# Title: Leveraging data science and machine learning for urban climate adaptation in Africa: a HE2AT CENTER STUDY protocol

Authors: Christopher Jack1\*, Craig Parker2\*, Yao Etienne Kouakou3,4, Bonnie R. Joubert5, Kimberly A. McAllister5, Maliha Ilias6, Gloria Maimela2, Matthew Francis Chersich2, Sibusisiwe Makhanya7, Stanley Luchters8,10, Prestige Tatenda Makanga8,11, Etienne Vos7 Kristie Ebi9, Brama Kone3, Akbar Waljee12, Gueladio Cisse3 on behalf of the HE2AT Center  
\*Equal first authors

HE2AT Center Group (alphabetical): Abdoulaye Tall, Adja Ferdinand Vanga, Christopher Jack, Craig Mahlasi, Iba Dieudonné Dely, James Mashiyane, Lisa van Aardenne, Madina Doumbia, Nicholas Brink, Pierre Kloppers, Piotr Wolski, Sibusisiwe Makhanya, Tamara Govindasamy, Toby Kurien

Author Affiliations:

1. Climate System Analysis Group, University of Cape Town
2. Wits RHI, University of the Witwatersrand, Johannesburg, South Africa
3. University Peleforo Gon Coulibaly, Korhogo, Côte d’Ivoire
4. Centre Suisse de Recherches Scientifique, Côte d’Ivoire
5. National Institute of Environmental Health Sciences, National Institutes of Health, Department of Health and Human Services, Durham, North Carolina, United States of America
6. National Heart Lung and Blood Institute, National Institutes of Health, Department of Health and Human Services, Bethesda, Maryland, United States of America
7. IBM Research Africa, Johannesburg
8. Centre for Sexual Health and HIV & AIDS Research (CeSHHAR), Zimbabwe
9. The University of Washington, Seattle, United States of America
10. Liverpool School of Tropical Medicine, Liverpool, UK; Department of Public Health and Primary Care, Ghent University, Belgium
11. Surveying and Geomatics Department, Midlands State University, Gweru, Zimbabwe
12. Department of Internal Medicine, Gastroenterology & Hepatology, University of Michigan Medical School, Ann Arbor, Michigan, United States of America

Correspondence to Dr Christopher Jack, [cjack@csag.uct.ac.za](mailto:cjack@csag.uct.ac.za)

(Word count: 3927 )

ABSTRACT

**Introduction:** African cities, particularly Abidjan and Johannesburg, face challenges of rapid urban growth, informality, and strained health services, compounded by increasing temperatures due to climate change. This study aims to understand the complexities of heat-related health impacts in these cities. The objectives are: 1) mapping intra-urban heat risk and exposure using health, socio-economic, geospatial climate, and satellite imagery data; 2) creating a stratified heat-health outcome forecast model to predict adverse health outcomes; and 3) establishing an Early Warning System for timely heatwave alerts. The ultimate goal is to foster climate-resilient African cities, protecting disproportionately affected populations from heat hazards.

**Methods and Analysis:** The research will acquire health-related datasets from eligible adult clinical trials or cohort studies conducted in Johannesburg and Abidjan between 2000 and 2022. Additional data will be collected, including socio-economic, geospatial climate datasets and satellite imagery. These resources will aid in mapping heat hazards and quantifying heat-health exposure, the extent of elevated risk, and morbidity. Outcomes will be determined using advanced data analysis methods, including statistical evaluation, machine learning, and deep-learning techniques.

**Ethics and Dissemination:** The study has been approved by the Wits Human Research Ethics Committee (reference no: 220606). Data management will follow approved procedures. Results will be disseminated through workshops, community forums, conferences, and publications. Data deposition and curation plans will be established in line with ethical and safety considerations.

(Word count: 228 )

**Keywords:** urban, heat, heatwaves, health, Early Warning Systems, intra-urban, socio-economics and environment, exposure mapping, hazard mapping, heat-related health impacts, African cities, data science and machine learning, temperature

Article Summary

Study strengths and limitations

1. Employs comprehensive data collection from clinical, socio-economic, and remote sensing sources, ensuring a multidimensional analysis of urban heat exposure.
2. Leverages state-of-the-art machine learning techniques for predictive modelling of heat-health outcomes, advancing the field of environmental health research.
3. A cross-disciplinary approach enriches the interpretation of data, linking climate science with public health implications.
4. Risk of sampling bias due to secondary data utilisation, which may influence the representativeness of findings.
5. The spatial resolution of datasets, particularly those capturing microclimatic urban variations, may limit the granularity of exposure assessments, affecting the precision in capturing heat stress metrics.

# Introduction

The HE2AT Center (HEat and HEalth African Transdisciplinary Center), a consortium spanning South Africa, Côte d'Ivoire, Zimbabwe, and the United States, embodies global collaboration. Funded through the United States NIH "Harnessing Data Science for Health Discovery and Innovation in Africa" (DS-I Africa) program, the Center amalgamates diverse expertise in pursuit of comprehensive urban climate resilience strategies[1].

This study emerges from the HE²AT Center as a Research Project (RP) aiming to interrogate the intricate relationships of urban spaces to heat-health impacts, emphasising the need for nuanced responses. It highlights the disproportionate risks borne by residents of impoverished areas, the elderly, those with pre-existing health conditions, children, outdoor workers, and inhabitants of densely populated or informal settlements—groups for whom the urban heat island effect is a daily lived reality[2-4].

Research on heat-related health risks in Africa, including seminal works in Abidjan and Johannesburg, reveals a critical need for localised interventions. Ncongwane et al. (2021), Pasquini et al. (2020), and Wright et al. (2019) lay the groundwork, explaining the socio-economic and infrastructural factors that exacerbate heat-health vulnerabilities[5-7].

Enhanced nighttime heatwaves over African urban clusters, as investigated by Eghosa Igun et al. (2022), underline the growing threat of heatwaves exacerbated by urban heat island effects [8]. Furthermore, an assessment of the health-related impacts of urban heat Islands (UHI) in Douala Metropolis, Cameroon, by Enete et al. (2017) provides insight into the localised health burdens of urban heat [9].

Building on this foundation, our study seeks to contribute to this burgeoning field by creating an effective, data-driven Urban Heat Health Early Warning System (EWS) tailored to the unique socio-demographic makeup of African metropolises. Integrating insights from recent studies, including Guillaume Thibaut Rohat et al.'s "Human Exposure to Dangerous Heat in African Cities" (2019), which assesses human exposure to extreme heat conditions [10], our research aims to offer a holistic understanding and innovative solutions to mitigate these escalating health risks.

The study is structured around three primary objectives(see Figure 1): (1) mapping intra-urban heat risks, (2) developing a heat-health outcome forecast model, and (3) establishing an EWS that empowers both policymakers and the public with actionable insights for preemptive action. These are inspired by the robust frameworks and pioneering methods established by Thiaw et al. (2022) and Chapman et al. (2022), who have significantly advanced the field of heat-health early warning systems[11, 12].

Our approach is grounded in the IPCC's hazard-vulnerability-exposure paradigm, as evidenced by the Key Concepts and Definitions in Heat Exposure Studies (Table 1). This alignment ensures consistency with the globally recognised framework and reinforces our research's applicability to the broader discourse on climate change and public health. The terms "exposure," "vulnerability," "hazard," and "adaptive capacity" are defined in Table 1, providing a clear conceptual framework for our study.

By integrating state-of-the-art machine learning techniques with comprehensive socio-economic and geospatial data as well as clinical trial/cohort health datasets, this study endeavours to provide stakeholders with a granular understanding of heat-health dynamics, ultimately aiding in the formulation of targeted interventions that can bolster the resilience of urban populations amidst the escalating challenges posed by global warming.

Table 1: Key Concepts and Definitions in Heat Exposure Studies (Aligned with IPCC Framework)

|  |  |
| --- | --- |
| **Concept** | **Description** |
| Exposure | The presence of people, livelihoods, species or ecosystems, environmental functions, services, resources, infrastructure, or economic, social, or cultural assets in places that heat could adversely affect. |
| Vulnerability | The propensity or predisposition to be adversely affected encompasses various concepts and elements, including sensitivity or susceptibility to harm and lack of capacity to cope and adapt to heat. |
| Hazard | The potential occurrence of a natural or human-induced physical event or trend that may cause loss of life, injury, or other health impacts, as well as damage and loss to property, infrastructure, livelihoods, service provision, ecosystems, and environmental resources. |
| Adaptive Capacity | The ability of a population to adjust to heat is linked to socioeconomic factors, resource access, institutional support, and social determinants of health (SDOH) and is often diminished in urban poor due to limited access to cooling resources and health services. |
| Risk | There is a potential for adverse consequences when hazards interact with vulnerable and exposed elements. It is often represented as the probability of occurrence of hazardous events or trends multiplied by the impacts if these events or trends occur. Risk results from the interaction of vulnerability, exposure, and hazard. In the context of heat, it refers to the likelihood and severity of negative outcomes due to heat exposure, considering the vulnerability and adaptive capacity of the affected population or system. |

# Study setting

Abidjan, located in Côte d'Ivoire, and Johannesburg, in South Africa, are cities experiencing rapid urbanisation—defined as the population shift from rural to urban areas along with the corresponding change in land use—compounded with stress on health services and increasing temperatures owing to climate change [13-15]. In Johannesburg, a diverse metropolis of 6.1 million people, HIV/AIDS, tuberculosis, and non-communicable diseases pose significant challenges. These are intensified by urbanisation, socio-economic disparities, and broader social determinants of health (SDOH) like education and employment [15-17]. Areas with less vegetation and higher levels of poverty face greater heat impacts, a reflection of the 'Green Apartheid' that characterises the city's urban forest and its accessibility. [18]​​. Similarly, in Abidjan, an economic centre with a population of 6.3 million, diseases such as malaria and non-communicable diseases are driven by urbanisation and wider SDOH [19-21].

Both cities represent Urban Heat Islands (UHI), a phenomenon where urban areas exhibit higher temperatures than their rural surroundings due to human activities[22] [23]. While Johannesburg's extensive urban forest offers some respite, Abidjan's Cocody district is increasingly experiencing the UHI effect due to accelerated urbanisation and land use modifications. These evolving urban landscapes underscore the requirement for holistic health strategies in both cities[24].

Abidjan and Johannesburg were selected for this study due to their unique characteristics and data availability. As cities with high population density and experiencing rapid urbanisation, Abidjan and Johannesburg represent the challenges facing many African cities in the context of climate change and heat-related health impacts. Additionally, these cities can access critical detailed health data from clinical trials and cohort studies. Both cities, therefore, enable a focused examination of heat-related health risks in urban African settings, potentially informing broader regional strategies for climate adaptation and public health.

# Methods

The study plans to combine datasets from many sources encompassing various fields—health, climate, environment, and SDOH. This multi-faceted approach will aid in building more thorough and locally pertinent models of heat-related health outcomes. These models will consider the diverse range of day-to-day realities and experiences encountered by inhabitants within each city, capturing how they impact their health in the context of heat [25]. In this study, 'lived experiences' refers to individuals' unique daily conditions, challenges, and opportunities shaped by their specific SDOH and environmental circumstances. Additionally, multiple datasets within a particular domain (e.g., multiple health trial datasets) both increase the statistical sample sizes for more robust modelling as well as enable a rigorous quantification of key uncertainties (e.g., multiple climate datasets) [26, 27].

* 1. Socio-economic and environmental data

This research will collect socio-economic geospatial data, which includes information on household economic conditions, service availability, and residential characteristics—referring to factors like housing type, construction materials used, and the quality and condition of living spaces [28]. The data will include national census records, specialised household, and demographic surveys and encompass details about individual and household income, education, occupation, living circumstances, and accessibility to healthcare, education, and transportation services. [29] For Johannesburg, the Gauteng City-Region Observatory (GCRO) datasets will provide key variables for the study. In the case of Abidjan, equivalent data will be sourced from the National Institute of Statistics (INS) of Côte d'Ivoire, which provides comprehensive socio-economic and demographic data [29, 30].

Remote sensing data will be retrieved from satellite sensors, including optical images and indicators of physical aspects such as land surface temperature, soil moisture, vegetation condition, and land use and coverage [31]. Where available, researchers will amalgamate data from current sensor networks with urban land use and building density details to create a model of urban land use heat [28, 29]. Although Landsat and MODIS data primarily measure land surface temperature (LST), statistical models can estimate air temperature from remotely sensed LST. However, it should be noted that LST may not fully capture heat stress experienced in urban areas. In this study, appropriate statistical models will be used to indirectly retrieve air temperature from the LST data provided by Landsat and MODIS, and where possible, we will incorporate humidity data to provide a more comprehensive assessment of heat stress [32].

Climate-associated data will be sourced from open data repositories, such as the Copernicus Climate Data Store (CDS) and Earth System Grid Federation (ESGF), offering observational-based datasets, historical re-analyses, and climate simulations While the Copernicus Climate Data Store (CDS) and Earth System Grid Federation (ESGF) provide valuable climate data, their spatial resolution may not be sufficient to distinguish different parts within the city[33]. To address this limitation, we will employ downscaling techniques to enhance the spatial detail of our geospatial climate data. Specifically, we will explore dynamic downscaling with high-resolution climate models such as the Weather Research and Forecasting (WRF) and UrbClim urban climate models. These models offer detailed results on heat stress for cities, allowing for a more precise analysis of intra-urban heat variations and can improve the accuracy of our heat risk assessments for Johannesburg and Abidjan [34, 35].

Additionally, the IBM-PAIRS platform will be employed as a source of climate data, including data from climate models, weather stations, and satellite observations[36]. To further enhance our analysis, we will integrate datasets from the European Space Agency's WorldCover portal and the Global Human Settlement Layer (GHSL), which provide detailed land cover and human settlement data, respectively[37, 38]. This will provide a comprehensive snapshot of Africa's past and future climate conditions, including the frequency, duration, and intensity of heat waves.

* 1. Health trials and cohort data

The health data for this study will be collected from clinical trials and cohort studies, such as HIV drug trials and COVID-19 vaccine trials. These studies typically involve many participants (hundreds to thousands) and are conducted over an extended period (multiple years) within a specific geographical area. They provide detailed longitudinal individual health data for building statistical models relating time-varying predictors to health outcomes. Potential outcomes of interest include cardiovascular events, respiratory issues, kidney conditions, and mental health impacts, which may be exacerbated by heat exposure in urban environments[39].

More specifically, the health cohort data integrated into the study will be identified based on the availability of three classes of variables within each study:

* Clinical variables: including vital signs (e.g., body temperature, blood pressure, and heart rate), indicators of heat-related illness (e.g., headache, dizziness, fatigue, and nausea), and details on pre-existing medical conditions (e.g., hypertension, diabetes, and cardiovascular disease) that could increase the risk of heat-related illness, and documentation of adverse events potentially related to heat exposure.
* Laboratory variables: including blood tests (e.g., electrolyte levels, liver and kidney function tests), markers of inflammation and oxidative stress, HIV tests, including viral load and CD4 count, and COVID-19 test results.
* Demographic and SDOH variables: involving basic demographic information (e.g., age, sex, race, and ethnicity), socio-economic factors (e.g., education, income, and occupation), and data on housing and urban infrastructure (e.g., air conditioning availability, ventilation, and shading) that could influence heat exposure and the degree to which individuals and households are at an increased risk.

In response to the shifts in mortality and morbidity during the 2020-2022 COVID-19 pandemic, we will analyse data separately for pre-pandemic, pandemic, and post-pandemic periods. Additionally, we will include COVID-19-related variables as covariates in our models to control for the pandemic's impact on health outcomes.

Table 2: Summary of Data Sources for Each Objective

|  |  |
| --- | --- |
| **Objective** | **Data Sources** |
| 1. Mapping intra-urban heat risk and exposure | - Socio-economic data (census, surveys, GCRO datasets)  - Geospatial data (land use, building density, OpenStreetMaps)  - Climate data (WRF, UrbClim models, downscaled CDS & ESGF data, IBM-PAIRS platform) |
| 2. Creating a stratified heat-health outcome forecast model | - Health data with clinical variables (e.g., vital signs, heat-related illness indicators)  - High-resolution urban temperature hazard maps (Landsat, MODIS data with statistical models for air temperature estimation)  - Remote sensing data (satellite imagery, land surface temperature, soil moisture, vegetation condition)  - Socio-economic and environmental data (household economic conditions, service availability, residential characteristics) |
| 3. Establishing an Early Warning System | - Integrated health and socio-economic data  - Geospatial heat hazard maps  - Health outcome forecast model outputs  - COVID-19 incidence and mortality rates (for pandemic period adjustment)  - Risk profile data (demographic groups, health conditions, locations, socio-economic statuses) |

* 1. Integration of Datasets

Our study relies on integrating socio-economic, clinical, environmental, and geospatial data to understand heat's impact on health in African cities. We will cross-reference health trial participant geolocations with socio-economic and environmental data, applying spatial jittering to protect privacy while retaining spatial trends. Additionally, we'll incorporate remote sensing and climate data to examine how environmental changes affect health outcomes related to heat exposure.

In pursuit of our research objective to explore the correlation between heat and health within the urban environments of Johannesburg and Abidjan, we have developed a comprehensive strategy to systematically identify relevant clinical trials and cohort studies. This strategy involves searching key databases using a combination of MeSH (Medical Subject Headings) and free-text terms, including study location, diseases of interest, the number of participants, study type, collected data, and the timeframe of study conduction. Our targeted search terms are designed to retrieve studies that provide robust clinical, laboratory, and demographic data relevant to the impact of heat on health outcomes.

To identify potentially relevant studies, a two-step dual independent review process will be employed. Initially, studies will be screened based on their titles and abstracts. Subsequently, potentially eligible studies will be procured in their full-text format for a more thorough assessment against our pre-defined selection criteria (Table 3).

Health researchers will evaluate the quality of the selected studies through a peer-reviewed tracking tool to ensure their scientific soundness and reliability. The data will be collated and synthesised, and any discrepancies will be addressed and resolved through consensus discussions among team members.

The following criteria outlined in Table 3 will be used to select research projects to be considered for inclusion in our study.

**Table 3: Eligibility Criteria for Research Project 2**

|  |  |
| --- | --- |
| Criteria | Description |
| Study type | Cohort or trial with at least 200 adult participants |
| Study location | Johannesburg or Abidjan, or both cities |
| Study design | Randomised or non-randomised clinical trial, or observational or interventional cohort with prospectively collected data |
| Data collected | At least two of the clinical or lab variables |
| Ethics approval | Local ethics approvals obtained |

Access to relevant trials and cohort data is crucial for this project's success. In the event of data unavailability or sharing restrictions, we have contingency plans to ensure the project's progression. These include exploring alternative data sources such as the National Health Laboratory Service (NHLS), adjusting the study's scope, and utilising synthetic data if necessary.

* 1. Managing bias

Managing potential biases is critical to ensuring our study's integrity and robustness, as outlined in the following strategy.

Primarily, our approach will involve carefully selecting health data sources, ensuring they meet established quality criteria and represent diverse demographic and geographic segments within our target cities of Johannesburg and Abidjan. This strategy will assist us in avoiding selection bias that could skew our findings [40].

We will adjust the analysis phase when potential biases are identified. Specific statistical methods like propensity score matching, inverse probability weighting, and stratification will be applied. These methods help to control for confounding variables and reduce bias in observational studies, increasing the validity of our outcomes [41].

* 1. Objective One: Assessing the degree of increased risk within cities

The primary goal of our research is to assess and illustrate the socio-economic and environmental factors that increase risk, in conjunction with exposure to heat hazards, as defined in Table 1, in two African cities. We will utilise the comprehensive datasets described earlier to accomplish this.[42-44]

In our study, we have opted for Principal Component Analysis (PCA) for dimensionality reduction due to its interpretability and computational efficiency, which are essential for our large datasets. PCA ranks new uncorrelated components by explaining variance and aiding in identifying key factors. Although PCA's linearity may miss nonlinear patterns, and its variance-based importance assumption may not align with the context of heat-related health risks, PCA's robustness makes it suitable for initial high-dimensional data exploration[45-49].

* 1. Objective Two: Creating a geographically and demographically stratified heat-health outcome forecast model

The second objective of this study is to construct a geographically and demographically stratified heat-health outcome forecast model designed to predict adverse health outcomes at varying temperature thresholds for different populations and neighbourhoods.

This involves creating high-resolution urban temperature hazard maps. We will use remote sensing, statistical downscaling, and combined modelling to derive near-surface air temperatures from Landsat and MODIS data [50]. For instance, Gudmundsson and Seneviratne (2021) demonstrated using such models to predict air temperature from LST data. Therefore, while Landsat and MODIS data are not direct measures of air temperature, they can be indirectly used for air temperature retrieval by applying an appropriate statistical model [32].

These temperatures will then be validated using weather station records and land-use maps. The resulting heat hazard maps will serve as a critical input for the subsequent stages of our machine learning pipeline.

Once generated, the temperature hazard maps will be integrated with health datasets. This combined dataset will then undergo feature engineering. Feature engineering is a crucial step in machine learning and involves selecting and transforming relevant predictors that better represent the underlying data patterns [51]. The features will be derived from the high-resolution temperature hazard fields and spatially disaggregated variables from the health datasets.

With the features engineered, we will apply various standard machine learning models, such as decision trees, linear and quantile regression trees, support vector machines, and logistic regression [52]. These models are chosen for their proven effectiveness in capturing relationships in complex datasets.

Additionally, we will explore deep recurrent neural networks, specifically Gated Recurrent Units (GRUs) and Long Short-Term Memory (LSTM) networks, due to their ability to model temporal dependencies in time-series data, essential for predicting heat-related health outcomes. While these models are state-of-the-art in computer science, their application in heat-health studies is still emerging, as demonstrated in a review of recent literature, including studies by Boudreault et al. (2023, 2024), Wang et al. (2023, 2021, 2020), Lee et al. (2022), and Nishimura et al. (2021) [53-59]. However, recognising that simpler statistical models may be effective, we plan to build on the work by Boudreault et al.(2024) to compare the performance of deep learning models with tree-based approaches and nonlinear statistical models in our analysis [55].

Throughout this process, we will assess the significance of predictors for different populations within the two cities. This will allow us to identify varying susceptibility levels to heat-induced health conditions based on demographics and risk factors. Potential health co-morbidities to be explored include cardiovascular disease, respiratory disease, renal disease, and HIV status [60].

We will use k-fold cross-validation to assess model performance and generalizability, train models on a designated set, and calibrate them with grid or random search techniques. Validation will occur on a separate set to evaluate generalisation, using metrics like accuracy, precision, recall, F1 score, MSE, and MAE. Special attention will be paid to model performance during heatwave periods to ensure effectiveness in predicting heat-related health outcomes.[61]

An iterative process of model refinement and validation will ensure the ongoing relevance of our model and enable us to continually improve the model's performance and maintain its applicability to the evolving urban heat-health landscape [62].

* 1. Sample size considerations for health datasets

To ensure our models' robustness and accuracy, a comprehensive, four-step approach will determine the required sample size for our health datasets. Based on documented best practices (see Riley et al. in BMJ (2020)), this approach will consider factors like the number of predictor parameters, event prevalence, anticipated model performance, and overfitting prevention strategies. Our methodology is adaptable to various outcome types, including continuous, binary, and time-to-event data, ensuring our predictive models are statistically robust and clinically relevant. By rigorously applying this tailored procedure to health data, we aim to build models that accurately capture the target populations' heat-health dynamics and risk profiles​.[63]

* 1. Objective Three: Develop an Early Warning System reflective of geospatial and individualised risk profiles

The third objective is to develop an Early Warning System (EWS) that integrates geospatial and individualised risk profiles of heat-related health impacts in Abidjan and Johannesburg, as depicted in Figure 2. The EWS aims to provide actionable insights to stakeholders, including community health workers, clinic managers, urban planners, and at-risk individuals. It combines high-resolution heat hazard maps and a forecast model to generate alerts for areas with predicted adverse heat-health outcomes. This involves refining the forecast model, merging it spatially with heat hazard maps, and generating timely alerts. The EWS also incorporates heat hazard predictions for proactive risk management, offering tailored guidance for at-risk individuals on hydration and activity scheduling. Inspired by the Ahmedabad Heat Action Plan, our system emphasises inter-agency coordination and community outreach for effective heat risk mitigation [64].

While our EWS aims to provide advanced warnings, we acknowledge the challenges of long-term forecasting. Prediction accuracy depends on data reliability, model complexity, and weather variability. Continuous model refinement is essential for improving predictive capabilities.

* 1. Patient and Public Involvement Statement

Public and patient input is integral to our study, especially informing our Early Warning System design: this input will guide risk mitigation strategies and the development of user-friendly, actionable digital tools.

* 1. Project timeline

The project is funded to run from 2022- 2026.

# Ethical approval and protection of human subjects

This research study received ethical approval from both the Wits Human Research Ethics Committee in Johannesburg (reference number 200606) on June 30, 2022, and the National Ethics Committee for Life and Health Sciences, Côte d'Ivoire, on November 25, 2022 (reference number 176-22/MSHPCMU/CNESVS-kp) and will follow the United States Department of Health and Human Services regulations for the protection of human subjects in research (45 CFR 46). Our research protocol has two critical ethical and legal considerations: informed consent for secondary data usage and the protection of potentially identifiable information.

Regarding informed consent for secondary data usage, we will critically examine the consent procedures intended for the original study. If a participant has previously provided "broad consent", permitting the use of their data in future research endeavours, we can share their data without additional ethical approvals. Careful deliberation is required for participants who have granted "narrow consent, " which restricts data sharing beyond the original study purpose. If obtaining renewed consent is unfeasible or involves a disproportionate effort, we will seek an informed consent waiver from the appropriate ethics committee.

To protect potentially identifiable information and minimise privacy risks (such as indirect identifiers like geographical data in the collected data), we will employ several protective measures, including the restriction of identifiable data and the non-use of real names or other identifying factors. Data will be stored on a password-protected server with limited access. Following data minimisation principles, we will retain only the data essential for achieving our study objectives. When applicable, we will anonymise data through geographical aggregation and jittering, especially when home addresses are used.

Finally, we acknowledge the specific legislative requirements for using health data in different countries, including the laws surrounding the cross-border transfer of such data. We will, therefore, require data providers to provide a contractual guarantee, as part of the data sharing agreement, that all original studies followed appropriate informed consent procedures and that the sharing of this data complies with all relevant data protection laws.

# Study oversight

Prof. Chersich, Prof. Luchters, and the Hub Administrator direct the HE2AT Center. Steering Committee members represent six South, East, and West African institutes. This study is led by Prof. Cisse of Ivory Coast's Peleforo Gon Coulibaly University and co-led by Dr Christopher Jack of the University of Cape Town.

# Dissemination

Prompt dissemination of research findings is crucial to the HE2AT Center's effectiveness. We devised a strategy detailing publication types, authors, and release dates. Our findings will be shared with research and relevant working partners to inform various levels of activities and update recommendations as needed. Timely dissemination is vital to the HE2AT Center's success and mission.

# Study status

Ongoing.

# Author Contributions**:**

# CJ, CP, SL, MFC, and GC were involved in the research's conception and design. CP, GM, and MFC obtained ethics approval. CP, MFC, and YEK were involved in data acquisition. CP, MFC and SL prepared the figures, and CP and CJ drafted the manuscript. AW was involved in the conception and design, reviewing the structure of the paper. All authors (CJ, CP, YEK, BRJ, KAM, MI, GM, MFC, SM, SL, PTM, EV, KE, BK, AW, and GC) were involved in the planning, conduct, and reporting of the work, editing and revising the manuscript, and approving the final version for submission.

# Funding statement:

Research reported in this publication was supported by the Fogarty International Center , the National Institute of Environmental Health Sciences (NIEHS), and OD/Office of Strategic Coordination (OSC) of the National Institutes of Health under Award Number U54 TW 012083. The content is solely the authors' responsibility and does not necessarily represent the official views of the National Institutes of Health.

# Competing interests

The authors declare potential competing interests: MF, GM, and CP have pension fund investments in the fossil fuel industry. The University of the Witwatersrand holds endowments and financial reserves invested in the same industry.

Data sharing statement:  In accordance with the NIH data sharing policy, data from the HEAT002 study will be made available to the scientific community. Researchers interested in accessing this data should submit a detailed request to Chris Jack at [cjack@csag.uct.ac.za](mailto:cjack@csag.uct.ac.za). The request will be reviewed, and data will be shared subject to the approval of the request, ensuring the purpose aligns with ethical standards and that the participants' privacy is protected.

# References

1.  *Harnessing Data Science for Health Discovery and Innovation in Africa (DS-I Africa). Retrieved from* [*https://commonfund.nih.gov/AfricaData*](https://commonfund.nih.gov/AfricaData)*.*

2. Johnson, D.P., J.S. Wilson, and G.C. Luber, *Socioeconomic indicators of heat-related health risk supplemented with remotely sensed data.* International Journal of Health Geographics, 2009. **8**(1): p. 57.

3. Jung, J., et al., *Heat illness data strengthens vulnerability maps.* BMC Public Health, 2021. **21**(1): p. 1999.

4. Xu, R., et al., *Socioeconomic level and associations between heat exposure and all-cause and cause-specific hospitalization in 1,814 Brazilian cities: A nationwide case-crossover study.* PLoS Medicine, 2020. **17**(10): p. e1003369.

5. Ncongwane, K.P., et al., *A Literature Review of the Impacts of Heat Stress on Human Health across Africa.* Sustainability, 2021. **13**(9): p. 5312.

6. Pasquini, L., et al., *Emerging climate change-related public health challenges in Africa: A case study of the heat-health vulnerability of informal settlement residents in Dar es Salaam, Tanzania.* Sci Total Environ, 2020. **747**: p. 141355.

7. Wright, C.Y., et al., *Socio-economic, infrastructural and health-related risk factors associated with adverse heat-health effects reportedly experienced during hot weather in South Africa.* Pan Afr Med J, 2019. **34**: p. 40.

8. Igun, E., et al., *Enhanced nighttime heatwaves over African urban clusters.* Environmental Research Letters, 2022. **18**.

9. Enete, I., *Assessment of Health Related Impacts of Urban Heat Island (UHI) in Douala Metropolis, Cameroon.* International Journal of Environmental Protection and Policy, 2014. **2**: p. 35.

10. Rohat, G., et al., *Projections of human exposure to dangerous heat in African cities under multiple socioeconomic and climate scenarios.* Earth's Future, 2019. **7**(5): p. 528-546.

11. Thiaw, W.M., et al., *Toward Experimental Heat–Health Early Warning in Africa.* Bulletin of the American Meteorological Society, 2022.

12. Chapman, S., et al., *Past and projected climate change impacts on heat-related child mortality in Africa.* Environmental Research Letters, 2022. **17**(7): p. 074028.

13. Lwasa, S., *Managing African Urbanization in the Context of Environmental Change.* Interactions, 2014. **2**.

14. Wang, Y.P. and K. Kintrea, *Urban Expansion and Land Use Changes in Asia and Africa.* Environment and Urbanization Asia, 2021. **12**: p. S13 - S17.

15. Abrahams, C. and D. Everatt, *City Profile: Johannesburg, South Africa.* Environment and Urbanization Asia, 2019. **10**: p. 255 - 270.

16. Rees, H.V., et al., *At the Heart of the Problem: Health in Johannesburg’s Inner-City.* BMC Public Health, 2017. **17**.

17. *Macrotrends. (2023). Johannesburg, South Africa Metro Area Population 1950-2023. Retrieved May 23, 2023, from* [*https://www.macrotrends.net/cities/22486/johannesburg/population*](https://www.macrotrends.net/cities/22486/johannesburg/population)*.*

18. Venter, Z.S., et al., *Green Apartheid: Urban green infrastructure remains unequally distributed across income and race geographies in South Africa.* Landscape and Urban Planning, 2020. **203**: p. 103889.

19. Granado, S., et al., *Appropriating "malaria": local responses to malaria treatment and prevention in Abidjan, Cote d'Ivoire.* Med Anthropol, 2011. **30**(1): p. 102-21.

20. Djomand, G., et al., *Virologic and immunologic outcomes and programmatic challenges of an antiretroviral treatment pilot project in Abidjan, Côte d'Ivoire.* Aids, 2003. **17 Suppl 3**: p. S5-15.

21. *World Population Review. (2023). Abidjan Population 2023. Retrieved May 23, 2023, from* [*https://worldpopulationreview.com/world-cities/abidjan-population*](https://worldpopulationreview.com/world-cities/abidjan-population)*.*

22. Souverijns, N., et al., *Urban heat in Johannesburg and Ekurhuleni, South Africa: A meter-scale assessment and vulnerability analysis.* Urban Climate, 2022. **46**: p. 101331.

23. *UN-Habitat. (n.d.). South Africa. Retrieved May 22, 2023, from* [*https://unhabitat.org/south-africa*](https://unhabitat.org/south-africa)*.*

24. Dongo, K., M. Kablan, and F. Kouamé, *Mapping urban residents’ vulnerability to heat in Abidjan, Côte d’Ivoire.* Climate and Development, 2018. **10**: p. 1-14.

25. Wolf, S.T., D.J. Vecellio, and W.L. Kenney. *Adverse heat-health outcomes and critical environmental limits (PSU HEAT Project)*. 2022.

26. Schubert, S. *An Update on Experimental Climate Prediction and Analysis Products Being Developed at NASA's Global Modeling and Assimilation Office*. 2011.

27. Riedel, M., S.E. Dosso, and L. Beran, *Uncertainty estimation for amplitude variation with offset (AVO) inversion.* Geophysics, 2003. **68**(5): p. 1485-1496.

28. Alonso, L. and F. Renard, *A Comparative Study of the Physiological and Socio-Economic Vulnerabilities to Heat Waves of the Population of the Metropolis of Lyon (France) in a Climate Change Context.* International Journal of Environmental Research and Public Health, 2020. **17**(3): p. 1004.

29. *Gauteng City-Region Observatory (2019). Quality of life in the Gauteng city-region: A report on key indicators. Retrieved from* [*https://www.gcro.ac.za/about/annual-reports/*](https://www.gcro.ac.za/about/annual-reports/)*.*

30. National Institute of Statistics of Côte, d.I., *National Institute of Statistics of Côte d'Ivoire Datasets*. INS.

31. Hofierka, J., M. Gallay, and K. Onačillová, *Physically-based land surface temperature modeling in urban areas using a 3-D city model and multispectral satellite data.* urban climate, 2020. **31**: p. 100566.

32. Hooker, J., G. Duveiller, and A. Cescatti, *A global dataset of air temperature derived from satellite remote sensing and weather stations.* Scientific Data, 2018. **5**(1): p. 180246.

33. Kershaw, P., et al. *Delivering resilient access to global climate projections data for the Copernicus Climate Data Store using a distributed data infrastructure and hybrid cloud model*. 2019.

34. *Copernicus Climate Data Store (CDS)*. 2024, Copernicus Climate Change Service (C3S).

35. *Earth System Grid Federation (ESGF)*. 2024, ESGF.

36. Albrecht, C.M., et al., *Pairs (Re)Loaded: System Design & Benchmarking For Scalable Geospatial Applications.* 2020 IEEE Latin American GRSS & ISPRS Remote Sensing Conference (LAGIRS), 2020: p. 488-493.

37. *10 m WorldCover 2020 v100*. 2021, European Space Agency (ESA).

38. *The Global Human Settlement Layer 2019 (GHSL 2019) public release*. 2021, Publications Office of the European Union, Luxembourg.

39. Arifwidodo, S.D., P. Ratanawichit, and O. Chandrasiri. *Understanding the Implications of Urban Heat Island Effects on Household Energy Consumption and Public Health in Southeast Asian Cities: Evidence from Thailand and Indonesia*. 2020.

40. Narod, S.A., *Countercurrents: The Bias of Choice.* Current Oncology, 2019. **26**(6): p. 712-713.

41. Schwartz, R., et al., *Towards a Standard for Identifying and Managing Bias in Artificial Intelligence.* 2022.

42. Zhou, M., et al., *Efficient Localisation Using Images and OpenStreetMaps.* 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2021: p. 5507-5513.

43. *European Space Agency. (n.d.). Sentinel Online: Sentinel Data Access. Retrieved from* [*https://sentinel.esa.int/web/sentinel/sentinel-data-access*](https://sentinel.esa.int/web/sentinel/sentinel-data-access)*.*

44. Ludwig, C., et al., *Mapping Public Urban Green Spaces Based on OpenStreetMap and Sentinel-2 Imagery Using Belief Functions.* ISPRS Int. J. Geo Inf., 2021. **10**: p. 251.

45. Abson, D.J., A.J. Dougill, and L.C. Stringer, *Using Principal Component Analysis for information-rich socio-ecological vulnerability mapping in Southern Africa.* Applied Geography, 2012. **35**(1): p. 515-524.

46. Friesen, C.E., P. Seliske, and A. Papadopoulos, *Using Principal Component Analysis to Identify Priority Neighbourhoods for Health Services Delivery by Ranking Socioeconomic Status.* Online J Public Health Inform, 2016. **8**(2): p. e192.

47. Liu, Y., A. Singleton, and D. Arribas-Bel, *A Principal Component Analysis (PCA)-based framework for automated variable selection in geodemographic classification.* Geo-spatial Information Science, 2019. **22**(4): p. 251-264.

48. Sera, F., et al., *How urban characteristics affect vulnerability to heat and cold: a multi-country analysis.* International journal of epidemiology, 2019. **48**(4): p. 1101-1112.

49. Yao, F., J. Coquery, and K.-A. Lê Cao, *Independent Principal Component Analysis for biologically meaningful dimension reduction of large biological data sets.* BMC Bioinformatics, 2012. **13**(1): p. 24.

50. Janatian, N., et al., *A statistical framework for estimating air temperature using MODIS land surface temperature data.* International Journal of Climatology, 2017. **37**(3): p. 1181-1194.

51. Kelleher, J.D. and B. Tierney, *Data science*. 2018: MIT Press.

52. Xu, J., et al., *Downscaling Aster Land Surface Temperature over Urban Areas with Machine Learning-Based Area-To-Point Regression Kriging.* Remote. Sens., 2020. **12**: p. 1082.

53. Usmani, R.S.A., et al., *Air pollution and cardiorespiratory hospitalization, predictive modeling, and analysis using artificial intelligence techniques.* Environ Sci Pollut Res Int, 2021. **28**(40): p. 56759-56771.

54. Boudreault, J., C. Campagna, and F. Chebana, *Machine and deep learning for modelling heat-health relationships.* Sci Total Environ, 2023. **892**: p. 164660.

55. Boudreault, J., C. Campagna, and F. Chebana, *Revisiting the importance of temperature, weather and air pollution variables in heat-mortality relationships with machine learning.* Environ Sci Pollut Res Int, 2024. **31**(9): p. 14059-14070.

56. Wang, C., L. Feng, and Y. Qi, *Explainable deep learning predictions for illness risk of mental disorders in Nanjing, China.* Environmental Research, 2021. **202**: p. 111740.

57. Wang, C., Y. Qi, and Z. Chen, *Explainable Gated Recurrent Unit to explore the effect of co-exposure to multiple air pollutants and meteorological conditions on mental health outcomes.* Environ Int, 2023. **171**: p. 107689.

58. Lee, W., et al., *Forecasting of non-accidental, cardiovascular, and respiratory mortality with environmental exposures adopting machine learning approaches.* Environ Sci Pollut Res Int, 2022. **29**(58): p. 88318-88329.

59. Nishimura, T., et al., *Social Implementation and Intervention with Estimated Morbidity of Heat-Related Illnesses from Weather Data: A Case Study from Nagoya City, Japan.* Sustainable Cities and Society, 2021. **74**: p. 103203.

60. Arsad, F.S., et al., *The Impact of Heatwaves on Mortality and Morbidity and the Associated Vulnerability Factors: A Systematic Review.* Int J Environ Res Public Health, 2022. **19**(23).

61. Hastie, T., R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Series in Statistics. 2009: Springer.

62. Lin, C.-Y., et al., *Impact of an improved WRF urban canopy model on diurnal air temperature simulation over northern Taiwan.* Atmospheric Chemistry and Physics, 2015. **16**: p. 1809-1822.

63. Richard, D.R., et al., *Calculating the sample size required for developing a clinical prediction model.* BMJ, 2020. **368**: p. m441.

64. Jaiswal, A. and S. Sarkar, *Climate Leadership: Ahmedabad’s 6th Heat Action Plan*. 2018, NRDC.

Figure Legend/Caption

**Figure 1: Development Stages of the Early Warning System (EWS) for Heat-Related Health Risks**

The figure illustrates the structured four-step process to establish an Early Warning System (EWS) for heat-related health risks.

* **Step 1** focuses on defining vulnerability and heat hazards, which includes quantifying social determinants of health (SDOH) and environmental factors (Aim 1a), and developing geospatial heat hazard maps (Aim 1b).
* **Step 2** integrates various data sources to define a heat-health-hazard model. This step involves developing a model that combines biomedical data, vulnerability, and heat hazard data from clinical trials and mortality data, including data from RP1 cohorts/trials and other DSI-Africa Hubs (Aim 2).
* **Step 3** is divided into app co-design for the Department of Health EWS and workplace EWS, including engaging multiple stakeholders to select risk temperature thresholds and commercialisation strategies (Aim 3a).
* **Step 4** involves implementing and testing the EWS, which entails monitoring the app’s performance through metrics such as the number of downloads, usage during heatwaves, symptom reports, and user feedback (Aim 3b).

Each step outlines specific objectives and strategies, aligning with the broader aim of reducing heat-related morbidity and mortality by leveraging advanced data integration and analysis, stakeholder collaboration, and targeted communication.

**Figure 2: Methodological Framework for the Stratified Heat-Health Outcome Forecast Model**

The figure illustrates the methodology for developing a forecast model that predicts heat-related health outcomes, stratified by demographic and geographic variables. It involves harmonizing clinical and cohort data with socio-economic and climatic factors, using machine learning methods such as GRU and LSTM for analysis. The outputs include a heat-health outcome model, scholarly publications, and advocacy tools, which lead to informed public health strategies and potential policy shifts