|  |
| --- |
|  |
| HE²AT Center Data Management Plan |
| Document version number 2.1 |

|  |
| --- |
| Developed by C Jack, C Parker  8-15-2024 |

Developed by Christopher Jack\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Name Date Signature

Reviewed by NIH Science officer tbd\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Name Date Signature

Approved by Matthew Chersich\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Name Date Signature

# 

Contents

[Acronyms 5](#_Toc174715342)

[Definitions 7](#_Toc174715343)

[Stakeholders 12](#_Toc174715344)

[1. Introduction and background 14](#_Toc174715345)

[1.1. Scope 14](#_Toc174715346)

[1.2. Target audience 14](#_Toc174715347)

[1.3. Purpose 14](#_Toc174715348)

[1.4. Background 15](#_Toc174715349)

[1.5. Overview of the Data Management Workflow 17](#_Toc174715350)

[2. Data acquisition 19](#_Toc174715351)

[2.1. Health-related data 19](#_Toc174715352)

[2.1.1. Categories of data, by harmonisation and deidentification status 19](#_Toc174715357)

[2.1.2. Original study data 19](#_Toc174715358)

[2.1.3. Consortium shared data 19](#_Toc174715359)

[2.1.4. RP1/RP2 De-identified data 20](#_Toc174715360)

[2.1.5. Inferential data 20](#_Toc174715361)

[2.2. Climate/weather data 21](#_Toc174715362)

[2.2.1. Remote Sensing Data (Subsection of Climate/Weather Data) 21](#_Toc174715363)

[2.3. Areal/geospatial socio-economic data 22](#_Toc174715364)

[3. Ethics Approval Process and Data Transfer Agreement 23](#_Toc174715365)

[3.1. Notification of New Studies 23](#_Toc174715366)

[3.2. Notification for Research Project 2 (RP2): 23](#_Toc174715367)

[3.3. Data Transfer Agreement (DTA): 23](#_Toc174715368)

[3.4. Ongoing Ethics Monitoring: 23](#_Toc174715370)

[4. Pre-processing and harmonisation 25](#_Toc174715371)

[4.1. Pre-processing 25](#_Toc174715372)

[4.2. Variable mapping 26](#_Toc174715383)

[4.3. Mapping validation 27](#_Toc174715385)

[4.4. Database population 27](#_Toc174715386)

[4.5. Final data cleaning and analysis dataset creation 28](#_Toc174715387)

[5. Integration and analysis interfaces 30](#_Toc174715388)

[5.1. Integration of Climate and Socio-Economic Variables 30](#_Toc174715389)

[5.2. Sourcing Pre-Processed Data 30](#_Toc174715390)

[5.3. Automated Data Retrieval 30](#_Toc174715391)

[5.4. Integration into Broader Dataset 30](#_Toc174715392)

[5.5. Collaboration with CSAG/UCT 31](#_Toc174715393)

[5.6. Alignment with Analysis Objectives 31](#_Toc174715394)

[6. Data analysis platforms 31](#_Toc174715404)

[6.1. Jupyter hub 31](#_Toc174715405)

[6.2. IBM data analysis platforms 32](#_Toc174715406)

[6.3. Integration approaches 32](#_Toc174715407)

[7. De-identification 33](#_Toc174715408)

[7.1. Geographic aggregation: 33](#_Toc174715409)

[7.2. Location jittering: 33](#_Toc174715411)

[8. Data sharing and open access 37](#_Toc174715412)

[8.1. Restrictions to Data Sharing 37](#_Toc174715413)

[8.2. Discoverability 37](#_Toc174715414)

[8.3. Levels of Data Access 37](#_Toc174715415)

[8.4. Procedure for Making RP1/RP2 De-identified Data Available to Bona Fide Researchers 38](#_Toc174715416)

[8.5. Post-Funding Data Management and Open Access 39](#_Toc174715417)

[9. POPIA compliance and protection of personal information 40](#_Toc174715418)

[10. Governance and compliance 41](#_Toc174715419)

[10.1. Data Governance 41](#_Toc174715420)

[10.2. Compliance Measures 41](#_Toc174715421)

[10.3. Data Indexing and Metadata Management 42](#_Toc174715422)

[10.4. Data Access Committees (DACs) 42](#_Toc174715423)

[10.5. Membership of the DAC 43](#_Toc174715424)

[10. Data retention 44](#_Toc174715468)

[10.1. Retention Periods: 44](#_Toc174715469)

[10.2. Ongoing Monitoring 44](#_Toc174715470)

[11. Restricted Data Access 45](#_Toc174715508)

[11.1. Encryption of Personally Identifiable Data 45](#_Toc174715509)

[11.2. Data Storage and Encryption 45](#_Toc174715510)

[11.3. Network Security 45](#_Toc174715511)

[11.4. Local Authentication and Authorization 45](#_Toc174715512)

[12. Roles and responsibilities 47](#_Toc174715513)

[13. Assessment and revision 51](#_Toc174715514)

[13.1. Assessment scope 51](#_Toc174715515)

[13.2. Revision process 52](#_Toc174715516)

[Annexe 1: Key data sources 1](#_Toc174715517)

[Annexe 2: Personal information processing agreement 0](#_Toc174715518)

[Annexe 3: Ethics notification letter 1](#_Toc174715519)

[Annexe 4: Data access committee (DAC) Terms of Reference for the HE](#_Toc174715520)[2](#_Toc174715520)[AT Center 4](#_Toc174715520)

[Annex 5: Data Access Committee: Data Request form 7](#_Toc174715521)

# 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 HEÂ²AT 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 HEÂ²AT Center |
| RP2 | Research Project 2 of the HEÂ²AT Center |
| Research Hubs | One of the seven NIH DS-I Africa Research hubs |
| SC | HEÂ²AT Center Steering Committee |
| SRTM | Shuttle Radar Topography Mission |
| TEC | Training and Engagement Core of the HEÂ²AT Center |
| TLS | Transport Layer Security |
| UCT | University of Cape Town |
| WWARN | Worldwide Antimalarial Resistance Network |
| sSA | sub-Saharan Africa |

# 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 HEÂ²AT Center's data processes, documenting Data Reference Syntax (DRS), and storing data management code. It ensures transparent, version-controlled data handling accessible to authorized team members. |
| Climate/Weather Data | 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 harmonisation and is shared among HEÂ²AT Center partners. This data is retained indefinitely or for five years post-project or DTA termination. |
| Controlled Access | A mechanism by which potentially personally identifiable research data is distributed to bona fide researchers upon submission of a data access request and supporting documentation. |
| 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 within the project. |
| Data Management and Analysis Core (DMAC) | The core component of the HEÂ²AT 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 legal entity responsible for authorizing access to a dataset. |
| 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 the provider and the recipient, addressing confidentiality, data use limitations, and compliance with relevant laws and guidelines. |
| Data Visiting | A modality of data sharing where data users access and analyse the data within the computing environment of the data provider. |
| Ethical Legal and Social Implications (ELSI) | Projects within DS-I Africa that focus on data science and research's ethical, legal, and social aspects. |
| Federal Information Processing Standards (FIPS) | Standards for federal computer systems that NIST develops. |
| 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 data on the quality of life in the Gauteng City-Region. |
| HE2AT Center Consortium | The consortium members jointly working on the HE2AT Project. |
| HE2AT Center Data Management Plan | The data management plan applicable to the RP1/RP2 Study as may be amended and updated from time to time. |
| HEÂ²AT Center | The Heat and Health in Africa Transdisciplinary Center, focused on heat-health research, capacity building, and engagement. |
| HEÂ²AT Center Steering Committee (SC) | The committee responsible for guiding the strategic direction and oversight of the HEÂ²AT Center. |
| Harmonization | Aligning various health datasets into a unified format. |
| Health-Related 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 a unified format. |
| NIH Data Science Initiative (DS-I) Africa | An NIH initiative aimed at enhancing data science capacity and collaboration across Africa. |
| National Institute of Standards and Technology (NIST) | A US federal agency that develops and promotes measurement standards. |
| 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 | Research data freely available in community-focused data repositories 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 data collected directly from various cohort studies and clinical trials. These data are owned by the data providers who conducted or commissioned the studies and is retained for five years following the termination of the project or the Data Transfer Agreement (DTA). |
| 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 (PI) | 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. |
| Quality of Service (QoS) | A measure of the overall performance of a service such as a survey or data collection process. |
| RP1 De-identified Data | Data with the following information deleted |
| 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 HEÂ²AT Center, focusing on individual participant data meta-analysis of heat-health impacts. |
| Research Project 2 (RP2) | The second major research project of the HEÂ²AT Center, focusing on urban heat vulnerability 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, health, finances, occupation, etc. |
| Shuttle Radar Topography Mission (SRTM) | A mission that obtained elevation data for most of the Earth using radar interferometry. |
| Social Media Data | Data representing activity on social media platforms, initially Twitter, and includes data like Google Search trends and Facebook mobility data. Used for analysing public perceptions. |
| Training and Engagement Core (TEC) | The core component of the HEÂ²AT 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. |
| Worldwide Antimalarial Resistance Network (WWARN) | A network focused on monitoring and understanding antimalarial drug resistance. |
| sub-Saharan Africa (sSA) | A geographical region of Africa located south of the Sahara Desert. |

# Stakeholders

* **CSAG/UCT Data Management Team:** The Climate System Analysis Group (CSAG) 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).
* **Compliance Officers:** Individuals responsible for monitoring and enforcing compliance with data security and privacy regulations, ensuring that all data management activities adhere to relevant laws and guidelines.
* **Core HE²AT Team:** A group of named vital personnel responsible for the initial processing and de-identifying 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 Providers:** Legal entities or organisations responsible for authorising access to their datasets and ensuring that data sharing complies with legal and ethical standards.
* **DSI 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.
* **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 who work on aligning and integrating various 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.
* **IBM Data Analysis Platforms Team:** The team responsible for managing IBM's suite of data analysis tools and platforms, such as IBM Watson Studio, which support advanced analytics and machine learning within the HE²AT Center.
* **OpenAI:** An artificial intelligence research organization that provides large language models (LLMs) to assist with variable mapping suggestions, data transformation recommendations, and other AI-driven insights.
* **Partner Researchers:** External researchers and scientists collaborating with the HE²AT Center and utilizing the integrated datasets for various research projects and analyses.
* **Post-Funding Data Management Team:** The team responsible for managing data after the funding period ends, ensuring the continued maintenance and availability of inferential data and transferring data to appropriate repositories as needed.
* **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).
* **Training Hubs:** Facilities that provide training and capacity-building activities for researchers, using shared data for educational purposes and contributing to the development of data management skills within the consortium.

# Introduction and background

## Scope

This Data Management Plan (DMP) applies to all data acquired and produced as part of the HE²AT Center project's activities, including pilot projects.

## Target audience

The scope encompasses various stakeholders, each with distinct roles and responsibilities as outlined below:

* **Data Providers**: These are the legal entities or organisations responsible for authorising access to their datasets. They ensure that data sharing complies with legal and ethical standards.
* **DSI Partners**: Partners within the NIH Data Science for Health Discovery and Innovation in Africa (DS-I Africa) Initiative contribute to and utilise the shared data resources. They collaborate on data management practices and ensure alignment with the broader DS-I Africa objectives.
* **Consortium Research Groups**: These groups, part of the HE²AT Center consortium, conduct research using the shared datasets. They adhere to the DMP data acquisition, transfer, processing, storage, and access guidelines.
* **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).
* **Training Hubs**: These hubs provide training and capacity-building activities for researchers. They utilise the shared data for educational purposes and contribute to developing data management skills within the consortium.

**Researchers Request to Acquire Data**: External researchers seeking access to the HE²AT Center's datasets must follow the data access procedures outlined in this DMP. They must comply with data transfer agreements, ethical guidelines, and any conditions set by the Data Access Committee (DAC).

## 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, security, and availability while facilitating effective data sharing and collaboration among all stakeholders.

Specifically, the DMP seeks to:

* **Standardize Data Management Practices**: Define uniform procedures and standards to streamline data handling processes across the 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 consortium, other DS-I Africa projects, and with external researchers.
* **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.

While this DMP outlines the procedures for data access by external parties, data sharing outside the consortium is further supported by associated Data Transfer Agreements between the HEAT Center and external partners, ensuring that all data management activities, both inside and outside the Consortium, align with the HE²AT Center's objectives and ethical commitments.

## Background

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. Such techniques have not yet been employed in climate change and health. 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]](#footnote-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 neonatal 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]](#footnote-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, socio-environmental factors and the characteristics of the built environment. This project, conducted in Abidjan, Ivory Coast, and Johannesburg, South Africa, will investigate heat exposure risks to develop 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 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. The project aligns with DS-I Africa objectives and may expand to include other Research Hubs and cities across Africa[[3]](#footnote-3).

## 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 access to open access—but also highlights the stakeholders involved at each stage. The diagram encapsulates the framework that will be unpacked and elaborated upon throughout the rest of this document.

As you progress through the sections that follow, you will see how each part of this framework is addressed in detail. This ensures that all processes adhere to the required ethical and legal standards while also promoting data accessibility and reuse. This diagram, therefore, acts as a roadmap for navigating the complexities of data management within the HE²AT Center.

A diagram of a company

Description automatically generated

Figure 1: HE²AT Center Data Management Workflow

# Data acquisition

**Data acquisition** is a foundational step in the HE²AT Center's research efforts, involving systematically collecting various 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 category uniquely analyses the intersections between health outcomes, environmental conditions, and socio-economic factors. By utilising both open and restricted-access repositories, the HE²AT Center ensures a comprehensive and integrative approach to data collection, supporting robust and multifaceted research across multiple disciplines.

## **2.1. Health-related data**

### Categories of data, by harmonisation and deidentification status

The categorisation of health data within the HE²AT Center ensures that the project adheres to ethical and legal standards while facilitating collaborative research. This structured approach enables the integration of new partners and maintains data privacy and protection throughout the project's lifecycle. Health data is classified into four distinct forms, ranging from raw, individual-level data to aggregated, inferential data. Each category serves a specific role, guiding the data's management, access, and usage to meet the project’s research objectives and privacy requirements.

### Original study data

This category includes raw, unprocessed health data collected directly through various cohort studies and clinical trials fulfilling the study eligibility criteria. This data is owned by the data providers who conducted or commissioned the studies. The original study data is retained for five years following the termination of the project or the Data Transfer Agreement (DTA). Access to this data is restricted to a small team of data managers within the HE²AT Center’s Core Team(see roles and responsibilities section for more detail). The source data is protected pseudonymous data, meaning it has been de-identified, but it still contains indirect identifiers such as dates and geolocation information. While these identifiers do not directly reveal personal identities, they could be used for identification if cross-referenced with other information.

### Consortium shared data

Once the source data undergoes harmonisation, it becomes consortium-shared data. The data is also integrated with climate and other data, as described below. Unless data providers state otherwise, the HE²AT Center and Data Access Committee now make decisions around data usage and access at this stage. This decision-making level stems from the team having performed significant alterations, harmonisation, and integration work on the data. However, original data providers may elect to stipulate different data use and access terms if desired. This data is retained indefinitely or for five years post-project or DTA termination, depending on the agreements in place. Access to consortium-shared data is granted to current HE2AT Center partners for the purpose of the research listed in Annexure B of the DTA and can be extended to new partners who join the Consortium.

This data is considered protected pseudonymous data as per POPIA guidelines, as several steps have been taken to enhance privacy protections. For example, age data will be reported in years rather than exact birthdates and geographic data will be jittered.

### RP1/RP2 De-identified data

Data that has been further processed and de-identified falls into this category. The HE²AT Center retains ownership of this de-identified data, which is stored indefinitely unless the Data Provider stipulates otherwise. Access is granted to data-sharing applicants who meet specific conditions and requirements set by the Data Access Committee. 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 personal identifiers. Age data will be reported in five-year age bands rather than an exact birthdate, and the geographic data will be reported in larger units, such as cities or districts.

### Inferential data

The final category is inferential data, aggregated and synthetic data derived from the analysis of the preceding data categories. The HE²AT Center owns these details and retains them indefinitely. Inferential data is made available for open access to support broader research initiatives. As aggregated anonymous data and synthetic data, it is classified as de-identified under POPIA, ensuring that individual privacy is fully protected. This category contains no direct or indirect identifiers, making it impossible to trace back to any individuals.

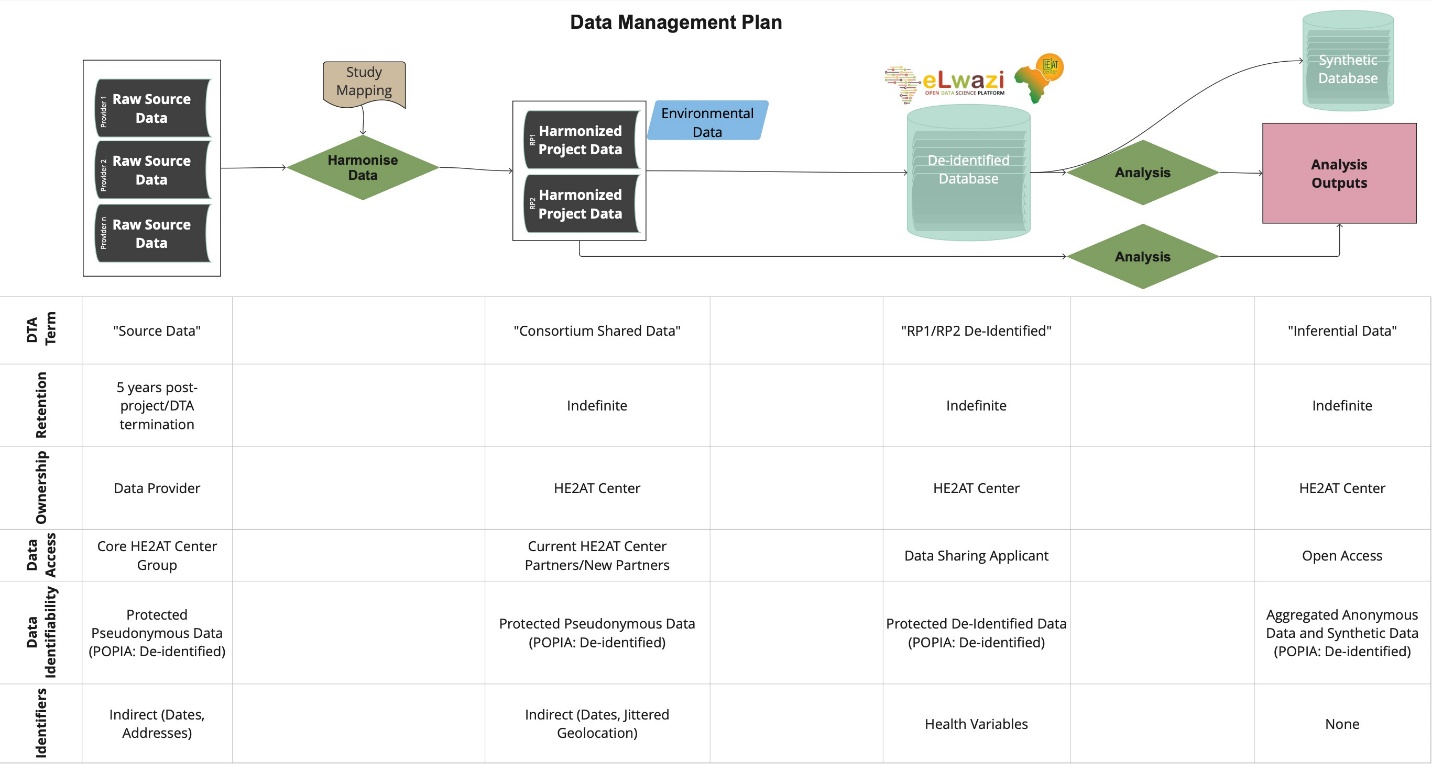


Figure 2: HE²AT Center Data Management Data Categories

## **2.2. Climate/weather data**

This category includes **observational-based datasets** (e.g., weather station observations, satellite proxy observations) and **processed/gridded data** from atmospheric re-analysis and climate simulations. These datasets will provide historical climate observations and forecasts. The data will primarily be sourced from open repositories such as the Copernicus Climate Data Store (CDS) and the Earth System Grid Federation. All climate-related data will be stored on CSAG/UCT data storage systems, which will manage and update the primary data index to ensure the consortium can access the latest information.

Most climate datasets are openly accessible with no restrictions on non-commercial research use, though proper citation of the data sources is required whenever these datasets are used.

### 2.2.1. Remote Sensing Data (Subsection of Climate/Weather Data)

Remote sensing data, primarily derived from **satellite sensors**, will focus on indicators like **optical imagery** (e.g., satellite images of urban centres) and **physical measures** such as land surface temperature, soil moisture, vegetation condition, and land use/cover. These data are used to estimate environmental variables and urban density but do not include any sensitive or personally identifiable information.

Similar to other climate-related data, remote sensing data will be accessed through open data repositories, such as the Copernicus Climate Data Store (CDS) and Sentinel data systems. The data will be stored on IBM 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.

## 2.3. 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. Still, versions may already exist or can be uploaded on IBM’s system to enable analysis through IBM analytical systems.

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.

# Ethics Approval Process and Data Transfer Agreement

In alignment with the HE²AT Center's commitment to ethical conduct across all research activities, the ethics approval process and Data Transfer Agreement (DTA) are essential for securing and maintaining oversight of studies that contribute data to the project. This section outlines the steps and considerations to meet all ethical guidelines throughout the research lifecycle.

## Notification of New Studies

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

## Notification for Research Project 2 (RP2):

Specific notifications for RP2 data activities will be sent to the Wits HREC to ensure compliance with ethical standards. This includes details of data acquisition, processing, and usage within the scope of RP2. Regular updates on RP2 activities will help maintain alignment with Wits HREC’s ethical requirements(notification template can be found in Annex 2).

## Data Transfer Agreement (DTA):

The DTA between each health data provider and UCT or WITS Planetary Health Research will outline the terms for transferring and processing health data. The DTA must align with national or international health data-sharing legislation. Depending on the DTA terms, data may be accessible solely on the institution's computing platform or made available for consortium-wide access. This agreement ensures that the handling, storage, and transfer of data adhere to ethical standards and protect the rights of study participants.

## Ongoing Ethics Monitoring:

After the data transfer process has begun, continuous monitoring ensures compliance with the agreed-upon ethical standards. This includes reviewing any concerns the ethics committee raises and ensuring that using original data for new analyses remains ethically sound.

By adhering to these procedures, the HE²AT Center ensures that all research activities are conducted ethically, safeguarding participants' rights and maintaining the highest standards of research integrity.

# Pre-processing and harmonisation

Data harmonisation is critical for integrating diverse health datasets into a unified format, enabling comprehensive analyses across various data sources. The workflow begins with transforming **Original Study Data** into **De-identified Data**, ensuring that datasets are compatible and ready for analysis.

## Pre-processing

The pre-processing stage is managed by the Core HE²AT team, which includes members with authorised access to incoming **Original Study Data**. This phase involves several key tasks that lay the foundation for subsequent harmonisation and analysis. First, the **Original Study Data** is reformatted into a standardised format, such as CSV or JSON, to ensure compatibility across different systems and tools. This standardisation follows established data formatting guidelines, such as the Open Data Standards or the OMOP Common Data Model, to promote consistency and interoperability across the consortium's datasets.

Once the **Original Study Data** is in a standardised format, the team extracts and labels key variables with standardised names and descriptions. This step involves aligning the variables with ontology frameworks, such as SNOMED CT or ICD-10, where applicable. This ensures that the data is consistent and properly categorised, making integrating with other datasets such as Consortium Shared Data easier.

In addition to variable extraction, the team generates synthetic data. This involves creating ten random samples of each variable, which are used to test data integrity and prepare for harmonisation. Synthetic data is particularly valuable for validating the integration process without exposing the Original Study Data, ensuring any potential issues can be addressed in a controlled environment.

Finally, the team identifies and reviews relevant study documentation. This includes study protocols, codebooks, and other essential documents linked to the dataset for further analysis. This documentation provides context for the **Original Study Data**, ensuring that the variables are correctly understood and utilised in subsequent stages.

Although the **Original Study Data** is not encrypted at this stage, access permissions are strictly limited to the Core group. This ensures that only authorised personnel can handle the data, maintaining its confidentiality. Additionally, periodic data backups are made to a secure and encrypted disk, separate from the wider HPC system, to ensure that the data remains protected and recoverable in case of any issues.

## Variable mapping

The variable mapping process begins with **Synthetic Data** generated during the pre-processing phase. This **Synthetic Data** is a valuable tool for testing and refining the variable mapping process without exposing the **Original Study Data**. By working with **Synthetic Data**, the team can identify potential challenges early on and address them in a controlled environment.

In addition to **Synthetic Data**, the team relies heavily on the **Metadata** and study documentation created during pre-processing. This **Metadata**, which includes detailed variable names and descriptions, provides essential context for understanding the variables and their original use. By carefully reviewing this information, the team ensures they fully understand the variables before mapping them to standard ontologies.

To further enhance the accuracy and efficiency of the mapping process, the team leverages AI tools, specifically OpenAI's language models (LLMs). These models assist in generating detailed descriptions of variables, suggesting possible mappings to standard ontologies, and recommending data transformations based on the available **Metadata**. The LLMs are particularly useful in filling in gaps in variable descriptions and offering initial mapping suggestions, which are then reviewed and validated by the harmonisation team.

Once the AI-generated suggestions are reviewed, the team maps the source variables from the **Original Study Data** to a standardised set of ontologies, such as SNOMED CT, ICD-10, or the OMOP Common Data Model. This process involves re-coding the variables to fit within a common data framework, ensuring that the data can be integrated and analysed across different datasets as **Consortium Shared Data**. Throughout this process, the team meticulously documents each mapping decision, capturing the rationale behind the choices and ensuring transparency for future data users.

Collaboration and feedback are essential components of the variable mapping process. The Core group and the extended HE²AT Center staff work closely together, sharing insights and refining the mappings as needed. Harmonisation team members complete standardised forms that document the mapping of source variables from the **Original Study Data** to the chosen ontologies and any re-coding performed. This documentation supports transparency and is a valuable reference for future data users.

## Mapping validation

The validation process begins with the Core HE²AT team revisiting the **Consortium Shared Data**, which has been reformatted from the **Original Study Data**. By comparing the original variables with the newly mapped versions, the team checks for consistency and accuracy. This step is essential for identifying any discrepancies or errors that may have occurred during the initial mapping phase. The goal is to ensure that the mapped variables retain their original meaning and can be effectively used in subsequent analyses. This ensures the data's integrity as it transitions to **De-identified Data**.

To further strengthen the validation process, a health expert is brought in to review the mappings. The involvement of a health expert is critical, as they possess the domain-specific knowledge needed to assess whether the variables have been accurately interpreted and whether the mappings align with established health ontologies, such as SNOMED CT or ICD-10. This expert review helps to ensure that the data remains clinically relevant and that no important nuances are lost during the mapping process. This step is crucial for ensuring that the **Consortium Shared Data** is aligned with the Center's research objectives and can be effectively used across various studies.

The Core HE²AT team verifies after the health expert's review. This team, which includes harmonisation specialists, data scientists, and other key personnel, performs a thorough check to ensure that the mappings are correct and that the data transformations have been appropriately applied. This multi-layered validation process helps to catch any remaining issues. It ensures that the final harmonised dataset is both accurate and reliable, ready for further transformation into **De-identified Data** if necessary.

An essential aspect of this validation process is version control. To maintain transparency and track changes, all mappings and transformation instructions are version-controlled using CSAG’s GitLab system. By employing a robust version control system, the team ensures that every modification to the dataset is documented, providing a clear audit trail. This allows for easy rollback to previous versions if any issues are identified later on and ensure that the final dataset is reproducible and trustworthy.

## Database population

The database population stage marks the final transformation of data into a harmonised and accessible format, ready for consortium-wide use. This step creates the Consortium Shared Data by transforming the validated **Original Study Data** into a standardised format. The Core HE²AT team meticulously manages this process, ensuring that the data aligns with the mappings and instructions established in the earlier harmonisation phases.

The population of the final database involves applying the validated mappings and transformations to the **Original Study Data**, converting it into **Consortium Shared Data** that can be utilised across the HE²AT Center. This transformation ensures that the data adheres to standardised ontologies and is compatible with the HE²AT Center’s analytical platforms.

An important aspect of this stage is the additional anonymisation of patient identifiers. Although the data has already undergone de-identification, the Core HE²AT team further anonymises any residual identifying information. This includes re-processing data to remove or generalise sensitive details, such as exact dates and specific geolocation information. For instance, location data is aggregated to a broader geographic level, excluding precise details like street addresses to mitigate any risk of re-identification.

Once the **Consortium Shared Data** has been populated into the final database, it becomes accessible to the HEAT Center Partners . This database serves as the primary resource for conducting research across the consortium, enabling the integration of **Climate/Weather Data** and **Areal/Geospatial Socio-Economic Data** for comprehensive analysis.

The database is then made available to all approved HE²AT Center partners, with clear guidelines for access and use established by the DMAC. This step ensures that the final dataset is secure and compliant with all relevant data protection regulations, including POPIA, and is ready for collaborative research across the consortium.

## Final data cleaning and analysis dataset creation

The final data cleaning step is essential for ensuring the quality and reliability of the **Consortium Shared Data** before it is used for analysis. During this stage, the Core HE²AT 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 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.

# Integration and analysis interfaces

## Integration of Climate and Socio-Economic Variables

Integrating climate and socio-economic variables within the HE²AT Center's data management workflow is streamlined to ensure efficiency while maintaining data quality. This process pulls relevant variables and indices from pre-existing datasets for the analysis period. The steps involved are as follows:

## 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.

## 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.

## 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.

## 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 ensures that the data retrieval process is seamless and that the analysis is grounded in the latest climate science.

## 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.

# Data analysis platforms

Data analysis will be facilitated through the CSAG/UCT Jupyter Hub and IBM platforms, providing robust and scalable environments for processing and analysing the HE²AT Center datasets.

## JupyterLab

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

* **Web-Based Platform**: Researchers can access Jupyter Hub 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**: 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.

## IBM data analysis platforms

IBM offers a suite of interfaces and tools for rapid and complex data analysis, which will be available to HE²AT Center analysts upon request.

## Integration approaches

To enhance the analytical capabilities, approaches to integrate data from IBM into the CSAG/UCT Jupyter Hub will be developed. This integration will allow analysts to leverage the strengths of both platforms:

* **Data Transfer Pipelines**: Establishing automated pipelines to securely transfer data between IBM and CSAG/UCT storage systems, ensuring data consistency and availability.
* **API Integrations**: APIs are utilised to enable seamless data access and manipulation across platforms, allowing Jupyter Hub notebooks to fetch and process data stored in IBM systems.
* **Unified Access**: Developing interfaces within Jupyter Hub that allow analysts to directly invoke IBM services.
* **Security and Compliance**: Ensuring that all data transfers and integrations comply with security standards and data governance policies, maintaining the confidentiality and integrity of the data.

By leveraging these platforms and integration approaches, the HE²AT Center will provide a comprehensive and flexible environment for data analysis, enabling advanced research and insights into heat and health impacts.

# De-identification

POPIA Section 10 prescribes the principle of Minimality, which means that only information relevant to the purpose of processing 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) guidelinesand informed by the findings in Zandbergen’s 2014 review on geographic masking strategies.[[4]](#footnote-4)

## Geographic aggregation:

Street addresses may be aggregated into larger geographical regions to prevent the derivation of individual residential locations. For instance, in RP2, where high spatial granularity is necessary to map urban heat-health outcomes, data will be aggregated at the level of census small areas or wards with spatial scales of 2 to 5 km. This approach ensures that multiple records map to the same region, mitigating the risk of re-identification.

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.

## 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)[[5]](#footnote-5), 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.

A diagram of different colors of a circle

Description automatically generated

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.

For both RP1 and RP2, **Gaussian displacement jittering** will be applied to shift latitude/longitude coordinates by tens of meters. This method effectively obscures individual residential locations while maintaining sufficient geographic detail for analysis. The displacement follows a Gaussian distribution, with adjustments based on local population density to ensure privacy. This approach will be implemented with input from UCT, IBM, and NIH experts, who will help balance data utility and privacy, ensuring that the jittering process meets both research and confidentiality requirements.

In this project, geographic aggregation and location jittering are employed based on the data's specific needs and the required level of privacy protection.

**Aggregation** is used when the research requires broader spatial patterns, such as RP2, where urban heat-health outcomes are analysed at a higher spatial level (e.g., census small areas or wards). This method reduces the risk of re-identification by grouping data into larger geographic areas, especially in low-density regions.

**Jittering** is applied when finer spatial detail is necessary, but privacy must still be protected. For instance, in RP1, jittering is used to slightly alter latitude/longitude coordinates by randomising them within a small radius, effectively obscuring exact locations while retaining sufficient accuracy for spatial analysis. **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.

**Expert Review and Validation**:

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.

**Assessing Re-identification Risk**:

The risk of re-identification will be quantified using spatial k-anonymity metrics, as described by Zandbergen (2014)[[6]](#footnote-6). 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.

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.[[7]](#footnote-7)

# Data sharing and open access

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).

## 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."

## 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 decision on whether the DS-I Africa ODSP or UCT's ZivaHub repository will serve as the primary mechanism for metadata search needs to be finalised.
* **Accessibility:** Data will be accessible openly through a public-facing component of the DMP data repository or via a data access request to the DAC (where a Data Transfer Agreement is required).
* **Interoperability:** Adherence to established data and metadata standards will ensure data interoperability (as outlined earlier in the document).
* **Reusability:** Rigorous data documentation will support reuse, including limitations and guidance for responsible reuse.

## 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:

1. **Original Study Data:** Raw, unprocessed health data collected directly from cohort studies and clinical trials. Access is restricted to the “Core Team” and retained for five years following project completion or DTA termination, with access limited to authorised personnel only.
2. **Consortium Shared Data:** Data that has undergone harmonization and integration, shared among approved Consortium partners. This data has been significantly altered, including applying privacy protections like pseudonymisation.
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.

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

The procedures for making RP1/RP2 De-identified Data available to qualified 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 Section 9.4 and Annexe 4, which contains the full terms of reference for the DAC.

1. **Development of Data Access Request Form:** The DAC will create a comprehensive Data Access Request Form for researchers outside the HE²AT Center. This form must be completed by those requesting access to RP1/RP2 De-identified Data and include:
   * **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 specific 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 or data visiting, 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 confidentiality and security of the data.
2. **Preliminary Screening:** Upon submission, the Data Access Request Form undergoes preliminary screening by the DAC to confirm completeness and basic compliance with the Consortium's requirements.
3. **Review by Data Access Committee:** The DAC evaluates the request based on the alignment with data use permissions outlined in the Data Transfer Agreement (DTA), potential privacy risks, and the avoidance of overlap with ongoing research. The full governance procedures are detailed in Section 9.4 and 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 and Encryption:** The data will be transferred upon approval and fulfilment of all conditions, including ethics approval. If the data contains personally identifiable information, it will be encrypted by the data provider before transfer using a secure service employing encrypted data transport protocols such as TLS (e.g., FileSender, Box.com).
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, as detailed in Section 9.4 and Annexe 4.

## Post-Funding Data Management and Open Access

Upon completing the HE²AT Center's research projects, inferential data will undergo a structured post-funding data management process. This involves two key pathways: maintaining the inferential data in-house within the HE²AT Center's infrastructure or transferring it to a designated external repository (Repository X) managed by an NIH-appointed Data Access Committee (DAC). The decision on the appropriate pathway will be guided by the data's relevance to ongoing or future research initiatives.

Once inferential data is managed according to these procedures, it will be available for open access. This final step ensures that the aggregated and anonymised data is publicly accessible, supporting broader research endeavours while safeguarding participant confidentiality. Inferential data in open-access repositories will be accompanied by comprehensive documentation to facilitate its reuse by the scientific community.

# 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 HE²AT 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.[[8]](#footnote-8)

# 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.

## 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 DSI Africa Consortium. The governance framework emphasises transparency, accountability, and compliance with relevant regulations, such as POPIA.

## Compliance Measures

To ensure that data management practices comply with both legal and ethical standards, the HE²AT Center has implemented several key measures:

* **Data Protection Impact Assessments (DPIAs):** These are required for personal information datasets. DPIAs assess privacy risks and implement mitigation strategies, such as encryption and limiting data exposure. If data is fully de-identified, a DPIA may not be necessary.
* **Data Transfer Agreements (DTAs):** Before data is shared, a DTA is executed between the provider and recipient, outlining the terms of use and prohibiting unauthorised sharing with third parties. This ensures that data transfers are legally and ethically sound.
* **Post-Funding Data Management:** After the project's conclusion, data management is sustained through either maintaining Inferential Data in-house or transferring it to Repository X, managed by an NIH-appointed DAC. In both cases, data is preserved for future scientific inquiry.

## 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 Consortium partners and external 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.

## Data Access Committees (DACs)

The DACs play a central role in the governance of data sharing. Their responsibilities include evaluating data access requests, ensuring alignment with ethical guidelines, and implementing protection measures, such as DPIAs. They also oversee DTAs to ensure compliance with legal and ethical standards.

For more detailed roles, responsibilities, and specific procedures, refer to **Annex 4: Terms of Reference (ToR) for the DAC.**

## Membership of the DAC

The DAC includes representatives from key stakeholders, such as:

* UCT Ethics Representative
* ELSI Team Representative
* eLwazi Representatives
* Study Principal Investigators
* Independent Scientists from outside the HE²AT Center

Refer to **Annexe 4** for a complete description of DAC membership and terms of reference.

# 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:

## 10.1. Retention Periods:

* **Original Study Data**: Retained for at least five years after the completion of the project. This includes raw, unprocessed data from cohort studies and clinical trials.
* **Consortium Shared Data**: Depending on the agreements in place, this data is retained indefinitely or for five years post-project or DTA termination. It has undergone initial harmonisation and is shared among HE²AT Center partners.
* **RP1/RP2 De-Identified Data**: Retained indefinitely. This data has been further processed and de-identified for specific analyses related to RP1 and RP2.
* **Inferential Data**: This is retained indefinitely. It includes aggregated and synthetic data derived from the analysis of the other data categories.

## **10.2.** Ongoing Monitoring

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

By adhering to these procedures, the HE²AT Center ensures that data sharing is conducted responsibly, ethically, and in a manner that maximises the utility of the data for future research while protecting the rights and privacy of data subjects.

# Restricted Data Access

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

## 11.1. Encryption of Personally Identifiable Data

Before transfer, the data provider encrypts any data containing personally identifiable information (PII) using strong encryption algorithms, such as AES-256. This ensures that the data is unreadable without a decryption key. Transport Layer Security (TLS) protocols are employed during transmission to maintain encryption and prevent interception. Both the data provider and recipient are notified upon successful transfer, and the UCT data platform verifies the data's integrity before authorized personnel decrypt it.

## 11.2. Data Storage and Encryption

Once transferred to UCT, any data identified as containing PII or specified by the DTA is encrypted for storage. The AES-256 encryption standard is applied, with encryption keys restricted to essential personnel responsible for anonymisation processes. Metadata, however, is stored separately to facilitate indexing and software development while maintaining security.

## 11.3. 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.

## 11.4. 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, 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.

# Roles and responsibilities

The table below details the various roles and responsibilities associated with the data management plan, as well as who is currently associated with each, their institution, and contact details

|  |  |  |  |
| --- | --- | --- | --- |
| **Role and responsibilities** | **People** | **Institution** | **Contact** |
| **DMAC PIs**  Responsible for ongoing (quarterly) assessment of data management and changes to the data management plan (annual) | Christopher Jack  Sibusisiwe Makhanya | UCT  IBM | [cjack@csag.uct.ac.za](mailto:cjack@csag.uct.ac.za)  sibusisiwe.makhanya@ibm.com |
| **Health data acquisition**  Identification of relevant health datasets, coordination and development of the DTA | Matthew Chersich for RP2  Craig Parker for RP2  Stanley Luchters for RP1 | WITS Planetary Health Research  The Centre for Sexual Health and HIV/AIDS Research Zimbabwe (CeSHHAR Zimbabwe) | [matthew.chersich@tcd.ie](mailto:matthew.chersich@tcd.ie)  [Craig.parker@witsphr.org](mailto:Craig.parker@witsphr.org)  stanley.luchters@ceshhar.co.zw |
| **Data processing and harmonization “Core Team” (including de-identification)**  De-identification, quality control, remapping, harmonisation and integration of all datasets  Note: These are the only individuals with access to encryption keys for original sensitive data | [Lisa van Aardenne](mailto:lisa.vanaardenne@uct.ac.za)  [Pierre Kloppers](mailto:pierre.kloppers@uct.ac.za)  Piotr Wolski  Peter Marsh  Nicholas Brink  Craig Parker | UCT  UCT  UCT  UCT  WITS Planetary Health Research  WITS Planetary Health Research | [lisa@csag.uct.ac.za](mailto:lisa@csag.uct.ac.za)  [pierre@csag.uct.ac.za](mailto:pierre@csag.uct.ac.za)  [wolski@csag.uct.ac.za](mailto:wolski@csag.uct.ac.za)  Peter.marsh.uct.ac.za  nicholas.brink@witsphr.org  [Craig.parker@witsphr.org](mailto:Craig.parker@witsphr.org) |
| **Managing access to the UCT data analysis platform** | Rodger Duffett | UCT | rodger@csag.uct.ac.za |
| **Managing access to the IBM platform** |  | IBM | sibusisiwe.makhanya@ibm.com |

# 

# Assessment and revision

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). These assessments will occur at least every six months to ensure the plan remains effective and up-to-date.

## 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.

## 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 HE²AT 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 HE²AT Center ensures that the Data Management Plan remains robust, effective, and aligned with best data management and analysis practices.

The DMAC co-PIs will periodically re-assess the Data Management Plan in consultation with the HE²AT Center Steering Committee, including the RP1 and RP2 leads. This will take place at least every six *months*.

The assessment will look at three aspects of the data management plan:

* Data process: Is the data processing workflow working effectively and resulting in data that is ready for analysis?

The Data Management Plan will be updated and changed based on the assessment. The DMAC co-PIs (see above) will propose revisions to the plan, which will be approved by the SC.

# Annexe 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). Data owners: ki platform and NICD repository | Preterm birth, pre-eclampsia, neonatal admission | African cohorts and trials conducted between 2000 and Oct 2020 | Research Project 1 |
| HIV Databases | Pooled health database from multiple large HIV trials conducted among adults in Johannesburg, South Africa (WHC studies) | Participants are followed up every 3-months for several years, 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. Humidity; Daily precipitation (total, rate); Dewpoint; 3-hrly Pressure Change | Spatial: Global coverage, 4km landmass and coastal waterways (hourly & daily res from 2015) | Determination of heat hazard; Thermal comfort metrics; combined 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 Classification); Aerosol Optical Thickness; NDVI sh layer. | 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 |
| 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 |

# Annexe 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 HEAT Center project. This includes the Core Team[Full name] hereby agrees 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 \_\_\_\_\_\_\_

# Annexe 3: Ethics notification letter

A logo of a university

Description automatically generated

**APPLICATION FOR A NOTIFICATION REGARDING AN APPROVED STUDY**

|  |  |
| --- | --- |
| **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  ……………………………………. |

# Annexe 4: Data access committee (DAC) Terms of Reference for the HE2AT Center

**1. Purpose**

The HE2AT Center's Data Access Committee (DAC) oversees and governs access to health-related data specifically processed and de-identified for Research Project 1 (RP1) and Research Project 2 (RP2). The DAC ensures that all data requests are reviewed and approved based on scientific merit, ethical considerations, 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.

**2. Scope of Data Types**

The DAC’s oversight focuses exclusively on the following data type:

**RP1/RP2 De-Identified Data**: Data that has been processed and de-identified specifically for analyses within Research Project 1 (RP1) and Research Project 2 (RP2). These projects focus on understanding the impacts of heat on health and urban vulnerability. Since this data is already de-identified, it can be shared in this form while ensuring compliance with relevant ethical and legal standards.

**3. Composition and Membership**

The DAC will comprise:

* A representative from the UCT Ethics Committee.
* A representative from the ELSI team.
* Representatives from the eLwazi platform.
* Principal Investigators (PIs) from relevant HEAT Center studies.
* Additional members with expertise in data science, ethics, and legal matters may be co-opted as needed​ (NIH HE2AT Center Data M…).

**4. Roles and Responsibilities**

* **Review of Data Requests**: The DAC is responsible for reviewing all data access requests for RP1/RP2 De-Identified Data. Each request will be assessed for alignment with the HE2AT Center's scientific goals, ethical considerations, and compliance with existing Data Transfer Agreements (DTAs) . Only requests that meet all 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. Incorporating Best Practices**

To enhance the process, the DAC will incorporate practices from the Vivli platform, particularly focusing on:

* **Independent Review**: An independent review panel may be established to provide impartial assessments of data requests, similar to Vivli’s Independent Review Panel.(maybe not necessary for us)
* **Accelerated Review Process**: An accelerated review process may be implemented for urgent data requests, allowing timely access to de-identified data.

**6. 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.

**7. Review Process**

* **Submission**: Data requests must be submitted using a standardized form that details the project objectives, required data, and ethical approvals. The form must be complete before it will be considered. This form is included in annex 4 of the DMP.
* **Preliminary Screening**: All requests will undergo preliminary screening to confirm completeness and compliance with the HEAT Center’s requirements.
* **Criteria for Approval**: The DAC will evaluate requests based on alignment with the HEAT Center's scientific agenda, scientific merit, feasibility, public health impact, and adherence to ethical and legal standards.

**8. Legal and Ethical Compliance**

* **Data Transfer Agreement (DTA)**: Before any data is shared, a DTA must be executed between the data provider and the data recipient. 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.

**9. Support and Collaboration**

The DAC will facilitate collaboration between data requesters and the HEAT Center’s resources to ensure that de-identified data is used effectively and accurately.

# Annex 5: Data Access Committee: Data Request form

**Requesting 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 at any time. The proposal will be reviewed by the xxx HE2AT Center Data Access Committee(DAC) using the following criteria:

* **Scientific merit**
  + Study is aligned with HE2AT Center scientific agenda and priorities
  + 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
* **Potential public health impact**
  + Study is relevant to one or more HE2AT Center study populations (infants, pregnant women, postpartum women)
  + Study will answer important public health questions or is in the critical pathway of research toward such answers
* **Feasibility**
  + Study population and variables required are available in the HE2AT Center data
  + Study is feasible within proposed timelines
* **Data management** 
  + The study proposal has outlined how data will be managed to ensure the protection of individual’s health data

 Please contact **[xxx](mailto:impaact.capsubmissions@fstrf.org)**with any questions.

All completed submissions should be sent to the HE2AT Center using this email: xxx.

**HE2AT Center Data Request Application Form**

Submit the completed form to the **HE2AT Center** for consideration using this email address: xxx

**GENERAL INFORMATION**

* + 1. **Request submitted by:**
* Name:
* Institution
* Email address:
  + 1. **Date submitted:**
    2. **Title:**
    3. **Lead investigator:**
* Name
* Institution
* Email address:
  + 1. **Co-investigator(s):** Include name(s), institution(s), email(s)
    2. **Relevant studies:** List all HE2AT Center studies from which data are being requested:
    3. **Relevant variables:** List all variables that are being requested:
    4. **Please indicate your preference regarding the type of file format you wish to receive.**

.csv (text)

.xlsx (Excel)

Other (please specify):

* + 1. **For access to HE2AT 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):

* + 1. **For access to HE2AT Center data, we will need to execute a data sharing agreement. Will you adhere to the terms and conditions of the data sharing agreement?**

Yes

No

**STUDY DESIGN**

1. **Background:** Include brief literature review, and any research gaps that the proposed analysis will fill (maximum 200 words)
2. **Study rationale:** (maximum 100 words)
3. **Study aims and objectives:**
4. **Study design and analysis:** (maximum 300 words)
5. **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)
6. **Timeline for completion:**
7. **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)

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*. [↑](#footnote-ref-1)
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. [↑](#footnote-ref-2)
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](https://doi.org/10.1136/bmjopen-2023-077529" \t "_new) [↑](#footnote-ref-3)
4. https://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/index.html [↑](#footnote-ref-4)
5. 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 [↑](#footnote-ref-5)
6. Zandbergen, P.A., *Ensuring Confidentiality of Geocoded Health Data: Assessing Geographic Masking Strategies for Individual-Level Data.* Adv Med, 2014. **2014**: p. 567049. [↑](#footnote-ref-6)
7. https://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification [↑](#footnote-ref-7)
8. 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 [↑](#footnote-ref-8)