M A D I V A

Data Access and Sharing Policy

Version 3.0.1

3 April 2025
Approved: Coordination Committee, October 2024
Final approval: TCPDC, 4 November 2024
Typographical corrections: 3 April 2025

1 Introduction

Most of the raw data for the MADIVA Research Hub was collected by MADIVA's principal collaborators in rural Agincourt, South Africa, and in urban informal settlements in Nairobi, Kenya. These collaborators have rich datasets from longitudinal population studies collected by the health and demographic surveillance systems based in the two locations, together with a nascent set of clinical health records and genomic data. Given the sensitive nature of this data, it is essential that access to and sharing of it are governed by up-to-date policies. We are also sensitive to the fact that the data is given to MADIVA by the original data providers for purposes of MADIVA only.

1.1 Definitions

- A *Coinvestigator* is either a person or organisation (depending on context) who/which is a formal member of the MADIVA Research Hub.
- A *Collaborator* is a person or organisation (depending on context) who/which is not a formal member of the MADIVA Research Hub but which we have decided to work with on a common project.

1.2 Scope

(a) MADIVA data is data that has (a) been provided by co-investigating or collaborating parties to MADIVA for the purpose of MADIVA research; or (b) acquired from third parties using MADIVA resources under such conditions specified by these third parties.

(b) This policy covers the conduct of members of the MADIVA team as well as the procedures for data access and sharing between collaborators and other external parties. Partners covered by this policy are the Wits Health Consortium (Pty), the African Population and Health Research Center (APHRC), the South African Medical Research Council (SAMRC) and IBM Research Africa. Over time additional partners will be recruited as pilot projects; however, this data sharing policy will not apply to them and ad hoc policies will be drawn up to support them.

(c) This document does not cover issues such as details of data protection (e.g., encryption of data transfer and managing security breaches) which will be in a separate SOP document.

1.3 Roles and responsibilities

(a) **Coinvestigators and data providers** will prepare the data and supply it to the MADIVA Hub. They are responsible for the integrity and quality of the data. These teams will remove primary personal identifiers, such as identity numbers, phone numbers, health insurance numbers, names, precise dates of birth, location data, etc. Precise location data will be removed and the data will be re-coded so that researchers can establish the geographical proximity of data subjects to one another. A unique identifier will be allocated to each data subject. Hence, the demographic and clinical data transferred to the MADIVA Hub will not identify data subjects and will not be directly identifiable data.

(b) **The MADIVA Coordination Committee** (CC) takes overall responsibility for the data; it is responsible for approving all analyses and Manuscript Concept Documents, and for determining what data is sharable and how it will be shared.

(c) **MADIVA PI and Project Manager** The MADIVA PI and project manager are responsible for coordinating access to data. They are responsible for reporting any breaches of security to the appropriate organisational authorities. They are also responsible for reporting any breaches of conditions with respect to the use of data to the co-PIs, project leads and organisational leads, and taking appropriate remedial action in consultation with these individuals.

(d) **Data Users**. Anyone using the data is expected to respect the confidentiality and privacy of individuals whose records they access; to observe any restrictions that apply to sensitive data; and to abide by applicable laws, policies, procedures and guidelines with respect to access, use, sharing or disclosure of information. The unauthorised storage, disclosure or distribution of MADIVA data in any medium or use of any such data for one's own personal gain is strictly prohibited and considered gross misconduct. Data may only be processed for approved purposes and analyses.

(e) **All parties** recognise that the same data may be used by multiple analyses within MADIVA and collaborating parties. All MADIVA researchers are responsible for identifying potential overlap of use of data to avoid duplication of effort and unhealthy competition. This is important both for intra-MADIVA relationships and for acting collegially with respect to third parties.

1.4 Data types and categories

We use different types of data

(a) Aggregated data: Data that has been combined and summarised from multiple individual records to provide statistical information without revealing personal identities.

(b) Anonymised data: Data that has been processed or modified to remove or alter any identifying information, making it impossible to link the data to an individual.

(c) Closed data: Data that is restricted in access and which can only be accessed by authorised personnel or specific stakeholders in an approved manner.

(d) Demographic data: Information related to the characteristics of populations, such as age, gender, race, ethnicity, education level, and socioeconomic status.

(e) Derived data: Data that has been processed, transformed, or derived from raw data through analysis or computation.

(f) Genotype data: Genetic information related to an individual's DNA sequence, often used in genetic research and precision medicine.

(g) Geospatial data: Data that includes geographic information, such as GPS coordinates, addresses, or geographic boundaries.

(h) Health records: Medical information about an individual's health history, diagnoses, treatments, medications, and test results.

(i) Metadata: Data that provides information about other data, such as data source, format, structure, and context.

(j) Open data: Data that is freely available to the public with minimal restrictions on access and use.

(k) Personal data: Information that identifies or can be used to identify an individual, such as name, address, email, phone number, social security number, or any other identifiable information.

(l) Phenotype data: Observable characteristics or traits of individuals, such as height, weight, blood pressure, medical conditions, and lifestyle factors.

(m) Proprietary data: Data owned by a specific organisation or individual and protected by intellectual property rights.

(n) Raw data: Unprocessed and unstructured data as originally collected.

(o) Sensitive data: Highly confidential information that requires special protection due to its potential for harm if disclosed or misused. This may include health records, financial information, biometric data, religious or philosophical beliefs, political affiliations, and ethnicity.

2 Principles of data access and sharing

We have two principles that are in tension with each other.

2(1) The MADIVA Research Hub recognises and adopts the FAIR principles accessing and sharing data which include maximising the availability of data collected by the Hub, in a timely and responsible manner, protecting the rights and privacy of human subjects

who participated in research studies, recognising the scientific contribution of researchers who generated the data, considering the nature and ethical aspects of proposed research whilst ensuring the timely release of data, promoting deposition of data in existing community data repositories whenever possible.

2(2) Reasonable period of exclusive time: MADIVA members should have sole access to the data for a limited and reasonable period of time, subject to section 5.

3 Guidelines for data access by members of MADIVA during the MADIVA research phase

3.1 Mode and procedure of sharing

(a) MADIVA is a multi-organisation collaboration and all MADIVA collaborators require access to common data sets.

(b) MADIVA data will be securely stored on the Wits Core Research Cluster for analysis by MADIVA collaborators.

(c) MADIVA collaborators will at minimum follow security procedures of the Wits Core Cluster (e.g., authentication and authorisation, reporting of breaches)

(d) Co-investigator components of MADIVA will sign Data Access Agreements with Wits Health Consortium committing them and their staff members to the principles of this agreement.

(e) MADIVA data that is stored on the Wits Core Research Cluster can only be used for MADIVA purposes in accordance with this document and the Publication Policy. Breaching of this is considered a serious matter.

(f) The following is explicitly permitted

  (a) Using data solely for Manuscript Concept Documents approved by the CC. (If in the course of an analysis, researchers identify further research questions then an MCD should be modified or a new MCD created).

  (b) The work of the DMAC solely for the purpose of creating and curating data sets.

(g) Any other use of the data internally must be approved by the PI in writing , who should report this to the next CC meeting.

3.2 Conditions of access

(a) Data may only be analysed for MADIVA projects. Data and results of analyses of data may only be used in papers which have been approved by the MADIVA Coordination Committee, and by team members named on that MCD.

(b) Reproduction of the data and/or sharing the data with any unauthorised individual(s) or institution(s) outside the Hub is strictly prohibited.

(c) MADIVA team members must, therefore, not release nor permit others to release data to any other person without written authorisation from the Coordinating Committee.

(d) Copying of data off the Wits Core may only be done with written permission of the PI in consultation with the DMAC lead. In some cases, this may require a DTA to be signed.

(e) Confidentiality pledge: MADIVA team members will not attempt to use or permit others to use data to ascertain the identity of any individual or household included in any data set. MADIVA team members also commit to refraining from using, or permitting others to use, data to report information that could identify individuals or households directly or by inference.

(f) Respondent identifiers: The Hub is committed to protecting the identity of all respondents who provided information in its databases. All analytical data sets (both qualitative and quantitative) released by the Hub MUST be anonymised by being stripped of all respondent identifiers. By accepting the use of MADIVA data, the user is pledging that he/she will not, under any circumstance, regenerate identifiers or try to identify respondents.

(g) Reporting of errors or inconsistencies: Researchers will promptly notify the DMAC lead and PI of any errors discovered at the earliest opportunity.

(h) Any suspected security breach should be reported according the security protocols of the organisation storing the data.

(i) Publications: All authors and co-authors will comply with the MADIVA publication policy on the use of MADIVA data.

(j) Security: MADIVA team members will ensure that the data in their possession is used in a secure environment where access to the storage medium is password-protected. This will prevent unauthorised access to the data.

(k) Loss of privilege to use data: Any violation of the conditions of use of MADIVA data must be reported to the reporting PI within 48 hours of knowledge of the event.

(l) Acknowledgement: Any work/reports drawing on MADIVA data must include an appropriate acknowledgement of the data source. The following acknowledgement is mandatory:

> This research was supported by the Fogarty International Center and National Institute of Biomedical Imaging and Bioengineering (NIBIB) and OD/Office of Data Science Strategy (ODSS) and OD/Office of Strategic Coordination (OSC) of the National Institutes of Health under Award Number U54 TW 012077. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

(m) Other acknowledgements of funding will be required for use of specified data sets. The DMAC Core Lead is responsible for obtaining the appropriate acknowledgements and they and the Project Manager will keep a list of required acknowledgments.

(n) Data rights after leaving MADIVA: Team members who leave MADIVA will be treated as external users of MADIVA data and will need written permission from the MADIVA PI to use MADIVA data following the guidelines for data access for external users.

(o) Team members MUST delete all MADIVA generated data in their possession before they leave MADIVA, unless they obtain prior written authorisation from the Coordinating Committee to continue using the data after leaving the Hub. In cases in which a team member who has left the Hub releases a publication that draws on MADIVA data without following this policy, the Hub reserves the right to ask the publisher or editor of the publication to retract the paper.

(p) Every member of MADIVA will sign a personal commitment to the above, and acknowledged that breaching conditions of the use of data deliberately or negligently may result in (a) removal of access to data and/or (b) removal from MADIVA and/or (c) disciplinary action by their home organisation.

(q) Further, each participating institution in MADIVA must bind MADIVA team members according to their disciplinary codes to observe confidentiality and follow MADIVA and institutional security policies.

(r) The reporting PI will consult the other PIs and make a decision about what actions should take place. This could include reporting the event to the senior executives of the institution(s) employing the team member and to appropriate regulatory or ethical authorities.

4 Procedures for sharing of MADIVA data with external collaborators

4(1) The Hub may grant permission for use of MADIVA data to external individuals or institutions who make a request for the data. External parties will be required to submit a data request form and the Coordinating Committee will consider such requests on a case-by-case basis. The Coordinating Committee will not be considered quorate for purposes of decisions to share data unless all partners who have contributed raw data to the Hub are represented at the Coordinating Committee meeting. If there is an irresolvable dispute in the Coordinating Committee and the parties in MADIVA wish to take the matter further, the MADIVA dispute resolution procedure will apply.

4(2) In making a decision, the CC should consider appropriate benefit sharing agreements (e.g., co-authorships); the scientific quality of the proposal; quality of the team; ethical and social risks; and benefit of long-term collaborations.

4(3) Any data transfer or access must be approved by the Wits HREC.

4(4) Once the CC has made a decision to share MADIVA data with an external user, the Wits Health Consortium (on behalf of MADIVA) and the external user will enter into a Data Transfer Agreement ("DTA").

4(5) In general a DAC is preferred to a DTA.

4(6) A DTA or DAC will

(a) protect the security of the data and prohibit any attempt to reidentify the data subjects or share data with other parties;

(b) require the recipient to securely store and process the data;

(c) explicitly describe (i) what the purpose of the transfer/access is; (ii) whether further onward transmission of data is allowed; (iii) what data and/or results will be published and how and when it will be published;

(d) forbid any processing or copying of data not permitted under point 4(6c) above;

(e) permit processing of MADIVA data only to the extent and in such a manner as is reasonably necessary for the purpose for which that data has been shared;

(f) only permit employees or students of the recipient to access to shared data;

(g) require the recipient to ensure anyone accessing MADIVA data to: be under an obligation of confidentiality; be sufficiently trained in data protection legislation to understand the obligations upon the external user in relation to the MADIVA data; treat all MADIVA data as strictly confidential; disclose any MADIVA data to any third party unless such use or disclosure is expressly authorised by the DTA or required by law.

(h) where appropriate require the recipient organisation to provide a copy of any paper or publication to the Coordinating Committee for review and assent prior to publication, where MADIVA will only refuse consent to publish if (i) there is a risk to participants or a significant risk of stigmatising communities or (ii) the research publishes results not covered by the purposes for which the request was made.

(i) require the recipient to immediately report any data breaches to the PI of MADIVA;

(j) require the recipient to indemnify the Hub from any liability arising from its use or sharing of the data.

4(7) In cases where an external user releases a publication that draws on MADIVA data while violating the Data Sharing Agreement, the Hub reserves the right to ask the publisher or editor of the publication to retract the paper. When MADIVA shares data with an external party, MADIVA may charge for data abstraction costs where necessary in order to recover its costs of abstraction.

4(8) Recording and reporting: An audit trail of data access requests and approvals or rejections shall be maintained by the CC. It is important to record in a tracking log that data have been requested and, if released when, to whom and for what proposal, and if not, why not.

5 Procedures for long-term accessibility of MADIVA data and compliance with DS-I Africa policies

It is a commitment of MADIVA to make its data accessible beyond the life of the Hub to fulfil its requirements to the NIH. In order to ensure such long-term access, Data Access Committees have or will be established at Wits/Wits Health Consortium and at the APHRC.

5.1 Responsibility of the MADIVA Coordination Committee

(a) The MADIVA CC will determine whether access to any specific data presents any ethical or legal risks.

(b) Many journals require data on which research findings are based to be available on publication. The publication policy requires that the MCD specify what data may need to be published and what type of access to the data (e.g., open or controlled) will be required. Approval of the MCD by the CC constitutes approval of publication and mode of publication of the data specified.

(c) The CC may decide that other data sets will be available and when. A general guideline is that such data could be made available within 24 months after quality control has completed, with a further 12 month embargo period. We will encourage the publishing of data description articles linked to released data sets.

(d) The CC will be responsible for negotiating with APHRC and/or Wits/WHC about procedures for storing and onward sharing, which should comply with the principles in §5.3

(e) The CC is NOT responsible for responding to any particular request for sharing (since the data may need to available after the period of the MADIVA grant)

5.2 Mode of sharing

(a) Wits data will be stored in the European Genome-Phenome Archive or a South African data repository.

  (i) Any data that is sensitive, personal or for which there is a risk that participants may be re-identified or there is some other ethical risk can only be stored in a repository approved by the Wits Human Research Ethics Committee (Medical) (Medical).

  (ii) Approval or rejection of data requests will be made by a data access committee (DAC) approved by the Wits Deputy Vice-Chancellor (Research) – and where data is sensitive – by the Wits HREC (Medical).

  (iii) The Data Access Committee should contain members from the Agincourt Research Unit and the SBIMB.

  (iv) At the point of submission of data, the team which submits the data may include a set of planned analyses on that data and a timeline for those analyses.

(b) APHRC data will be stored in the APHRC microdata portal. The decision of how the data is shared and data requests must be approved by the APHRC Director of Research subject to the APHRC procedures and Kenyan ethical and regulatory policies.

(c) Aggregate open data that contains both Kenyan and South African data will be shared in a way agreed by CC. No data set that is to be shared shall contain data of individual Kenyan and South African participants.

(d) If other participating parties provide data to MADIVA (e.g., pilot projects), one of the above methods can be used. If it is not possible to use the above methods, this policy will need to be changed.

(e) In the final preparation of the files that should be shared to a repository, the lead of the DMAC and the champion of the paper or data set will jointly and individually be responsible for approving data.

(f) The DMAC is responsible for the submission of data and for technical implementation.

(g) All data sets should contain suitable metadata and be structured in such a way as assist in making the data FAIR.

(h) Any relevant ethics certificates and related information should be included in the submission with any restrictions on data use.

5.3 Conditions of sharing

(a) Any data access committee shall impose the following minimum requirements on access to sensitive data or other data that is controlled.

Data recipients[1]:

(a) shall process MADIVA data only to the extent and in such a manner as is reasonably necessary for the purpose for which that data has been shared;

(b) will be under an obligation of confidentiality;

(c) must be sufficiently trained in data protection principles to understand the obligations upon the external user in relation to the MADIVA data;

(d) treat all MADIVA data as strictly confidential, and shall not disclose any MADIVA data to any third party unless such use or disclosure is expressly authorised by each Data Access Committee, or required by law;

(e) shall only publish results of analyses in accordance with the approved purposes of sharing;

(f) shall not otherwise disclose any MADIVA data or any analytical output deriving from, or relating to, MADIVA data to any third party (including, without limitation, the media, or any subcontractors, partners, donors or affiliates) without prior written consent of each Data Access Committee;

(g) shall not attempt to identify from any dataset, and by any means whatsoever, any data subject, nor claim to have done so,

(h) obtains any ethics certifications required by the Data Access Committee;

(i) report any data breaches to the Wits Health Consortium and/or the APHRC;

(j) shall indemnify the Wits Health Consortium and/or the APHRC from any liability arising from its use or sharing of the data.

(b) In making a decision, a Data Access Committee must ensure ethical and legal requirements on sharing of data. The DAC should also take into account the following:

(a) The scientific merit of the proposal,

(b) The training (ethical and legal) of the team and the institutions where the work will be done,

[1]This would typically be done in a DTA – this documents spells out the principles and leaves the details to the DTA as Wits/WHC/APHRC have sound processes in place and we note that new draft DTAs are being produced.

(c) The inclusion of African scientists in the proposal. As a general principle, African scientists should be involved in a substantive manner in projects initiated within 3 years of data deposit.

(d) The recency of the data, and the planned anayses/timelines in terms of §5.2.

(c) Recording and reporting: An audit trail of data access requests and approvals or rejections shall be maintained by the access committee. It is important to record in a tracking log that data have been requested and, if released when, to whom and for what proposal, and if not, why not.

6 Amendments

Amendments to this document may be made by the CC after consultation with the TCPDC. Any changes should be circulated to the MADIVA consortium for at least three weeks.

A Key terms and concepts

(a) Multimorbidity: The co-existence of two or more long-term health conditions in an individual.

(b) Data Science: The interdisciplinary field that uses scientific methods, processes, algorithms, and systems to extract knowledge and insights from data.

(c) Data Integration: The process of combining and linking different data sets from various sources to create a unified dataset.

(d) Data Visualisation: The representation of data in graphical or visual formats to aid understanding and analysis.

(e) Analytic Tools: Software tools and algorithms used for data analysis and modeling.

(f) Longitudinal Data: Data collected over an extended period, allowing the study of trends and changes over time.

(g) Electronic Health Records (EHR): Digital records of patient's health information stored in electronic format.

(h) Primary Healthcare: The first level of contact with the healthcare system, providing essential care and promoting health.

(i) Polygenic Risk Scores: A numerical score that predicts an individual's risk of developing a disease based on multiple genetic factors. Precision Medicine: Medical treatments tailored to an individual's lifestyle characteristics.

(j) Public Health Interventions: Strategies and actions implemented to improve public health and prevent diseases.

(k) Sub-County-level Health Management: Health management and decision-making at a regional level smaller than a county.

(l) Training and Capacity Development: Activities aimed at improving the knowledge and skills of individuals involved in the project.

(m) Nairobi Urban Health and Demographic Surveillance System (NUHDSS): A research site run by the African Population and Health Research Center (APHRC) in Nairobi, Kenya, for tracking health and demographic data.

(n) Machine Learning Techniques: Algorithms that enable computers to learn from data and improve their performance without being explicitly programmed.

(o) Data Privacy and Security: Policies and measures to protect the confidentiality and integrity of the shared data.

(p) Data Transfer Agreement: A formal agreement outlining the terms and conditions for sharing and using data among project collaborators.