
























Research Article

Developing Clinical Phenotype Data Collection Standards for Research in Africa

Lyndon Zass ¹, Katherine Johnston ¹, Alia Benkahla ², Melek Chaouch ²,
Judit Kumuthini ³, Fouzia Radouani ⁴, Liberata Alexander Mwita ⁵, Nihad Alsayed ⁶,
Taryn Allie ¹, Dassen Sathan ⁷, Upendo Masamu ⁵,
Milaine Sergine Seuneu Tchamga ⁸, Tsaone Tamuhla ¹, Chaimae Samtal ⁹,
Victoria Nembaware ¹⁰, Zoe Gill ^{1,11}, Samah Ahmed ⁶, Yosr Hamdi ^{12,13},
Faisal Fadlelmola ⁶, Nicki Tiffin ^{1,3,14} and Nicola Mulder ^{1,14}

¹Computational Biology Division, Department of Integrative Biomedical Sciences, IDM, University of Cape Town, Cape Town, South Africa

²Laboratory of Bioinformatics, BioMathematics and BioStatistics LR16IPT09, Institut Pasteur de Tunis, Tunis, Tunisia

³South African National Bioinformatics Institute (SANBI), Life Sciences Building, University of Western Cape, Bellville, Cape Town, South Africa

⁴Chlamydiae & Mycoplasmas Laboratory Research Department, Institut Pasteur du Maroc, 20360 Casablanca, Morocco

⁵Muhimbili Sickle Cell Program, Department of Hematology and Blood Transfusion, Muhimbili University of Health and Allied Sciences, Dar-es-Salaam, Tanzania

⁶Kush Centre for Genomics & Biomedical Informatics, Biotechnology Perspectives Organization, Khartoum 11111, Sudan

⁷Software Information Systems Department, FOICDT, University of Mauritius, Reduit, Mauritius

⁸Department of Mathematics and Physics, Cape Peninsula University of Technology, Bellville, Cape Town, South Africa

⁹Laboratory of Biotechnology, Environment, Agri-Food and Health, Faculty of Sciences Dhar El Mahraz-Sidi Mohammed Ben Abdellah University, Fez 30000, Morocco

¹⁰Division of Human Genetics, Department of Pathology, Faculty of Health Sciences, University of Cape Town, Cape Town, South Africa

¹¹Department of Molecular Biology, Johannes Gutenberg University, Mainz, Germany

¹²Laboratory of Biomedical Genomics and Oncogenetics, Institut Pasteur de Tunis, University of Tunis El Manar, Tunis, Tunisia

¹³Laboratory of Human and Experimental Pathology, Institut Pasteur de Tunis, Tunis, Tunisia

¹⁴Wellcome Centre for Infectious Disease Research in Africa, Institute of Infectious Diseases and Molecular Medicine, Faculty of Cape Town, University of Cape Town, Cape Town, South Africa

Correspondence should be addressed to Lyndon Zass; lyndon.zass@uct.ac.za

Received 30 November 2022; Revised 30 June 2023; Accepted 21 July 2023; Published 19 September 2023

Academic Editor: Gerald Mboowa

Copyright © 2023 Lyndon Zass et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Modern biomedical research is characterised by its high-throughput and interdisciplinary nature. Multiproject and consortium-based collaborations requiring meaningful analysis of multiple heterogeneous phenotypic datasets have become the norm; however, such analysis remains a challenge in many regions across the world. An increasing number of data harmonisation efforts are being undertaken by multistudy collaborations through either prospective standardised phenotype data collection or retrospective phenotype harmonisation. In this regard, the Phenotype Harmonisation Working Group (PHWG) of the Human Heredity and Health in Africa (H3Africa) consortium aimed to facilitate phenotype standardisation by both promoting the use of existing data collection standards (hosted by PhenX), adapting existing data collection standards for appropriate use in low- and middle-income regions such as Africa, and developing novel data collection standards where relevant gaps were identified. Ultimately, the PHWG produced 11 data collection kits, consisting of 82 protocols, 38 of which were existing protocols, 17 were adapted, and 27 were novel protocols. The data collection kits will facilitate phenotype standardisation and harmonisation not

only in Africa but also across the larger research community. In addition, the PHWG aims to feed back adapted and novel protocols to existing reference platforms such as PhenX.

1. Introduction

Biomedical data are increasingly used in high-throughput and interdisciplinary approaches, with data generation and knowledge discovery occurring at a faster pace than ever before, particularly in the areas of genomics and bioinformatics [1, 2]. As a result, there has been an increase in multiproject, consortium-based collaborations and meta-analyses. Such efforts have the potential to expedite knowledge discovery in resource-limited research communities, particularly low- and middle-income countries, where funding for large sample collections is limited [3]. Advances in genomics research have raised the need to match large-scale phenotypic data which incorporate social, environmental, and clinical factors, with genetic information. However, researchers still struggle to collect complete and valuable data in a standard manner, which hinders data integration and collaborative efforts [4]. Meaningful analysis of multiple heterogeneous phenotypic datasets is difficult, and at times, impossible, without standardisation of the data elements prior to collection or retrospective harmonisation of the collected phenotypic data.

Phenotype standardisation involves harmonisation of the way researchers define and collect information about clinical phenotypes and environmental exposures. This is opposed to retrospective phenotype harmonisation, which involves the integration of phenotype data collected or defined in various ways and is often employed in meta-analyses and collaborative research efforts. Phenotype standardisation may be employed for various reasons. For example, to promote operability and the interoperability of data coming into a biomedical database, collaborative and multisite studies need to ensure that their collection methods and dataset formats are as similar as possible to reduce bias and assist with overall study management and quality control. One way to achieve this is by employing ontologies to standardise the definitions for trait and case classification as previously shown [5]. In doing so, phenotype standardisation can also improve the quality of research outputs, as has been shown in pharmacogenetics and kidney disease research [6–8]. More recently, the NIH-funded PhenX (consensus measures for Phenotypes and eXposures) Toolkit (<https://www.phenxtoolkit.org/>) has emerged as a leading resource for well-established standard protocols for the collection of phenotype and exposure information. This publicly available online catalogue contains standard and recommended protocols to extract the maximum value from data collected for genomics research, although these protocols are also applicable to other fields [9–11]. Measures hosted by PhenX have been successfully employed to identify opportunities for cross-study collaboration through established databases such as the database of Genotypes and Phenotypes (dbGaP) and the Cancer Data Standards Registry and Repository (caDSR) [12]. Despite these advances, some protocols within PhenX may lack regional validity since they were largely developed from a Western

perspective. For example, protocols related to the concepts of diet, nuclear family, and extended family may need to be adapted since these vary across the world.

The Human Heredity and Health in Africa (H3Africa) consortium was established to drive novel and innovative genomics research in Africa and build capacity on the continent [13]. As part of the consortium, the H3Africa Bioinformatics Network (H3ABioNet) was formed to support these efforts, with a particular focus on the production and sharing of FAIR (Findable, Accessible, Interoperable, and Reproducible) data [14]. With the goal of facilitating the standardisation and interstudy sharing of phenotypic data within H3Africa and beyond, these initiatives established the H3Africa Phenotype Harmonisation Working Group (PHWG) in 2014. The PHWG aimed to build a core set of phenotypes to be collected within the consortium and develop protocols to facilitate the collection of these phenotypes. This work was later expanded to cover specific research or disease domains, which were represented in clusters across the consortium [15, 16]. Here we describe the development and dissemination of standard phenotype data collection kits for genomics research, which are adapted to ensure their suitability for use in African settings.

2. Methods

2.1. Core Phenotypes. The PHWG oversaw the establishment of a standard set of CORE PHENOTYPES, relevant to the study of the diseases or traits explored in H3Africa. Given the diversity of diseases and populations studied within H3Africa (covering the study of neonatal respiratory diseases to adult renal diseases to the transmission of zoonotic tuberculosis in rural pastoral populations), the initial set of phenotypes was extremely broad. The PHWG catalogued a set of CORE PHENOTYPES based on the number of H3Africa studies that measured them (each phenotype had to be measured in more than 3 studies). As a result, a set of 26 core phenotypes and 10 discretionary phenotypes were identified. These were selected based on their general interest and applicability in most study populations, relative ease of collection, and whether they were a primary phenotype in one of the H3Africa studies (to facilitate identification of common controls). To formulate these CORE PHENOTYPES, the working group chose to use the PhenX Toolkit to format the selected phenotypes, which initiated the development of the H3Africa Standard Case Report Form (Standard CRF). H3Africa grantees were encouraged, where possible, to measure the CORE PHENOTYPES in each research study so that the community could assess the power of the opportunity to measure a common set of phenotypes across a maximum number of projects. Following the initial development of the Standard CRF, the Working Group (WG) revised it to ensure its application to both paediatric and adult participants.

2.2. Domain-Specific Toolkits. Building on the development of the H3Africa Standard CRF, the PHWG expanded the scope of development to cover domain-specific research group needs within H3Africa, to facilitate potential phenotype standardisation and harmonisation within and across these groups, should they wish to expand their collaborative efforts. As with the CORE PHENOTYPES, consensus mapping between H3Africa projects of a specific group was first conducted by cataloguing the domain-specific phenotypes within the groups and identifying common overlaps. Thereafter, phenotypes were mapped to PhenX to identify the appropriate protocols associated with the phenotypes. Where relevant protocols were identified but not suited to implementation in Africa, the protocols were adapted for such use. Where relevant protocols were not identified, novel protocols were developed. Thereafter, guidelines to facilitate data collection were established for both the Standard CRF and domain-specific kits. The overall development process for these two components is illustrated in Figure 1.

2.3. Technical Development. Since the platform was widely employed by H3Africa for clinical data collection, REDCap (Research Electronic Data Capture) was selected to develop the data collection kits. REDCap is a secure, web-based software platform designed to support data capture for research studies. It provides (1) an intuitive interface for validated data capture; (2) audit trails for tracking data handling and export procedures; (3) automated export procedures for seamless data downloads to common statistical packages; and (4) procedures for data integration and interoperability with external sources [17]. PhenX also provides protocols that generate REDCap.xml files, and the software itself lends itself to easy design and sharing of data collection instruments.

Before building the kits, some functional elements had to be considered in terms of the variables and project structure within REDCap:

- (1) Each domain-specific kit needed to be able to stand alone with the CORE PHENOTYPES or be incorporated with any other combination of kits. To this end, all kits were built in a single REDCap project to eliminate the possibility of duplicate variable names before splitting them into separate domain-specific kits. In addition, because several kits have branching logic based on participants' biological sex and age (collected in the CORE PHENOTYPES), each domain-specific kit was packaged alongside the CORE PHENOTYPES but also made available as separate data collection instruments for import.
- (2) Consistent coding for missing data must be applied across data collection kits; therefore, each module has a preset list of missing data codes, which can be modified by users.
- (3) Variable naming conventions needed to be useful to the end-user; therefore, variable names relevant to what was being collected were employed to avoid nonsensical indexed naming.
- (4) Consistent coding was an absolute necessity, so basic codes for common responses and formats were applied throughout.
- (5) Cosmetically, all the forms were formatted with the same style, and the basic REDCap forms were maintained, without the use of external kits. Field embedding was limited to a few fields necessary to accommodate mobile data collectors.

The REDCap instruments for every kit are intended to allow studies to include additional data elements and can be adapted to suit the needs of any study. The only caution for developers is to consider the elements already included as recommended for collection and, when making changes, to consider the branching logic already in place (this is primarily with respect to the display of age or biological sex questions).

2.4. Ontology Mapping. Once the data dictionaries were finalised, each variable in the data dictionaries was mapped to a relevant ontology code, where possible, to facilitate interoperability and reproducibility. Ontology mapping was conducted using the Ontology Lookup Service (OLS) and Zooma developed by the European Bioinformatics Institute (EBI). Domain-dedicated and well-maintained ontologies were preferred during this mapping. Ontologies were checked and revised to ensure complete correspondence of mapped variables, and the ontology code was incorporated into the data dictionary file for machine readability.

2.5. Review. The final set of data collection kits (CORE PHENOTYPES and Domain-specific Kits) underwent multiple rounds of review, both internal and external, to ensure quality and usability. The external review involved distributing a survey to African experts in a particular field to review the selected phenotypes and the collection protocols. The experts included in the survey comprised of researchers in leadership positions in a particular field and professionals involved in data collection and management. The survey allowed experts to review each element and associated protocol included in a toolkit, designate its applicability or lack thereof, and make additional recommendations to improve the toolkit. The internal review involved the revision of ontology mappings by a dedicated team, two rounds of quality and consistency checks using the established guidelines, and a review of the structure and organisation of the data dictionaries to ensure technical validity and usability.

3. Results

3.1. Data Collection Kits. In total, the PHWG produced 11 data collection kits, including the CORE PHENOTYPES and 10 domain-specific kits. All kits contain protocols adapted for use in both paediatric and adult participants. In total, they included 82 protocols, of which 38 were existing

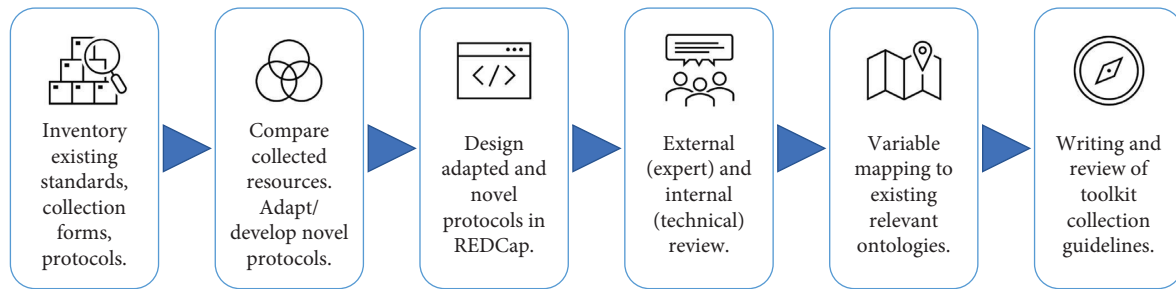


FIGURE 1: Development process of core phenotypes and domain-specific kits.

protocols (from PhenX or other existing initiatives, e.g., WHO, CIDRI-Africa, and the HIV Data Exchange Initiative), 17 were adapted, and 27 were novel protocols. Adapted protocols are protocols that are based on existing protocols but were edited (e.g., addition/removal of fields or rephrasing/restructuring of protocols) to improve their application in low- and middle-income regions and Africa, in particular. Examples of adapted protocols include those used to capture information on diet and household characteristics. Novel protocols are those for which it was not possible to identify appropriate existing protocols and were thus newly developed, such as asthma, pregnancy history, birth history, and family history protocols. An overview of the established data collection kits is illustrated in Figure 2 and summarised in Table 1.

As illustrated in Figure 3, each kit is composed of multiple components, including a REDCap implementation file, a data dictionary to be implemented on the platform of choice, a case report form or data collection document, and a guideline, to facilitate the use of the kit.

The CORE PHENOTYPES were first released internally within the H3Africa consortium in 2017/2018, then publicly thereafter. The additional domain-specific kits were finalised and released in 2022. All kits are currently available on various public platforms, including the H3ABioNet website (<https://h3abionet.org/data-standards/phenotype-data-collection-standard>), H3ABioNet GitHub (<https://github.com/h3abionet/h3aphenstds>), and ZivaHub/Figshare (https://zivahub.uct.ac.za/projects/H3ABioNet_H3Africa_Phenotype_Standards_Project/149305). Since its release on ZivaHub/Figshare, all related pages have been viewed for a total of 1795 times and the kits have been downloaded 288 times (as of May 2023).

3.2. External Review. In total, the data collection kits were reviewed by 45 individuals (experts) through a survey, with the majority of respondents coming from South Africa (16), Tunisia (5), and Nigeria (4). In addition, the majority of respondents were classified as non-H3Africa experts (17), while most of the H3Africa respondents were associated with the Phenotype Harmonisation (15), Cardiovascular Disease, and Rare Diseases H3Africa WGs. The majority of respondents identified themselves as experts in rare and developmental disorders (13), infectious diseases (13), and family history (12). Figure S1 (A and B) shows the complete breakdown of survey respondents.

3.3. External Use. The CORE PHENOTYPES have been disseminated internally for an extensive period of time (since 2018) and, to our knowledge, have been implemented by at least 9 H3Africa projects, mainly the projects initiated in the second round of H3Africa funding. These include studies involving exome sequencing and genome-wide association and epigenetic studies conducted on study populations from South Africa, Nigeria, Rwanda, Ethiopia, and Botswana and focused on the genetic underpinnings of infectious diseases and cardiovascular, mental, and developmental disorders [18–23]. The CORE PHENOTYPES served as a point of reference to facilitate retrospective phenotype harmonisation by the Cardiovascular H3Africa Innovation Resource (CHAIR) [24] and the Cardiometabolic Disorders in African-Ancestry Populations (CARDINAL) initiatives and to generate synthetic phenotype data by the Common Infrastructure for National Cohorts in Europe, Canada, and Africa (CINECA) initiative [25]. Since the majority of the Domain-specific Kits were only finalised and disseminated in 2022, we do not yet have reports on validation for these kits, although, to our knowledge, the Genome Tunisia Collaborative Alliance has used a slightly modified and translated version of the CORE PHENOTYPES and Family History kits.

4. Discussion

In this report, we have described the efforts of the H3Africa PHWG to establish standard data collection protocols for relevant biomedical and health research in Africa and in resource-limited regions globally. As far as we are aware, this is the first such standardisation effort at the African level. It has been carried out within the framework of the H3Africa projects with a vision to enable African researchers to access and use them for newly developed projects, thus facilitating future data integration, phenotype harmonisation, and meta-analyses efforts, while also building a solid foundation for future collaborative efforts, both within the continent but also cross-continently. We encourage use and feedback throughout the continent, to ensure continuous improvement on the existing recommendations.

As indicated, the established protocols not only encourage and promote standard data collection from the ground up but may also be useful in integrating phenotypic data from various sources retrospectively, i.e., harmonising phenotypes. Indeed, the kits are based on commonly collected phenotypes from specific research domains. When

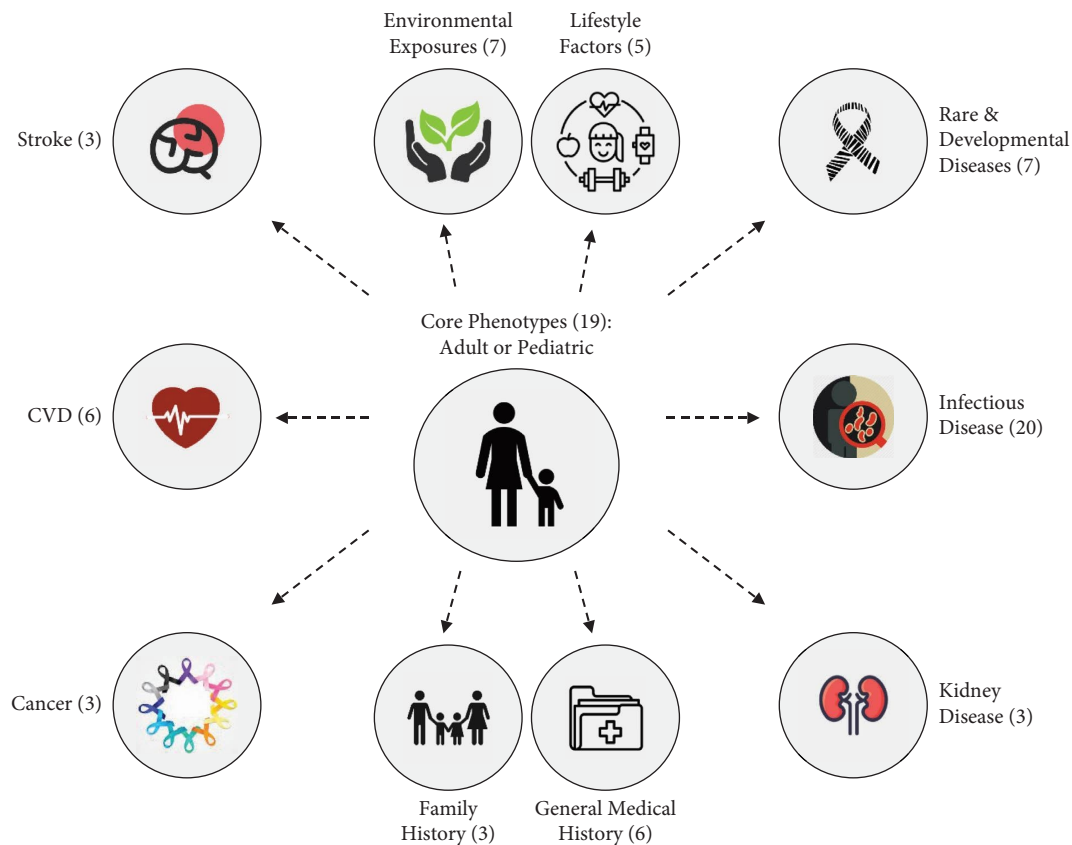


FIGURE 2: Overview of data collection kits developed by H3Africa. The number of protocols indicated per kit is given in brackets; CVD: cardiovascular disease.

developing the protocols, the PHWG sought to find a balance between simplicity and completeness, and to cover various scenarios. The first step in many retrospective phenotype harmonisation efforts is typically identifying overlapping or common phenotypes across various projects and then to develop a data dictionary structured to allow transformation of data from different sources into a harmonised data structure [26]. In so doing, the data collection kits, and specifically standard data dictionaries, can facilitate two of the key steps in phenotype harmonisation.

The data collection kits were developed and released in adherence to FAIR principles and also promote FAIR principles in research projects which employ them. The kits were released on multiple platforms, including the H3ABioNet website, GitHub, and ZivaHub/Figshare, making them freely accessible to any interested user and allocating them unique identifiers to facilitate their findability. In addition, the kits provide a REDCap-based XML file which can be implemented directly on REDCap, which is broadly available for academic use, along with an EXCEL file which facilitates viewing for novice users and enables users to transform the file for the data management platform of their choice. The kits encourage interoperability and reproducibility in that they are standard recommendations based on existing standards and, additionally, employ

ontology codes for machine-readable interoperability. They are also released with associated guidelines for reproducibility in practice.

Regarding the assessment of the validity of data collection kits, we had limited opportunity to measure the validity of the domain-specific kits in practice within the scope of H3Africa, given the time of development and final release of these kits. To counter this, we sought to assess and ensure the validity of the kits in other ways; therefore, we encouraged collaboration and worked alongside field experts to develop each of the kits, including, as previously illustrated, the Kidney Disease and Stroke kits [15, 16]. We also sought additional external feedback on the elements and data collection protocols included in the kits by surveying field experts. Finally, where existing standard protocols were found and applicable, like protocols observed in PhenX, we encouraged the use of these standards, which have previously undergone broad validity assessments. This was to avoid the development of new but overlapping protocols. Despite these measures, we recognise that the kits may have practical limitations not considered during the development and thus encourage feedback on the kits from the user community. Feedback may be provided through multiple channels, including submitting issues on GitHub and the H3ABioNet Helpdesk [27]. These also serve as an important channel

TABLE 1: Overview of structure of data collection toolkits developed by H3Africa.

Toolkit	Protocols		
Core phenotypes	Demographics ⁺ Anthropometrics ⁺ Blood pressure [#] Urine results [#] Household characteristics ⁺ Water resource ⁺ Physical activity [#] Vaccinations [#] Allergies ⁺ Family composition [±]	Drug use [#] Medication log ⁺ Diabetes history ⁺ Asthma ⁻ Occupational exposures [#] Air contaminants [#] Diet ⁺ Last pregnancy outcome [±] Medical history [±] Family history of traumatic life events [±]	CVD* Core [#] HIV core [#] Infectious disease core ⁺ Dyslipidemia [#] Cancer core ⁺ Kidney disease core [#] Stroke history core [#] Personal care products [#] Caffeine intake [#] Sleep habits [#]
Environmental factors	Sanitation ⁺ Occupational history [#] Dietary supplements [#] Current pregnancy status [±] Pregnancy history [±]	Occupational exposures [#] Air contaminants [#] Diet ⁺ Last pregnancy outcome [±] Medical history [±]	Personal care products [#] Caffeine intake [#] Sleep habits [#]
Lifestyle factors	Physical activity [#] Vaccinations [#] Allergies ⁺ Family composition [±]	Diet ⁺ Last pregnancy outcome [±] Medical history [±] Family history of traumatic life events [±]	Sleep habits [#]
General medical history	Family history [±]	Family history of traumatic life events [±]	Sleep habits [#]
Family history	Family history [±]	Family history of traumatic life events [±]	Sleep habits [#]
Cardiovascular disease	Angina [#] Heart attack [#]	Cardiac blood flow measurements [±] Cardiovascular disease blood biomarkers [±]	Sleep habits [#]
Stroke	Stroke characterisation [±]	Poststroke risk reduction strategies [±]	Sleep habits [#]
Cancer	Cancer prognosis and treatment ⁺	Cancer screenings [#]	Sleep habits [#]
Kidney disease	Kidney treatment status [#]	Kidney function assay [#]	Sleep habits [#]
Rare and developmental disorders	Biological parents and grandparents [±] Birth history [±]	Developmental milestones [±] Neurodevelopmental assessment [±]	Rare disorders [±]
Infectious diseases	Malaria exposure and treatment [±] Malaria testing [#] Malaria signs and symptoms ⁺ Trypanosomiasis testing [#]	Trypanosomiasis signs and symptoms ⁺ TB history ⁺ TB testing [#] Current TB ⁺	Liver function assay [#] HIV signs and symptoms ⁺ STDs [±] Complete blood cell count ⁺

*CVD: cardiovascular disease; HIV: human immunodeficiency virus; TB: tuberculosis; STD: sexually transmitted disease; ⁺adapted protocols; [±]novel protocols; [#]existing protocols.

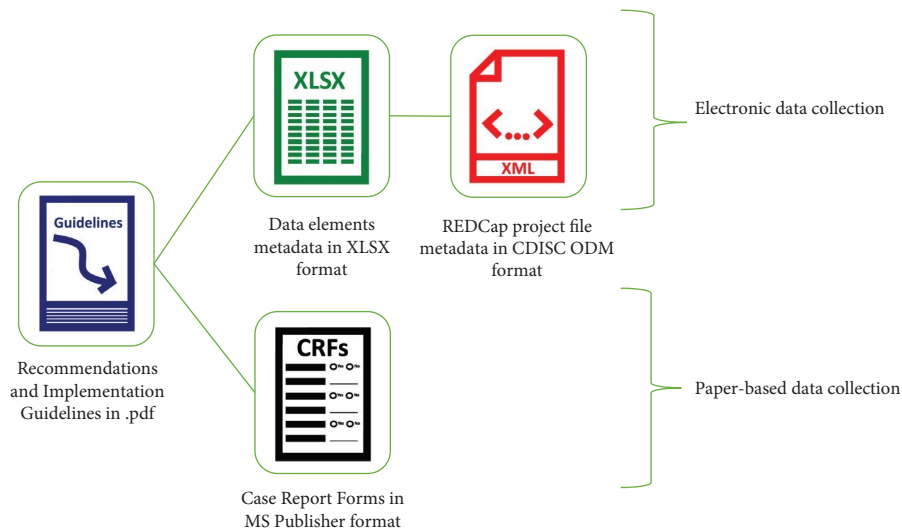


FIGURE 3: Data collection kit components.

through which we can track implementation of the kits. Although we encourage citation of kits when employed, many projects may not reference data collection forms in their methods section, representing a tracking barrier.

In contrast to the domain-specific kits, the CORE PHENOTYPES were released to H3Africa for a much longer period. H3Africa grantees were encouraged, where possible, to use the Standard CRF, as a guide to collect a common set of phenotypes to be measured in each research participant, although it was recognised that this was not always realistic, particularly because (1) there are costs in terms of time and effort to collect additional phenotypes, (2) recipients are not funded to collect phenotypes other than those in the original grant application, (3) some phenotypes have little or no relevance to certain populations and collection settings, and (4) some grants had already finalised their CRFs or were already recruiting subjects, which limits the possibilities for adding or revising measures. Despite the abovementioned considerations, several H3Africa projects did implement the CORE PHENOTYPES within their research projects and data collection processes, particularly the projects which were funded during H3Africa's second cycle (once the CORE PHENOTYPES were already established and released). Feedback from these projects was extremely positive regarding simplifying the CRF design process, the simplicity of use, and the comprehensive nature of design. As a by-product of collecting the CORE PHENOTYPES, various groups found great collaborative potential with regard to simple retrospective phenotype harmonisation, including the Mental Health and Cardiovascular Disease working groups, the latter of which developed a database for phenotypes harmonised to many of the CORE PHENOTYPES [24].

Although the kits have been successfully implemented by numerous studies, we recognise that existing barriers may prevent the more extensive implementation of the data collection kits. One such barrier is a lack of technical capacity to implement the kits electronically. This is one of the

main reasons why the kits also include a CRF which promotes paper-based collection, the other being that research studies in low-income settings may often still rely on paper-based collection. To address such capacity gaps, we also aim to provide training materials exhibiting implementation of the kits on technical platforms. The other foreseeable barrier is that research initiatives based on existing cohorts are often hesitant to switch data collection methods once data collection has previously been conducted in a different manner, as this presents data integration issues downstream. To address this barrier, we need to also promote the kits as tools for retrospective phenotype harmonisation, as it has previously been used for by the Mental Health and Cardiovascular Disease working groups.

In the future, the primary efforts of the PHWG will be focused on promoting the use of the developed data collection kits on a broader scale, particularly in Africa. This may be achieved through endorsement by local research and funding bodies but also endorsement from previous users. In addition, the PHWG will be maintaining the data collection kits so that they remain relevant for future. One of the key goals for the project will be to feed back the adapted and novel protocols developed to PhenX for incorporation into the catalogue. In this manner, a broader set of users may be reached through an established body, which will, ultimately, facilitate the broad goals of the project. In addition, we will also investigate integration and interoperability with existing common data models such as the Observational Medical Outcomes Partnership (OMOP) model to increase broad usability [28].

5. Conclusion

The PHWG has successfully developed a series of easy-to-use data collection kits that cover a range of biomedical research fields. These kits should facilitate phenotype standardisation and harmonisation efforts on the African continent and the larger user community. The standards can form the basis for data models for post hoc data

harmonisation. In addition to the abovementioned benefits, the data collection kits will also promote FAIR data principles, ultimately enabling data integration and interoperability. Finally, as mentioned previously, several novel data collection protocols were developed during these efforts, where relevant gaps were identified. These, along with protocols which were adapted from already established protocols, will be submitted for (re)incorporation into the PhenX platform.

6. Resource Availability

The resources discussed in the publication can be found on multiple platforms as listed below:

- (1) Website (<https://www.h3abionet.org/component/spagebuilder/page/259>)
- (2) GitHub (<https://github.com/h3abionet/h3aphenstds>)
- (3) Figshare (https://zivahub.uct.ac.za/projects/H3ABioNet_H3Africa_Phenotype_Standards_Project/149305)

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request. To request survey data, researchers can contact the first or second author.

Disclosure

The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

The authors would like to acknowledge all survey respondents who provided valuable input in the development of the research outputs, Dr Jeff Struewing, for guidance in the original development of the Standard CRF, all PHWG members for contributions via feedback and discussions, and Dr Emma Wiener for contributions to the rare and development disorders data collection kit. In addition, the authors would like to acknowledge that all research reported in this publication was supported by the National Human Genome Research Institute of the National Institutes of Health through the NIH Common Fund under Award No. U24HG006941. H3ABioNet is a member of the Human Health and Heredity in Africa Consortium (H3Africa). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Supplementary Materials

Figure S1: overview of survey respondents by (A) country; (B) H3Africa Working Group; and (C) expertise. (*Supplementary Materials*)

References

- [1] Y. Hamdi, L. Zass, H. Othman et al., “Human OMICs and computational biology research in Africa: current challenges and prospects,” *OMICS: A Journal of Integrative Biology*, vol. 25, no. 4, pp. 213–233, 2021.
- [2] J. Luo, M. Wu, D. Gopukumar, and Y. Zhao, “Big data application in biomedical research and health care: a literature review,” *Biomedical Informatics Insights*, vol. 8, pp. BII.S31559–10, 2016.
- [3] J. Dusdal and J. J. W. Powell, “Benefits, motivations, and challenges of international collaborative research: a sociology of science case study,” *Science and Public Policy*, vol. 48, no. 2, pp. 235–245, 2021.
- [4] NHLBI Trans-Omics for Precision Medicine, “WGS-2017 TOPMed DCC analysis workshop,” 2022, <https://topmed.nhlbi.nih.gov/2017-topmed-dcc-analysis-workshop>.
- [5] G. M. Volk, “Advantages for the use of standardized phenotyping in databases,” *HortScience*, vol. 45, no. 9, pp. 1310–1313, 2010.
- [6] R. L. Mehta, L. Awdishu, A. Davenport et al., “Phenotype standardization for drug-induced kidney disease,” *Kidney International*, vol. 88, no. 2, pp. 226–234, 2015.
- [7] M. Pirmohamed, G. P. Aithal, E. Behr, A. Daly, and D. Roden, “The phenotype standardization project: improving pharmacogenetic studies of serious adverse drug reactions,” *Clinical Pharmacology & Therapeutics*, vol. 89, no. 6, pp. 784–785, 2011.
- [8] G. P. Aithal, P. B. Watkins, R. J. Andrade et al., “Case definition and phenotype standardization in drug-induced liver injury,” *Clinical Pharmacology & Therapeutics*, vol. 89, no. 6, pp. 806–815, 2011.
- [9] T. Hendershot, H. Pan, J. Haines et al., “Using the PhenX toolkit to add standard measures to a study,” *Current Protocols in Human Genetics*, vol. 86, pp. 1.21.1–1.21.17, 2015.
- [10] P. J. Stover, W. R. Harlan, J. A. Hammond, T. Hendershot, and C. M. Hamilton, “PhenX: a toolkit for interdisciplinary genetics research,” *Current Opinion in Lipidology*, vol. 21, no. 2, pp. 136–140, 2010.
- [11] C. M. Hamilton, L. C. Strader, J. G. Pratt et al., “The PhenX Toolkit: get the most from your measures,” *American Journal of Epidemiology*, vol. 174, no. 3, pp. 253–260, 2011.
- [12] H. Pan, K. A. Tryka, D. J. Vreeman et al., “Using PhenX measures to identify opportunities for cross-study analysis,” *Human Mutation*, vol. 33, no. 5, pp. 849–857, 2012.
- [13] N. Mulder, A. Abimiku, S. N. Adebamowo et al., “H3Africa: current perspectives,” *Pharmacogenomics and Personalized Medicine*, vol. 11, pp. 59–66, 2018.
- [14] N. J. Mulder, E. Adebisi, R. Alami et al., “H3ABioNet, a sustainable pan-African bioinformatics network for human heredity and health in Africa,” *Genome Research*, vol. 26, no. 2, pp. 271–277, 2016.
- [15] J. Kumuthini, C. van Woerden, A. Mallett et al., “Proposed minimum information guideline for kidney disease—research and clinical data reporting: a cross-sectional study,” *BMJ Open*, vol. 9, no. 11, Article ID e029539, 2019.
- [16] J. Kumuthini, L. Zass, M. Chaouch et al., “Proposed guideline for minimum information stroke research and clinical data reporting,” *Data Science Journal*, vol. 18, no. 1, p. 26, 2019.
- [17] E. F. Patridge and T. P. Bardyn, “Research electronic data capture (REDCap),” *Journal of the Medical Library Association*, vol. 106, no. 1, pp. 142–144, 2018.
- [18] M. Ramsay, N. Crowther, E. Tambo et al., “H3Africa AWI-Gen Collaborative Centre: a resource to study the interplay

- between genomic and environmental risk factors for cardiometabolic diseases in four sub-Saharan African countries,” *Global Health, Epidemiology and Genomics*, vol. 1, Article ID e20, 2016.
- [19] D. J. Stein, T. Lehner, Z. Lombard, B. Pringle, G. Senthil, and M. Uddin, “Mental health delivery and neurogenetics discovery in Africa,” *The Lancet Psychiatry*, vol. 7, no. 6, pp. 473–474, 2020 Jun.
- [20] A. H. Wani, A. E. Aiello, G. S. Kim et al., “The impact of psychopathology, social adversity and stress-relevant DNA methylation on prospective risk for post-traumatic stress: a machine learning approach,” *Journal of Affective Disorders*, vol. 282, pp. 894–905, 2021 Mar 1.
- [21] A. Akpalu, F. S. Sarfo, B. Ovbiagele et al., “Phenotyping Stroke in sub-saharan Africa: Stroke investigative research and education network (SIREN) phenomics protocol,” *Neuroepidemiology*, vol. 45, no. 2, pp. 73–82, 2015.
- [22] G. Mboowa, S. Mwesigwa, E. Katagirya et al., “The Collaborative African Genomics Network (CAfGEN): applying Genomic technologies to probe host factors important to the progression of HIV and HIV-tuberculosis infection in sub-Saharan Africa,” *AAS Open Res*, vol. 1, p. 3, 2018 Jun 21.
- [23] A. Alelign, A. Zewude, T. Mohammed, S. Tolosa, G. Ameni, and B. Petros, “Molecular detection of *Mycobacterium tuberculosis* sensitivity to rifampicin and isoniazid in South Gondar Zone, northwest Ethiopia,” *BMC Infectious Diseases*, vol. 19, no. 1, p. 343, 2019 Apr 25.
- [24] M. O. Owolabi, O. M. Akpa, F. Made et al., “Data resource profile: cardiovascular H3Africa innovation resource (CHAIR),” *International Journal of Epidemiology*, vol. 48, no. 2, pp. 366–367g, 2019 Apr 1.
- [25] M. Mbiyavanga and N. Mulder, “CINECA synthetic cohort Africa H3ABioNet v1 (Version V1) [Data set],” *Zenodo*, 2021.
- [26] I. Fortier, P. Raina, E. R. Van den Heuvel et al., “Maelstrom Research guidelines for rigorous retrospective data harmonization,” *International Journal of Epidemiology*, vol. 46, no. 1, pp. 103–105, 2017.
- [27] J. Kumuthini, L. Zass, S. Panji et al., “The H3ABioNet helpdesk: an online bioinformatics resource, enhancing Africa’s capacity for genomics research,” *BMC Bioinformatics*, vol. 20, no. 1, p. 741, 2019 Dec 30.
- [28] I. Reinecke, M. Zoch, C. Reich, M. Sedlmayr, and F. Bathelt, “The usage of ohdsi OMOP—a scoping review,” *Studies in Health Technology and Informatics*, vol. 283, pp. 95–103, 2021.