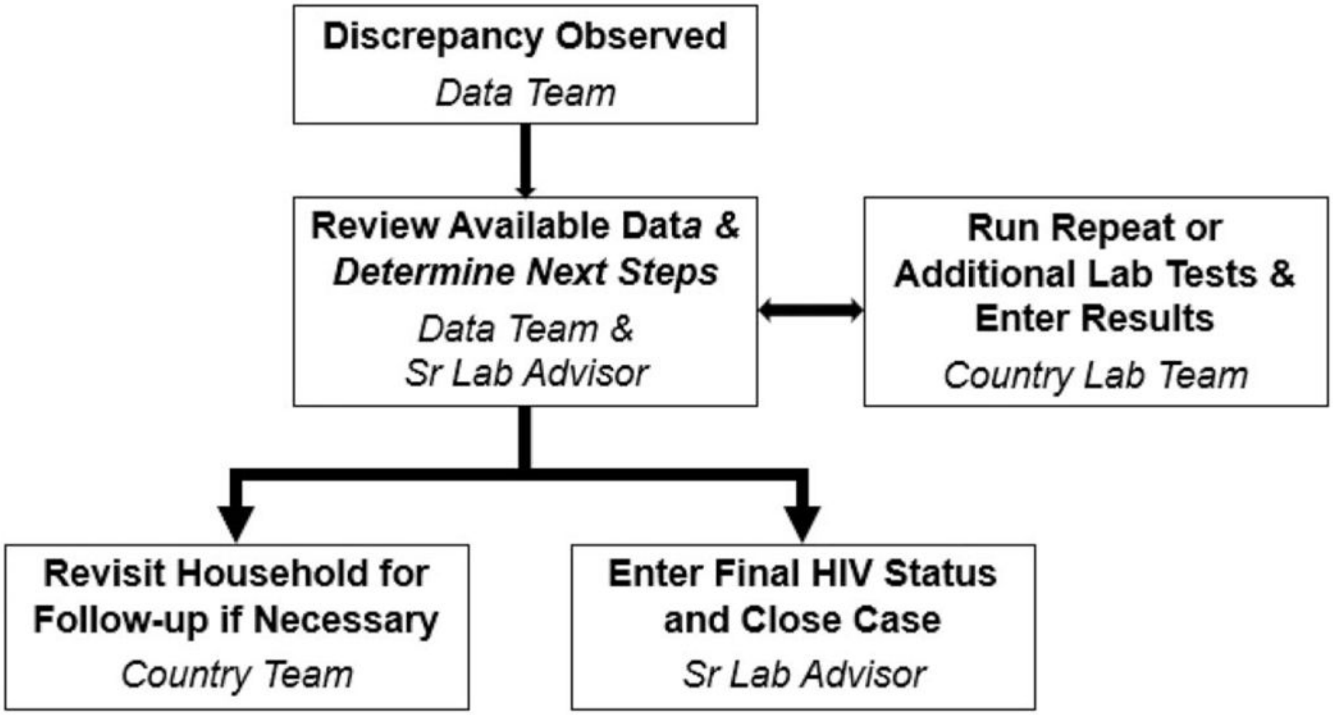
Discrepancy Observed
Data Team
Review Available Data & Determine Next Steps
Data Team & Sr Lab Advisor
Run Repeat or Additional Lab Tests & Enter Results
Country Lab Team
Revisit Household for Follow-up if Necessary
Country Team
Enter Final HIV Status and Close Case
Sr Lab Advisor

FIGURE 2.

HIV Sero Status Discrepancy Workflow for Population-Based HIV Impact Assessment Surveys (2015–2020).

TABLE 1.

Number of Records Processed and Timeline for the Population-Based HIV Impact Assessment (PHIA; 2015–2020)

Country (Survey Name)	Date Range	Household Questionnaires*	Individual Questionnaires*	Blood Draws†
Zimbabwe (ZIMPHIA)	October 18, 2015:August 7, 2016	11,717	33,156	30,082
Malawi (MPHIA)	November 27, 2015:August 26, 2016	11,386	28,952	23,697
Zambia (ZAMPHIA)	March 1, 2016:August 31, 2016	10,957	31,488	27,866
Uganda (UPHIA)	August 26, 2016:March 31, 2017	12,386	40,046	40,152
Eswatini (SHIMS2)	August 30, 2016:March 12, 2017	5185	15,453	14,634
Tanzania (THIS)	October 27, 2016:June 15, 2017	14,811	43,145	42,459
Lesotho (LePHIA)	November 25, 2016:April 25, 2017	8824	17,449	15,982
Namibia (NAMPHIA)	June 7, 2017:November 28, 2017	9315	26,254	23,800
Cameroon (CAMPHIA)	June 28, 2017:February 10, 2018	11,623	35,034	33,812
Côte d’Ivoire (CIPHIA)	August 12, 2017:March 27, 2018	8983	34,720	22,680
Ethiopia (EPHIA)	October 4, 2017:April 25, 2018	10,529	25,556	23,981
Kenya (KENPHIA)	May 26, 2018:February 11, 2019	16,918	39,274	36,403
Rwanda (RPHIA)	October 15, 2018:March 6, 2019	11,219	39,328	39,259
Haiti‡ (HAPHIA)	July 1, 2019:November 15, 2020	12,497‡	34,217‡	30,754‡

*Count of all questionnaires administered to eligible households and individuals, including refusals.

†Total blood draws include specimens that were not analyzed from de jure participants and specimens that were not linked to interview records.

‡Targets for HAPHIA, which was undergoing final data cleaning as of this writing.

SHIMS2, Swaziland HIV Impact Measurement Survey; THIS, Tanzania HIV Impact Survey.

TABLE 2.

Data Sources by Country, Population-Based HIV Impact Assessments (2015–2020)

Country	Survey Data Server	HIV Rapid Test Algorithm	QA Results	Plasma VL Instrument	DBS VL Instrument	EID Instrument	Review of VL
Zimbabwe	Multicountry	3 tests: Determine, First Response, and STAT-PAK	Excel workbook	CAP/CTM48 through LIS	NucliSENS easyMAG/easyQ	CAP/CTM48	In-country laboratory adviser
Malawi	Multicountry	2 tests: Determine and Uni-Gold	Excel workbook	m2000	m2000	m2000	In-country laboratory adviser
Zambia	Multicountry	2 tests: Determine and Uni-Gold	Excel workbook	CAP/CTM48	m2000	CAP/CTM48 at reference laboratory	CDC
Uganda	Multicountry	3 tests: Determine, STAT-PAK, and SD Bioline	TREM; QA performed at CL	CAP/CTM48, m2000	m2000 at reference laboratory	CAP/CTM96	CDC
Eswatini	Multicountry	3 tests: Determine, Uni-Gold, and Clearview	TREM	CAP/CTM96	CAP/CTM96, m2000	CAP/CTM96	CDC
Tanzania	Multicountry	2 tests: SD Bioline and Uni-Gold	TREM	CAP/CTM96, CAP/CTM48	CAP/CTM96, CAP/CTM48	CAP/CTM96, CAP/CTM48	CDC
Lesotho	Multicountry	2 tests: Determine and Uni-Gold	TREM	CAP/CTM96 through LIS	CAP/CTM96 through LIS	CAP/CTM96 through LIS	CDC
Namibia	Namibia-specific	3 tests: Determine, Uni-Gold and Sure Check	TREM	CAP/CTM96	CAP/CTM96	GeneXpert at SL with CAP/CTM96 as confirmatory	CDC
Cameroon	Multicountry	2 tests: Determine and OraQuick	TREM	m2000	m2000	m2000 at reference laboratory	CDC
Côte d’Ivoire	Multicountry	3 tests: Determine, SD Bioline, and STAT-PAK	TREM	CAP/CTM96	CAP/CTM96	CAP/CTM96	CDC
Ethiopia	Multicountry	3 tests: Wantai, Uni-Gold, and VIKIA	TREM	CAP/CTM96	m2000	CAP/CTM96	CDC
Kenya	Kenya-specific	2 tests: Determine and First Response	TREM	CAP/CTM96	CAP/CTM96	GeneXpert at SL with CAP/CTM96 as confirmatory	CDC
Rwanda	Rwanda-specific	2 tests: Alere Combo and STAT-PAK	TREM	CAP/CTM96	CAP/CTM96	CAP/CTM96	CDC
Haiti	Multicountry	2 tests: Determine and Uni-Gold	TREM	m2000	m2000	m2000	CDC

Alere Combo, Alere Determine HIV-1/2 Ag/Ab Combo; Clearview, Clearview HIV 1/2; DBS, dried blood spot; Determine, Determine HIV-1/2; EID, early infant diagnosis; First Response, First Response HIV-1-2; STAT-PAK, Chembio HIV 1/2 STAT-PAK; OraQuick, OraQuick HIV-1/2 rapid antibody test; SD Bioline, SD BIOLINE HIV-1/2; Sure Check, Chembio SURE CHECK HIV 1/2; Uni-Gold, Uni-Gold Recombigen HIV-1/2; VIKIA, VIKIA HIV1/2; Wantai, Wantai Rapid Test for Antibody to HIV.

TABLE 3.

Select Process Indicators for Population-Based HIV Impact Assessments (2015–2020)

Questionnaires	
Total households	173,174
Total interviews	453,613
Household questionnaires without errors	98.5%
Individual questionnaires without errors	99.2%
Data points per record	1500
Blood handling and testing	
Total blood draws	373,882
LDMS inventory records[*]	4,255,480
PIMA results	285,321
Return of results	
No. of VL returned	25,664
No. of VL returned with return dates	23,398
% VL on time[†]	91.50%
Secure file server	
No. of files	31,341
No. of folders	840
Files incorporated into biomarker	17,450
Staff accounts	600+
Discrepancies	
Discrepant cases reviewed and resolved	2192

[*] Includes up to 5 plasma aliquots and 2 dried blood samples cards from each blood draw and records from each laboratory where specimens were handled.

[†] Among those with return dates.