The SickleInAfrica Consortium - Promoting Data Sharing and Collaborative Sickle Cell Disease Research for Existing and Prospective Studies Using Ontology-Driven Data Standards and Tools

Target Journal - Blood Advances

Ambroise Wonkam, Gaston Mazandu, Raphael Z. Sangeda, Jade Hotchkiss, Annemie Stewart, Victoria Nembaware, Mario Jonas, Nchangwi S. Munung, Khuthala Mnika, Valentina Ngo Bitongui, Katherine Johnston, Daniel Ansong, Daniel Kandonga, Evans Xorse Amuzu, Isaac Nyanor, Isaac Olanrewaju, Reuben Chianumba, Vivian Painstil, Obiageli E. Nnodu, Kofi A. Anie, Julie Makani, Nicola Mulder; *Sickle Cell Disease Ontology Working Group and SickleIn Africa Consortium

*List of contributing members is in the appendix.

Summary

Background: Sickle Cell Disease (SCD) is a growing global health concern, highlighting an urgent need for data-sharing and collaborative international research. However, SCD research is largely conducted *in silos*, compromising the depth required for effective data-driven public health decisions. SickleInAfrica is addressing these needs by developing ontology-driven (SCDO) standardized data elements, standard operating procedures (SOPs) and processes for improved data quality and tools for data sharing and harmonization of existing and prospective data. These resources are developed collaboratively by an international group of SCD stakeholders.

Methods: SickleInAfrica data elements were created iteratively by integrating measures and data elements from multiple resources. The data elements were developed in Research Electronic Data Capture (REDCap) and mapped to SCDO top-level classes. An SCDO-based meta-data platform was developed to capture data from existing and ongoing studies, and to support new members joining the SCDO working group (SCDO-WG). An SCDO-based research database schema was designed.

Results: A total of 27 instruments with 1,514 data elements were developed from 2,580 pooled data elements and are accessible via online platforms including (www.sickleinafrica.org). To our knowledge, the associated core data elements mapped to SCDO terms are the most comprehensive and globally developed SCD data elements. Standard operating procedures (SOPs); and tools for harmonizing data are presented, as well as a database schema which uses the SCDO to cover SCD concepts in a consistent and unambiguous way. We highlight an online interactive SCD metadata platform and a membership registration platform for people who want to be part of the SCDO-WG.

Conclusion: The SCDO-WG, through SickleInAfrica, promotes data quality, data sharing and potential collaborative SCD research through community-developed tools, data standards, database schemas and a meta-data platform.

Introduction

Sickle cell disease (SCD), one of the most common monogenic diseases in humans (Rees et al, 2010), is caused by a single amino acid substitution in the beta-subunit of haemoglobin - the principal oxygen transporter in red blood cells (Piel et al., 2017). This seemingly simple Mendelian change has severe clinical consequences characterized by complex combinations of phenotypes involving multiple organ systems. The manifestations and management of SCD are further impacted by environmental, socio-economic and cultural factors, which, in turn, affect clinical management and patients' quality of life. SCD has its highest incidence and prevalence rates in sub-Saharan African (SSA) countries, in countries with populations of African descent (Diallo and Guindo, 2014) and in some Asian countries. SCD research data is currently collected in an *ad hoc* and uncoordinated manner, in project silos and mostly from healthcare and research facilities that are geographically dispersed, with differing availability of resources and distinct data capture systems. These factors limit the consistency and broad applicability of the data collected. Complex components of the disease require capturing information in a standardized and consistent manner to facilitate data sharing and cross-study comparisons and collaborations. The SCD community has long recognised the need to develop a standardized database and related tools to negate challenges such as under-powering of studies, data incompatibility, and irreproducibility within and across sites (Marques et al., 2015; Eckman et al., 2017; Sears et al., 2017; DiMartino et al., 2018).

Recent advances in the application of artificial intelligence (AI) and other information technologies in Big Data analytics have benefited from the use of ontology models to represent knowledge- and information-based systems (Martinez-Cruz et al., 2011; Kuiler, 2014). An ontology is useful in establishing a common and controlled vocabulary system, describing key concepts, properties, and hierarchical relationships between terms (Gruber, 1993), with precise definitions, for clear and unambiguous communication. An ontology can enable knowledge acquisition, sharing, reuse, verification and validation of data. This makes the application of a well-defined ontology within the conceptualized domain an important endeavour for data harmonization, quality and assurance processes. We developed the SCD ontology (SCDO) in a community effort involving ontologists, clinicians and researchers from a diverse cross-section of geographical locations (Mulder et al., 2016; Sickle Cell Disease Ontology Working Group, 2019). The SCDO includes 1,476 standard terms, capturing multiple aspects of the disease with a “Hemoglobinopathy” central class which is linked to phenotypes, genetic and other disease modifiers, therapeutics, diagnostics, quality of life and care and some other aspects (Sickle Cell Disease Ontology Working Group, 2019). The SCDO represents the most comprehensive compilation of knowledge of SCD and other hemoglobinopathies to date.

A comprehensive disease-specific ontology such as the SCDO can model relational databases of integrated clinical, psycho-social and environmental information about research participants. Such modelling eases the implementation of clinical and research databases as the SCDO substantially covers concepts used in the database in a consistent and unambiguous way. Ultimately, ontologically-enriched methods for classifying and stratifying research participants may enable seamless data harmonization and integration (Hoehndorf et al., 2015) across different studies and for multi-site collaborative research initiatives with different data element structures and coding systems. While multi-site SCD research initiatives such as the Globin Regional Network for Data and Discovery (GRNDaD), the SCD consensus measures for Phenotypes and Exposures (PhenX) toolkits and the sickle cell disease implementation consortium exist, they lack ontology frameworks to guide the design of data elements, related tools and databases (Sears et al., 2017; Eckman et al., 2017; Glassberg et al., 2020). Furthermore, they are limited in scope and global applicability as both these SCD resources have limited input from SSA countries where SCD is most prevalent.

To address the limited coverage of potential SCD research areas and poor international representation in the development of publicly available SCD measures and data elements, the SCDO-WG presents a comprehensive warehouse of SickleInAfrica SCD data collection instruments and elements for research and clinical data collection, capturing and databasing. SickleInAfrica (Makani et al., 2020) consists of three initiatives, namely Sickle Pan-African Research Consortium (SPARCo), Sickle Cell Pan-African Network (SPAN), and Sickle Africa Data Coordinating Center (SADaCC) (Makani et al., 2017). SPARCo is a three-country consortium (Ghana, Nigeria, Tanzania), coordinated from a hub in Tanzania. It aims to develop research capacity for SCD in Africa through a multidimensional approach, which addresses infrastructure, education, training, provision of longitudinal research data and the translation of research into practices. A key deliverable for SPARCo is the SCD registry. SPAN is a network involving multiple sites in 17 countries, consisting of researchers, clinicians, funders and centers working on SCD, which aims to foster research and infrastructure development. SADaCC is the administrative, data standardization and coordinating center, whose core mandate in collaboration with the SPARCo Hub (Makani et al., 2020) is to support ethical data acquisition, storage, management and analyses, as well as coordinating communications for SPARCo and other SCD stakeholders. SADaCC also coordinates the SCDO-WG (http://scdontology.h3abionet.org/) (Mulder et al., 2016) and plays a major role in defining the ethical sharing of data in SickleInAfrica (Munung et al., 2019).

This report describes the SickleInAfrica data elements and their creation using an iterative process of integrating case report forms (CRFs) and measures from multiple sources, sites, and countries, including resource-limited settings. This was a result of collaborative efforts by over 40 multi-disciplinary experts from four continents (Africa, North America, Europe and South America) and 14 countries. These data elements were mapped to SCDO concepts and developed into a Research Electronic Data Capture (REDCap) template (Harris et al., 2009; 2019). Furthermore, we present a SickleInAfrica metadata platform for SCD stakeholders to register as SCDO-WG members and to share metadata from existing, ongoing, and future research. This metadata platform is searchable and viewable as a table or via an interactive global map.

Materials and Methods

SCDO-based framework for designing data elements

Figure 1 summarizes the SCDO-based framework used to guide the development of data elements, data capturing processes, and database design; all geared at formalizing and easing the management of SCD-related datasets from research and clinical records in a scalable manner. We considered four key steps in designing this framework:

Step 1: This involved collecting data elements from numerous sources, merging these, and using them to create a non-redundant SickleInAfrica database of coreSCD data elements. The community agreed-upon data elements are beingmapped to the standardized terminology within the SCDO.

Step 2: Involved developing SCDO-guided tools to support data flow paths into the SickleInAfrica database for prospective research and existing data from participating sites. The SickleInAfrica data elements are publicly available via numerous platforms at no cost. However, existing data from distinct sites were largely incomparable as research-specific structured data elements could not be coded, or if they were, it would be in a site-specific manner. An SCDO-based tools and processes

SOP was developed to facilitate SickleInAfrica data harmonization.

Step 3: In the database application, REDCap tables or data collection instruments or measures were mapped to upper-level concepts close to the root of the SCDO. Such a database structure of storing clinical or research participants as instances, is necessary as the number of participants may increase significantly.

Step 4: This step involved creating a web-based platform to allow new SCDO-WG members to register. In addition, there is a form for researchers working in SCD and other hemoglobinopathies to upload and search research metadata from past, ongoing, and future research.

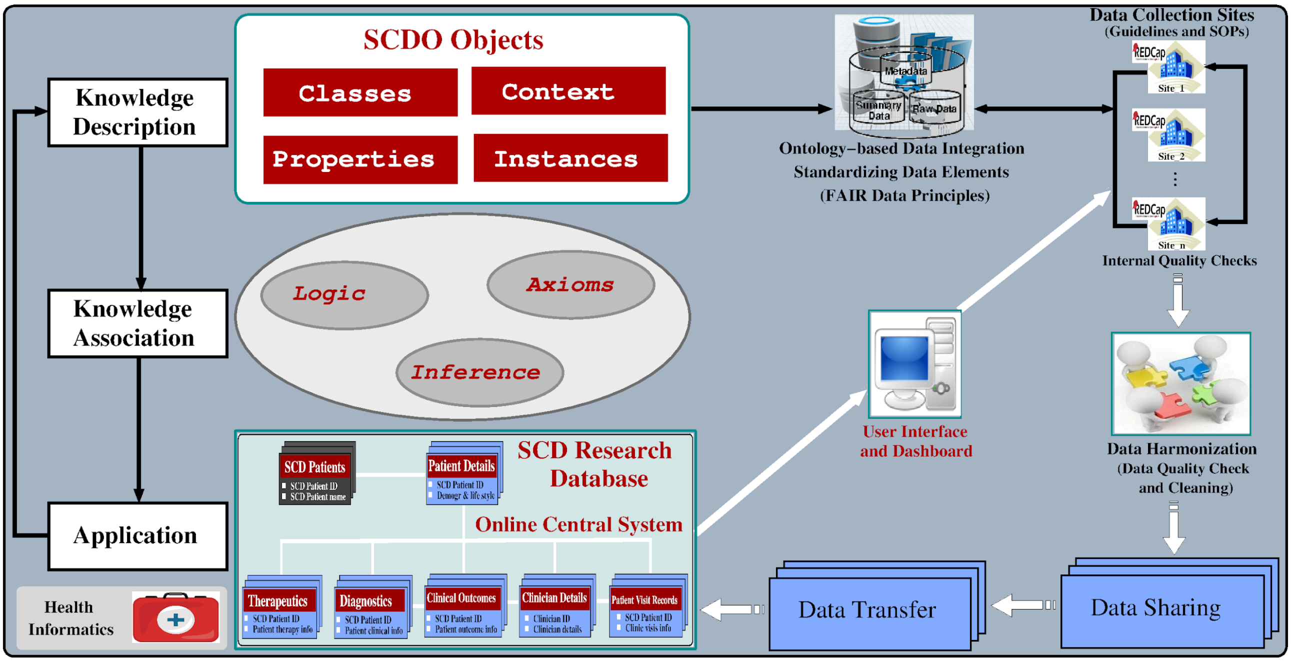


Figure 1: A summary of how the SCDO framework is being applied for data collection (through clinical systems, self-report and other means), standardization and databasing tools for SCD research.

Development of standardized measures and SCD data elements

An initial set of 2,580 SCD data elements was created by combining SCD measures and CRFs obtained from publicly available sources and SickleInAfrica members. Measures for Smoking, Alcohol and Lifetime Drug Use were obtained from the Human Heredity and Health in Africa (H3Africa) Consortium (https://h3africa.org/). These data elements are a small set of the H3Africa standardised minimal set of phenotypic data elements which were identified as commonly collected in research projects for both patient health care and genomic research (https://www.h3abionet.org/data-standards/datastds). PhenX: Sickle Cell Tier 1 and 2 measures were downloaded from this toolkit (https://www.phenxtoolkit.org/). Additional data elements were obtained from the three SPARCo sites: Ghana, Tanzania and Nigeria, and a SickleInAfrica member from Cameroon.

Non-redundant and universal questions/measures were retained based on an iterative reviewing process during face-to-face SCDO-WG meetings (Nembaware et al., 2020), SickleInAfrica consortium meetings and SADaCC-supported data management training workshops. Data elements and measures were chosen based on SCD expert and researcher reviews on their relevance, and ability to address scientific questions related to SCD. Texts describing PhenX protocols were moved to separate SOP documents and guidelines, retaining the PhenX protocol ID in order to maintain links between SOPs, guidelines and the relevant data elements. To obtain a minimal set of data elements for the SickleInAfrica registry, existing CRFs from the three SPARCo sites (Munung et al., 2019; Nembaware et al., 2020; Makani et al., 2020), listed in **Table 1**, were analysed and merged, avoiding redundancy and maintaining high coverage.

Table 1: Summary of measures from SPARCo sites.

Consortium sites	CRF Features
Ghana	● Record of Routine Visit ● Record of First Visit ● Record of Acute Care Visit ● Summary of Hospitalization
Nigeria	● Blood Transfusion ● SCD Test ● Problems/Vital signs ● Medications
Tanzania	● Follow-up ● Entry, Control ● Acute ● Inpatient ● SCD Passport

During a data management training workshop held in March 2018, SickleInAfrica members evaluated the data elements based on the following criteria: 1) Present in at least one SickleInAfrica site CRF, 2) Essential data element for SPARCo, 3) May be required in other SCD studies, 4) Remove: not required in SCD studies, 5) Cannot say, i.e., a case where the group or individual did not have an opinion about the inclusion/exclusion or the importance of a data element. Elements were removed if two of the sites agreed to the removal and the third site had "Cannot Say" or "Maybe" as an option. Elements were also removed if one site selected removal and the other two sites selected "Maybe". In addition, these core SickleInAfrica data elements were mapped to the EQ-5D and EQ-5D-Y measures (Balestroni and Bertolotti, 2012), where possible, after obtaining the necessary permissions for these measures. These measures can be collected at minimal or no cost to the patients and data collectors. A resulting consolidated SCD REDCap data dictionary from the activities described above was then mapped to broad/upper level

SCDO concepts.

Harmonization of existing multi-site research data elements

To facilitate retrospective data element harmonization from multiple sites, a Data Migration SOP was developed which is centered around the SCDO-mapped SickleInAfrica data elements (see Supplementary File). This SOP required manual mapping of study-specific data elements to the SickleInAfrica data elements. After the manual mapping, an in-house python code script was developed (Nembaware et al., 2020) to produce the following files for each site/study data elements: (1) The new harmonized SCD registry, (2) The new site SCD-harmonized data dictionary, (3) The standardized SCD and site variable map file and (4) The site and standardised variable mapping annotation file providing site variables mapped or not, with the harmonized variable name if mapped.

Joining the SCDO Working Group and/or Populating the SickleInAfrica metadata platform

An ontology-based meta-data platform was created to promote registration of members into the SCDO-WG and to indicate their interest in sharing data from existing, ongoing or upcoming data collection by uploading key characteristics of their studies using an SCDO s form accessible via https://www.sickleinafrica.org/registries-list. Once accepted, members are required to make ethical and legal documents/policies available to SickleInAfrica. A preliminary scan of consents, data access, and IP policies is conducted to assess the potential for each study to share the actual data as done previously (Munung et al., 2019).

Results

SickleInAfrica instruments and data elements

We designed a comprehensive set of 27 data collection instruments for SCD, to capture SCDO concepts like disease outcomes and associated clinical factors, including phenotypes, diagnostics and treatment practices necessary for SCD research and for enhancing clinical practice and patients' management. This work integrated 1,476 data elements from three different sources to reflect a wide range of research needs. A multi-disciplinary team of SCD experts, researchers and clinicians checked the relevance of these variables during the two data workshops. It is expected that the number of data elements will decrease even further in revised versions. The proposed data collection instruments are summarized in **Table 2** and a more complete list of instruments and variables (date elements) is accessible at https://sadacc.org/SIA_data_elements. The proposed

SickleInAfrica data elements are intended to form an extensive and dynamic data collection instrument and, as such, contain many more data elements than what is currently captured by the SickleInAfrica consortium for their registry, the SickleInAfrica Core Instrument.

Table 2: Summary of some proposed SickleInAfrica SCD data instruments in the REDCap database.

Data Collection Instrument	**Description and Features**
SickleInAfrica Core Instrument	Captures the minimal/mandatory set of data elements determined by SickleInAfrica to be sufficient for addressing their current research questions. Information captured includes basic demographics, diagnosis details, current medication management, consent details and selected laboratory results. An optional set of measures to assess the mental health of patients is also added as part of these core data elements.
Optional demographics	This section captures additional demographics. This includes region or province of residence of the patient and parents, the type of study the patient was enrolled in, the date registration was done, the age at which the patient was first seen under the protocol, gender, family income, marital status, relationship of primary caregiver to patient, consanguinity between parents and grandparents, and employment status. For income, an item was added to define the currency of the country of residence as well as items stratifying different income levels based on the poverty scale aligned to PhenX measures. Due to the complexity and potential ethnicity based conflicts of definitions, discussions on how to best expand the existing proposed ethnic categories are still ongoing. For “education level”, the education system promised by PhenX measures was adapted.
Anthropometrics	Captures general anthropometric measures, such as height and weight.
Medical History	Has the following sub-classes: 1. Pain episodes (6 and 12 month): The number of painful crises varies between patients hence this instrument captures, among others, event-date information as well as hospital treatment information. The simplicity of this instrument allows it to become a repetitive instrument for a varying number of painful crises. 2. Stroke: The instrument asks an explicit question on whether a patient has had a stroke confirmed by a physician. Because of the difficulties patients, guardians or even health professionals have in diagnosing stroke, this instrument furthermore captures finely grained information to establish whether a stroke occurred in the absence of medically proven stroke. 3. Arterial and Venous Disease. 4. Kidney Disease. 5. Epilepsy/Seizures. 6. Transfusion. 7. SCD Complications.
Blood Pressure	This collects information about hypertension; initial age of hypertension diagnosis and related medication; and collects a blood pressure measurement.
Alcohol Use	This contains variables that describe lifetime alcohol consumption, age of first use of alcohol, 30-day quantification and frequency of use of alcohol and maximum drinks taken in a single day.
Smoking Status	This instrument captures variables that include the smoking status during one’s entire life; the frequency of smoking, the number of cigarettes smoked per day, the age of initiation, smoking regularity, the age of offset (if any), the type of tobacco, the number of cigarettes smoked over the last 30 days, and how many pockets per day each year.
Lifetime Drug Use	This includes a list of substances consumed during the last 30 days, the age of first

	substance (sedatives, tranquilizers, painkillers, stimulants, marijuana, cocaine, crack cocaine, hallucinogens, inhalants or solvents, heroin, methamphetamines, non-prescribed medications) use and their frequencies of use during the last 30 days, also substance abuse and dependence during the last 12 months.
Laboratory Results	This explicitly includes the following components: 1. Complete Blood Count. 2. Kidney Function Assay. 3. Liver Function Assay, Renal Function Tests, Reticulocyte count, Peripheral film, Extended phenotyping for transfusion products and Iron studies. 4. Additional laboratory results.
Laboratory Equipment Used	This includes diagnostic instruments, laboratory assay or tests, or method of patient examination used for differential diagnosis.
Therapeutics	This subdivides into the following sub-instruments: 1. Reception variables include medications patients have brought to their appointment: "are these all the medications that you have taken in the past two weeks?" 2. Prescription Medication variables include prescribed medication (name of the drug), strength, record the units of strength, number prescribed, circle cycle (day, week, month), Pro Re Nata (PRN) Medicine (when necessary), on average in the last two weeks how many pills did the patient take in a day/week/month (circle: day/week/month), number unable to transcribe. 3. Over the Counter Medication variables include (a) Over-the-Counter Medications: name of the drug, strength (mg, IU, etc), units of strength, number prescribed, circle (day/week/month). (b) PRN Medicine–"how many pills did you take per day/week/month, on average during the last two weeks"? circle (day/ week/month), number unable to transcribe; comments about medications. 4. Therapeutics: Past and Present Pain Medications variables include the following information: (a) Do you currently take pain-relieving medication regularly (at least once a week)? (total tablets per week), "Baby" or low-dose aspirin; Aspirin or Aspirin-containing products; Ibuprofen; Naproxen; Ketoprofen or other non-steroidal; Cox-2 inhibitor; Acetaminophen. (b) Did you stop the regular use of any of the following medications during the past 3 years? Why did you stop regular use? ("Baby" or low-dose Aspirin, Aspirin or Aspirin-containing product/ Ibuprofen/ Naproxen, Ketoprofen or other non-steroidal/ Cox-2 inhibitor/ Acetaminophen). (c) In the past 3 years please indicate if you have taken either of the following: Stain medications (Lovastatin, Atorvastatin, Rosuvastatin, Pravastatin, Simvastatin, Fluvastatin), Steroid medication in pill form such as Prednisone, Dexamethasone, Solumedrol.
Quality of Life and Care	This captures variables which include general well-being, state of health (excellent, good, fair, poor, not sure). It also includes variables to determine physical and mental health that are specifically divided into three groups: adult, adolescent and paediatric. Social functioning impact, sleep impact, stiffness impact, pain impact, pain episodes, self-efficacy, stroke impact scale, age of initiation of first cigarette use, substance abuse problems (alcohol, drugs and tobacco in adults and adolescents), age of initiation of use (alcohol, drugs and tobacco), migraine, paediatric school performance, sleep apnea (adult protocol), sleep apnea (child protocol).
Phenotype	This instrument includes observable features of SCD patients, such as signs, symptoms and other possible disease-associated clinical manifestations.
Gross Motor Skills - Child	This instrument contains general physical actions, ranging from self-care (activities of daily living) to more complex activities that require a combination of skills, often within a social context.

The SickleInAfrica Core or minimal set of data elements, determined by SickleInAfrica for their registry, captures basic demographics, diagnosis details, current medication management, consent details and selected laboratory results, which are mapped to the EQ measures, where applicable. Data elements related to ethnicity caused a heated debate due to the complex ethnic-based conflicts and genocides that have plagued some regions. The group prioritized inclusion of ethnic groups for Nigeria, Ghana and Tanzania as these are SPARCo sites; Cameroon and South Africa are key sites for SADaCC. The proposed ethnic groups are an expansion of data from PhenX, the Demographic and Health Survey Program (DHS - https://dhsprogram.com/) and other sources. Nigeria had the highest number of ethnic groups with over 300 ethnic groups retrieved from DHS. Discussions are still ongoing about including additional ethnicities. Finally, several data elements are collected under different instruments, as highlighted in **Table 2**, including Demographics, Anthropometrics, Medical History, Lifestyle including Alcohol Use and Smoking Status, Lifetime Drug Use, Laboratory Results, Therapeutics, Quality of life and SCD Phenotypes.

Assessing the relevance of different variables

Different retained variables were subjected to SCD expert and researcher judgement during the 3rd SCDO meeting held in June 2018, to check their relevance and ability to address scientific questions related to SCD. As an illustrative example, members of the Quality of Life and Care working group identified 587 data elements that captured potentially duplicated information. This refinement process decreased the initial set of data elements (variables) to 1,476 (see **Figure 2**).

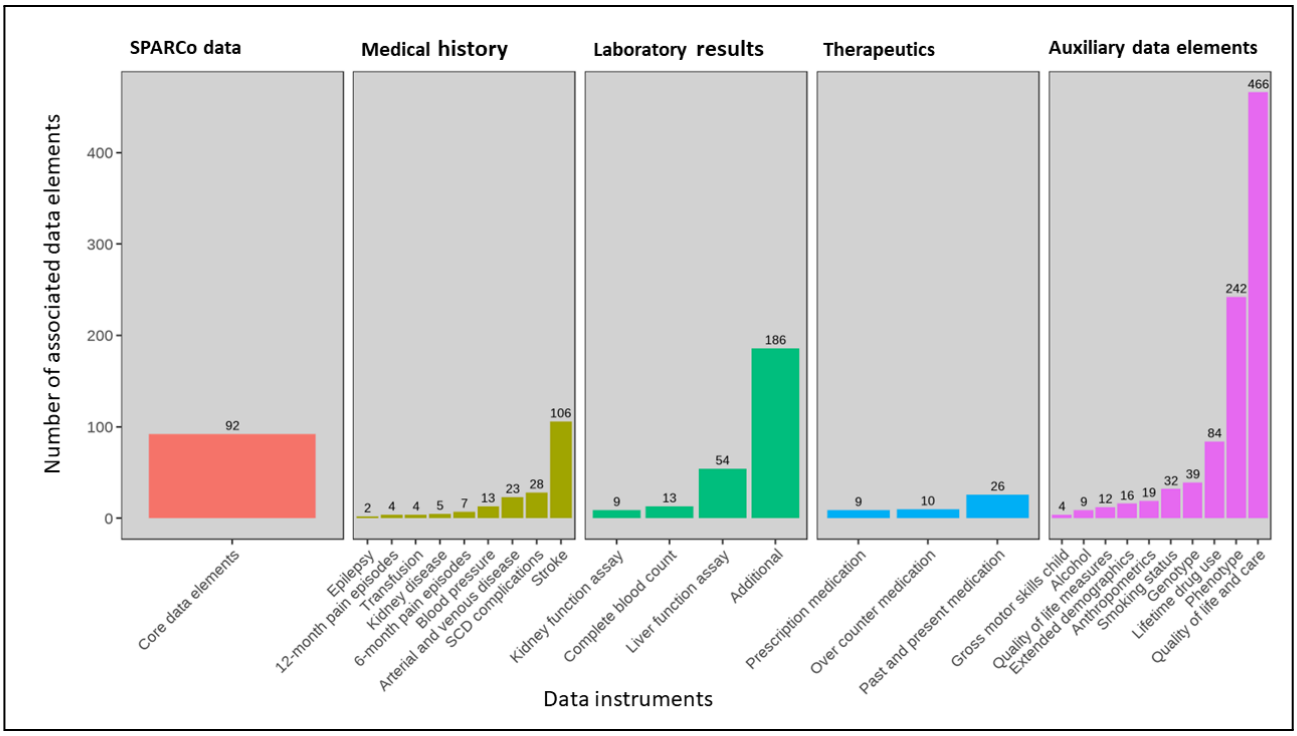


Figure 2: Different REDCap data instruments with the associated number of data elements.

Mapping data elements to SCDO identifiers

The SCDO (https://bioportal.bioontology.org/ontologies/SCDO) provides consistent and controlled vocabularies (Bodenreider, 2005) and describes key concepts and properties (Mazandu and Mulder, 2011; Mazandu et al, 2018) that establish hierarchical relationships between concepts and axioms (evidence or truths) associated with SCD. These data are collated into a human- and machine-readable format in order to help process, reuse and re-apply knowledge in biomedical research and healthcare systems. The SCDO builds systems around the key component (hemoglobinopathy), linking it to phenotypes, therapeutics, diagnostics, disease modifiers, and other environmental and behavioural information for patients (personal attributes, quality of life and care).

The SCDO upper-level concepts (https://bioportal.bioontology.org/ontologies/SCDO) cover different REDCap data collection instruments listed in **Table 2**. Although the data elements were initially mapped to SCDO upper-level terms, we are in the process of mapping them to more specific SCDO terms to enable optimal ontology-driven data capture and quality control. This more fine-grained mapping process includes assessing whether new ontology terms should be added to cater for data elements that could not be mapped to the SCDO.

Data Harmonization

The SOP and tools for data harmonization are accessible via the SickleInAfrica website (https://www.sickleinafrica.org/sops_booklet). As with most existing guidelines, this SOP relies heavily on manual tasks. In this context, these tasks were distributed within the SickleInAfrica database team to manually map, where possible, site data to SickleInAfrica data elements. **Figure 3** highlights the number of variables that could be mapped between each site and SickleInAfrica data dictionaries, as well as those that were not comparable, as they were data elements specific to a site.

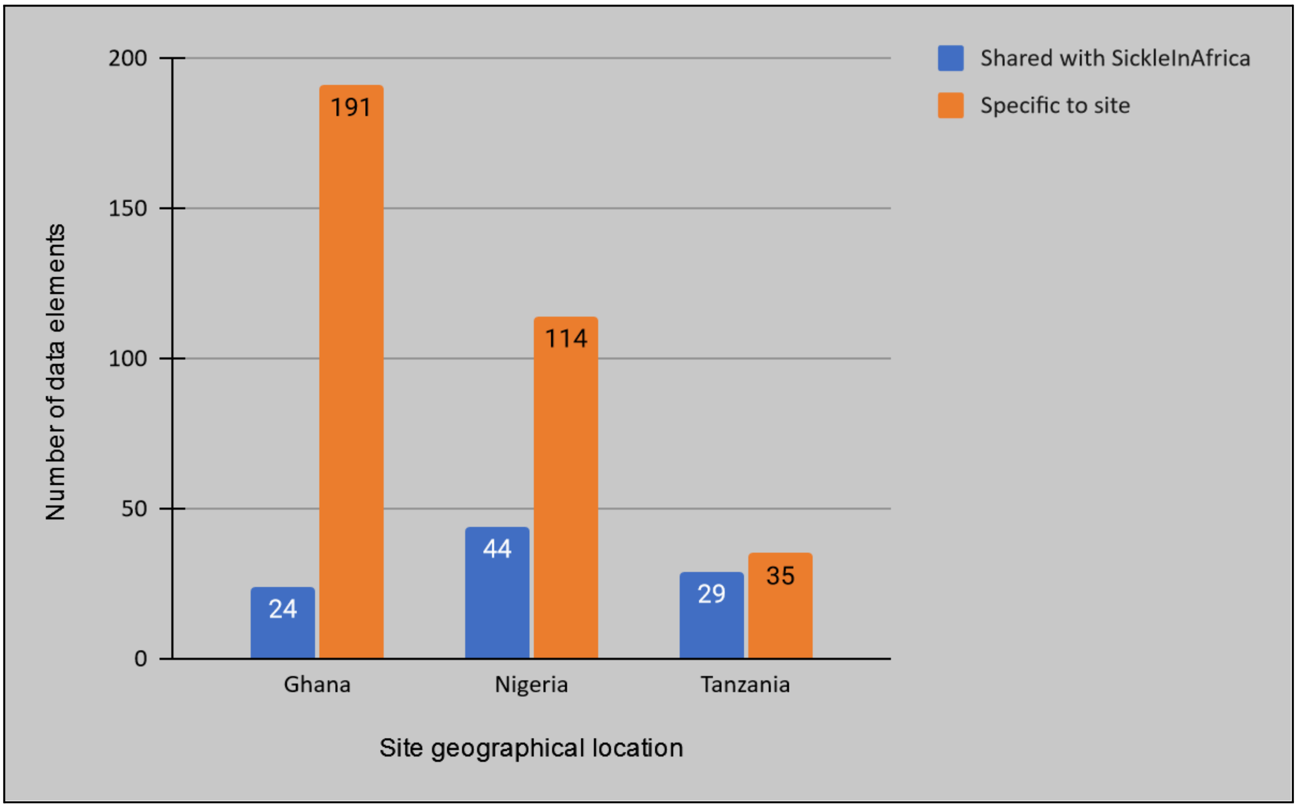
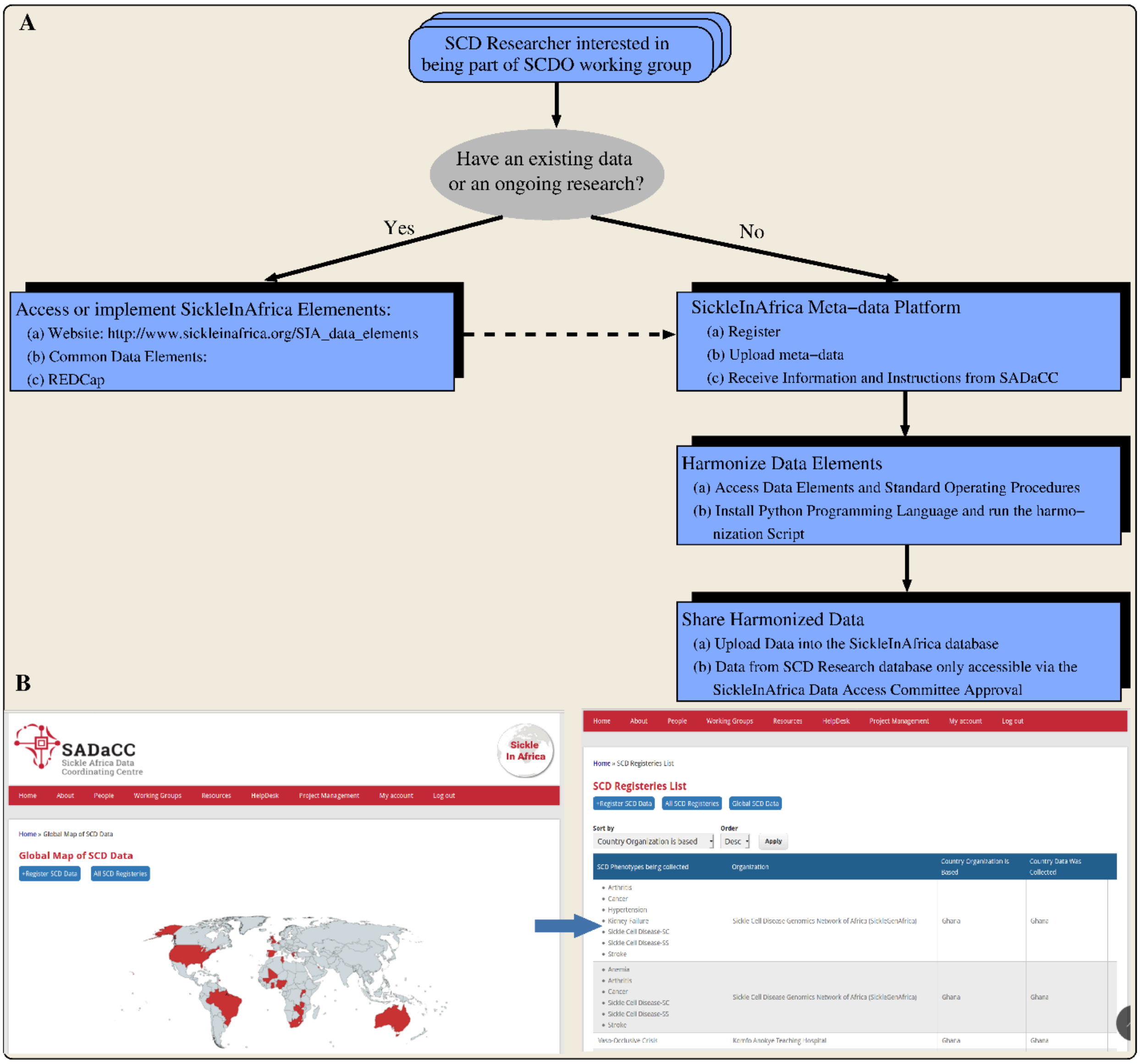


Figure 3: Numbers of data elements specific to a site and shared between a site and SickleInAfrica after the harmonization process.

Data elements shared between sites and SickleInAfrica data elements constitute the SickleInAfrica Core Data, a standardized minimal set of variables suggested to be commonly collected across sites in SPARCo research projects.

SCDO working group membership, resources & tools

A scalable database schema is presented, which facilitates cataloguing of retrospective and prospective data using an ontology-based metadata interface (illustrated in **Figure 4**). Stakeholders are invited to register as members of the SCDO-WG which is governed by the Open Science principles (Gallagher et al., 2020). SCD stakeholders with or without data can join the SCDO-WG as stated on the SCDO website (). **Figure 4(A)** illustrates how a stakeholder could pool their metadata into the SickleInAfrica database and **Figure 4(B)** provides the interactive map that shows existing SCD-related data. The registration form is accessible via https://www.sickleinafrica.org/scd-registry-form and allows participants to register as SCDO Working Group members even when they do not have data.


A
SCD Researcher interested in being part of SCDO working group
Have an existing data or an ongoing research?
Yes
No
Access or implement SickleInAfrica Elemenents:
(a) Website: http://www.sickleinafrica.org/SIA_data_elements
(b) Common Data Elements:
(c) REDCap
SickleInAfrica Meta–data Platform
(a) Register
(b) Upload meta–data
(c) Receive Information and Instructions from SADaCC
Harmonize Data Elements
(a) Access Data Elements and Standard Operating Procedures
(b) Install Python Programming Language and run the harmo–nization Script
Share Harmonized Data
(a) Upload Data into the SickleInAfrica database
(b) Data from SCD Research database only accessible via the SickleInAfrica Data Access Committee Approval
B
SADaCC
Sickle Africa Data Coordinating Centre
Sickle In Africa
Home
About
People
Working Groups
Resources
HelpDesk
Project Management
My account
Log out
Home » Global Map of SCD Data
Global Map of SCD Data
+Register SCD Data
All SCD Registeries
Home
About
People
Working Groups
Resources
HelpDesk
Project Management
My account
Log out
Home » SCD Registeries List
SCD Registeries List
+Register SCD Data
All SCD Registeries
Global SCD Data
Sort by
Country Organization is based
Order
Desc
Apply

SCD Phenotypes being collected	Organization	Country Organization is Based	Country Data Was Collected
• Arthritis • Cancer • Hypertension • Kidney Failure • Sickle Cell Disease-SC • Sickle Cell Disease-SS • Stroke	Sickle Cell Disease Genomics Network of Africa (SickleGenAfrica)	Ghana	Ghana
• Anemia • Arthritis • Cancer • Sickle Cell Disease-SC • Sickle Cell Disease-SS • Stroke	Sickle Cell Disease Genomics Network of Africa (SickleGenAfrica)	Ghana	Ghana
Vaso-Occlusive Crisis	Komfo Anokye Teaching Hospital	Ghana	Ghana

Figure 4: Joining SCDO-WG process and Meta-data platform access. (A) Flow diagram illustrating how new members can join the SCDO working group and upload meta-data of existing, ongoing and future research. Access to SCD data elements is also highlighted. (B) Illustration of information that a meta-data platform user could obtain from the platform if, for example, they clicked on Ghana in the interactive Global Map of SCD Data.

All resources developed in the SickleInAfrica project are freely available on the internet. However, various stakeholders of SCD and related research can become an SCDO-WG member based on their skill sets and willingness to harmonize their data elements and share their existing or prospective data (on the assumption that they have acquired appropriate ethics approvals).

The data elements are accessible via the SickleInAfrica website (www.sickleinafrica.org) and the REDCap Core Data Elements mapped to the SCDO specific terms will be submitted to the data element public repository, namely Common Data Elements (CDE) – National Institutes of Health: (https://cde.nlm.nih.gov/home).

Discussion

Remarkable progress has been made towards alleviating the SCD burden (Makani et al., 2017; Piel et al., 2017; Ngo-Bitoungui et al., 2021; Bukini et al., 2021; 2021; Nkya et al., 2020; Oron et al. 2020), however, efforts are often weakened by several factors, including lack of well-structured and standardized SCD research data. The latter has hampered SCD data sharing, integration and more generally, collaborative research. To promote collaborative research and facilitate cross-cohort information retrieval and analysis of large-scale SCD patient data, there was an urgent need to develop globally acceptable data standards in SCD unlike the existing data elements which were developed mainly by stakeholders from developed countries (Eckman et al., 2017; Sears et al., 2017). Through international researchers, clinicians and domain experts, we have developed standardized SCD data collection instruments and associated data elements accessible at https://sadacc.org/SIA_data_elements, by integrating data elements from PhenX data dictionaries, H3Africa and SickleInAfrica sites. Such a resource that has the potential to facilitate inter-study comparison of datasets from multiple sources can be utilized and adapted as needed in multinational cohort studies to capture SCD information on hemoglobinopathies, phenotypes, therapeutics, diagnostics, as well as quality of life and care, and the environment.

To our knowledge, the SickleInAfrica data elements are the most comprehensive African community-developed SCD data elements to-date which were mainly selected based on utility from several sites. They are being mapped to the current SCDO (the most comprehensive knowledge portal for SCD) to make SCD data elements findable, accessible, interoperable and reusable (FAIR) compliant, enabling their curation within the ontology and easing data harmonization across sites. The SickleInAfrica core data elements, capturing the minimal set of data elements determined by the SickleInAfrica consortium to be essential for the registry and addressing their current research questions, are currently being used by the SPARCo consortium to collect data for their registry. A flexible data capturing model is in use, which includes the use of mobile phones or tablets, or even hard paper printed case report forms (CRFs), if necessary, where sites do not have access to computers with the internet to ensure uninterrupted data collection. This is likely to yield large research datasets with reduced heterogeneity and the ability to achieve the required statistical power for large-scale association analyses (Makani et al., 2017; Makani and Wonkam et al., 2019). The SickleInAfrica data elements are a blueprint for building potentially the largest multinational SCD registry in the world (Makani et al., 2017; Makani and Wonkam et al., 2019).

In addition to increased coverage in comparison to different SPARCo site data elements and measures (Eckman et al., 2017; Sears et al., 2017), the SickleInAfrica data elements are also tailored for low-resourced regions and for Africa where the SCD burden is highest. Inclusion of different African ethnic groups provides the means to collect key data which would not have been possible with existing published data elements which had limited input from Africans. In addition, the data elements support capturing of consent data which is key for secondary use of the information. The SickleInAfrica data elements also permit collection of data from equipment that is most commonly used in low-income regions, as indicated above, whereas such data elements are largely missing from other freely available SCD data elements.

‘Post’ or ‘retrospective’ data element harmonization

Large sample sizes are needed to guarantee statistical power in elucidating associations between disease and factors influencing disease status, including genetic, clinical and environmental factors. In most instances, research results are silo-based and include disconnected data items or variables across multiple sites. It is therefore essential to standardize and harmonize such project-specific variables before analysing merged datasets. However, retrospective harmonization is tedious, not generally preferred, and likely to yield inconclusive results (Fortier et al., 2011) due to several reasons. This is mainly because data items are collected by individual studies which may lead to an increased heterogeneity and compromise data compatibility. For effective harmonization, a common data model that enables efficient data representation across sites is required (McCarty et al., 2014). While the SickleInAfrica data elements support retrospective data elements, we envisage that this resource will avoid such retrospective harmonization in the future since the data elements should preferentially be used for prospective data analyses. This should also enable the resolution of SCD inter-geographical variations and advance knowledge of phenotypic manifestations in relation to ethnic and regional variations and contribute to improving clinical practice and patients' management. An added advantage is that this proposed set of data elements is compatible with existing resources, such as the PhenX SCD measures that it builds on (Eckman et al., 2017).

SCDO-enabled databasing

As an ontology-guided database application, the associated ontology (SCDO) is expected to cover concepts used in the database in a consistent and unambiguous way, and to facilitate seamless data sharing and collaborations, including meta-analysis within the SCD community, and to support the development and curation of databasing and clinical informatics in SCD. Mapping different data elements to SCDO concepts allows one to set the ontology in the SickleInAfrica electronic system as an interface that enables database access and guides information capture, formalizing and easing management of the large research participants’ records. Some of the fundamental functionalities for consideration in ontology-driven database management are specified below.

a) Data collection: Predefined permissible values determine which values are allowable in a specified domain (i.e., a field in a form). In this case, the SCDO may be used to ensure data quality in the database by optimizing data capture and to expand data concepts and improve re-usability of the data.

b) Query of data: This is performed to retrieve information from the database. However, the semantic description of the database is often unavailable and using SCDO with proper relationships can produce more enriched results.

c) Data integration across studies or sites: The SCDO is a formal model of this domain (has standardized terms and relationships between the terms). Context-dependent concepts and relations with explicitly specified semantics are included to facilitate data integration across studies.

Availability and maintenance

SickleInAfrica built a sickle cell disease REDCap database model, which includes personal attributes, quality of life, diagnostics, phenotype, therapeutics and other clinical aspects related to SCD. This model utilizes REDCap, a secure web-based application, hosted at the SADaCC project, University of Cape Town. It requires minimal hardware requirements to operate, is portable and not dependent on the server environment system. This database is centrally stored and backed up daily. Currently, the set of harmonized data elements and associated documentation are accessible at https://sadacc.org/SIA_data_elements.

Future work - Proposed Implementation of the SCDO in REDCap

The conceptual framework of implementing an ontology-driven database in RedCap is built on the incorporation of an ontology-variable mapping in the database back-end. The mapping table would link data elements/variables included in the database to a 'parent' or more specific 'child' ontology term that can then be accessed during electronic data collection to display the name, description and synonyms of the ontology term to the data collector. For instances where data is collected electronically by medical specialists who can identify the specific ontology term, this could be available in an ontology lookup field to be entered. The final selection of ontology terms for each element of data collected would be determined through an optimized searching script processing the data retrospectively where a specific ontology term is not captured. To the best of our knowledge, the actual implementation of this process has not been carried out in REDCap yet. This implementation process may follow the steps below:

1. A set of finalized CRFs with harmonized data elements is established.
2. These data elements get mapped to an ontology term in the SCDO.
3. The SCDO version is imported into the SQL back-end of the REDCap database (or a standardised established method of regular version updating would need to be identified that could be run on the SCDO imported into the database).
4. A mapping table in the SQL database would need to be included which links each data element to an SCDO parent term and description.
5. Based on the content of the data collected, for some data elements an SCDO term would be pre-applied but for elements which may not be identifiable with an ontology term immediately, an AI script would need to be run retrospectively on the data collected to store the correct SCDO term in a related field for each record of data collected.
6. Descriptions and synonyms from the ontology need to be accessible in real-time via a REDCap application program interface (API) used during electronic data collection.
7. Properties from the SCDO should also be automatically imported into REDCap, e.g., laboratory ranges could be used to restrict or inform data input.
8. REDCap provides the ability to choose terms from an ontology in a text field that then provides a drop-down list of matching terms (see example: https://redcap.h3abionet.org/redcap/surveys/?s=8ERYNFH8TK). However, this is restricted to identical text matches whereas we should be allowed to choose the main class we want to select a term from - not only children in that class should appear in this list.
9. Once the ontology loading and searching process has been executed to store an ontology term for each data element in each record, curators would have to review the machine-selected terms initially to approve the correct ontology term selected. This would have to be an established quality check run regularly on the data in the database/s.

Clinical data

The SCDO will be used for SCD clinical data in ways synthesized in a recent review (Haendel et al., 2018). This includes clinical decision support, knowledge discovery, information retrieval across electronic medical records once the medical records have been annotated and indexed using the

SCDO. In the future, we intend to explore implementation of a graph database.

Conclusion

We present a comprehensive and community-developed SCDO-based REDCap data element measure for SCD research which is applicable to both low-resource and high-income countries. We also propose a database schema that integrates data elements from existing multiple-site research. This schema is flexible, scalable and dynamic, which facilitates efficient query and retrieval of research information. The new SCD system together with related SOPs and guidelines are currently being tested by different sites to build multi-site SCD research with improved quality and accuracy assurance. Future work includes operative integration of the REDCap system with the SCDO and application of the SCDO in clinical care.

Acknowledgements

We acknowledge the National Heart, Lung, and Blood Institute of the National Institutes of Health (NIH) for grant number U24HL135600–Sickle Africa Data Coordinating Centre (SADaCC) and U24HL13045881–Sickle Pan-African Research Consortium (SPARCo), as well as the National Institutes of Health Common Fund grant U24HG006941 (H3ABioNet).

The success of this work was greatly impacted by Dr. Gaston K Mazandu (deceased September, 2021), who enormously contributed to the design and implementation of the SickleInAfrica data elements, data harmonization and maintenance of a high quality database.

References

Rees DC, Williams TN, and Gladwin MT. Sickle-cell disease. Lancet 2010; 376:2018-2031.

Piel FB, Steinberg MH, Rees DC. Correspondence: Sickle Cell Disease. N Engl J Med 2017;377:302-305.

Diallo DA, Guindo A. Sickle cell disease in sub-Saharan Africa: stakes and strategies for control of the disease. Curr Opin Hematol 2014; 21:210-214.

Eckman JR, Hassell KL, Huggins W, Werner EM, Klings ES, Adams RJ, Panepinto JA, Hamilton CM. Standard measures for sickle cell disease research: the PhenX Toolkit sickle cell disease collections. Blood Adv. 2017;1(27):2703-2711.

Marques MB, Lorenz RG, Williams LA. High Percentage of Evanescent Red Cell Antibodies in Patients with Sickle Cell Disease Highlights the Need for a National Antibody Database. Blood 2015; 126:3572.

Hoehndorf R, Schofield PN, Gkoutos GV. The role of ontologies in biological and biomedical research: a functional perspective. Briefings in Bioinformatics 2015;16(6):1069–1080.

Harris PA, Taylor R, Thielke R, Payne J, Gonzalez N, Conde JG. Research Electronic Data capture (REDCap) – A metadata-driven methodology and workflow process for providing translational research informatics support, J Biomed Inform. 2009;42(2):377–381.

Harris PA, Taylor R, Minor BL, et al. The REDCap consortium: Building an international community of software platform partners. *J Biomed Inform*. 2019;95:103208.

Martinez-Cruz C, Blanco IJ, Vila MA. Ontologies versus relational databases: are they so different? A comparison. Artificial Intelligence Review 2011;38(4):271–290.

Kuiler EW. From Big Data to Knowledge: An Ontological Approach to Big Data Analytics. Review of Policy Research 2014;31(4).

Gruber TR. Towards Principles for the Design of Ontologies Used for Knowledge Sharing. In N. Guarino and R. Poli, editors, Kluwer Academic Publishers. Formal Ontology in Conceptual Analysis and Knowledge Representation, 1993.

Bodenreider O, Mitchell JA, Mccray AT. Biomedical ontologies. Pac Symp Biocomput 2005;76–78.

Mazandu GK, Mulder NJ. A topology-based metric for measuring term similarity in the Gene Ontology. Adv Bioinformatics 2012, 2012: Article ID 975783, 17 pages.

Mazandu GK, Chimusa ER, Mulder NJ: Gene ontology semantic similarity tools: survey on features and challenges for biological knowledge discovery. Briefings in bioinformatics 2018;18(5):886-901.

Ngo-Bitoungui VJ, Belinga S, Mnika K, Masekoameng T, Nembaware V, Essomba RG, Ngo-Sack F, Awandare G, Mazandu GK, Wonkam A. Investigations of Kidney Dysfunction-Related Gene Variants in Sickle Cell Disease Patients in Cameroon (Sub-Saharan Africa). Front Genet. 2021 Mar 15;12:595702.

Bukini D, Mbekenga C, Nkya S, Malasa L, McCurdy S, Manji K, Makani J, Parker M. Influence of gender norms in relation to child's quality of care: follow-up of families of children with SCD identified through NBS in Tanzania. J Community Genet. 2021;12(1):143-154.

Bukini D, Nkya S, McCurdy S, Mbekenga C, Manji K, Parker M, Makani J. Perspectives on Building Sustainable Newborn Screening Programs for Sickle Cell Disease: Experience from Tanzania. Int J Neonatal Screen. 2021 Feb 26;7(1):12.

Nkya S, Mwita L, Mgaya J, Kumburu H, van Zwetselaar M, Menzel S, Mazandu GK, Sangeda R, Chimusa E, Makani J. Identifying genetic variants and pathways associated with extreme levels of fetal hemoglobin in sickle cell disease in Tanzania. BMC Med Genet. 2020 Jun 5;21(1):125.

Oron AP, Chao DL, Ezeanolue EE, Ezenwa LN, Piel FB, Ojogun OT, Uyoga S, Williams TN, Nnodu OE. Caring for Africa's sickle cell children: will we rise to the challenge? BMC Med. 2020;18(1):92.

Gallagher RV, Falster DS, Maitner BS, et al. Open Science principles for accelerating trait-based science across the Tree of Life. Nat Ecol Evol 2020; 4:294-303.

Sears M, Ewing C, Lanzkron S, et al. A Prospective Multi-Centered Registry Is Feasible in Sickle Cell

Disease: Globin Regional Network for Data and Discovery (GRNDaD). Blood 2017;130(Supplement 1):3375.

DiMartino LD, Baumann AA, Hsu LL, et al. The sickle cell disease implementation consortium: Translating evidence-based guidelines into practice for sickle cell disease. *Am J Hematol*. 2018;93(12):E391-E395.

Glassberg, J.A., Linton, E.A., Burson, K. *et al.* Publication of data collection forms from NHLBI funded sickle cell disease implementation consortium (SCDIC) registry. *Orphanet J Rare Dis* **15,** 178 (2020).

Sickle Cell Disease Ontology Working Group. The Sickle Cell Disease Ontology: enabling universal sickle cell-based knowledge representation. *Database (Oxford)*. 2019;2019:baz118.

Mulder NJ, Nembaware V, Adekile A, Anie KA, et al. Proceedings of a Sickle Cell Disease Ontology workshop – Towards the first comprehensive ontology for Sickle Cell Disease. Appl Transl Genom 2016;9:23–29.

Nembaware V, Mazandu GK, Hotchkiss J, et al. The Sickle Cell Disease Ontology (SCDO): Enabling Collaborative Research and Co-Designing of New Planetary Health Applications. OMICS-A Journal of Integrative Biology 2020;24(10): 559-567.

Fortier I, Doiron D, Little J, Ferretti V, et al. Is rigorous retrospective harmonization possible? Application of the DataSHaPER approach across 53 large studies. Int J Epidemiol 2011; 40(5):1314–28.

McCarty CA, Huggins W, Aiello AE, Bilder RM, et al. PhenX RISING: real world implementation and sharing of PhenX measures.BMC Med Genomics 2014; 7:16.

Makani J, Sangeda RZ, Nnodu O, et al. SickleInAfrica. The Lancet Haematology 2020;7(2):e98-e99.

Makani J, Ofori-Acquah SF, Tluway F, Mulder N, Wonkam A. Sickle cell disease: tipping the balance of genomic research to catalyse discoveries in Africa. Lancet. 2017;389(10087):2355-2358.

Munung NS, Nembaware V, de Vries J, et al. Establishing a Multi-Country Sickle Cell Disease Registry in Africa: Ethical Considerations. Front Genet. 2019;10:943.

Wonkam A, Makani J. Sickle cell disease in Africa: an urgent need for longitudinal cohort studies. Lancet Glob Health 2019;7(10):e1310-e1311.

Haendel AM, McMurry JA, Relevo R, et al. A Census of Disease Ontologies. Annual Review of Biomedical Data Science 2018;1:305-331.

Balestroni G, Bertolotti G. L'EuroQol-5D (EQ-5D): uno strumento per la misura della qualità della vita [EuroQol-5D (EQ-5D): an instrument for measuring quality of life]. Monaldi Arch Chest Dis. 2012;78(3):155-159.

SCDO Consortium members

Name	Surname	Title	Qualification	Affiliation	Email Address
Kofi	Anie	Dr	PhD	London North West Healthcare University NHS Trust and Imperial College London	kofi.anie@nhs.net
Deogratias	Munube	Dr	MBChB, Mmed	Mulago National Referral Hospital/Makerere University , Kampala, Uganda	deomunube@gmail.com
Marsha	Treadwell	Prof	PhD	University of California San Francisco Benioff Children’s Hospital Oakland	marsha.treadwell@ucsf.edu
Obiageli	Nnodu	Prof	BMBCH, FWACP(Lab Med),	Centre of Excellence for Sickle Cell Disease Research & Training, University of Abuja, Abuja, Nigeria and Department of Haematology and Blood Transfusion, University of Abuja, Abuja, Nigeria	obiageli.nnodu@uniabuja.edu.ng ORCID ?
Kais	Ghedira	Dr	PhD	Laboratory of Bioinformatics, Biomathematics and Biostatistics (LR20IPT09), Pasteur Institute of Tunisia, 1002, University of Tunis El Manar, Tunis, Tunisia.	kais.ghedira@pasteur.tn
Miriam V.	Flor-Park	Dr	MD, PhD	Hospital das Clínicas da Faculdade de Medicina da Universidade de São Paulo, Instituto da Criança, São Paulo-Brazil	parkmiriam0@gmail.com
Neil	Hanchard	Dr.	MBBS, DPhil	Dept. of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX *Current Address: National Human Genome Research Institute, National Institutes of Health, Bethesda, MD, USA	hanchard@bcm.edu
Charmaine	Royal	Dr.	PhD	Departments of African & African American Studies, Biology, Global Health, and Family Medicine & Community Health Duke University, Durham, NC, USA	charmaine.royal@duke.edu
Irene	Kyomugisha				
Arthemon	Nguweneza				

