# H3ABioNet Phenotype Standardisation: Project Documentation

## Table of Contents

Introduction
History of the H3Africa Phenotype Harmonisation Working Group
Project Methodology
Project Outputs
Concluding Remarks - Value & Benefit
Acknowledge/Cite Us
Contact Us
Contributors

# Introduction

Recent advances in genomics studies have raised a need for matching large-scale phenotypic data which incorporate social, environmental and clinical factors together with genetic information. With increased data generation worldwide, increasing statistical power, data integration and meta-analyses have become central components of all genetic studies. Meaningful analysis of multiple small heterogeneous phenotypic datasets is a challenge, and at times, impossible, without standardising the data elements prior to collection or retrospectively harmonising the collected phenotypic data. To encourage harmonisation and data sharing in biomedical research and to increase the scientific impact thereof, study CRFs should use common measures and terminology; and promote the collection of high-quality standardised data to facilitate cross-study analysis.

H3ABioNet, a Pan African Bioinformatics Network which forms part of the Human Heredity and Health in Africa (H3Africa) consortium, was established to develop bioinformatics capacity in Africa and to enable genomics data analysis by H3Africa researchers across the continent. The H3ABioNet Phenotype Standardisation Project originated from the desire to standardise the collected phenotype and exposure information across the H3Africa consortium and within research-specific domains. The project aims to raise awareness and adoption of data collection standards within the African biomedical and genomics research communities, and to facilitate the harmonisation of the dissimilarly collected H3Africa data.

In collaboration with the H3Africa Phenotype Harmonisation Working Group (PHWG), a set of essential phenotype data elements that are applicable to all projects in the consortium were identified and referred to as the Core Phenotypes. To ensure collection of all relevant information for the core phenotypes, a standardised data collection instrument known as the H3Africa Standard Case Report Form (CRF) was developed. As the project progressed, the core phenotypes were expanded to cover domain-specific modules with field-specific clinical data elements relevant to African biomedical and genomics research. This documentation provides a brief history of the project's origins and initial developments, describes the methods by which the core phenotypes and domain-specific modules were developed, and also summarises the outputs developed from the project.

# History of the H3Africa Phenotype Harmonisation Working Group

H3Africa aims to improve human health in Africa by i) facilitating genomics research which prioritises diseases relevant to the continent, such as cardiovascular and infectious diseases, ii) building capacity in genomic research expertise on the African continent, and iii) promoting data sharing. Within the consortium, there was an initial need to streamline the consortium information capture by H3Africa Biorepositories. However, the phenotypic elements were not readily available, and out of the need to facilitate the standardisation and sharing of phenotype data within the H3Africa consortium and beyond, the Phenotype Harmonization Working Group (PHWG) was born. The work conducted by this PHWG can be subdivided into three phases, detailed in this document.

In the first phase, the PHWG members agreed upon a set of 24 Core Phenotypes to be standardised during H3Africa data collection. These Core Phenotypes included general demographics, anthropometrics, smoking status, alcohol and drug use and more. They were selected as they were highly represented in the initial H3Africa projects CRF drafts, and standardised using selected **PhenX**[1] protocols. PhenX protocols were used because they are standardised and facilitate cross-study analysis. These Core Phenotypes formed the initial primary foundation for data sharing across the various H3Africa studies, allowing for studies to incorporate any additional phenotypes as needed by their study objectives. Following this agreement, a meeting was organized to ensure that the methods of collection of the Core Phenotypes were suitable for the African context.

By the time that the Core Phenotype protocols were finalised, however, some projects had already started data collection, thus the PHWG anticipated the need to develop methods by which to harmonise dissimilarly collected data as well. The PHWG worked alongside other H3Africa WGs (e.g. CVD WG) throughout to enable such harmonisation. In the second phase, the PHWG identified issues in implementing the Core Phenotypes protocols in practice, and reworked these protocols into the H3Africa Standard CRF. Additionally, in an effort to move towards using **REDCap**[2] for phenotype data collection, H3ABioNet created standardised REDCap data dictionary templates for the Core Phenotypes with consideration for both digital and paper-based data collection. REDCap was recommended for use because it is cost-effective; allows for secure data collection both online and offline; and allows for the easy dissemination and sharing of standardised database templates. A set of CRFs for paper-based data collection (which is still heavily relied upon in the African continent), were designed in Microsoft Publisher alongside the REDCap data collection templates for the H3Africa Core Phenotypes and made

[1] The PhenX Toolkit (consensus measures for Phenotypes and eXposures) is a NIH-funded effort which provides recommended standard data collection protocols for conducting biomedical research. The protocols are selected by Working Groups of domain experts using a consensus process. Using protocols from PhenX facilitates data harmonisation and cross-study analysis.

[2] REDCap (Research Electronic Data Capture) is a secure web application for building and managing online surveys and clinical or translational research databases.

available publicly on the H3ABioNet website. To facilitate implementation, the PHWG developed recommendations for CRF design and for using the H3Africa Standard CRF locally.

In the final phase, the PHWG recognised a need to expand the Core Phenotypes to cover additional domain-specific phenotypes. The initial request for expansion came from projects with pediatric participants as the Standard CRF was insufficient and not comprehensive for use in pediatric research studies. To remedy this, the Core Phenotypes were revisited from a pediatric perspective, and a pediatric alternative was developed. An additional expansion was implemented by the H3ABioNet Minimum Data Dictionaries project, who identified a need for specific harmonisation between groups of projects within specific research domains. The Minimum Data Dictionaries project subsequently developed modules (collections of existing standard protocols) for the stroke and kidney disease research domains. These parallel efforts were later combined and streamlined in the Phenotype Standardisation Project, expanding the coverage of domains as illustrated in **Figure 1**. These modules can be used alongside the Core Phenotypes (adult OR pediatric), allowing comprehensive standardised data collection for any research study. In some cases, existing standard protocols were not sufficient for implementation in the African context, requiring adjustments, or entirely new protocols to be developed.

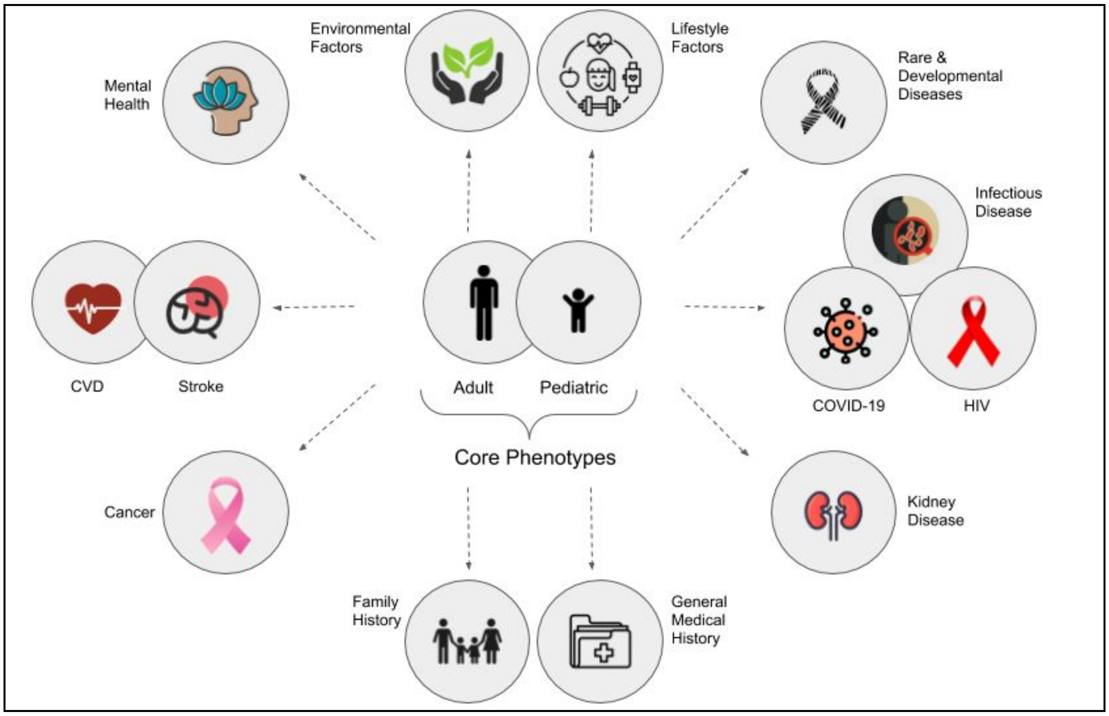

**Figure 1.** Modules developed by the Phenotype Standards Project.

## Project Methodology

Although the H3Africa Core Phenotypes and the H3ABioNet Phenotype modules were developed on separate occasions, they share a common development process, summarised in **Figure 2** and further discussed below.

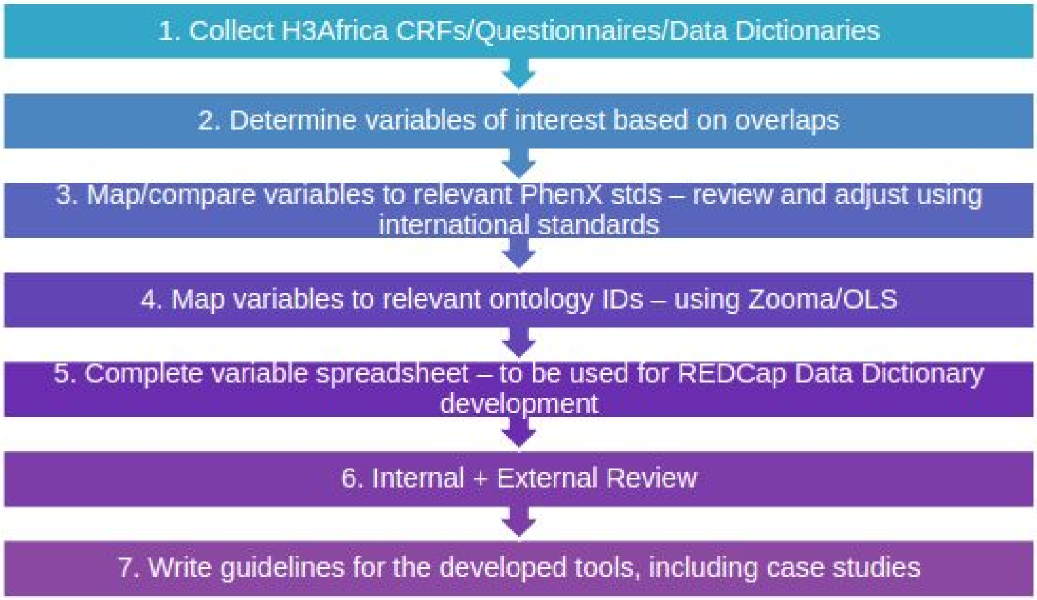
1. Collect H3Africa CRFs/Questionnaires/Data Dictionaries
2. Determine variables of interest based on overlaps
3. Map/compare variables to relevant PhenX stds – review and adjust using international standards
4. Map variables to relevant ontology IDs – using Zooma/OLS
5. Complete variable spreadsheet – to be used for REDCap Data Dictionary development
6. Internal + External Review
7. Write guidelines for the developed tools, including case studies

**Figure 2:** Harmonised workflow for modules development.

To begin, a collection of H3Africa CRFs, Data Dictionaries and Questionnaires were compiled along with publicly available African data collection forms in order to identify the common phenotypes collected within a specific domain of research. These phenotypes were identified as essential to be collected within a given research domain module. A module is defined as a collection of standard protocols for the collection of domain phenotypes. Therefore, once the domain phenotypes were identified, we explored the internet for existing publicly available data collection standards associated with the phenotype, with a particular focus on standards hosted on PhenX. If a particular standard did not exist for a given phenotype, new data collection standards were developed from scratch. These are developed based on domain knowledge and the originally compiled African data collection forms. If a standard does exist, the standard is reviewed in order to determine whether it is appropriate for implementation in Africa, and subsequently adapted if needed. A comprehensive spreadsheet was created to compile the data collection standards for all the phenotypes within a given module. Once all variables were determined, these variables were mapped to existing ontology IDs using Ontology Lookup Service (OLS) and Zooma. Specific and well-maintained ontologies were prioritised during this mapping.

Next, the modules were reviewed both externally and internally. For external review, a survey was developed and distributed to domain experts within Africa. The survey allowed respondents to comment on the applicability of the suggested phenotypes and the associated data collection standards within a specific domain, and whether these standards can be appropriately used in Africa. Respondents could also suggest the inclusion of additional phenotypes within a given module. During the internal review, the final format of the module is reviewed by a data manager to identify any additional issues with regards to data flow and capture. Once reviews are completed, the associated modules are created, consisting of a data dictionary, REDCap Template, CRF and implementation guideline.

# Project Outputs

The project's primary outputs are domain-specific modules relevant for use in Africa (as previously illustrated in **Figure 1**. Illustrated in **Figure 2**, each module consists of a paper-based CRF (.pdf), a data dictionary (.xlsx), a project template (.xml) and an associated implementation guideline (.pdf).

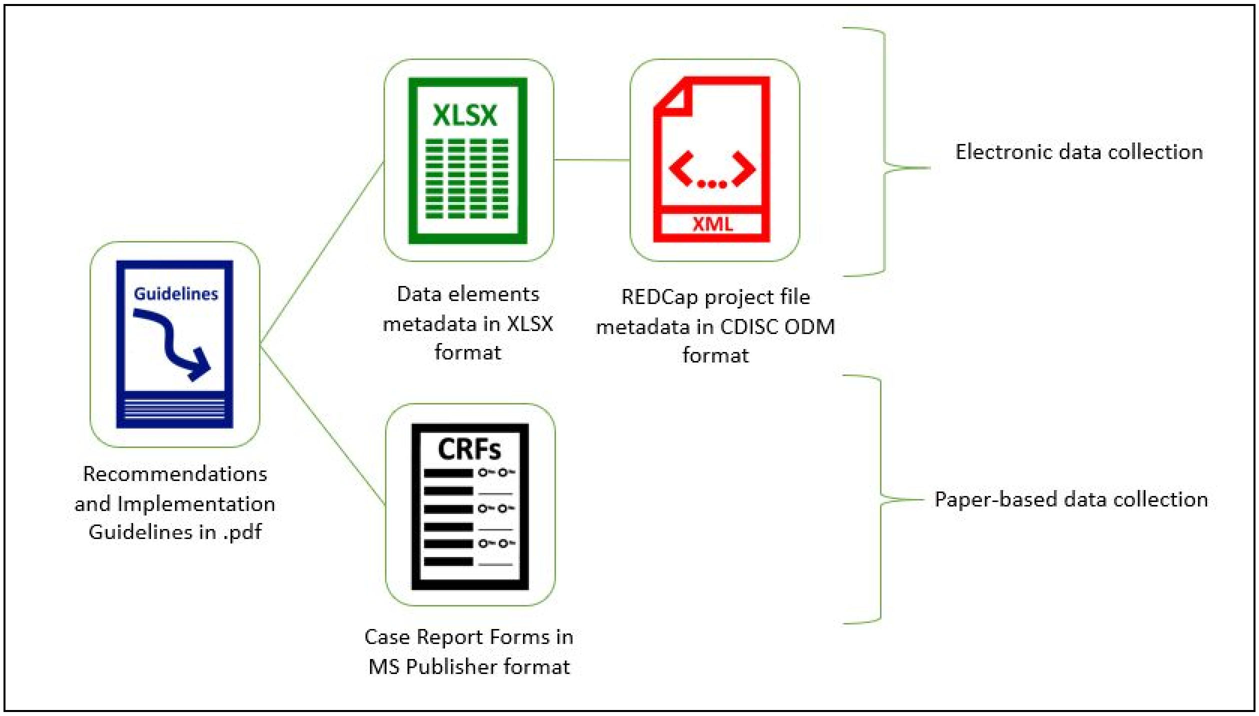
XLSX
XML
Electronic data collection
Guidelines
Data elements metadata in XLSX format
REDCap project file metadata in CDISC ODM format
CRFs
Recommendations and Implementation Guidelines in .pdf
Paper-based data collection
Case Report Forms in MS Publisher format

**Figure 2.** Module components and use.

H3ABioNet aims to share these modules on multiple online platforms, including the H3ABioNet Website (https://www.h3abionet.org/data-standards/datastds), GitHub (https://github.com/h3abionet), the REDCap Shared Library and FAIRSharing (https://fairsharing.org/). In addition, H3ABioNet aims to feedback all newly developed and(or) adapted protocols to PhenX, as well as gain endorsement from genomic research bodies in order to promote wide-scale adoption and implementation in Africa. H3ABioNet welcomes feedback and recommendations on the developed standards as well, such feedback can be provided via the project issue tracker on the H3ABioNet GitHub. The project will also have additional secondary outputs, including technical guides and published manuscripts.

**Table 1** summarises the modules that have been developed or are currently being developed and their associated release status with version.

**Table1:** The H3ABioNet Phenotype Data Collection Modules

| **Phenotype Standards** | **Version** | **Release Date** |
|---|---|---|
| Core Phenotypes - Adult | 1.0 | Sept 2018 |
| Core Phenotypes - Pediatrics | In Development | |
| Cancer | In Development | |
| Cardiovascular Diseases | In Development | |
| COVID | In Development | |
| Environmental Exposures | In Development | |
| Family History | In Development | |
| General Medical History | In Development | |
| HIV | In Development | |
| Infectious Diseases | In Development | |
| Kidney Diseases | 1.0 | Nov 2019 |
| Lifestyle Factors | In Development | |
| Mental Health | In Development | |
| Rare & Developmental Diseases | In Development | |

| | | |
|---|---|---|
| Stroke | 1.0 | Jun 2019 |

## Concluding Remarks - Value & Benefit

To encourage phenotype harmonisation and data sharing in biomedical research and to increase the scientific impact thereof, studies should employ standard data collection protocols; and promote the collection of high-quality standardised data to facilitate cross-study analysis.

The H3ABioNet Phenotype Standardisation project has established a rich and comprehensive suite of domain-specific data collection modules (collection of data collection standards) which are compatible and can be used for standardised and comprehensive phenotype data collection in any given biomedical or genomic research project in Africa.

The **benefits** of using these modules are multi-fold:

1. The modules are all aligned with existing, global data collection standards.
2. The modules are specifically adapted for biomedical phenotype data collection in Africa.
3. The modules are highly interoperable, with variables mapped to existing and maintained ontologies.

## Acknowledge/Cite Us

We would like to request that users of the project outputs, acknowledge our work in their related publications. You can do so, by either citing the module DOI (TBA) or related publication (TBA). The templates below are recommended:

1. In methods section of publications:
   “The data collection modules (DOI) used were developed by H3ABioNet/H3Africa and were compiled using a variety of resources described in “standards github link.“

2. In acknowledgements section:
   “We acknowledge the use of data collection modules (DOI) from H3ABioNet/H3Africa which can be found, including associated references, at https://github.com/h3abionet, funded by the H3Africa NIH grant U24HG006941.”

# Contact Us

For more information on how to use the modules developed by the H3ABioNet Phenotype Standardisation Project, please consult the H3ABioNET Helpdesk, and log a query/ticket in the Phenotype Standardisation Modules queue.

# Contributors

Many thanks to all our contributors!

- H3Africa Phenotype Harmonisation WG
- H3Africa Principal Investigators, Study Coordinators & Data Managers
- H3Africa CVD WG
- H3Africa Mental Health WG
- H3Africa Rare Diseases WG
- H3ABioNet Standard CRF Project Team
- H3ABioNet Minimum Data Dictionaries Project Team

A comprehensive list of planners and developers to be added in the near future.