Data Sharing and Access Guideline for DS-I Africa

Responsible Entity: DS-I Africa Consortium steering committee

Version: 1.0

Originally Issued: November 21, 2024

Effective Date: November 21, 2024

Find this Guideline at: TBC

Contents

1 What is the purpose of this Guideline?

The purpose of this Guideline is to create a unified and comprehensive framework for data sharing among Consortium Research Groups? both within and outside the Consortium. By supplementing existing data management plans and the NIH data sharing policy, this Guideline aims to harmonise data-sharing practices across the Consortium. Although adherence to this Guideline is voluntary, it fosters a shared commitment to the harmonisation of data sharing, so promoting a culture of collaboration and advancing the global public good.

2 Why do we need this Guideline?

We need this Guideline to help Consortium Research Groups to:

- plan to share their data
- understand how to structure their data to facilitate sharing
- set out how a request to share data should be considered
- explain what to do when a data-sharing request is received
- differentiate between different types of data-sharing requests.

3 NIH values that this Guideline promotes[1]

Governance:

- Governance mechanisms should ensure protection from any intentional or unintentional unauthorised access, use, disclosure, or re-identification of data; and proper identification, management, and mitigation of breaches.
- Researchers and other data users should be informed of and be subject to consequences for failure to adhere to all rules developed in furtherance of these principles.

Transparency:

- Information should be communicated to participants clearly and conspicuously concerning the privacy and security measures that are in place to protect participant data, including notification plans in the event of a breach.
- Participants should be notified promptly following discovery of a breach of their personal information. Notification should include, to the extent possible, a description of the types of information involved in the breach; steps that individuals should take to protect themselves from potential harm, if any; and steps being taken to investigate the breach, to mitigate losses, and to protect against further breaches.

[1] Precision Medicine Initiative: Privacy and Trust Principles
https://allofus.nih.gov/protecting-data-and-privacy/precision-medicine-initiative-privacy-and-trust-principles

4 What is this Guideline about?

The DS-I Africa Consortium (hereafter referred to as "the Consortium") is committed to advancing scientific research and knowledge dissemination through the open and ethical sharing of African health-related data. These guidelines are designed to facilitate the sharing of and access to health related and associated data among Consortium Research Groups and the wider scientific community. This Guideline aims to standardise practices, and ensure a clear understanding of responsibilities and expectations.

5 Who needs to read this Guideline?

- Consortium Research Groups
- DS-I Africa SC
- Data and material access committees (DMACs)
- DACs
- Data requestors
- Data collectors
- Health researchers.

6 What terms do you need to know in this Guideline?

In this Guideline, the following terms have the following meanings:

Bona Fide Researcher: An individual or entity engaged in legitimate scientific research with the objective of advancing knowledge in the field of health data science, including genomics or associated disciplines. Bona fide researchers operate within the ethical, legal, and professional frameworks of academic and scientific research. This includes, but is not limited to, academics, clinicians, and researchers affiliated with educational institutions, research institutions, or non-profit organisations focused on health and disease research.

Consortium Research Groups: Research groups in the DS-I Africa Consortium engaged in scientific research using demographic, surveillance, clinical/phenotype, genomic (human/pathogen), image, geospatial and other data as captured in the eLwazi catalogue. These groups may work with both source data and inferential data for their research endeavours. A full list of the research groups can be found here: https://dsi-africa.org/infographics

Data Access: The mechanism through which bona fide researchers are able to retrieve data in the manner approved by an organisation holding the data, and can include controlled or open access.

Controlled Access: A mechanism by which to control the distribution of research data that is potentially personally identifiable and which requires a bona fide researcher to submit a data access request which is accompanied by supporting documentation such as the specific research questions the data will be used for, and institutional ethical clearances for conducting the study.

Open Access: Research data that is freely available in community-focused data repositories and which can be accessed, downloaded and used without any major restrictions.

Data Access Committee (DAC): A Data Access Committee designated by a Consortium Research Group responsible for reviewing and approving requests for access to data. Such a DAC ensures, inter alia, that requests meet the Consortium's ethical, legal, and scientific standards, in particular focusing on the legitimacy of the requester and the intended use of the data.

Data Dictionary: A data dictionary is a document that outlines the structure, content, and meaning of a given variable. This includes what type of data is being collected (e.g., free text, numerical, categorical or group data), the full wording of a question, what values are allowable (e.g., numeric ranges, multiple choice codes), and what those values mean (e.g., 0 = no high blood pressure diagnosis, 1 = borderline high blood pressure, 2 = high blood pressure). A data dictionary is a critical tool for data analysis and reproducibility.

The term 'codebook' is often used interchangeably with 'data dictionary', although the data dictionary can contain more information about the structure of a database. In the widely used data collection tool, REDCap, the data dictionary is a CSV file containing information on the variables and the structure of the REDCap database, while the codebook is a human readable document that provides information on each data element.[2]

Data Downloads: A modality of data sharing where researchers are permitted to directly download datasets to their local computing environments.

Data Sharing: Generally, data sharing is where one or more persons, the data provider(s), provide access to data to another person(s) – the data user(s). In the context of this Guideline, data sharing specifically refers to the sharing of data by Consortium Research Groups with each other and with external parties.

Data Sharing Applicant: An individual or entity requesting access to data. This term encompasses both Consortium Research Groups and external parties, including bona fide researchers.

Data Transfer Agreement (DTA): A legal document outlining the terms and conditions under which data is shared between the provider and the recipient. The DTA addresses issues such as confidentiality, data use limitations, and compliance with relevant laws and guidelines.

Data Visiting: Generally, data visiting is a modality of data sharing where data user(s) access and analyse the data in the computing environment(s) of the data provider(s). In the context of this Guideline, data visiting specifically refers to data sharing where the data sharing applicant accesses and analyses the data in the computing environment(s) of a Consortium Research Group.

DUOS: A Data Use Oversight System enables researchers to submit a single request for multiple datasets and the corresponding data access committees to review these requests and provide a response to the researchers. In this process, a DUOS also codifies researchers' requests and datasets' data use limitations using the Data Use Ontology (DUO) developed by

[2] https://www.nnlm.gov/guides/data-glossary/data-dictionary

the Global Alliance for Genomics and Health (GA4GH). This makes requests more decipherable both to DACs and the DUOS's matching algorithm, so expediting the data access request turnaround time.

Embargo: A temporary restriction placed on the sharing of certain data to protect the proprietary interests of Consortium Research Groups. Embargoes are typically applied when there is a reasonable likelihood that immediate data sharing could compromise the results or the integrity of ongoing research.

eLwazi ODSP: eLwazi Open Data Science Platform facilitates the storage, retrieval and processing of data for health research purposes.

Inferential Data: Data that arises not merely from the cleaning, ordering, or reformatting of the source data, or the combination thereof with other data, but from analysis of the source data that generates new knowledge or hypotheses that were not explicitly contained in the source data or its combination with other data.

Metadata: Data that provide additional information intended to make scientific data interpretable and reusable (e.g., date, independent sample and variable construction and description, methodology, data provenance, data transformations, and any intermediate or descriptive observational variables).

Public Statement: Information made available by Consortium Research Groups on their websites or other public fora describing the data they hold. This includes details necessary for understanding the potential utility of the data for external research projects.

Primary Data Steward: This is the key individual who is responsible for management, oversight and governance of data in an organisation.

Secondary Data Steward: This is an individual who supports the primary data steward and has a more targeted role, such as supporting data quality initiatives or training.

Source Data: Data and its associated metadata in its unprocessed or processed form, including but not limited to, imaging, electronic health records, geolocation, model training, raw genomic sequence data, genomic variants, data and associated metadata. It serves as the foundational material for analyses by Consortium Research Groups, sourced from source data providers both within and outside the DS-I Africa Consortium.

Source Data Providers: Organisations, groups, or entities that supply source data to the Consortium. These can include members within the Consortium as well as external entities.

7 Are there other relevant documents and resources?

- 2023 NIH Data Management and Sharing Policy
- Writing a Data Sharing and Management Plan
- NIH Sharing Policies
- List of other NIH-related sharing policies
- United Nations Trade and Development Data Protection and Privacy Worldwide

- Data protection legislation in Africa and pathways for enhancing compliance in big data health research[3]
- Data Privacy Impact Assessment Guideline [to be developed].
- Data Protection and Privacy Laws of participating countries (please refer to the country specific laws related to the source of the data)

Guidelines and policies are frequently updated or amended. Be sure to refer to most recent versions of any documents used or referenced.

8 What are the modalities of data sharing?

The DS-I Africa Consortium facilitates two primary data-sharing modalities: **Data Downloads** and **Data Visiting**, which are designed to meet the diverse needs of health data research while upholding data privacy and security. Each Consortium Research Group is tasked with selecting the most appropriate modality or combination thereof for sharing its data. This decision is informed by several factors, including the sensitivity of the dataset, legal and ethical obligations, specific research requirements, and the Consortium Research Group's technical capacity.

A key consideration for Consortium Research Groups, especially when contemplating the **data visiting** modality, is their technical expertise and the resources to establish a secure analytical workspace for researchers. For Consortium Research Groups lacking this capability, the Consortium recommends leveraging the **eLwazi Platform**. This platform provides a robust, secure environment for making inferential data accessible, so ensuring that researchers can conduct their work effectively in a controlled setting. Opting for the **eLwazi Platform** allows Consortium Research Groups to circumvent the technical challenges of setting up and maintaining a **data visiting** infrastructure, thereby ensuring that their data remains accessible and usable in alignment with the Consortium's standards for data sharing, privacy, and security.

9 How should I describe my data?

To enhance transparency and accessibility for data science health research, the DS-I Africa Consortium mandates that Consortium Research Groups adhere to specific guidelines for publicly sharing and describing their datasets. DS-I Africa research projects are responsible for completing descriptions of project datasets to be included in the DS-I Africa Data Catalogue being developed by the eLwazi Open Data Science Platform. The eLwazi ODSP will make available to each DS-I Africa research project a set of dataset description collection instruments to capture metadata and locations of the DS-I Africa project datasets.

[3] Munung, N.S., Staunton, C., Mazibuko, O. et al. Data protection legislation in Africa and pathways for enhancing compliance in big data health research. *Health Res Policy Sys* 22, 145 (2024).

9.1 Public statements of data holdings

9.1.1 Accessibility

DS-I Africa research groups will work with **eLwazi ODSP** to increase the findability and accessibility of DS-I Africa project data through a DS-I Africa Data Catalogue being developed by eLwazi ODSP. Data reuse conditions for each individual DS-I Africa dataset should be made explicit and, where possible, mapped to the Data Use Ontology. Information about the location of the datasets and their access mechanisms must be included. DS-I Africa research groups must deposit DS-I Africa datasets in domain-specific data repositories that provide dataset accession numbers, or they must work with eLwazi ODSP to determine a suitable data repository where none exist in order to obtain Digital Object Identifiers for their datasets.

Each Consortium Research Group must maintain a dedicated website section for data descriptions, accessible to all visitors, including researchers and policymakers, or provide the necessary information to the eLwazi ODSP to host this information.

9.1.2 Comprehensiveness

DS-I Africa research projects must nominate primary and, if required, secondary data stewards to work with the eLwazi ODSP to ensure that the correct current data and metadata descriptions are captured and, when the DS-I Africa project is satisfied, these dataset descriptions are then made available in the DS-I Africa data catalogue. The dataset information captured by eLwazi ODSP should facilitate deposition of the datasets to regional or community-specific repositories in the DS-I Africa Research Consortium.

9.2 Guidelines for DS-I Africa dataset descriptions

To ensure transparency and to facilitate the effective use of datasets, descriptions of individual datasets will be provided by Consortium Research Groups to nominated primary and secondary data stewards who must complete the elements listed in Annexure A. Note that these are subject to change by eLwazi ODSP as the DS-I Africa Data Catalogue is updated. When changes occur in the DS-I Africa Data Catalogue, eLwazi ODSP will notify the DS-I Africa Research Consortium via the coordinating centre.

Please refer to Annexure A for the full metadata sets

9.2.1 DS-I Africa dataset citation requirements

An output of the DS-I Africa Consortium that is of value to the scientific community is the high quality, diverse African datasets. To ensure that generators of these datasets are correctly acknowledged and to measure the citations and impact of DS-I Africa-derived datasets over time, each dataset must carry a citation with the following elements:

- Authors (creators of the dataset)
- Title (of the dataset)
- Publication date (of the dataset)
- Publisher (name of repository where the data is published)

- Version number (of the dataset, if any)
- Location (URL/DOI to access the dataset).

These must include the statement “As part of the Data Science for Health Discovery and Innovation in Africa (DS-I Africa) Initiative, grant number: XYZ” at the end of the citation – for tracking purposes.

9.2.2 Findable, Accessible, Interoperable and Reusable (FAIR) data

DS-I Africa research projects will make their derived and analytical datasets conform to the FAIR principles (https://www.go-fair.org/fair-principles/). DS-I Africa research projects, together with eLwazi ODSP will determine the current FAIR status of project datasets, set their high level FAIR goals using the FAIRPlus Dataset Maturity Model (https://fairplus.github.io/Data-Maturity/), and implement recipes from the FAIR Cookbook (https://faircookbook.elixir-europe.org/content/home.html). DS-I Africa projects will provide a description of:

- the data collection methods, including types of genomic sequencing
- data sources, and any collaborative efforts
- steps taken to clean, order, reformat, and analyse the source data to generate inferential data
- other relevant information relating to the dataset's creation and refinement.

9.2.3 Ethical considerations

By their nature, all datasets in the DS-I Africa Research Consortium must have obtained ethical approval from one or more organisations, and this ethical approval may affect various aspects of data sharing. In order to deal with this, the requirements imposed by ethics committees must be provided to indicate issues such as:

- Name of ethics committee
- Date when ethics approval was provided
- Consent procedures
- Anonymisation practices
- Other requirements/restrictions on data use imposed by the ethics committee.

9.2.4 Personal data indication

Clearly state whether the dataset contains personal data and, if so, detail the nature of this data and the measures taken to ensure individual confidentiality. This may include de-identification, anonymisation and restrictions on data use to prevent re-identification. Informing users about the presence of personal data and privacy protections is crucial for maintaining ethical standards and legal compliance.

In certain cases, DS-I Africa projects may wish to conduct a risk assessment of their data in their institution, and have in place data transfer agreements such as the one created in the DS-I Africa Law project (https://doi.org/10.5281/zenodo.7110269), which must be signed before any data are shared. Sensitive data should be stored and shared in an encrypted format only with authorised researchers.

By adhering to these guidelines, Consortium Research Groups will ensure that dataset descriptions are thorough, transparent, and aligned with best practices in data sharing and research ethics. This approach facilitates informed and responsible use of datasets, so enhancing the integrity and impact of research conducted in the DS-I Africa Consortium.

If there are any bespoke data protection requirements, such as a prohibition on any attempts to reidentify the data or safeguards to protect personal information, then such requirements should be disclosed.

9.3 Examples

Two examples of dataset descriptions are attached:

9.3.1 Example 1: Population-Wide Study (No Personal Data)

9.3.2 Example 2: Individual-Level Study (Contains Personal Data)

9.4 Maintenance and updates

Consortium Research Groups must regularly review and update their public statements to reflect any changes in their data holdings – including how often data updates occur. This will ensure ongoing accuracy and relevance through the dataset description forms created by eLwazi ODSP and the DS-I Africa Data Catalogue.

10 What are the requirements for a data access request?

The DS-I Africa Consortium facilitates access to its datasets under a framework designed to uphold ethical standards, privacy, and data protection. These guidelines detail the requirements for researchers seeking access to datasets.

To request access to datasets, applicants must provide the following information:

10.1 Consortium membership status

Indicate whether the applicant is a member of the DS-I Africa Consortium. For non-members, detailed information about the applicant and their organisational affiliation is required.

10.2 Applicant information for non-members

Information about the applicant's institution, including name, address, and the nature of its research activities.

10.3 Dataset

Identification of the specific datasets to which access is sought.

10.4 Purpose

The purpose of the intended research with the requested data.

10.5 Data-sharing modality preference

Specify whether data access is sought through download or data visiting, with a justification based on the research and data sensitivity.

10.6 Additional requirements for data download of inferential data containing personal data

When requesting inferential data containing personal data through data download, additional safeguards must be detailed to ensure data protection. The additional requirements are:

10.6.1 Country and institution

Information on the country and institution where the data will be received and held, considering the legal and ethical data handling standards applicable.

10.6.2 Protection measures

Detailed description of statutory, organisational, and technical measures in place at the receiving institution to safeguard the confidentiality and safety of the inferential data. This includes data encryption, access controls, secure data storage, and policies on data sharing and disposal, which demonstrate adherence to best practices and legal requirements for data protection.

10.6.3 Data Protection Impact Assessment (DPIA)

Provide either a justification for why a Data Protection Impact Assessment is not required, or provide information on how the DPIA will be conducted. In this regard, please refer to clause 13.1.

11 How will data access requests be processed?

The DS-I Africa Consortium employs a flexible and structured approach to managing requests for access to health datasets. This accommodates the diverse needs and preferences of Consortium research groups while ensuring adherence to ethical and legal standards. This system outlines the procedures for the submission, review, and approval of data access requests tailored to the type of data being requested.

11.1 Request submission options

Each Consortium Research Group must adopt one of the following request submission options:

11.1.1 Centralised request portal

The DS-I Africa Consortium and eLwazi ODSP will investigate various solutions for the hosting of data access committees and data access requests such as the Data Use Oversight System (DUOS) currently implemented by the Broad Institute for suitability in the DS-I Africa context. This option is designed to streamline the application process and to provide a uniform experience for applicants. Consortium Research Groups may opt to use this portal for handling requests.

11.1.2 Research group-specific processes

Alternatively, Consortium Research Groups may choose to manage data access requests through their own systems. Groups opting for this approach are responsible for ensuring that

their process is transparent, equitable, and consistent with the Consortium's data sharing principles.

11.2 Differentiation Between source data and inferential data

The data access system distinguishes between requests for source data and inferential data to ensure appropriate handling and review:

11.2.1 Source data requests

Requests for source data are forwarded to the respective source data provider, whether within or outside the Consortium. The requesting party is provided with the source data provider's contact details and any relevant information required to pursue the request directly. Consortium Research Groups are encouraged to facilitate the introduction or notification process to ensure a smooth transition.

11.2.2 Inferential data requests

Every Consortium must have a Data Access Committee (DAC) and requests for inferential data are processed internally by the Consortium Research Group's DAC. The review process considers the applicant's membership status in the Consortium and the specific nature of the proposed research. Special attention is given to ensuring that the requested data's use aligns with ethical standards and the Consortium's objectives.

11.3 Review process

11.3.1 Preliminary screening

Regardless of the submission method, all requests undergo a preliminary screening to confirm completeness and basic compliance with the Consortium's requirements.

11.3.2 Review and Approval

When it comes to source data, the review process is managed directly by the source data provider, who evaluates the data access requests in alignment with their own established policies. In this scenario, the Consortium Research Group plays a supportive role, and facilitates communication between the requester and the source data provider as necessary to ensure a smooth review process. On the other hand, requests for inferential data are meticulously reviewed by the DAC based on the provisions of this document and any other guidelines that the DAC has adopted. This committee conducts a comprehensive evaluation of each request, considering both the overarching guidelines of the Consortium and the specific characteristics of the inferential data in question.

12 What protocol should a Data Access Committee follow?

The DACs of each of the projects in the DS-I Consortium play a crucial role in the governance of data access in the Consortium.

Each DAC evaluates and decides whether and on what conditions to make data available to researchers who apply for data from its project. Each DAC ensures that data access requests are evaluated fairly and in alignment with the Consortium's ethical guidelines, privacy standards, and research objectives.

12.1 Steps in the protocol

Each DAC follows a systematic protocol in reviewing data access requests, which at a minimum should include the steps below:

12.1.1 Step 1: Verification of Consortium membership

A DAC begins the review process by verifying the membership status of the applicant in the Consortium. For Consortium members, the DAC checks against the current membership roster to confirm the applicant's affiliation and standing in the Consortium. This step streamlines the review process for internal members by leveraging existing agreements and understandings related to data sharing and use.

12.1.2 Step 2: Confirmation of bona fide researcher status

External applicants, i.e., those not affiliated with the Consortium, undergo a thorough verification to confirm their status as bona fide researchers. A DAC reviews the provided credentials, affiliations and the research track record to ensure that applicants are engaged in legitimate scientific research and that they adhere to ethical research practices. This may involve requesting additional documentation or references to substantiate the applicant's bona fide status.

12.1.3 Step 3: Assessment of research purpose

A DAC applies three criteria for access:

- First, that the data request must be aligned with the Data Use Permission (see clause 14.4.5 of Annexure A).
- Second, that potential risks to the privacy of research participants associated with the proposed use of the data, where relevant, have been identified and addressed satisfactorily.
- Third, that the proposed research does not overlap with or pre-empt the results of the ongoing Consortium projects. Should the latter be the case, the DAC should impose a temporary embargo on the data sharing, instead of declining access. An embargo period allows Consortium members to complete their research and publish results before the data is made available to external parties. The length of the embargo should be determined based on the specific circumstances and projected timelines of the involved Consortium projects.

12.1.4 Step 4: Recording and communication of decision reasons

Once a decision has been made, the DAC meticulously records the reasons for its approval, conditional approval, or denial of the data access request. This documentation ensures transparency and provides valuable feedback to applicants. The decision, along with the reasons and any conditions attached (such as an embargo period), is communicated to the applicant in a clear and timely manner. For approved requests, the communication includes detailed instructions for accessing the data and any requirements or conditions that must be met by the applicant.

12.2 Legal and ethical compliance for personal data

12.2.1 Special consideration for personal data

When the requested dataset contains personal data, the DAC's permission to access is always contingent upon strict legal and ethical compliance, as detailed in the Consortium's Legal and Ethical Considerations section. Examples of bodies that can provide ethical approval include: the National Research Ethics Committees of the country of data origin and registered Institutional Review Boards of academic institutions that approved the primary data collection. You may also need to refer to the Data Protection Officer in the country of data origin. This additional scrutiny ensures that all necessary protections are in place to safeguard individual privacy and to ensure adherence to applicable regulations.

12.2.2 Compliance review

This involves a detailed review of the dataset in question to ensure that it meets all relevant privacy laws, ethical guidelines, and the stipulations of any Data Transfer Agreement (DTA) to be executed. The DAC works closely with legal and ethics experts to ascertain that the proposed data use respects the rights and expectations of the data subjects and complies with the Consortium's ethical standards.

12.3 Protocol compliance

The DAC protocols are designed to uphold the highest standards of data governance, so ensuring that data sharing in the DS-I Africa Consortium is conducted responsibly, ethically, and in a manner that promotes scientific advancement. The DAC remains committed to a transparent and fair review process, by balancing the need for open science with the imperative to protect the interests of data providers and research participants.

13 Notes on the legal and ethical considerations

In the context of health research and data sharing, adherence to legal and ethical standards is paramount. The DS-I Africa Consortium is committed to ensuring that all data sharing activities respect the privacy, dignity, and rights of individuals whose data are shared, and comply with applicable laws and ethical guidelines. This commitment is underpinned by strict protocols for DTAs and special considerations for datasets containing personal data.

13.1 Data Protection Impact Assessment

Not all applications to get access to data require a Data Protection Impact Assessment (DPIA). For example, where the data being provided has been de-identified and there is credible evidence that it could not be re-identified to belong to an individual, then it is not necessary to conduct a DPIA. Given the variable nature of projects and data requests there is no set format that a DPIA must follow, although there are numerous international tools to assist.[4]

[4] UCISA Privacy Impact Assessment Toolkit, available at https://www.ucisa.ac.uk/Resources/Privacy-Impact-Assessment-Toolkit; Article 29 Data Protection Working Party Guidelines on Data Protection Impact Assessment (DPIA) and determining whether processing is 'likely to result in a high risk' for the purposes of Regulation 2016/679 (4 October 2017); ICO Sample DPIA Template, available at

While the approach set out below is not mandatory, a decision not to follow it should be formally noted in writing, which would include the reasons for deviating from this Guideline. This process can be assisted by data protection lawyers.

DPIA steps:[5]

1. Determine whether a DPIA is required
   1.1. At a high level determine whether personal information is going to be used. In general, if personal information is present then a DPIA will be required.
2. Perform an inherent risk-rating: determine the depth to which a DPIA should go
   2.1. This high-level risk-rating is designed to look at overall factors. Questions that should be asked are:
      2.1.1. What is the volume of the personal information?
      2.1.2. Is there special personal information involved (like genomic information)?
      2.1.3. Is there any further processing (i.e., processing which was not anticipated when the personal information was collected)?
      2.1.4. Is there any profiling or automated decision-making (artificial intelligence)?
      2.1.5. Will the data subject be aware of the processing?
      2.1.6. Does the processing require prior authorisation from the Information Regulator before it can happen?
      2.1.7. What is the value of the personal information (how much would it cost to replace it)?
      2.1.8. What would happen if the personal information was not used/was unavailable?
      2.1.9. How valuable would the personal information be to a bad actor?
3. Describe the processing activity
   3.1. Map out the personal data in a precisely, including all the purposes that it will be used for and various processing points including:
      3.1.1. Collection
      3.1.2. Use
      3.1.3. Transfer
      3.1.4. Archiving
      3.1.5. Destruction.
4. Identify privacy risks
   4.1. Based on the mapping of the personal data, identify privacy risks for:
      4.1.1. data subjects
      4.1.2. the organisation
      4.1.3. The DS-I Africa Consortium.
5. Identify and evaluate privacy solutions
   5.1. These solutions are unique to the project, but examples of solutions/mitigation strategies are:

https://gdpr.eu/wp-content/uploads/2019/03/dpia-template-v1.pdf.

[5] Adapted from De Stadler, E., Hattingh, I. L., Esselaar, P., & Boast, J. (2021). *Overthinking the Protection of Personal Information Act*. Juta.

5.1.1. Contracts with parties protection data protection rights and ensuring accountability
5.1.2. Obtaining further consent from participants
5.1.3. Introducing a procedure to reduce the risk, such as a security breach procedure similar to that suggested by the DS-I Africa Guideline on Security Breaches
5.1.4. Reducing the personal information that is shared
5.1.5. Encrypting personal information when it is transferred
5.1.6. Reducing the amount of processing/inferential data done by the project.

6. Sign off and record outcomes
   6.1. Ensure that the outcomes are reduced to writing and confirmed by all stakeholders.
7. Integrate outcomes into project plan/agreements.
8. Agree on monitoring plan
   8.1. Whether there will be any checks that the recipient is complying with the agreement and how this will be monitored.

For more detailed guidance, please refer to the Data Privacy Impact Assessment Guideline [to be developed].

13.2 Data Transfer Agreement (DTA)

13.2.1 Mandatory DTA

Before any data is shared, a DTA must be executed between the data provider (Consortium Research Group or source data provider) and the data recipient. The DTA serves as a legally binding document that outlines the terms and conditions under which data is shared, used, and protected. The original application should be included as an Annexure to the DTA as should the ethical approval for the project.

- Example of a DTA
- GA4GH DTA Example
- Explanatory Memorandum on how the DTA works

13.2.2 Prohibition of on-sharing

A crucial provision in the DTA is the explicit prohibition of the on-sharing of data by the recipient. This clause ensures that the shared data cannot be passed on or made available to third parties, therefore maintaining a DAC's control over the distribution and use of the data.

13.2.3 Compliance and accountability

The DTA includes stipulations for compliance with applicable legal and ethical standards, ensuring that both parties are aware of their responsibilities and the consequences of non-compliance. It establishes a framework for accountability, so safeguarding the interests of data subjects, researchers, and the institutions involved.

13.3 Considerations for datasets containing personal data

13.3.1 Pre-transfer review

If the dataset to be shared contains personal data, the Consortium Research Group is obligated to conduct a thorough review to ensure that all relevant legal and ethical requirements are satisfied before entering into a DTA.

13.3.2 Privacy protections

This review process includes an assessment of privacy protections in place for the dataset, such as anonymisation or de-identification techniques and whether these measures comply with privacy laws and ethical guidelines specific to the jurisdiction(s) of the data provider and recipient. If a DPIA was performed, then this would be included in this section.

13.3.3 Ethical oversight

The Consortium Research Group must also consider any ethical approvals or oversight requirements that apply to the research from which the data originated, as well as any ethical approval obtained by the applicant.

13.3.4 Legal frameworks

Compliance with international, national, and local data protection laws is essential. The DTA must reflect the legal frameworks governing the collection, storage, and use of personal data, such as the General Data Protection Regulation (GDPR) in the European Union, so ensuring that data sharing practices are legally sound and ethical.

13.4 Implementation

13.4.1 Guidance and support

The Consortium provides guidance and support to research groups in drafting and negotiating DTAs, and offering templates and resources that encapsulate the necessary legal and ethical provisions.

13.4.2 Ongoing monitoring

After the execution of a DTA, there is ongoing monitoring as agreed in the DTA (note that not all DTAs will require ongoing monitoring) to ensure that the terms of the agreement are adhered to throughout the data sharing process. This could include periodic reviews and audits, as necessary, to maintain compliance and to protect the rights of all parties involved.

14 How will the data be managed post-funding?

The DS-I Africa Consortium recognises the critical importance of sustainable data management practices, in particular following the completion of NIH funding. To this end, Consortium Research Groups are provided with two principal options to ensure the long-term preservation and accessibility of inferential data. These options are designed to accommodate varying capabilities and resources among research groups, while upholding the Consortium's commitment to data sharing and scientific advancement.

14.1 Option 1: Maintain inferential data

Consortium Research Groups may choose to continue maintaining inferential data in-house. This option involves several key responsibilities:

14.1.1 Data maintenance

The main host institution of the Research Group assumes ongoing responsibility for the storage, preservation, and security of inferential data. This includes implementing measures to prevent data loss, corruption, and unauthorised access.

14.1.2 Access requests and DAC operations

The Research Group continues to manage access requests for the inferential data, ensuring that such requests are reviewed and approved in accordance with the Consortium's guidelines. The DAC remains operational, overseeing the evaluation of applications for data access, and ensuring compliance with ethical and legal standards.

14.1.3 Data visiting infrastructure

If the Research Group offers data visiting as a mode of data access, it must also maintain the necessary infrastructure to support this functionality. This includes secure computing environments where approved researchers can analyse data without the need to transfer data outside of the controlled environment.

14.2 Option 2: Transfer to Repository X

Alternatively, research groups may opt to transfer their inferential data to Repository X, a free data repository designated by the NIH. This option entails:

14.2.1 Data handover

The inferential data is transferred to Repository X, which is managed by an NIH-appointed DAC. This transfer includes all necessary metadata and documentation to ensure the data remains useful and accessible to future researchers.

14.2.2 Managed access

Once the data is transferred, Repository X's DAC assumes responsibility for managing access requests, so relieving the original Consortium Research Group of this duty. The NIH-appointed DAC will review and approve requests based on criteria aligned with those of the DS-I Africa Consortium.

14.3 Flexibility to switch

Consortium Research Groups retain the flexibility to switch from Option 1 to Option 2 at any stage in the future if they determine that they are unable to continue maintaining the data or the DAC operations. This ensures that inferential data remain accessible to the scientific community even if the original managing entity can no longer sustain its commitments.

14.4 Prohibition of data deletion

Regardless of the chosen post-funding data management option, it is imperative that inferential data is not deleted. The Consortium prioritises the preservation of data to ensure its availability for future scientific inquiry and advancement, in line with the principles of open science and data sharing.

Annexure A: Metadata sets

14.4.1 Project metadata

Project-level metadata include metadata fields relevant to describing the project as a whole, and all possible datasets pertaining to the project. This information is collected to organise various datasets according to their related project and is only completed once.

Field label (variable)	Definition/guide
Organisation (organization)	Specify whether the project is part of a listed Consortium or network.
DS-I Africa project title (p_title)	Specify the project title. List of DS-I Africa project titles provided.
(Non-DS-I Africa affiliated) Project Title (p_title_other)	Specify project title. Non-DS-I Africa projects only.
Project acronym (p_acronym)	Specify project acronym (if applicable).
Project website (p_website)	Specify project website (if applicable).
Project description (p_description)	Provide a brief project description, and highlight the scope and aims.
Project keywords (p_keywords)	Provide some keywords which describe the primary components of the project.

14.4.2 DS-I Africa project data stewards

Primary and secondary data stewards contact details are to be provided from each DS-I Africa research project to determine the primary people responsible for project metadata management, and to ensure future communications related to such management. The primary and secondary metadata stewards will ensure that information about project datasets are kept current, including version updates, and that they are sent to the eLwazi ODSP DS-I Africa data catalogue. Only email addresses will be displayed in the catalogue. Other contact information can be provided when required as a result of a legitimate request.

Field label (variable)	Definition/guide
Primary + Secondary Metadata Steward	
First Name (firstname)	Specify the primary and secondary metadata stewards' first names.
Last Name (surname)	Specify the primary and secondary metadata stewards' first name.
Primary email address (email)	Specify the primary and secondary metadata stewards' primary email address for metadata management-related communications.

Primary affiliation (institution)	Specify the primary and secondary metadata stewards' primary affiliation.

14.4.3 Dataset access contact

Should there be a need to provide a link between the catalogue and the person responsible for data access-related queries, for example, a DAC, these contact details are also collected. All personal data will be protected according to the South African Protection of Personal Information Act (POPIA), and only an email will be displayed in the catalogue.

Dataset access contact details	
Does the project have a Data Access Committee? (dac)	Specify whether the project's data access is managed via a Data Access Committee.
First name (dac_firstname)	Specify the first name of the person who should be contacted for data access-related queries.
Last name (dac_surname)	Specify the last name of the person who should be contacted for data access-related queries.
Primary email address (dac_email)	Specify the primary email address of the person who should be contacted for data access-related queries.
URL for data access requests (dac_url)	Specify the URL for data access-related queries or requests, if available.

14.4.4 Dataset metadata

Dataset-level metadata include various general fields that should be completed for all dataset metadata records submitted by the projects. This form should be completed once for each individual dataset record being submitted to the archive.

Field label (variable)	**Definition/guide**
Dataset name (d_name)	Specify a unique name or identifier for the dataset for which information is being submitted.
Dataset category (d_category)	Specify the category pertaining to the submitted dataset.
Dataset description (d_description)	Provide a brief description of the data, including the scope of the data.
Sample size (sample_size)	Specify the sample size of the given dataset (if applicable).
Country(ies) from which data is sourced/collected (d_countries)	Specify the country(ies) from which the data in the dataset was generated or collected.
Type of dataset (d_type)	Specify whether the dataset is a primary or derived dataset.
Dataset version (dataset_version)	Specify the dataset version.
Dataset status	Specify the dataset status.

(d_status)	
Last updated (last_updated)	Specify the last date the metadata record related to the dataset was updated.
First date in dataset (d_startdate)	Earliest date listed in dataset (whether this is date of birth or date of visit/collection).
Last date in dataset (d_lastdate)	Most recent date listed in dataset (whether this is date of birth or date of visit/collection).
Dataset storage location (Description) (storage_description)	Provide a brief description of the dataset storage location, e.g. institution.
Dataset storage location (URL, if available) (storage_link)	Provide a link to the dataset storage location, if available.
Dataset format (d_format)	Specify the format in which the dataset is stored.
If secondary/derived data, provide references to original source publications (d_original_source)	Provide references to the original data sources from which the secondary dataset was derived.
Should this dataset be listed on the catalogue? (catalogue_active)	This option is provided for the metadata steward to specify when the metadata record is complete and ready to display on the catalogue. This option should be considered carefully.
[Optional] Do you have a data dictionary/codebook associated with this dataset that you want to upload? (dataset_codebook)	An optional field to provide a data dictionary or codebook related to the submitted dataset record.
[Optional] Any other dataset-associated documentation (dataset_zip_files)	An optional field to provide documentation related to the submitted dataset record, e.g. case report forms, protocols, standard operating procedures.

14.4.5 Data reuse conditions

This indicates the secondary data use conditions and modifiers pertaining to the submitted dataset, as per the Data Use Ontology (DUO). These conditions and modifiers are typically based on data use agreements, policies, and consent forms. Definitions for each condition and modifier can be explored here: https://www.ebi.ac.uk/ols4/ontologies/duo. In addition, eLwazi can support the mapping of data use agreements, policies and consent forms to DUO. Should you require assistance from us, please see the presentation linked on our website, or contact us via our helpdesk - https://helpdesk.elwazi.org/. This form should be completed once for each individual dataset record being submitted to the archive.

Field label (variable)	**Definition/guide**
Data Use Permission (data_use_permission)	A metadata item that is used to indicate permissions for datasets and/or materials and relates to the purposes for which datasets and/or material might be removed, stored or used.

Data use modifier (data_use_modifier)	Data use modifiers indicate additional conditions for use, which modify the data use permission specified above.
Specify geographical restriction (d_use_country)	If geographical restriction is selected as a data use modifier, specify the country(ies) to which secondary use is restricted.

14.4.6 Additional information for specific types of datasets

Air quality datasets

General metadata elements pertaining to air quality data, defined as data pertaining to the assessment of air pollution. This form should be completed once for each individual air quality dataset record being submitted to the catalogue.

Field label (variable)	Definition/guide
Origin (organisation where data was produced) (aq_origin)	Specify the organisation which produced the dataset.
Method of data production (aq_method)	Specify the method used by the organisation which produced the dataset.
Reference (publication/web documentation) (aq_reference)	Provide a reference to documentation regarding the production of the dataset.
Collection period (in years) (collection_year)	Specify the production period associated with the data in the dataset, in years.
Number of monitoring stations (aq_nr_stations)	Specify the number of monitoring stations involved in the production of the data.
Type of monitoring stations (aq_type_stations)	Specify the type of monitoring stations involved in the production of the data.
Monitored city/town(s) (aq_city)	Specify the city(ies) or town(s) from which the data was produced.

Climate datasets

General metadata elements pertaining to climate data, defined as data which helps to specify the climate of a specific location or region, such as precipitation, temperature, wind speed, and humidity parameters. This form should be completed once for each individual climate dataset record being submitted to the catalogue.

Field label (variable)	Definition/guide
Origin (organisation where data was produced) (c_origin)	Specify the organisation which produced the dataset.
Source (model instrument & version) (c_source)	Specify the instrument and version employed to produce the dataset.
Reference (publication/web documentation) (c_reference)	Provide a reference to documentation regarding the production of the dataset.

Area type (c_area_type)	Specify the area types assessed during the production of the dataset.

Demographics and health datasets

General metadata elements pertaining to demographic and health-related datasets. Demographic data is information about groups of people according to certain attributes such as age, sex, and place of residence. Health data is defined as any information relating to a person's state of health. This form should be completed once for each individual demographic and health-related dataset record being submitted to the catalogue. For all indicators specified in this form, the format in which those indicators are included in your dataset is not relevant, only whether there is an indicator related to that specification.

Field label (variable)	**Definition/guide**
Demographics (dh_demographics)	Specify the demographic-related indicators which are included in the dataset record being submitted.
Anthropometrics (dh_anthropometrics)	Specify the anthropometric-related indicators which are included in the dataset record being submitted.
Disease elements (dh_clinical)	Specify the disease-related indicators which are included in the dataset record being submitted.
Lab tests (dh_labs)	Specify the lab-related indicators which are included in the dataset record being submitted.
Vital signs (dh_vitals)	Specify the vital-related indicators that are included in the dataset record being submitted.
Lifestyle factors (dh_lifestyle)	Specify the lifestyle-related indicators which are included in the dataset record being submitted.
Other (dh_other)	Specify any other indicators which are included in the dataset record being submitted, an not included above.

Genomics datasets

General metadata elements pertaining to genomics data, defined as data related to the structure and function of an organism's genome. This form should be completed once for each individual genomics dataset record being submitted to the catalogue.

Field label (variable)	**Definition/guide**
Organism (g_organism)	Specify the organism on which the genomics experiment was conducted.
Genotyping method (g_method)	Specify the genotyping method employed to produce the dataset.
Experiment design (g_design)	Specify the experimental design employed to produce the dataset.
Sequencing technology used (g_tech)	Specify the sequencing technology employed to produce the dataset.

Image datasets

General metadata elements pertaining to image data, defined as data produced by scanning a surface with an optical or electronic device. This form should be completed once for each individual image dataset record being submitted to the catalogue.

Field label (variable)	Definition/guide
Imaging method (i_method)	Specify the imaging method employed to produce the dataset.
Organism (i_organism)	Specify the organism to which the image data pertains.
Image type (i_type)	Specify the types of images contained in the dataset.
Organ system (i_organ_system)	Specify the organ system pertaining to the images contained in the dataset.

Mobility datasets

General metadata elements pertaining to mobility data, which, in a geospatial context, is aggregated anonymised information of people's movements surrounding points of interest or neighbourhoods. This form should be completed once for each individual mobility dataset record being submitted to the catalogue.

Field label (variable)	Definition/guide
Origin (organisation where data was produced) (m_origin)	Specify the organisation which produced the dataset.
Method of tracking (m_method_of_tracking)	Specify the method employed to produce the dataset.
Reference (publication/web documentation) (m_reference)	Provide a reference to documentation regarding the production of the dataset.
Duration of coverage (m_duration)	Specify the production period associated with the data in the dataset, in years.

Example 1: Population-Wide Study (No Personal Data)

Project Metadata

- **Organisation**: DS-I Africa Consortium
- **DS-I Africa Project Title**: Population Health Metrics in Sub-Saharan Africa
- **Project Acronym**: PHM-SSA
- **Project Website**: www.phmssa.org
- **Project Description**: The Population Health Metrics in Sub-Saharan Africa (PHM-SSA) project aims to collect and analyse health data from various regions across sub-Saharan Africa to better understand population health trends, disease prevalence, and the impact of health interventions. This project focuses on non-personal aggregated data to ensure privacy and ethical standards.
- **Project Keywords**: Population health, sub-Saharan Africa, Health metrics, Disease prevalence, Health interventions

DS-I Africa Project Data Stewards

- **Primary Metadata Steward**
  - **First Name**: Thabo
  - **Last Name**: Mokoena
  - **Primary Email Address**: thabo.mokoena@phmssa.org
  - **Primary Affiliation**: PHM-SSA Research Group
- **Secondary Metadata Steward**
  - **First Name**: Zintle
  - **Last Name**: Ndlovu
  - **Primary Email Address**: zintle.ndlovu@phmssa.org
  - **Primary Affiliation**: PHM-SSA Research Group

Dataset Access Contact

- **Does the project have a Data Access Committee?** Yes
- **First Name**: Lerato
- **Last Name**: Nkosi
- **Primary Email Address**: lerato.nkosi@phmssa.org
- **URL for Data Access Requests**: www.phmssa.org/data-access

Dataset Metadata

- **Dataset Name**: PHM-SSA_2024_HealthMetrics
- **Dataset Category**: Population Health Data
- **Dataset Description**: This dataset contains aggregated health metrics collected from various regions in sub-Saharan Africa. The data includes information on disease prevalence, health intervention outcomes, and general health trends from 2020 to 2024. The dataset is designed to support research on population health and to inform public health policies.
- **Sample Size**: Data aggregated from approximately 1 000 000 individuals
- **Country(ies) from which data are sourced/collected**: Kenya, Nigeria, South Africa, Ghana, Ethiopia
- **Type of Dataset**: Primary
- **Dataset Version**: 1.0
- **Dataset Status**: Active

- **Last Updated**: 2024-06-01
- **First date in dataset**: 2020-01-01
- **Last date in dataset**: 2024-05-31
- **Dataset Storage Location (Description)**: Stored at PHM-SSA Research Group, Sub-Saharan Africa Division
- **Dataset Storage Location (URL, if available)**: www.phmssa.org/data-storage
- **Dataset Format**: CSV, JSON
- **Should this dataset be listed on the catalogue?** Yes
- **[Optional] Do you have a data dictionary/codebook associated with this dataset that you want to upload?** Yes, available at www.phmssa.org/data-dictionary
- **[Optional] Any other dataset-associated documentation**: Case report forms, protocols, standard operating procedures, available at www.phmssa.org/documentation

Data Reuse Conditions

- **Data Use Permission**: Research Use Only (DUO:0000014)
- **Data Use Modifier**: No Commercial Use (DUO:0000016)
- **Specify Geographical Restriction**: Data use restricted to sub-Saharan Africa (DUO:0000018)

Additional Information for Specific Types of Datasets

Demographics and Health Datasets

- **Demographics**: Age, gender, region
- **Anthropometrics**: Height, weight
- **Disease Elements**: Prevalence of malaria, HIV/AIDS, tuberculosis
- **Lab Tests**: N/A
- **Vital Signs**: Blood pressure, heart rate
- **Lifestyle Factors**: Smoking status, physical activity levels
- **Other**: N/A

DS-I Africa Dataset Citation Requirements

- **Authors**: Thabo Mokoena, Zintle Ndlovu, Lerato Nkosi
- **Title**: PHM-SSA_2024_HealthMetrics
- **Publication date**: 2024-06-01
- **Publisher**: PHM-SSA Data Repository
- **Version number**: 1.0
- **Location**: www.phmssa.org/data/PHM-SSA_2024_HealthMetrics
- **Statement**: “As part of the Data Science for Health Discovery and Innovation in Africa (DS-I Africa) Initiative, grant number: XYZ”

FAIR Data

- **Findability**: The dataset is indexed in the PHM-SSA Data Catalogue and searchable via keywords and metadata fields.
- **Accessibility**: Accessible through the PHM-SSA Data Repository with appropriate permissions.
- **Interoperability**: Metadata follows standard ontologies and is compatible with various data analysis tools.
- **Reusability**: Detailed metadata and documentation provided to ensure effective reuse.

Ethical Considerations

The dataset does not contain personal data. All data is aggregated to ensure participant privacy and compliance with ethical standards.

Personal Data Indication

- **Contains Personal Data**: No
- **Risk Assessment**: N/A
- **Data Transfer Agreements**: N/A

Example 2: Individual-Level Study (Contains Personal Data)

Example Dataset Description for a Hypothetical Individual-Level Study

Project Metadata

- **Organisation**: DS-I Africa Consortium
- **DS-I Africa Project Title**: Longitudinal Study of Diabetes in East Africa
- **Project Acronym**: LSDEA
- **Project Website**: www.lsdea.org
- **Project Description**: The Longitudinal Study of Diabetes in East Africa (LSDEA) aims to track the health outcomes and progression of diabetes among individuals in East Africa over a ten-year period. This study collects individual-level data, including health metrics, lifestyle factors, and genetic information, to understand the interplay between genetics, environment, and disease progression.
- **Project Keywords**: Diabetes, East Africa, Longitudinal study, Health outcomes, Genetic data

DS-I Africa Project Data Stewards

- **Primary Metadata Steward**
  - **First Name**: Sipho
  - **Last Name**: Mkhize
  - **Primary Email Address**: sipho.mkhize@lsdea.org
  - **Primary Affiliation**: LSDEA Research Group
- **Secondary Metadata Steward**
  - **First Name**: Elize
  - **Last Name**: Van der Merwe
  - **Primary Email Address**: elize.vandermerwe@lsdea.org
  - **Primary Affiliation**: LSDEA Research Group

Dataset Access Contact

- **Does the project have a Data Access Committee?** Yes
- **First Name**: Thandi
- **Last Name**: Dlamini
- **Primary Email Address**: thandi.dlamini@lsdea.org
- **URL for Data Access Requests**: www.lsdea.org/data-access

Dataset Metadata

- **Dataset Name**: LSDEA_2024_DiabetesCohort
- **Dataset Category**: Individual Health Data
- **Dataset Description**: This dataset contains individual-level health data collected from participants in the Longitudinal Study of Diabetes in East Africa. The data includes demographic information, clinical measurements, genetic data, and lifestyle factors collected from 2015 to 2024. Personal identifiers have been removed to protect participant privacy.
- **Sample Size**: 10 000 individuals
- **Country(ies) from which data is sourced/collected**: Kenya, Tanzania, Uganda
- **Type of Dataset**: Primary
- **Dataset Version**: 2.0
- **Dataset Status**: Active
- **Last Updated**: 2024-06-01
- **First date in dataset**: 2015-01-01

- **Last date in dataset**: 2024-05-31
- **Dataset Storage Location (Description)**: Stored at LSDEA Research Group, East Africa Division
- **Dataset Storage Location (URL, if available)**: www.lsdea.org/data-storage
- **Dataset Format**: CSV, JSON
- **Should this dataset be listed on the catalogue?** Yes
- **[Optional] Do you have a data dictionary/codebook associated with this dataset that you want to upload?** Yes, available at www.lsdea.org/data-dictionary
- **[Optional] Any other dataset-associated documentation**: Case report forms, protocols, standard operating procedures, available at www.lsdea.org/documentation

Data Reuse Conditions

- **Data Use Permission**: Research Use Only (DUO:0000014)
- **Data Use Modifier**: No Commercial Use (DUO:0000016)
- **Specify geographical restriction**: Data use restricted to East Africa (DUO:0000018)

Additional Information for Specific Types of Datasets

Demographics and Health Datasets

- **Demographics**: Age, gender, ethnicity, region
- **Anthropometrics**: Height, weight, body mass index (BMI)
- **Disease Elements**: Diabetes type, duration of diabetes, comorbid conditions
- **Lab Tests**: HbA1c levels, blood glucose levels, lipid profile
- **Vital Signs**: Blood pressure, heart rate
- **Lifestyle Factors**: Diet, physical activity, smoking status
- **Other**: Medication adherence, health service use

Genomics Datasets

- **Organism**: *Homo sapiens*
- **Genotyping Method**: Whole genome sequencing
- **Experiment Design**: Longitudinal cohort study
- **Sequencing Technology Used**: Illumina NovaSeq 6000

DS-I Africa Dataset Citation Requirements

- **Authors**: Sipho Mkhize, Elize van der Merwe, Thandi Dlamini
- **Title**: LSDEA_2024_DiabetesCohort
- **Publication date**: 2024-06-01
- **Publisher**: LSDEA Data Repository
- **Version number**: 2.0
- **Location**: www.lsdea.org/data/LSDEA_2024_DiabetesCohort
- **Statement**: "As part of the Data Science for Health Discovery and Innovation in Africa (DS-I Africa) Initiative, grant number: XYZ"

FAIR Data

- **Findability**: The dataset is indexed in the LSDEA Data Catalogue and searchable via keywords and metadata fields.
- **Accessibility**: Accessible through the LSDEA Data Repository with appropriate permissions.
- **Interoperability**: Metadata follows standard ontologies and is compatible with various data analysis tools.
- **Reusability**: Detailed metadata and documentation provided to ensure effective reuse.

Ethical Considerations

The dataset contains personal data. Consent procedures and anonymisation practices have been implemented to ensure ethical compliance and participant privacy. All data has been de-identified to protect individuals' identities.

Personal Data Indication

- **Contains Personal Data**: Yes
- **Risk Assessment**: Conducted in the LSDEA Research Group, ensuring all data transfers are secure and complying with data protection regulations.
- **Data Transfer Agreements**: In place, following the DS-I Africa Law project guidelines (https://doi.org/10.5281/zenodo.7110269). Sensitive data are stored and shared in an encrypted format with authorised researchers only.