HE^{2}AT Center Data Management Plan

Document version number 2.1

Developed by C Jack, C Parker

8-15-2024

Approved by Christopher Jack____ 28/10/2024 ____________

Name Date Signature

Approved by Gueladio Cisse____ ____ 28.08.2024 ______ _________

Name Date Signature

Approved by Matthew Chersich___ ____01.09.24_________________

Name Date Signature

Approved by Sibusiswe Makhanya____ __________ ______________

Name Date Signature

Approved by Stanley Luchters____ _____28 Oct '24__________

Name Date Signature

Contents

Acronyms

Acronym	Description
AOT	Aerosol Optical Thickness
BMGFKi	Bill and Melinda Gates Foundation Ki repository
CSAG	Climate System Analysis Group
DAC	Data Access Committee
DAP	Data Analysis Platform
DMAC	Data Management and Analysis Core of the HE²AT Center
DMP	Data Management Plan
DS-I Africa	NIH Data Science Initiative Africa
DTA	Data Transfer Agreement
DUOS	Data Use Oversight System
ELSI	Ethical Legal and Social Implications Projects of DS-I Africa
FAIR	Findable, Accessible, Interoperable, and Reusable
FIPS	Federal Information Processing Standards
GCRO	Gauteng City-Region Observatory
HPC	High-Performance Computing
HSS	US Department of Human and Health Services
LDAP	Lightweight Directory Access Protocol
NDVI	Normalized Difference Vegetation Index
NIH	US National Institute of Health

NIR	Near-Infrared
NIST	National Institute of Standards and Technology
ODSP	Open Data Science Platform
PI	Principal Investigator
QoS	Quality of Service
RP1	Research Project 1 of the HE^{2}AT Center
RP2	Research Project 2 of the HE^{2}AT Center
SC	HE^{2}AT Center Steering Committee
SRTM	Shuttle Radar Topography Mission
TEC	Training and Engagement Core of the HE^{2}AT Center
TLS	Transport Layer Security
UCT	University of Cape Town
WWARN	Worldwide Antimalarial Resistance Network
sSA	sub-Saharan Africa
WHC	Wits Health Consortium Pty (Ltd)
Wits PHR	Wits Planetary Health Research a division of Wits Health Consortium (Pty) Ltd
UPGC	University Peleforo Gon Coulibaly Korhogo Côte d'Ivoire

Definitions

Term	Definition

Areal/Geospatial Socio-Economic Data

Represents socio-economic conditions such as household economic status, access to services, and dwelling type. Sourced from national census data and focused household and demographic surveys.

Bona Fide Researcher

An individual or entity engaged in legitimate scientific research to advance knowledge in health data science, operating within the ethical, legal, and professional frameworks of academic and scientific research.

CSAG GitLab

The version control and collaboration platform used by the Climate System Analysis Group (CSAG) at the University of Cape Town (UCT) for managing the HE^{2}AT Center's data processes, documenting Data Reference Syntax (DRS), and storing data management code ensures transparent, version-controlled data handling accessible to authorized team members.

Climate/Weather Data

This includes observational-based datasets such as weather station observations, satellite proxy observations, and gridded climate data produced from atmospheric re-analysis and climate simulations.

Consortium Shared Data

Data that has undergone initial processing, harmonisation and integration and includes, amongst other variables, a limited set of indirect personal identifiers that are required for the purposes of conducting the HE^{2}AT Center project analysis. This data is only shared amongst the HE^{2}AT Center Consortium partners through the Consortium Data Sharing Agreement.

Consortium Data Sharing Agreement

The Data Sharing Agreement executed between the HE^{2}AT Center Consortium.

Data Access Committee (DAC)

A committee responsible for reviewing and approving data access requests, and ensuring adherence to ethical, legal, and scientific standards.

Data Analysis Platform (DAP)

The platform used for conducting data analysis, typically including tools like JupyterHub.

Data Downloads

A modality of data sharing where researchers can download datasets to their local computing environments directly.

Data Management Plan (DMP)

A document outlining the procedures and standards for data acquisition, transfer, processing, storage, and access for the HE^{2}AT Center Project.

Data Management and Analysis Core (DMAC)

The core component of the HE^{2}AT Center responsible for overseeing data management and analysis activities.

Data Protection Legislation

Any data protection or data privacy laws as may be applicable including but not limited to POPIA, the Electronic Communications and Transactions Act 26 of 2005, the Consumer Protection Act 68 of 2008, and the General Data Protection Regulation (GDPR).

Data Provider

The party that owns, controls or otherwise possesses the rights to transfer data and is responsible for ensuring it has the legal authority to transfer such data and that such data is provided in accordance with any applicable laws, regulations or contractual obligations

Data Subject

The individuals whose personal information is captured in health datasets.

Data Transfer Agreement (DTA)

A legal document outlining the terms and conditions under which data is shared between a data provider and data recipient, addressing confidentiality, data use limitations, and compliance with relevant laws and guidelines.

Ethical Legal and Social Implications (ELSI)

Projects within DS-I Africa that focus on data science and research's ethical, legal, and social aspects.

Findable, Accessible, Interoperable, and Reusable (FAIR)

Principles that aim to improve the discovery, accessibility, interoperability, and reuse of data.

Gauteng City-Region Observatory (GCRO)

A research institute producing socio-economic data for the Gauteng City-Region, including the Quality of Life Survey.

HE^{2}AT Center

The Heat and Health in Africa Transdisciplinary Center, operating under a U54 Cooperation agreement with the NIH (2021-2026), focused on heat-health research, capacity building, and engagement

HE^{2}AT Center Consortium

The HE^{2}AT Center consortium partners funded through the HE^{2}AT Center grant jointly working on the HE^{2}AT Center Project who conduct research using the shared datasets. They adhere to the DMP data acquisition, transfer, processing, storage, and access guidelines, as well as the Consortium Data Sharing Agreement.

HE^{2}AT Center Data Management Plan The data management plan applicable to the HE^{2}AT Center Project as may be amended and updated from time to time.

HE^{2}AT Center Project *The research project titled: "Developing data science solutions to mitigate the health impacts of climate change in Africa"* and comprised of Research Project 1 and Research Project 2. .

HE^{2}AT Center Steering Committee (SC) The committee responsible for guiding the strategic direction and oversight of the HE^{2}AT Center consisting of representatives from each consortium partner as well as representatives from the NIH

Harmonization Aligning various health datasets into a unified format according to a common code book of variable names and definitions and common units and categories.

Individual Participation Data Includes data collected from previous clinical cohort and trial studies, generally considered personal data due to its association with individual medical records and health events.

Integration Aligning and integrating health and other various datasets into an integrated dataset.

Inferential Data Aggregated and anonymised data derived from analyses.

NIH Data Science Initiative (DS-I) Africa An NIH initiative aimed at enhancing data science capacity and collaboration across Africa.

Near-Infrared (NIR) A region of the infrared spectrum of light used in remote sensing applications.

Normalized Difference Vegetation Index (NDVI) An index of plant 'greenness' or photosynthetic activity.

Open Access data Research data freely available through a data repository without major restrictions.

Open Data Science Platform (ODSP) A platform facilitating the storage, retrieval, and processing of data for health research.

Operator A person who processes Personal Data for a Responsible Party in terms of a contract or mandate without coming under the direct authority of that party.

Original Study Data Raw, unprocessed Individual Participant Data collected directly from various cohort studies and clinical trials. This data is provided by the Data Provider who conducted or commissioned the relevant study and/or clinical trial.

Personal Data Any information relating to an identifiable living natural person and where it is applicable an identifiable existing juristic person.

Personally Identifiable Data Data variables that enable the identification of an individual either directly through names, ID numbers, etc., or indirectly through combining other variables such as locations (GPS, street address), age, gender, and medical information.

Principal Investigator The lead researcher responsible for the conduct of a research project.

Processing Any operation or set of operations which is performed upon Personal Data whether or not by automatic means such as collection, recording, organization, storage, adaptation or alteration, retrieval, consultation, use, disclosure by transmission, dissemination or otherwise making available, alignment or combination, blocking, erasure or destruction.

RP1 De-identified Data Data with the following information deleted; (1) information that directly identifies the Data Subject, (2) information that can be used or manipulated by a reasonably foreseeable method to identify the Data Subject or, (3) information that can be linked by a reasonably foreseeable method to other information that identifies the Data Subject.

RP2 De-identified Data Data with the following information deleted; (1) information that directly identifies the Data Subject, (2) information that can be used or manipulated by a reasonably foreseeable method to identify the Data Subject or, (3) information that can be linked by a reasonably foreseeable method to other information that identifies the Data Subject.

Re-analysis A dynamical model simulation of historical climate evolution continuously nudged by observations to provide an approximate historical representation of the climate system.

Remote Sensing Data Data derived from satellite sensors, including optical imagery and indicators of physical measures like land surface temperature, soil moisture estimates, and vegetation condition.

Research Hubs Seven NIH DS-I Africa U54 grants that are called “Research Hubs” and who contribute to and utilize shared data resources.

Research Project 1 (RP1) The first major research project of the HE^{2}AT Center Project focused on individual participant data meta-analysis of heat-health impacts. It is entitled: “Individual Participant Data meta-analysis to quantify the impact of high ambient temperatures on maternal and child health in Africa”

Research Project 2 (RP2) The second major research project of the HE^{2}AT Center Project, focusing on urban heat health impacts and Early Warning Systems.

Responsible Party A public or private body or any other person which alone or in conjunction with others determines the purpose of and means for Processing Personal Data.

Sensitive data Data that pertains to an individual's personal information, such as, but not limited to health, finances, occupation.

Shuttle Radar Topography Mission (SRTM) A mission that obtained elevation data for most of the Earth using radar interferometry.

Training and Engagement Core (TEC) The core component of the HE^{2}AT Center responsible for training and capacity-building activities.

Transport Layer Security (TLS) A cryptographic protocol designed to provide secure communication over a computer network.

US National Institute of Health (NIH) The primary agency of the United States government responsible for biomedical and public health research.

sub-Saharan Africa (sSA) A geographical region of Africa located south of the Sahara Desert.

1. Stakeholders and target audience

- **CSAG/UCT Data Management Team:** The Climate System Analysis Group (CSAG) in the Faculty of Science at the University of Cape Town (UCT) manages data storage, harmonisation, and integration processes within the HE²AT Center.
- **Climate Data Providers**: Organizations or agencies that supply data related to climate and weather, such as weather station observations, satellite data, and gridded climate data. Examples include national meteorological services and international climate data repositories like Copernicus Climate Data Store (CDS).
- **Core Data Team:** A group of vital personnel responsible for the initial processing and de-identifying of the original study data.
- **Data Access Committee (DAC):** This committee evaluates and approves requests for data access and ensures that data sharing complies with ethical, legal, and scientific standards.
- **Data Provider:** The party that owns, controls, or otherwise possesses the rights to transfer data and is responsible for ensuring it has the legal authority to transfer such data and that such data is provided in accordance with any applicable laws, regulations or contractual obligations
- **DS-I Partners:** Partners within the NIH Data Science for Health Discovery and Innovation in Africa (DS-I Africa) Initiative who contribute to and utilize the shared data resources, collaborating on data management practices to ensure alignment with the broader DS-I Africa objectives.
- **eLwazi**: As a data management platform, eLwazi facilitates the indexing, storage, and sharing of data within the consortium. It ensures that data is discoverable, accessible, interoperable, and reusable (FAIR principles).
- **Ethics Committees**: Institutional bodies that review and approve the ethical aspects of research projects, ensuring that studies comply with ethical standards and protect participants' rights and welfare.
- **Harmonization Team Members**: Researchers and data scientists within the HE²AT Center Consortium who work on harmonisation and integrating datasets into a unified format, ensuring consistency and usability for analysis.
- **Health Data Providers**: Entities responsible for collecting and providing biomedical data, including clinical trial coordinators, cohort study administrators, hospitals, and research institutions involved in health-related studies.
- **Health Experts:** Specialists in health-related fields who validate the harmonization of biomedical data and ensure the accuracy and reliability of the integrated datasets.
- **HE²AT Center Consortium**: The HE²AT Center consortium partners funded through the HE²AT Center grant jointly working on the HE²AT Center Project who conduct research using the shared datasets. They adhere to the DMP data acquisition, transfer, processing, storage, and access guidelines, as well as the Consortium Data Sharing Agreement.
- **IBM Research Africa Geospatial and Climate Analysis Team:** The HE²AT Center Consortium team at IBM Research Africa, responsible for leveraging and developing geospatial artificial intelligence tools and models in the context of the HE²AT Center. Most of the tools and models

leveraged are open source and access to IBM Research proprietary prototype tools will be made available to the HE²AT Center through relevant software licenses that facilitates access to those resources for the collaboration.

- **External *Bone Fide* Researchers:** collaborate with the HE²AT Center and utilize the integrated datasets for various research projects and analyses. They follow the data access procedures outlined in this DMP and comply with data transfer agreements, ethical guidelines, and any conditions set by the Data Access Committee.
- **Socio-Economic Data Providers:** Organizations that provide data on socio-economic conditions, such as national census bureaus, household survey agencies, and specialized observatories like the Gauteng City-Region Observatory (GCRO).

2. Background and overview of Data Management Plan

2.1. Background to the HE²AT Center

The HEat and HEalth Africa Transdisciplinary Center (HE²AT Center), is a U54 Cooperation agreement with the NIH (2021-2026). The HE²AT Center aspires to become a Center of Excellence in heat-health research, capacity building and engagement, using population health science and applying data science methodologies to improve the health of populations in Africa and beyond. The goal of the HE²AT Center is to advance the development of new health knowledge and human capacities by reusing existing data to generate and then disseminate heat-health knowledge and innovations.

RP1 description

Research project 1 is an Individual Participant Data (IPD) meta-analysis to assess the size and nature of associations between exposure to high ambient temperatures and selected health outcomes in pregnant women and children within the first two years of life. The IPD approach has not yet been employed in climate change and health research. An IPD can overcome many limitations of traditional analyses of individual datasets and biases in classic systematic review methodology. The project has systematically identified potentially eligible African cohort studies or trials through a systematic mapping of studies on pregnant women and children in Africa[1]. Data are being harmonised by re-coding raw individual participant data into a standard set of variables. Subsequently, all the individual participant's data from each eligible study will be pooled. Analyses that include a range of traditional statistical and novel machine-learning approaches will quantify associations between exposure to high temperatures and adverse maternal and child health outcomes. The study may provide robust, definitive evidence on the impacts of heat on maternal and child health and allow for estimation of the burden of rises in temperatures and other climate change manifestations on maternal and neonatal health[2].

RP2 description

Rapid growth in urban populations, geographical extent of cities and over-burdened health services in African cities, coupled with rising temperatures and Urban Heat Island phenomenon, pose significant public health challenges. High ambient temperatures cause considerable morbidity and mortality in urban areas, influenced by temperature gradients across a city, socio-environmental factors and the characteristics of the built environment. This project, conducted in Abidjan, Ivory Coast, and Johannesburg,

[1] Solarin, I., Dumbura, C., Lakhoo, D. P., Chande, K., Maimela, G., Luchters, S., & Chersich, M., for the HE²AT Center. (under review). Characteristics of longitudinal maternal health studies in sub-Saharan Africa: A systematic mapping of literature between 2012 and 2022. *International Journal of Gynecology and Obstetrics*.

[2] Lakhoo, D. P., Chersich, M. F., Jack, C., Maimela, G., Cissé, G., Solarin, I., Ebi, K. L., Chande, K. S., Dumbura, C., & Tatenda, P. (2024). Protocol of an individual participant data meta-analysis to quantify the impact of high ambient temperatures on maternal and child health in Africa (HE²AT IPD). *BMJ Open, 14*(1), e077768. https://doi.org/10.1136/bmjopen-2023-077768.

South Africa, will investigate heat exposure risks to inform development of an Early Warning System for vulnerable groups.

We will use data science methods, including natural language processing and predictive geospatial analysis, to integrate diverse data streams. Potential mediating and confounding factors, such as urban form, differentials in socio-economic status and vegetation indices, will be included in exposure-response analyses involving high-resolution climate data and health outcomes. The project aims to estimate historical and future heat hazards and develop a predictive model linking heat exposure to health outcomes under different emission and development scenarios. Health data will come from cohort and clinical trials in the target cities. Various dissemination methods for the Early Warning System will be explored, including a web-based application.[3].

2.2. Scope of Data Management Plan

This Data Management Plan (DMP) applies to all data acquired and produced as part of the HE²AT Center Project activities.

2.3. Purpose

The Data Management Plan (DMP) establishes comprehensive procedures and standards for transferring, processing, storing, and accessing data within the HE²AT Center Project. This plan ensures data integrity and security while facilitating effective data sharing and stakeholder collaboration. The DMP outlines the procedures for data access by External Researchers, ensuring that all data management activities, both inside and outside the HE²AT Center Consortium, align with the HE²AT Center's objectives and ethical commitments.

Specifically, the DMP seeks to:

- **Standardize Data Management Practices**: Define uniform procedures and standards to streamline data handling processes across the HE²AT Center Consortium, ensuring consistency and reliability.
- **Enhance Data Security and Privacy**: Implement robust measures to protect sensitive and personally identifiable data, complying with relevant legal and ethical standards.
- **Facilitate Data Sharing and Access**: Clear guidelines and DTAs promote transparency and collaboration and support data sharing within the HE²AT Center Consortium, with other DS-I Africa projects, and with External Researchers.

[3] Jack, C., Parker, C., Kouakou, Y. E., Joubert, B., McAllister, K. A., Ilias, M., Maimela, G., Chersich, M., Makhanya, S., Luchters, S., Makanga, P. T., Vos, E., Ebi, K. L., Koné, B., Waljee, A. K., Cissé, G., Tall, A., Vanga, A. F., Mahlasi, C., Dely, I. D., … Kurien, T. (2024). Leveraging data science and machine learning for urban climate adaptation in two major African cities: a HE²AT Center study protocol. *BMJ Open, 14*(6), e077529. https://doi.org/10.1136/bmjopen-2023-077529

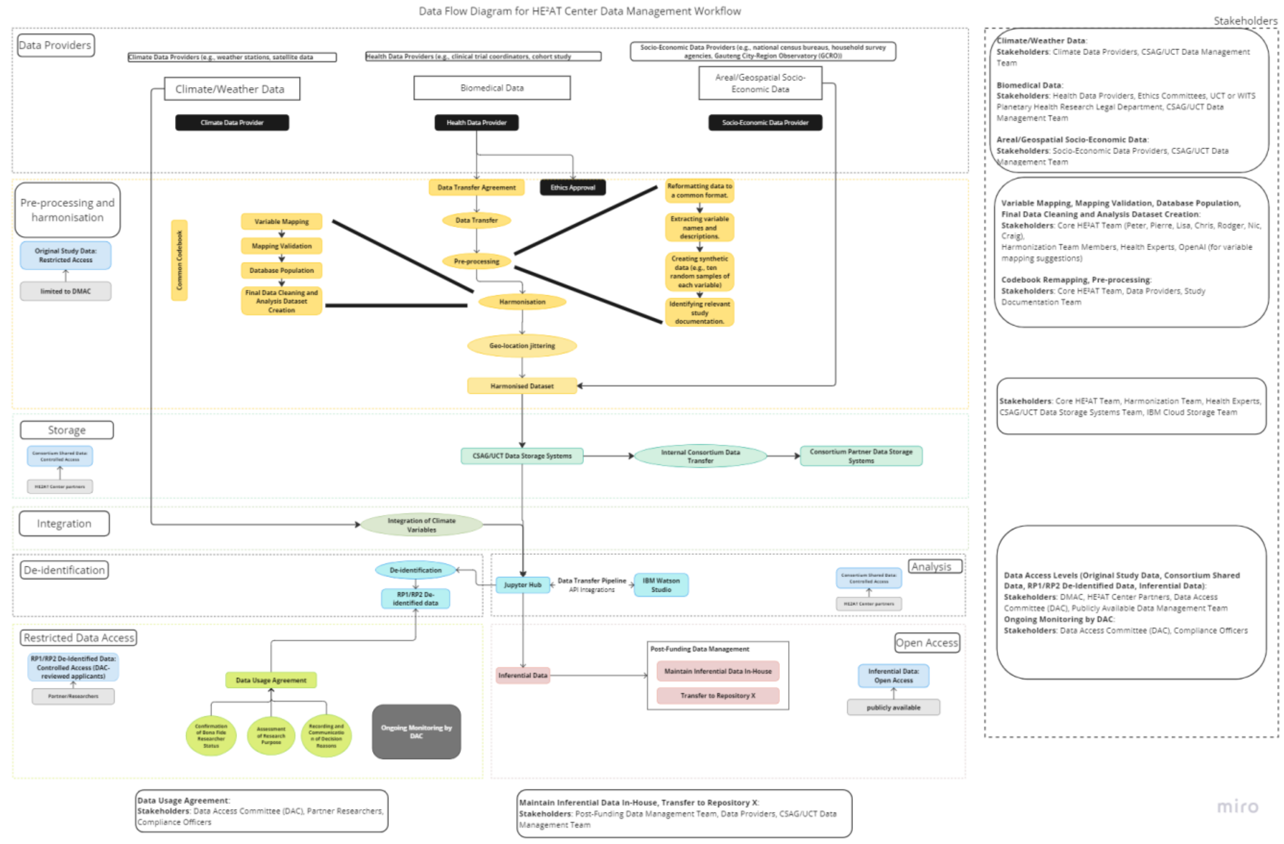
- **Support Data-Driven Research**: Provide a framework that enables researchers to efficiently access and utilise data for innovative research, contributing to the HE²AT Center's mission of advancing heat-health knowledge and interventions.
- **Ensure Compliance with Ethical Guidelines**: Maintain adherence to ethical standards and guidelines, including obtaining necessary approvals and consents for data use.

2.4. Overview of the Data Management Workflow

The HE²AT Center's data management process is structured around several critical stages, from data acquisition to sharing and open access. The following diagram (Figure 1) provides an overview of this comprehensive workflow, illustrating the relationships between data types, processing steps, and access levels.

This diagram not only serves as a guide to understanding the flow of data through various phases—such as pre-processing, harmonisation, integration, de-identification, and the transition from restricted data to open access—but also highlights the stakeholders involved at each stage. The diagram presents the framework that will be unpacked and elaborated upon throughout the rest of this document.

Each of the sections that follow addresses one part of this framework in detail, especially concerning the processes around ethical and legal standards to follow and steps to promote data accessibility and reuse. This diagram, therefore, acts as a roadmap for navigating the complexities of data management within the HE²AT Center.


Data Flow Diagram for HE²AT Center Data Management Workflow
Data Providers
Climate Data Providers (e.g., weather stations, satellite data
Health Data Providers (e.g., clinical trial coordinators, cohort study
Socio-Economic Data Providers (e.g., national census bureaus, household survey agencies, Gauteng City-Region Observatory (GCRO))
Climate/Weather Data
Biomedical Data
Areal/Geospatial Socio-Economic Data
Climate Data Provider
Health Data Provider
Socio-Economic Data Provider
Pre-processing and harmonisation
Original Study Data: Restricted Access
limited to DMAC
Common Codebook
Variable Mapping
Mapping Validation
Database Population
Final Data Cleaning and Analysis Dataset Creation
Data Transfer Agreement
Ethics Approval
Data Transfer
Pre-processing
Harmonisation
Geo-location jittering
Harmonised Dataset
Reformatting data to a common format.
Extracting variable names and descriptions.
Creating synthetic data (e.g., ten random samples of each variable)
Identifying relevant study documentation.
Storage
Consortium Shared Data: Controlled Access
HE2AT Center partners
CSAG/UCT Data Storage Systems
Internal Consortium Data Transfer
Consortium Partner Data Storage Systems
Integration
Integration of Climate Variables
De-identification
De-identification
RP1/RP2 De-identified data
Jupyter Hub
Data Transfer Pipeline
API Integrations
IBM Watson Studio
Analysis
Consortium Shared Data: Controlled Access
HE2AT Center partners
Restricted Data Access
RP1/RP2 De-identified Data: Controlled Access (DAC-reviewed applicants)
Partner/Researchers
Data Usage Agreement
Confirmation of Bona Fide Researcher Status
Assessment of Research Purpose
Recording and Communicatio n of Decision Reasons
Ongoing Monitoring by DAC
Inferential Data
Post-Funding Data Management
Maintain Inferential Data In-House
Transfer to Repository X
Open Access
Inferential Data: Open Access
publicly available
Stakeholders
miro

Climate/Weather Data:
Stakeholders: Climate Data Providers, CSAG/UCT Data Management Team

Biomedical Data:
Stakeholders: Health Data Providers, Ethics Committees, UCT or WITS Planetary Health Research Legal Department, CSAG/UCT Data Management Team

Areal/Geospatial Socio-Economic Data:
Stakeholders: Socio-Economic Data Providers, CSAG/UCT Data Management Team

Variable Mapping, Mapping Validation, Database Population, Final Data Cleaning and Analysis Dataset Creation:
Stakeholders: Core HE²AT Team (Peter, Pierre, Lisa, Chris, Rodger, Nic, Craig),
Harmonization Team Members, Health Experts, OpenAI (for variable mapping suggestions)

Codebook Remapping, Pre-processing:
Stakeholders: Core HE²AT Team, Data Providers, Study Documentation Team

Stakeholders: Core HE²AT Team, Harmonization Team, Health Experts, CSAG/UCT Data Storage Systems Team, IBM Cloud Storage Team

Data Access Levels (Original Study Data, Consortium Shared Data, RP1/RP2 De-identified Data, Inferential Data):
Stakeholders: DMAC, HE²AT Center Partners, Data Access Committee (DAC), Publicly Available Data Management Team
Ongoing Monitoring by DAC:
Stakeholders: Data Access Committee (DAC), Compliance Officers

Data Usage Agreement:
Stakeholders: Data Access Committee (DAC), Partner Researchers, Compliance Officers

Maintain Inferential Data In-House, Transfer to Repository X:
Stakeholders: Post-Funding Data Management Team, Data Providers, CSAG/UCT Data Management Team

Figure 1: HE²AT Center Data Management Workflow

3. Data types

Data acquisition is a foundational step in the HE²AT Center's research efforts, involving systematically acquiring a collection of datasets critical to achieving the project's objectives. These datasets are broadly categorised into health-related, climate/weather, and areal/geospatial socio-economic data, each discussed in turn below.

3.1. Health-related data

3.1.1. Summary of data categories

The categorisation of health data within the HE²AT Center ensures that the HE²AT Center adheres to ethical and legal standards while facilitating collaborative research among the HE²AT Center Consortium and External Researchers with controlled access. Health data is classified into four distinct forms, ranging from original, individual-level data, aggregated to inferential data. This structured approach aims to maintain data privacy and protection throughout the HE²AT Center Project lifecycle.

In all activities involving health data, the HE²AT Center Consortium contractually undertakes not to use the data to attempt to re-identify any Data Subjects.

3.1.2. Original Study Data

This category includes original, unprocessed health data collected directly through the various previously completed cohort studies and clinical trials that fulfil the study eligibility criteria. Ownership of this data is retained by the Data Provider who conducted or commissioned the studies and/or clinical trial. The Original Study Data is acquired on the basis of a Data Transfer Agreement, and is retained for a period of five years after the completion of the HE²AT Center Project for the purposes of concluding and correcting any analysis and publications resulting from the Data. Any retention of Data after this five-year period will be further agreed with the Data Provider. The Data provider may elect to terminate the Data Transfer Agreement and data retention will then be dealt with in a manner requested by the Data Provider.

The Original Study Data is held under strong access control on servers hosted at UCT and access to this data is restricted to a small team of data managers within the Core Data Team (see roles and responsibilities section for more detail). The Original Study Data has all direct identifiers removed prior to transfer to the remainder of the HE²AT Center Consortium (e.g., names and contact numbers), but may still contain indirect identifiers such as dates and geolocation information.

3.1.3. Consortium-shared data

Once the **Original Study Data** is processed, harmonised, and integrated, it becomes **Consortium Shared Data**. The data includes, amongst other variables, a limited set of indirect identifiers, such as absolute dates and geolocation, required for conducting the **RP1/RP2 study analyses**.

The data may be transferred between **HE²AT Center Consortium** members to conduct the **RP1/RP2 study analyses**. Only the following **HE²AT Center Consortium** members will have access to the **Consortium Shared Data**: **WHC (South Africa)**, **UCT (South Africa)**, **IBM Research Africa (US)**, **University of Peleforo Gon Coulibaly (Côte d'Ivoire)**, and **CeSHHAR (Zimbabwe)**. Access may be extended to new partners who join the **HE²AT Center Consortium** and adhere to the conditions set out in the **Consortium Data Sharing Agreement** and this **Data Management Plan (DMP)**. The **Data Providers** will be given written notice of any new member(s) joining the **HE²AT Center Consortium**.

Unless **Data Providers** state otherwise in the agreed **Data Transfer Agreement (DTA)**, the **HE²AT Center** establishes ownership of this **Consortium Shared Data** and can make decisions regarding data usage at this stage. Where possible, requests from **Data Providers** for alternative arrangements to those described here will be accommodated. This ownership stems from the **Core Data Team** having performed significant **initial processing**, **harmonisation**, and **integration** work on the **Original Study Data** once made available. The **Consortium Shared Data** is retained indefinitely unless agreed otherwise in the DTA with the **Data Provider**. Access to the **Consortium Shared Data** is granted solely for the purposes of the **RP1/RP2 studies**.

Several steps have been taken to reduce the risk of identifiability. For example, the date of birth will be removed for pregnant or postpartum women in **RP1** and for adolescents and adults in **RP2**. For these groups, age will be reported in years (as a whole number). The date of birth of the infant or of clinical events will be retained in this **Consortium Shared Data**, as it is required to link climate/weather and areal/geospatial socio-economic data. To minimise the risk of identifiability, geographic data will be jittered or aggregated as described in **Section 8: Deidentification**.

The final data cleaning step is essential for ensuring the quality and reliability of the **Consortium Shared Data**. During this stage, the **Core Data Team** reviews the dataset to confirm that all values fall within expected ranges and that the data is free from errors or inconsistencies. Specific provisions for identifying twins and multiple pregnancies are critical for accurate analysis in maternal health studies. Additionally, the **Core Data Team** works to identify and resolve any duplication errors that may have arisen during the data harmonisation process.

A sanity check ensures data integrity as part of this final cleaning process. This involves cross-referencing the dataset with the original sources and performing logical checks to verify that the data is consistent and accurate. The goal is to confirm that the **Consortium Shared Data** is ready for analysis and that no significant errors remain.

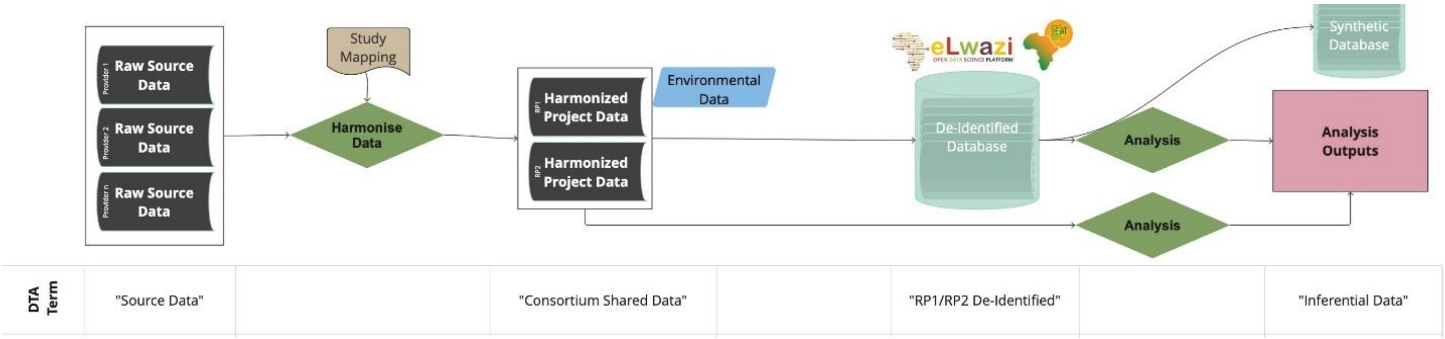
3.1.4. RP1/RP2 De-identified data

Data that has been further processed and de-identified falls into this category. The HE²AT Center owns this de-identified data. The de-identified data is retained indefinitely unless agreed otherwise in the DTA with the Data Provider. Where possible, requests from Data Providers for an arrangement other than those described here shall be accommodated. . Access is granted to External Researchers outside the HE²AT Center Consortium who have met specific conditions and requirements set by the HE²AT Center Data Access Committee. The de-identified data may be harmonised health data, or harmonised health data integrated with climate and other environmental data, depending on available resources and type of data request.

The HE²AT Center will apply the principles of de-identification through two complementary approaches. The Safe Harbour approach, is utilised where feasible, where 18 identifying variables are removed, including all dates, and high-resolution geolocation information (please add a footnote: U.S. Department of Health and Human Services. (2012)[4]. Additionally, the expert determination approach, which relies on an external expert to certify the minimal risk of re-identification, is used in conjunction with, and in certain cases, in lieu of the Safe Harbour approach, where appropriate. This data type is categorised as de-identified data under POPIA, meaning it has been thoroughly de-identified to prevent any reasonable possibility of re-identification. The identifiers in this dataset are limited to health variables without any direct or indirect personal identifiers. For example, age data for pregnant or postpartum women in RP1, and for adolescents and adults in RP2 will be reported in five-year age bands rather than as age whole numbers, and the geographic data will be reported in larger units, such as cities or districts.

3.1.5. Inferential data

The final category is Inferential Data, which is aggregated or synthetic data derived from the analysis of the preceding data categories. The HE²AT Center owns the Inferential Data and will retain this data indefinitely. Inferential Data is made available for open access to support broader research initiatives. Aggregated anonymous and synthetic data are classified as de-identified under POPIA, ensuring that



[4] Guidance regarding methods for de-identification of protected health information by the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule. Available at: https://www.hhs.gov/sites/default/files/ocr/privacy/hipaa/understanding/coveredentities/De-identification/hhs_deid_guidance.pdf).

individual privacy is fully protected. Aggregated data contains no individual-level records from Data Subjects and no direct or indirect identifiers, making it impossible to re-identify any individuals. Synthetic data is, by definition, generated in a random manner and possesses only the statistical properties of the underlying data with no link to the individual data points, thus rendering it impossible to re-identify any individuals.

Figure 2: HE^{2}AT Center Data Management Data Categories

3.2. Climate/weather data

Climate and weather data are critical components of the **HE^{2}AT Center's** research, providing the environmental context for understanding heat exposure and its impact on health outcomes. These data can be broadly categorised into three types:

Observational data: This includes data collected from weather stations located at ground level. Observational datasets provide real-time or historical measurements of various meteorological parameters, such as temperature, precipitation, humidity, and wind speed. These datasets are often considered the most accurate for specific locations and serve as the foundation for validating and calibrating other types of climate data. Examples of sources for observational data include national meteorological services and local weather stations.

Remote sensing data: These datasets are collected via satellite-based sensors that capture a wide range of environmental variables. Remote sensing allows for large-scale and continuous monitoring of climate variables such as land surface temperature, soil moisture, vegetation condition, and cloud cover. Satellite-derived data offer broader geographic coverage and enable the analysis of areas where ground-based observations are sparse or unavailable. Common remote sensing datasets used include those from satellites like **MODIS**, **Landsat**, and the **Sentinel missions**.

Re-analysis data: Re-analysis datasets integrate observational and remote sensing data into numerical models, producing a comprehensive and continuous representation of past and present climate conditions. These models use a combination of historical measurements from weather stations, satellites, and other sources to provide spatially and temporally complete datasets. Re-analysis products are particularly valuable for studying climate trends over time and for filling in gaps in observational records. Examples of widely used re-analysis datasets include **ERA5** and **MERRA-2**.

Each of these data types provides unique insights into the climate variables relevant to the **HE^{2}AT Center's** research. Together, they allow for comprehensive analyses of heat exposure and its effects on health outcomes in urban environments.

All climate related-data will be accessed through open data repositories, such as the Copernicus Climate Data Store (CDS), Earth System Grid Federation and Sentinel data systems. The data will be stored on IBM Research Africa and CSAG/UCT systems, with CSAG/UCT responsible for managing and updating the relevant data indexes. These datasets also follow open data policies, typically requiring citation when used for non-commercial research.

3.3. Areal/geospatial socio-economic data

These data represent measures of socio-economic and related conditions, such as household economic status, access to services such as water and sanitation, and dwelling type. Typical sources include national census data and more focused household and demographic survey data.

Socio-economic data will be sourced from open and restricted-access repositories (e.g., South African census data and GCRO Quality of Life Surveys). Primary copies will be indexed and stored on CSAG/UCT data storage. The IBM Research Africa team will acquire and have access to some of the geo-reference Quality of Life survey data from the GCRO and will undertake processing of this data from their own internal computing systems. Outputs and indices calculated from the data by IBM Research Africa will be shared with the rest of the HE²AT Center Consortium through the HE²AT Center's DAC at UCT.

South African census data is already available through the UCT DataFirst data repository. GCRO Quality of Life Survey data is available through the GCRO open data platform, which directs queries to GCRO.

South African census data is aggregated into small areas and does not constitute personally identifiable data. Likewise, GCRO Quality of Life survey data is aggregated to small areas and does not constitute personally identifiable sensitive data.

4. Data transfer, and ethics approvals and notification

In alignment with the HE²AT Center's commitment to ethical conduct across all research activities, the ethics approval and notification process, and Data Transfer Agreement (DTA) are essential for securing and maintaining oversight of studies that contribute data to the HE²AT Center Project.

The RP1 and RP2 study activities have been reviewed and approved by the Wits Human Ethics Committee, with Ethics Reference Numbers of **220605** for RP1 and **220606** for RP2. As data from additional studies are received, the HE²AT Center notifies the relevant Ethics Committee of such and notes any concerns with a study, if relevant.

4.1. Data Transfer Agreement:

The DTA between each Data Provider and UCT or WHC will outline the terms for transferring and processing health data. Where WHC is the Data Recipient, the Original Study Data is transferred and received from the Data Provider directly by the Core Data Team on UCT Data servers. The DTA must align with applicable national or international health data-sharing legislation. The DTA aims primarily to ensure that the handling, storage, and transfer of data adhere to relevant legislation and ethical standards and protects the rights of study participants and Data Providers.

The Wits Human Research Ethics Committee (HREC) (Medical) is notified of each new study that contributes health data. The HE²AT Center team reviews the consent and ethic approvals of each of the studies that agree to share data. This section below outlines the steps and considerations relating to ethical procedures and these notifications.

4.2. Ethics committee notification for new studies:

After the data transfer process has begun, additional steps are required for the ethics committee to monitor compliance with the agreed-upon ethical standards on an ongoing basis. This includes reviewing any concerns the Ethics Committee raises, ensuring that using Original Study Data for the intended analyses remains ethically sound, and appraising progress and any concerns with the study.

Taken together, all the study procedures around ethics ensure that the research activities are conducted ethically, safeguarding participants' rights and maintaining the highest standards of research integrity.

For every new study that contributes data to the HE²AT Center, the Wits Human Research Ethics Committee (HREC) (Medical) will be notified in writing. Notifications detail the study's name, acronym, contact details of the data owners, and relevant ethics approval information, including consent parameters.

Notifications occur on a six-monthly basis in RP1, and in real-time with RP2 given that the studies in RP2 were done at WHC, and all studies had already received approval from the University ethics committee who also serve as the ethics committee for the HE²AT Center (notification template can be found in Annex 2).

5. Pre-processing and harmonisation of health data

Data harmonisation is critical for integrating diverse health datasets into a unified format, enabling comprehensive analyses across various data sources.

5.1. Pre-processing

The Core Data Team, a small group of named personnel responsible for the initial handling of the Original Study Data, manages the pre-processing stage. This team has exclusive authorisation to access incoming data securely stored on UCT data servers. The **Core Data Team** is responsible for preparing the data for further analysis by other HE²AT Center Consortium members. First, they reformat the **Original Study Data** into standardised formats, such as CSV or JSON, ensuring compatibility with various tools and systems. This step follows established guidelines like **Open Data Standards** or the **OMOP Common Data Model**, which help to promote consistency and interoperability across datasets.

Once the data is in a standardized format, the **Core Data Team** extracts and labels key variables, ensuring each variable is named and described consistently. They align these variables with ontology frameworks like **NCIT**, **SNOMED CT**, or **ICD-10**, making the metadata easier to integrate with other datasets during the harmonisation process led by the **Harmonization Team Members**.

Additionally, the **Core Data Team** generates synthetic data from the **Original Study Data** to test data integrity and ensure the harmonisation process can proceed without exposing the real data. This synthetic data simulates the characteristics of the original data, allowing potential recoding or cleaning processes to be tested safely.

Lastly, the **Core Data Team** reviews key documentation related to the data, such as study protocols and codebooks. This documentation is essential for providing context to the **Harmonization Team Members** during later data integration and analysis stages. Although the **Original Study Data** is not yet encrypted, access is strictly limited to the **Core Data Team**, who ensure the confidentiality of the data and perform regular backups to a secure, encrypted system.

5.2. Variable mapping

Once the pre-processing is completed by the **Core Data Team**, the **Harmonization Team Members**, who are researchers and data scientists from the **HE²AT Center Consortium**, take responsibility for the next stage: variable mapping. This involves mapping the variables from the **Original Study Data** to a standardized set of ontologies, such as **NCIT**, **SNOMED CT**, or **ICD-10**.

The **Harmonization Team Members** start by working with synthetic data created during the pre-processing stage. This allows them to evaluate the accuracy of the mapping process by identifying where the synthetic data does not align with expected values. The team also relies heavily on metadata and study documentation provided by the **Core Data Team** during pre-processing. This ensures that they fully understand the variables before mapping them to the appropriate ontology.

To enhance the efficiency of this process, the **Harmonization Team Members** use AI tools, such as **OpenAI's language models (LLMs)**, to generate descriptions of variables and suggest mappings to standardized ontologies. These suggestions are then reviewed and refined by the **Harmonization Team Members** in collaboration with the **Core Data Team** to ensure accuracy and alignment with the project's objectives.

Throughout this process, the **Harmonization Team Members** document every mapping decision, including the rationale behind their choices. This documentation is essential for transparency and serves as a reference for future users of the data.

5.3. Mapping validation

The validation process is led by the **Harmonization Team Members**, with close collaboration from the **Core Data Team**. During this stage, the **Harmonization Team Members** revisit the mapped variables and compare them with the original data provided by the **Core Data Team** to ensure consistency and accuracy. The **Core Data Team** plays a supportive role, helping to cross-check data integrity and flag any discrepancies or errors that may have arisen during the variable mapping.

An additional layer of review is conducted by a health expert, who assesses whether the mapped variables align with the health ontology frameworks such as **SNOMED CT** or **ICD-10**. This expert review is crucial to ensure that the data remains clinically relevant and useful for analysis across the project's research objectives.

After validation by the health expert, the **Harmonization Team Members**, in collaboration with the **Core Data Team**, perform a final check of the mapped data. This thorough review ensures that all mappings are correct and that the data transformations have been applied appropriately. Once these checks are complete, the data is ready for further transformation into **De-identified Data** if required.

To maintain transparency, all versions of mappings and data transformations are controlled through **CSAG's GitLab** system. This version control ensures that any changes to the data are documented, and previous versions can be restored if necessary.

5.4. Database population

Once mapping and validation are complete, the **Core Data Team** takes responsibility for transforming the validated data into a harmonized format, ready for broader use by the **HE²AT Center Consortium**. This stage involves the population of the final database, where the **Core Data Team** applies the validated mappings and transformations to the **Original Study Data**, turning it into **Consortium Shared Data**.

At this point, additional de-identification steps are taken by the **Core Data Team** to further anonymize the dataset. Any residual personal identifiers are removed or generalized, with location data being aggregated to broader geographic levels to reduce the risk of re-identification. These steps ensure that the data can be shared safely among consortium members without compromising participant confidentiality.

Once the database is fully populated and de-identified, it is made available to approved **HE²AT Center Consortium** partners. This database serves as the primary resource for conducting research, allowing for the integration of health, climate, and socio-economic data. The **DMAC** ensures that all access and use of the database comply with relevant data protection regulations, including **POPIA**.

5.5. RP1/RP2 De-Identified Dataset Creation

The final stage involves creating a de-identified dataset that can be shared with External Bone Fide Researchers. The Core Data Team continues with the de-identification process, ensuring that the dataset meets the highest privacy standards. The Safe Harbour method is supplemented by expert determination, supervised by the Data Access Committee, to ensure that the dataset complies with the HE²AT Center's ethical and legal guidelines. Once these steps are complete, the dataset can be shared under the terms outlined in Section 8.

6. Integration and analysis interfaces

6.1. Integration of climate and socio-economic variables

Integrating climate and socio-economic variables within the HE²AT Center's data management workflow pulls relevant variables and indices from pre-existing non-health related datasets for the analysis period. The steps involved are as follows:

6.2. Sourcing pre-processed data

Climate and socio-economic data are sourced from previously cleaned and harmonised datasets, such as those available through the Climate System Analysis Group (CSAG) at UCT and national data repositories (e.g., census data, GCRO Quality of Life Surveys). These datasets have already undergone rigorous quality checks, reducing the need for extensive pre-processing at this stage.

6.3. Automated data retrieval

For climate data, a script-based system automates the retrieval of relevant variables and indices from the CSAG system. These scripts are designed to pull data specific to the analysis period, ensuring the dataset is tailored to the study's needs. The retrieval process may include variables such as temperature, precipitation, humidity, and indices like heat waves or drought conditions.

Similarly, socio-economic data is accessed through predefined queries that extract relevant indicators for the analysis period. These indicators may include household economic status, access to services, and other socio-demographic factors.

6.4. Integration into broader dataset

Once retrieved, the climate, socio-economic variables, and health data are integrated into the broader dataset. This integration occurs during the **Integration and Analysis** step, where the different data types are aligned, based on common temporal and spatial attributes. The integration process is relatively straightforward, leveraging the pre-existing alignment of these datasets to minimise the need for additional harmonisation. Integration requires the use of indirect identifiers in the Consortium Shared Data, in particular geolocation information and dates (e.g. date of birth) in order to align the non-health data temporally and spatially with the health data. These indirect identifiers are not carried through into the de-identified dataset ensuring that it is fully de-identified.

6.5. Collaboration with CSAG/UCT

The CSAG team at UCT plays a crucial role in managing and updating the climate data system, ensuring that the variables and indices used in the analysis are up-to-date and relevant. This collaboration aims to ensure that the data retrieval process is seamless and that the analysis is grounded in the latest climate science.

6.6. Alignment with analysis objectives

Finally, the integrated dataset is prepared for analysis, with the climate and socio-economic variables aligned to match the study's objectives. This step ensures the data is ready for statistical and modelling exercises exploring the relationships between climate, socio-economic conditions, and health outcomes.

7. Data analysis platform

Data analysis involving the Consortium Shared Data and RP1/RP2 de-identified data will be facilitated through the CSAG/UCT Jupyter Hub platform, providing robust and scalable environments for processing and analysing the HE²AT Center datasets.

Jupyter Hub is a collaborative, web-based Python coding environment that allows analysts to develop and execute analysis code using a browser interface. While web based, access is controlled through username and password. Key technical details include:

- **Web-Based Platform**: Researchers can access Jupyter Hub securely through a web browser, providing a user-friendly interface for coding and data analysis.

- **Python Environment**: This environment supports the development of analysis code in Python, leveraging a wide range of libraries and frameworks for data science, machine learning, and statistical analysis.

- **CSAG High-Performance Computing (HPC) Integration**: Analysis code executed in Jupyter Hub runs on the CSAG HPC platform, providing high computational power for processing large datasets.

- **Data Accessibility**: Consortium Shared Data, RP1/RP2 de-identified data, climate, and socio-economic datasets stored on CSAG storage servers are directly accessible within the Jupyter Hub environment, allowing seamless data retrieval and manipulation.

- **Collaboration**: Jupyter Hub supports collaborative work, enabling multiple analysts to share and work on the same notebooks, fostering teamwork and knowledge sharing.

8. De-identification

Personal information as it pertains to POPIA, can be considered de-identified provided the stipulated de-identification process is undertaken. To ‘de-identify’’, in relation to personal information of a data subject, means to delete any information that—

(a) identifies the data subject;

(b) can be used or manipulated by a reasonably foreseeable method to identify the data subject; or

(c) can be linked by a reasonably foreseeable method to other information that identifies the data subject.

There exists no clear guidance on reasonably foreseeable methods of re-identification, and thus it can be considered that the identifiability of personal data can be considered on a spectrum of risk of re-identification. Although the Original Study Data and Consortium-Shared Data have minimal risk of re-identification, as well as contractual limits on any attempts to re-identify any individuals, we commit to further reducing the risk of re-identification, in accordance with POPIA Section 10, in the context of sharing information outside the **Core Data Team** when processing the Original Study Data, or the broader **HE^{2}AT Center Consortium** when producing the RP1/RP2 De-Identified Datasets.

POPIA Section 10 prescribes the principle of “Minimality”, which means that only information relevant to the purpose of the study should be processed. Where personal information is acquired that is required to fulfil the research purposes described by the relevant research project protocols, de-identification will be implemented according to the following steps, which are guided by US Department of Human and Health Services (HSS)[5] guidelines and informed by the findings in Zandbergen’s 2014 review on geographic masking strategies.[6]

8.1. Safe Harbour and/or expert determination

We apply the principles of de-identification through two complementary approaches, for production of the RP1/RP2 de-identified datasets described above. The Safe Harbour approach is utilised where feasible, where 18 identifying variables are removed, including all dates, and high-resolution geolocation information. Additionally, the expert determination approach, which relies on an external expert to certify the minimal risk of re-identification, is used in conjunction with, and in certain cases, in lieu of the Safe Harbour approach, as appropriate. External expert advice and inputs from the Data Access Committee is sought to determine risk of re-identification before sharing with external parties. For example, sensitive dates will be aggregated to calendar years as whole numbers, and ages will be reported in five-year age bands, rather than as whole numbers, to further reduce the risk of re-identification.

8.2. Geographic aggregation

Street addresses may be aggregated into larger geographical regions to prevent the derivation of individual residential locations. Population density provides a good guide to the spatial granularity required. For instance, in RP2, where high spatial granularity is necessary to map urban heat-health outcomes, consortium shared data will be aggregated at the level of census small areas or wards with spatial scales of 2 to 5 km which would typically expand the population of potential data subjects into the range of 1000s of individuals.

Larger geographical areas will be used in the aggregation process if an area has a low population density or contains sensitive locations that might make identification easier. For example, in sparsely populated regions, aggregation might occur at a municipal or district level instead of a smaller area like a ward. This ensures that even in areas with fewer individuals, privacy is maintained by preventing identifying any individual within the dataset. The aggregation process will also account for the number of records that map to the same geographical area, adjusting the aggregation level accordingly to ensure privacy is preserved.

[5] https://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/index.html#protected

[6] https://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/index.html

8.3. Location jittering

Latitude/longitude coordinates may be "jittered" by adding random values to each coordinate to obscure the exact location while retaining sufficient geographical information to support analysis. As detailed by Zandbergen (2014)[7], jittering can involve various methods:

One method is **Random Direction and Fixed Radius**, where points are displaced randomly within a fixed radius around the original location. Another method, **Random Perturbation within a Circle**, places locations within a circular area with displacement following a uniform or normal distribution. **Gaussian Displacement** involves random direction but with distances following a Gaussian distribution, adjusted based on local population density. **Donut Masking** sets minimum and maximum displacement levels, ensuring locations are neither too close nor too far from the original points. Finally, **Bimodal Gaussian Displacement** is a variation of Gaussian masking, achieving effects similar to donut masking but with less uniform placement probability.

[7] Zandbergen, P. A. (2014). Ensuring confidentiality of geocoded health data: Assessing geographic masking strategies for individual-level data. *Advances in Medicine, 2014*, Article 567049. https://doi.org/10.1155/2014/567049

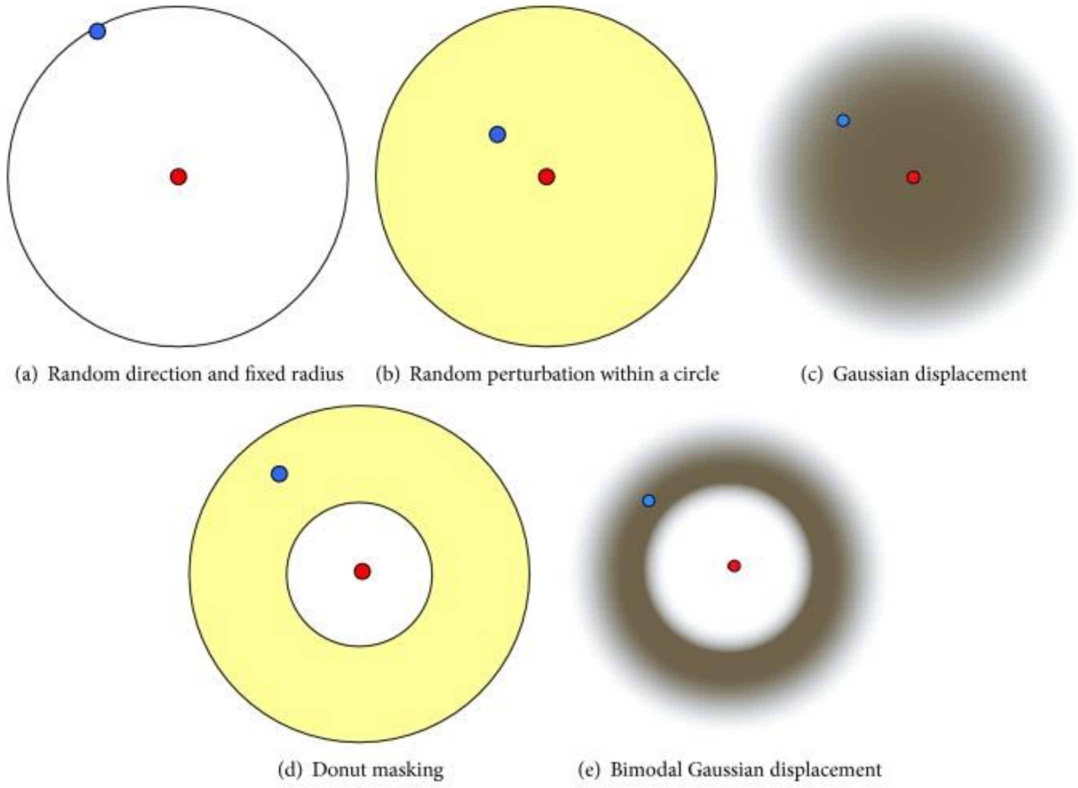

(a) Random direction and fixed radius
(b) Random perturbation within a circle
(c) Gaussian displacement
(d) Donut masking
(e) Bimodal Gaussian displacement

Figure 3: Geographic Masking Techniques: Different geographic masking techniques as described by Zandbergen (2014). (a) Random direction and fixed radius; (b) Random perturbation within a circle; (c) Gaussian displacement; (d) Donut masking; (e) Bimodal Gaussian displacement.

Jittering is applied when finer spatial detail is necessary, but privacy must still be protected, such as RP2, where urban heat-health outcomes are analysed at a high-resolution (e.g., census small areas or wards).

Gaussian displacement jittering is specifically used when more precise geographic information is required, balancing data utility and privacy by adjusting displacement based on population density.

The risk of re-identification will be quantified using spatial k-anonymity metrics, as described by Zandbergen (2014)[8]. This involves ensuring that each masked location is indistinguishable from at least k-1 other locations within a specified distance. The displacement required for adequate masking will be inversely proportional to the local population density to maintain high spatial k-anonymity.

[8] Zandbergen, P.A., *Ensuring Confidentiality of Geocoded Health Data: Assessing Geographic Masking Strategies for Individual-Level Data.* Adv Med, 2014. **2014**: p. 567049.

8.4. Expert review and risk assessment

Geo-location masking/jittering and aggregation techniques will be reviewed through expert determination involving UCT, IBM, and NIH experts. This process will involve assessing the risk of re-identification and ensuring that the applied techniques sufficiently protect participant confidentiality while maintaining the integrity of spatial analyses.

In summary, by incorporating these enhanced de-identification techniques, we aim to ensure compliance with POPIA, protect participant privacy, and maintain the data's utility for research purposes.[9]

9. Data sharing

According to UCT's Research Data Management Policy, "publicly funded research data are a public good, produced in the public interest, which should be made openly available with as few restrictions as possible in a timely and responsible manner." Data is, therefore, open by default and closed by exception (e.g., privately funded research or research with commercialisation possibilities).

9.1. Restrictions to Data Sharing

According to Section 4.6 of the UCT Research Data Management policy: *"[n]ecessary constraints on the availability of data include the protection of personal data; the protection of intellectual property; the protection of commercial interests of project partners; and security concerns."*

9.2. Discoverability

The HE²AT Center and DMAC will implement FAIR principles to ensure that:

- **Findability:** Data will be discoverable through publicly accessible and searchable metadata indexes. The DS-I Africa ODSP and UCT's ZivaHub repository both offer platforms for metadata searching.

- **Accessibility:** De-Identified data will be accessible via a data access request to the DAC, which, if approved, will require a Data Transfer Agreement.

- **Interoperability:** Adherence to established data and metadata standards will ensure data interoperability (as outlined earlier in the document).

[9] https://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification

- **Reusability:** Rigorous data documentation will support reuse, including limitations and guidance for responsible reuse.

9.3. Levels of Data Access

The HE²AT Center categorises data into distinct access levels, each governed by specific rules and protocols to ensure protection and ethical use throughout the data lifecycle. These access levels include, in summary (see section above for more details):

1. **Original Study Data:** Raw, unprocessed health data collected directly from cohort studies and clinical trials. Access is restricted to authorised personnel in the **Core Data Team** and retained for five years following the **HE²AT Center** Project completion. A Data Provider may elect to terminate the DTA at any point and request data to be returned or destroyed.

2. **Consortium Shared Data:** Data that has undergone initial processing, harmonization and integration, shared amongst the HE²AT Center Consortium . This data has been significantly altered, including applying privacy protections like initial de-identification in accordance with the principle of minimality.

3. **RP1/RP2 De-identified Data:** Data that has been further processed and de-identified to prevent re-identification. Available to External Researchers under conditions set by the Data Access Committee (DAC), this data adheres to strict privacy guidelines.

4. **Inferential Data:** Aggregated and anonymised data derived from analyses. This data is made available for open access and has no direct or indirect identifiers, ensuring complete confidentiality.

9.4. Procedure for Making RP1/RP2 De-identified Data Available to Bona Fide Researchers

The procedures for making RP1/RP2 De-identified Data available to qualifying External Researchers are outlined below. These procedures ensure compliance with ethical, legal, and scientific standards, while maintaining data integrity and security.

Detailed governance and oversight of these procedures, including DAC responsibilities, are covered in Annex 4, which contains the full terms of reference for the DAC.

1. **The Data Access Request Form:** This form must be completed by those requesting access to RP1/RP2 De-identified Data and includes:

- **Applicant Information:** Details about the applicant and their institution, including name, address, and the nature of the research activities.

- **Consortium Membership Status:** Indication of whether the applicant is a member of the DSI-Africa Consortium. Non-members must provide detailed information about their organizational affiliation.

- **Dataset Identification:** Specification of the datasets for which access is sought.

- **Research Purpose:** A proposal outlining the intended research, including objectives and significance.

- **Data Sharing Modality Preference:** Specification of whether data access is sought through download, with a justification based on the research and data sensitivity.

- **Ethics Approval:** Proof of ethics approval from the applicant's institutional ethics committee.

- **Data Protection Measures:** A detailed description of statutory, organizational, and technical measures in place at the receiving institution to safeguard the security of the data.

2. **Preliminary Screening:** Upon submission of the Data Request Form, the request undergoes preliminary screening by the HE^{2}AT Center SteerCo to confirm completeness and basic compliance with the **HE^{2}AT Center** Consortium's requirements as well as resources available in the HEAT Center for preparing the De-Identified database.

3. **Review by Data Access Committee:** The DAC evaluates the request based on criteria such as potential privacy risks, and the avoidance of overlap with ongoing research. The full governance procedures are detailed in Annexe 4.

4. **Recording and Communication of Decision:** The DAC records the reasons for its decision (approval, conditional approval, or denial). This documentation ensures transparency and provides valuable feedback to applicants. Approved requests will include detailed instructions for accessing the data and any conditions the applicant must meet.

5. **Data Transfer:** The data will be transferred upon approval and fulfilment of all conditions, including ethics approval.

6. **Ongoing Monitoring:** The DAC will continue to monitor compliance with the terms of the Data Transfer Agreement, including periodic reviews and audits, if necessary.

10. POPIA compliance and protection of personal information

The use of health datasets requires careful consideration of data security and confidentiality, guided by relevant legislation specific to each dataset, including country-specific laws on personal/sensitive data and cross-border data transfer. The DMAC manages the development and negotiation of these DTAs in conjunction with research projects, as they are the primary interfaces with the data sources.

The Protection of Personal Information Act (POPIA) of South Africa (2013) regulates the processing of personal information, providing a legal basis for its use in scientific research. POPIA works alongside other South African legislation, such as the Constitution, the National Health Act No. 61 of 2003, and the Department of Health guidelines on Ethics in Health Research (2015). The law offering the most comprehensive protection for individuals' rights takes precedence.

POPIA’s Section 6 states that the Act does not apply if personal information has been de-identified to the extent that re-identification is virtually impossible. Many health databases for the HE²AT Project will meet this criterion. For those that do not, the following sections of POPIA provide a basis for processing health data:

- **Section 10 (Minimality):** Personal information may only be processed if it is adequate, relevant, and not excessive for the purpose of the research.
- **Section 15(1) (Further Processing):** Further processing of personal information must be aligned with the purpose of its collection.
- **Section 15(3)(e) (Research Exception):** Allows processing for historical, statistical, and research purposes, regardless of the original purpose of collection. This is crucial for HE²AT, as health datasets were collected before the project began.
- **Section 18(1) (Notification):** Requires informing data subjects about the processing of their personal information.
- **Section 18(4)(f) (Research Exception):** Provides an exemption for informing data subjects if the information is used for historical, statistical, or research purposes.
- **Section 14(2) (Retention of Records):** Allows retention of personal information for research as long as safeguards prevent its use for other purposes.
- **Section 16 (Information Quality):** Mandates reasonable measures to maintain the accuracy and quality of the data.
- **Section 17 (Documentation):** Requires clear documentation of all processing activities.

- **Section 19 (Security Safeguards):** Requires security measures to prevent unlawful access to or processing of personal information.
- **Section 20 (Processing by Operators):** Specifies requirements for individuals processing personal information. This includes maintaining a continually updated list of authorized personnel with restricted access to personal information through passwords and other security measures.
- **Section 21 (Operator Contracts):** Requires a written contract between the responsible party and operators implementing processing. This contract mandates that operators inform the responsible party if unauthorized access to personal information is suspected.[10]

[10] Republic of South Africa. (2013). *Protection of Personal Information Act 4 of 2013*. Government Gazette No. 37067. Retrieved from https://www.gov.za/documents/protection-personal-information-act

11. Governance and compliance

The HE²AT Center has established a comprehensive governance and compliance framework to ensure that all data management activities adhere to ethical standards, legal requirements, and best practices. This framework is designed to protect participant privacy, ensure data quality, and facilitate responsible data sharing within the Consortium and beyond.

11.1. Data Governance

Data governance within the HE²AT Center involves various activities that ensure the responsible management of data. This includes the ethical oversight of data collection, establishing protocols for data sharing, and ensuring that data access is aligned with the research objectives of the DS-I Africa Consortium. The governance framework emphasises transparency, accountability, and compliance with relevant regulations, such as POPIA.

11.2. Data Indexing and Metadata Management

Effective governance also includes a robust system for data indexing and metadata management, ensuring that datasets are discoverable and accessible to the HE²AT Center Consortium andExternal Researchers. The indexing process includes:

1. **Metadata Standards**
   - **eLwazi Integration:** Data shared outside the HE²AT Consortium utilizes the eLwazi platform, ensuring compliance with broader data-sharing protocols.
   - **Data Reference Syntax (DRS):** Implemented by CSAG, the DRS provides structured mapping from metadata elements to directory and file naming syntax, standardizing climate and remote sensing datasets.
   - **Health Data Indexing:** Health data is indexed using a codebook with relevant ontologies, ensuring consistency and discoverability.
2. **Documentation and Integration**
   - **CSAG GitLab Wiki:** The DRS is documented on the CSAG GitLab Wiki, ensuring consistency and serving as a guide for indexing processes.
   - **DSI-Africa Open Data Science Platform (ODSP):** Integration with the ODSP metadata index ensures that metadata propagates to the ODSP system, making datasets discoverable through metadata queries.

11.3. Data Access Committee

The DAC is an independent committee that plays a central role in the governance of data sharing. The committee's responsibilities include evaluating data access requests and overseeing DTAs that are signed between the HEAT Center Consortium (Data Provider) and the Bone Fide External Researcher (Data Recipient) to ensure compliance with legal and ethical standards. The DTA will outline the terms for data use, confidentiality, and compliance with relevant laws and guidelines. The agreement will also prohibit the on-sharing of data, without explicit DAC approval.

11.4.1 Review Process

Submission: Data requests must be submitted using a standardized form that details the project objectives, required data, resource implications and ethical approvals. The form must be complete before the request will be considered. This form is included in **Annex 5** of the DMP.

Preliminary Screening: All requests will undergo preliminary screening by the HEAT Center SteerCo to confirm completeness of Data Access Requests forms.

Evaluation of data request: The DAC will evaluate requests based on criteria such as research credentials of the applicants (only applications from bone fide researchers will be considered), scientific merit, feasibility, potential privacy risks, resources available in the HEAT Center for preparing the De-Identified database, the avoidance of overlap with ongoing research, potential public health impact, and adherence to ethical and legal standards outlined in the application.

11.4.2 Membership of the DAC

The DAC will comprise of independent experts who may include Ethics Committee members, representatives from the DS-I Africa ELSI team or the eLwazi platform and people with expertise in data science, ethics, and legal matters. The HEAT Center Scientific Advisory Board and the HEAT Center SteerCo will select the members of the DAC.

For a more detailed description of the roles, responsibilities and specific procedures relating to the DAC, refer to **Annex 4: Terms of Reference for the DAC.**

12. Data retention

Participant data will be retained according to the following guidelines, ensuring compliance with the Protection of Personal Information Act (POPIA) and maximising its utility for future research:

12.1. Retention Periods:

- **Original Study Data**: Retained for at least five years after the completion of the **HE²AT Center Project**. This includes raw, unprocessed data from cohort studies and clinical trials. The Data Provider may elect to terminate the DTA prior to the completion of the HE²AT Project. On early termination of the DTA, the HEAT Center will immediately discontinue use of the Original Study Data and depending on the Data Provider's instructions, either return all copies of the data to the Data Provider, destroy all copies of the Original Study Data, or deal with the Original Study Data in any other manner, as requested by the Data Provider.

- **Consortium Shared Data**: Depending on the agreements in place, this data may be retained indefinitely.

- **RP1/RP2 De-Identified Data**: Depending on the agreements in place, this data may be retained indefinitely.

- **Inferential Data**: This is retained indefinitely. It includes aggregated and synthetic data derived from the analysis of the other data categories.

12.2. Ongoing Monitoring of Data Transfer Agreements

After executing a Data Transfer Agreement (DTA), ongoing monitoring, including periodic reviews and audits if necessary, ensures compliance with the agreement's terms.

13. Restricted data access

To safeguard personal information, the HE²AT Center implements robust encryption and security measures:

13.1. Data transfer, storage and encryption

Data is transferred Transport Layer Security (TLS) protocols are employed during transmission to maintain encryption and prevent interception We use WeTransfer modality, which is encrypted. Once transferred to UCT, any data identified as containing Personal Identifiers or specified by the DTA is encrypted for storage. The AES-256 encryption standard is applied, with encryption key access restricted to authorised personnel by Clause 2.10 of the Data Transfer Agreement, ensuring that the Data Recipient does not attempt to re-identify any Data Subjects. This complies with privacy and data protection legislation, including the Protection of Personal Information Act (POPIA). Metadata, however, is stored separately to facilitate indexing and software development while maintaining security.

13.2. Network security

The CSAG compute infrastructure benefits from UCT's comprehensive security policies. Key measures include:

- **Firewall Protection:** UCT’s Cisco firewall safeguards against external threats, ensuring only authorized access is permitted.
- **VPN Access:** A Cisco VPN service encrypts all traffic, enabling secure remote access to the UCT intranet and maintaining confidentiality.
- **Access Control:** Access to CSAG servers and services is carefully managed, with strict limits on authorized users.

13.3. Local authentication and authorization

Beyond UCT’s broader security measures, the CSAG/UCT platform employs additional authentication and authorization protocols. User identities are verified through a Linux filesystem and Lightweight Directory Access Protocol (LDAP), with access to restricted datasets managed through UCT’s authentication protocols and internal CSAG Data Management Plan mechanisms. All activities comply with UCT's information security policies, ensuring adherence to institutional standards.
By implementing these comprehensive encryption modalities, network security and authentication measures, the HE²AT Center ensures the protection and confidentiality of sensitive data throughout its lifecycle, maintaining compliance with ethical and legal standards.

14. Roles and responsibilities

The table below details the various roles and responsibilities associated with the data management plan and who is currently associated with each, their institution, and contact details. Personnel may change over time.

Role and responsibilities	People	Contact
DMAC PIs	**Christopher Jack (UCT)**	**cjack@csag.uct.ac.za**
	Sibusisiwe Makhanya (IBM)	**sibusisiwe.makhanya@ibm.com**
Responsible for ongoing (quarterly) assessment of data management and changes to the data management plan (annual).		
Health Data Acquisition	**Craig Parker for RP2 (Wits PHR)**	**Craig.parker@witsphr.org**
	Stanley Luchters for RP1 (CeSHHAR)	**stanley.luchters@ceshhar.co.zw**
Identification of relevant health datasets, coordination, and development of the DTA.		
Data Processing and Harmonization: Core Data Team De-identification, quality control, remapping, harmonization, and integration of all datasets. Note: Only these individuals have access to encryption keys for original sensitive data.	**Lisa van Aardenne (UCT)**	**lisa@csag.uct.ac.za**
	Pierre Kloppers (UCT)	**pierre@csag.uct.ac.za**
	Piotr Wolski (UCT)	**wolski@csag.uct.ac.za**
	Peter Marsh (UCT)	**Peter.marsh@uct.ac.za**
	Nicholas Brink (Wits PHR)	**nicholas.brink@witsphr.org**
	Craig Parker (Wits PHR)	**Craig.parker@witsphr.org**
Harmonization Team Members	**Members of the Core Data Team and**	**Same as Core Data Team**

	additional researchers from the consortium are working on harmonisation and integration tasks. **Access to metadata and synthetic health datasets for mapping on JupyterHub only.**	**contacts plus additional members where needed**
Managing Access to the UCT Data Analysis Platform	**Rodger Duffett (UCT)**	**rodger@csag.uct.ac.za**
Managing Access to the IBM Platform	**Sibusisiwe Makhanya (IBM)**	**sibusisiwe.makhanya@ibm.com**

15. Assessment and revision of the Data Management Plan

The Data Management and Analysis Core (DMAC) co-Principal Investigators (co-PIs) will conduct periodic reassessments of the Data Management Plan (DMP) in consultation with the HE²AT Center Steering Committee (SC), including the leads of Research Project 1 (RP1) and Research Project 2 (RP2). The SC may request assessment and review of specific aspects of the plan. These assessments will occur at least every six months to ensure the plan remains effective and up-to-date.

15.1. Assessment scope

The reassessment will focus on three key aspects of the data management plan:

1. **Data Process Efficiency**:
   - **Workflow Evaluation**: Assess whether the data processing workflow is functioning effectively and producing data ready for analysis. This includes evaluating each step, from data acquisition and harmonisation to storage and indexing.
   - **Error Identification and Resolution**: Identify any issues or bottlenecks in the current workflow and propose solutions to enhance efficiency and data quality.
2. **Compliance and Security**:
   - **Compliance Check**: Ensure that all data management activities comply with relevant legal and ethical standards, including POPIA and the DSI-Africa Data Sharing Guideline guidelines.
   - **Security Measures**: Review and update data security measures to protect personally identifiable information and ensure the integrity and confidentiality of the data.
3. **Usability and Accessibility**:
   - **Data Accessibility**: Evaluate whether the data is easily accessible to authorised users, including partner researchers and members of the HE²AT Center.
   - **Usability for Analysis**: Ensure that the data is in a usable format for analysis, with appropriate metadata and documentation to support effective use by researchers.

15.2. Revision process

Based on the findings from the assessment, the DMAC co-PIs will propose revisions to the Data Management Plan. The proposed revisions will undergo the following process:

1. **Proposal Development**:
    - The DMAC co-PIs will draft detailed proposals for necessary updates and changes to the DMP, addressing any identified issues and incorporating feedback from the assessment.
2. **Review and Approval**:
    - The proposed revisions will be presented to the HE^{2}AT Center Steering Committee (SC).
    - The SC, including the RP1 and RP2 leads, will evaluate the proposed changes and provide their approval or request further modifications as needed.
3. **Implementation**:
    - Upon approval, the DMAC team will implement the revised Data Management Plan.
    - All relevant stakeholders will be informed of the changes, and any necessary training or guidance will be provided to ensure smooth implementation.

By conducting regular assessments and making necessary revisions, the HE^{2}AT Center ensures that the Data Management Plan remains robust, effective, and aligned with best data management and analysis practices.

Annex 1: Key data sources

Name and Source of Dataset	Description	Key Variables*	Spatio-Temporal Coverage	Relevance
Biomedical Data				
Individual Participant Data Platform	Collation of prospectively collected high-quality data from pregnant women & and/or neonates (PROSPERO: CRD42020214637).	Preterm birth, pre-eclampsia, neonatal admission	African cohorts and trials	Research Project 1: pregnant women and young children are high-risk populations in Africa
HIV Databases	Pooled health database from cohorts and trials conducted among adolescents and adults in Johannesburg, South Africa (WHC studies)	Participants are followed up over time, with a multitude of physical measurements, laboratory tests, images and health questionnaires		Research Project 2: The study population has high rates of co-morbidities and adverse health outcomes
Climate/Weather Data				
European Centre for Medium-Range Weather Forecasts (ECMWF)	Outputs from a numerical weather prediction system, run twice daily, designed to produce state-of-the-art medium (10 days) global forecasts (contains only the latest forecast).	Temperature (Ground, Min, Max) at 2 m above ground; Solar irradiance; Wind speed (toward east, north) at 10 m above ground; Daily precipitation (total, rate); Dewpoint; Pressure	Spatial: Global coverage, 0.065536 deg. Temporal: 3 – 6 hourly & daily res.; Jan 2014 – Oct 2019	Determination of heat hazard; Thermal comfort metrics; combined climate exposures (forecasts)
IBM TWC (The Weather Company)	Current and historical weather Data layers from The Weather Company, an IBM Business	Temperature (Change, Min, Max, Feels like); Solar irradiance; Wind (speed, gust & dir.); Rel.	Spatial: Global coverage, 4km landmass and coastal waterways	Determination of heat hazard; Thermal comfort metrics; combined

		Humidity; Daily precipitation (total, rate); Dewpoint; 3-hrly Pressure Change	(hourly & daily res from 2015)	climate exposures (historical)
Fifth-generation ECMWF high-res. Reanalysis (ERA5)	A global reanalysis dataset combining observed data with the output of meteorological models.	Temperature (2 m above ground, Min, Max); Wind speed (toward east, north); Daily precipitation (total, rate, type); Atmospheric water/water vapour content; Thermal radiation; Soil temperature; Vegetation types and cover (high, low)	Spatial: Global coverage, 0.131072 degrees PAIRS resolution (raw: 0.25 deg.) Temporal: hourly; coverage from Jan 1980 – Jun 2019	Determination of heat hazard; Thermal comfort metrics; combined climate exposures (historical)
Fifth-generation ECMWF high-res. Reanalysis ERA5-Land	A global reanalysis dataset combining observed data with the output of meteorological models. Contains hourly data from 1950 to present.	Includes a range of surface and near-surface variables including: 2m temperature and dewpoint temperature, surface skin temperature, precipitation, near-surface winds, surface net thermal radiation.	Spatial: Global coverage, 0.1 deg. Temporal: hourly 1950 - present	Determination of heat hazard; Thermal comfort metrics; combined climate exposures (historical)
WATCH Forcing Data methodology applied to ERA5 (WFDE5)	A global bias-corrected reconstruction of near-surface meteorological variables derived from the ERA5.	Includes a range of surface and near-surface variables including: near surface air temperature, specific humidity, rainfall, wind speed, air pressure and surface longwave and shortwave radiation	Spatial: Global land Temporal: Hourly 1979 - 2019	Determination of heat hazard; Thermal comfort metrics; combined climate exposures (historical)

Temperature and precipitation gridded data for global and regional domains derived from in-situ and satellite observations	Temperature and precipitation from different datasets including: GISTEMP, Berkeley Earth, CPC and CPC-CONUS, CHIRPS, IMERG, CMORPH, GPCC and CRU	Precipitation, maximum, mean and minimum temperature	Spatial: Global, quasi-global, Africa depending on the dataset. Temporal: daily or monthly depending on the dataset	Determination of heat hazard; Thermal comfort metrics; combined climate exposures (historical)
Copernicus S2S seasonal forecast data	Model outputs forecasting climate conditions over the three months following the forecast initialization	Temperature 2m above ground (min, max), Daily precipitation (total)	Temporal: daily	Seasonal (weeks to 3 months) time horizon forecasting of relevant weather conditions (heat hazard) for early warning
CP4-A (NERC JASMIN)	Very high resolution (4km) simulations of historical and future climate over Africa	Temperature 2m above ground (min, max), daily precipitation (total), multi-level circulation	Temporal: hourly	Dynamical downscaling to support sub-urban temp hazard mapping
CORDEX Africa (ESGF)	Ensemble of dynamically downscaled simulations of African climate to 50km, 25km, and 10km resolution	Temperature 2m above ground (min, max), daily precipitation (total), multi-level circulation fields	Temporal: daily and sub-daily (6 hourly)	Dynamical downscaling of climate to support sub-urban temperature hazard mapping
GHCN station data (NOAA GHCN)	Global archive of daily weather station data	Temperature 2m above ground, daily precipitation (total)	Temporal: daily, station locations	To support statistical downscaling of temperature hazard
Remote Sensing Data				
30 m res Elevation (SRTM) (NASA)	Global elevation data from the Shuttle Radar Topography Mission (SRTM)	Elevation	Released in 2013	Determination of heat hazard; Urban heat Island Effect
High res imagery (ESA Sentinel 2)	Images from the Sentinel 2 satellite pair which view land surface regions in 13 spectral bands.	Urban land cover – vegetation coverage, morphological features, possibly pollution levels (AOT). Bands 4 (red), 8 (NIR) and SCL (Scene	Spatial: Global coverage; 0.000064 deg res Temporal: every 5 days or faster; from Aug 2015 – Nov 2020.	If there is a requirement to control for pollution effects or to look at combined heat-pollution exposures

		Classification); Aerosol Optical Thickness; NDVI sh layer.		
Aqua MODIS Land Surface Temperature (MYD21A1D & MYD21A1N)	Satellite derived day and night time, high resolution (1KM) land surface temperature dataset.	Land surface temperature	Spatial: Global land surface coverage; 0.00983 deg res. Temporal: daily; 2002/07/04 to present	
Areal/Geospatial Socio-Economic Data				
Gauteng City-Region Observatory	GIS raster and shapefiles for the Gauteng City-Region area	Demographics, economics, environmental, spatial structure, spatial change and transport	Spatial: Gauteng city-region. Temporal: various depending on the variable	Research Project 2: provides information on socio-economic circumstances and attitudes of residents within the Gauteng City-Region.
General Household Surveys, Statistics South Africa	Annual household Survey	Living circumstances of South African households: education, health, social development, housing, access to services and facilities, food security and agriculture.	Sample survey data, units are households and individuals	Research Project 2: provides information on socio-economic circumstances of residents within the Gauteng City-Region.
Quality of Life Surveys, Gauteng City-Region Observatory (GCRO)	Household Survey	Quality of life, socio-economic circumstances, attitudes to service delivery, psycho-social attitudes, value-base and other characteristics of residents of the Gauteng City-Region.	Sample survey data, units are households and individuals	Research Project 2: provides information on socio-economic circumstances and attitudes of residents within the Gauteng City-Region.

Global Population (SEDAC) - Gridded Population of the World (GPW), v4	Distribution of human population (counts and densities) on continuous global raster surface. Input data are extrapolated to produce population estimates for 5-year intervals	Population counts and density estimates	Spatial: Global coverage, 1km grid res Temporal: 5-yearly; coverage from Jan 2000 to Jan 2020	Accounting for the population exposed
News Coverage (GDELT)	Portion of news coverage about specific area and time related to Covid-19	Global events derived from worldwide news coverage.	Spatial: Global coverage, 0.008192 deg. res	Example for production of spatial data layer for news events

^Examples of relevant databases (the study may draw on additional sources). *Examples of relevant variables.

Annex 2: Personal information processing agreement

The following agreement will be signed by each person (Operator under POPIA definitions) involved in processing personal information used by the HE²AT Center Project. This includes the **Core Data Team:**.

I, [Full name] hereby agree to comply with the requirements of the POPIA Act of South Africa as regards the processing of personal information. These requirements include:

1. Only processing personal information for the purposes described in the HE²AT Center research protocols;
2. Only processing personal information that is required for these purposes;
3. Not enabling or allowing access to personal information to anyone who does not have authorization for such access;
4. Notifying the HE²AT, Steering Committee as the responsible party, if there is any reason to believe that personal information has been accessed or made available to an unauthorized person.

I further note that I have received appropriate training on my responsibilities and I am subject to professional obligations of confidentiality.

Signed _______________________ at _________________ on this ___ day of __ in the year _______

Annex 3: Ethics notification letter

UNIVERSITY OF THE
WITWATERSRAND,
JOHANNESBURG

APPLICATION FOR A NOTIFICATION REGARDING AN APPROVED STUDY in RP2

PART 1: ADMINISTRATIVE *(Blocks will expand to contain the information required, no extra references or pages should be added)*	
Ethics Reference Number:	
Study Title:	
Phase of trial:	
Protocol/Project/Study Number:	
Approved Version/No. and Date:	
Amended Version/No. and Date:	
Health product being studied:	
Sponsor/Funder/Donor:	
Applicant:	
Contact Person:	
Address:	
Cell No.:	
E-mail address:	
Date of Application:	

PART 2: DETAILS OF NOTIFICATION

Briefly provide:

1. Motivation / Background:

2. Study Plan: I, the undersigned, agree to conduct/manage the above-mentioned study under the conditions as stated in this application	
Applicant/Principal Investigator: Signature:	Date ....

Annex 4: Data Access Committee Terms of Reference

1. Purpose of the Data Access Committee

The HE²AT Center's DAC oversees and governs access to health-related data in Research Project 1 (RP1) and Research Project 2 (RP2) that has been de-identified. The DAC ensures that all data requests are reviewed and approved based on scientific merit, ethical considerations, feasibility, and compliance with legal and institutional requirements. The goal is to facilitate high-quality research, while maintaining the integrity, confidentiality, and security of the data, and ensuring compliance with relevant ethical and legal standards.

2. Roles and Responsibilities

- **Review of Requests for Data Access**: The DAC is responsible for reviewing all data access requests for RP1/RP2 De-Identified Data. Each request will be assessed against scientific, feasibility, and ethical considerations. The burden of work that required by the HEAT Center Consortium to prepare the De-Identified datasets will be considered. Funds may be requested if database preparation is onerous. Only requests that meet these criteria will be approved.

- **Decision-Making Process**: Decisions on data access requests will be made by a majority vote of DAC members. In case of a tie, the Chair will have the deciding vote. All decisions will be documented, including the rationale behind approvals or rejections.

- **Transparency and Accountability**: The DAC will ensure that all decisions are documented and communicated transparently to the applicants, maintaining compliance with POPIA and other relevant regulations.

5. Meeting frequency

The DAC will meet monthly or as needed, depending on the volume and urgency of data requests. Special meetings can be convened by the Chair to address urgent requests.

6. Oversight of legal and ethical compliance

- **The DAC will confirm Data Transfer Agreement procedures are complied with.** Before any data is shared, a DTA must be executed between the HEAT Center Consortium (who serves as the data provider in this instance) and the data recipient (the researcher(s) who made the data access request).

Annex 5: Data Request Form

Request for Access to De-identified HE²AT Center Data

Proposals for new studies that utilize data collected in the HE²AT Center may be submitted by external investigators in the form of a study proposal. The proposal will be reviewed by the HE²AT Center Steering Committee and the Data Access Committee (DAC) using the following criteria:

1. Scientific Merit:
 - Research question is scientifically sound and can be tested with the proposed study design and methodologies.
 - Appropriate skills and expertise available in the proposed investigators.

2. Potential Public Health Impact:
 - Study is relevant to one or more HE²AT Center study populations (pregnant women, postpartum women, children in RP1, and adolescents or adults in RP2).
 - Study will answer important public health questions or is in the critical pathway of research toward such answers.

3. Feasibility:
 - Study population and variables required are available in the HE²AT Center data.
 - Study is feasible within proposed timelines.

4. Data Management:
 - The study proposal has outlined how data will be managed to ensure the data security and protection of individual's health data.

Please contact Elizabeth Frederick (Elizabeth.Frederick@witsphr.org) with any queries.

HE^{2}AT Center Data Request Application Form

Submit the completed form to the HE^{2}AT Center for consideration to Elizabeth Frederick (Elizabeth.Frederick@witsphr.org)

GENERAL INFORMATION

Request submitted by:

Name: ____

Institution: ____

Email address: ____

Date submitted: ____

Title: ____

Lead investigator:

Name: ____

Institution: ____

Email address: ____

Nature of research activities:

____

____

____

Co-investigator(s): Include name(s), institution(s), email(s):

____

____

____

Consortium Membership Status: Indication of whether the applicant is a member of the DSI-Africa Consortium. Non-members must provide detailed information about their organizational affiliation.

____

____

____

Relevant studies: List all HE^{2}AT Center studies from which data are being requested:

____

____

____

Relevant variables: List all variables that are being requested:

____

____

____

____

____

____

____

____

____

____

Please indicate your preference regarding the type of file format you wish to receive.

☐ .csv (text)

☐ .xlsx (Excel)

☐ Other (please specify): __________________________

For access to HE^{2}AT Center study data, you must have your institution's ethics approval or waiver. Please confirm that you have attached a copy of the ethics approval or waiver from your institution.

☐ Confirmed

☐ Pending (please provide further details): ______________________

For access to HE^{2}AT Center data, we will need to execute a Data Transfer Agreement. Will you adhere to the terms and conditions of the data sharing agreement?

☐ Yes

☐ No

For access to HE^{2}AT Center data, you will need to agree not to ever make an attempt to re-identify a data subject using the data provided. Will you agree to, and adhere to this condition?

☐ Yes

☐ No

STUDY DESIGN

Background: Include brief literature review, and any research gaps that the proposed analysis will fill (maximum 200 words):

____

____

____

____

____

____

____

____

Study rationale (maximum 100 words):

____

____

____

____

Study aims and objectives:

Study design and analysis (maximum 300 words):

Data management: Please include information here about your procedures to safely manage and store the data, as well as how the data will be shared between study investigators (maximum 200 words):

Timeline for completion:

Data Protection Measures: A detailed description of statutory, organizational, and technical measures in place at the receiving institution to safeguard the confidentiality and security of the data.

Dissemination plan/impact: Please indicate meetings, conferences, and/or journals where you are planning to submit this work, with anticipated dates. Please include any other planned dissemination activities (maximum 100 words):