BMJ Open

Leveraging data science and machine learning for urban climate adaptation in two major African cities: a HE²AT Center study protocol

Christopher Jack ,[1] Craig Parker ,[2] Yao Etienne Kouakou ,[3,4] Bonnie Joubert,[5] Kimberly A McAllister,[5] Maliha Ilias,[6] Gloria Maimela ,[7] Matthew Chersich,[2,8] Sibusisiwe Makhanya,[9] Stanley Luchters,[10,11] Prestige Tatenda Makanga,[10,12] Etienne Vos,[9] Kristie L Ebi,[13] Brama Koné,[3,4] Akbar K Waljee,[14,15] Guéladio Cissé,[3,4] HE²AT Center Group

To cite: Jack C, Parker C, Kouakou YE, *et al*. Leveraging data science and machine learning for urban climate adaptation in two major African cities: a HE²AT Center study protocol. *BMJ Open* 2024;**14**:e077529. doi:10.1136/bmjopen-2023-077529

► Prepublication history for this paper is available online. To view these files, please visit the journal online (https://doi.org/10.1136/bmjopen-2023-077529).

CJ and CP contributed equally.

CJ and CP are joint first authors.

Received 16 July 2023
Accepted 03 May 2024

For numbered affiliations see end of article.

Correspondence to
Dr Christopher Jack; cjack@csag.uct.ac.za

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 intraurban heat risk and exposure using health, socioeconomic, climate and satellite imagery data; (2) creating a stratified heat–health 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 socioeconomic, 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. The 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.

STRENGTHS AND LIMITATIONS OF THIS STUDY

- ⇒ Our study collects comprehensive data from clinical, socioeconomic and remote sensing sources, ensuring a multidimensional analysis of urban heat exposure.
- ⇒ It leverages state-of-the-art machine learning techniques for modelling of heat–health outcomes, advancing the field of environmental health research.
- ⇒ A cross-disciplinary approach enriches the interpretation of data, linking climate science with public health implications.
- ⇒ Presents a risk of sampling bias due to secondary data utilisation, which may influence the representativeness of findings.
- ⇒ 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 HEat and HEalth African Transdisciplinary Center (HE²AT Center), a consortium spanning South Africa, Côte d'Ivoire, Zimbabwe and the USA, embodies global collaboration. Funded through the US NIH 'Data Science for Health Discovery and Innovation in Africa' (DS-I Africa) programme, the centre amalgamates diverse expertise in pursuit of comprehensive urban climate resilience strategies.[1]

This study emerges from the HE²AT Center as a research project 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 (UHI) 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*, Pasquini *et al* and Wright *et al* lay the groundwork,

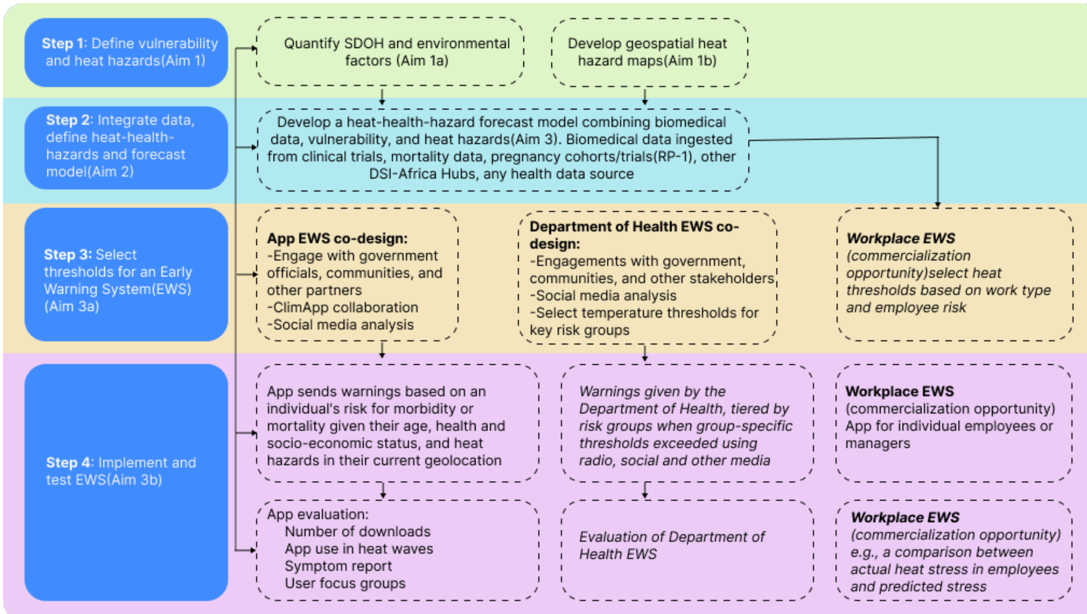

Figure 1 Development stages of the early warning system (EWS) for heat-related health risks. This illustrates the structured four-step process to establish an EWS for heat-related health risks. Step 1 focuses on defining vulnerability and heat hazards, which includes quantifying social determinants of 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 Data Science for Health Discovery and Innovation in Africa (DS-I Africa) Hubs (Aim 2). Step 3 is divided into app codesign 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.

explaining the socioeconomic and infrastructural factors that exacerbate heat–health vulnerabilities.[5–7]

Enhanced night-time heatwaves over African urban clusters, as investigated by Igun *et al*, underline the growing threat of heatwaves exacerbated by UHI effects.[8] Furthermore, an assessment of the health-related impacts of UHIs in Douala metropolis, Cameroon, by Enete *et al* 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 sociodemographic makeup of African metropolises. Integrating insights from recent studies, including 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 intraurban 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 pre-emptive action. These are inspired by the robust frameworks and pioneering methods established by Thiaw *et al*and Chapman *et al*, who have significantly advanced the field of heat–health EWSs in the African context.[11 12]

Our approach is grounded in the Intergovernmental Panel on Climate Change (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 socioeconomic 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.

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

Table 1 Key concepts and definitions in heat exposure studies aligned with the Intergovernmental Panel on Climate Change (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 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.

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, socioeconomic disparities and broader social determinants of health (SDOH) such as 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 present UHIs, a phenomenon where urban areas exhibit higher temperatures than their rural surroundings due to human activities. 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.[22]

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 as summarised in table 2. This multifaceted 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.[23] 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 (eg, multiple health trial datasets) both increase the statistical sample sizes for more robust modelling and enable a rigorous quantification of key uncertainties (eg, multiple climate datasets).[24 25]

Socioeconomic and environmental data

This research will collect socioeconomic geospatial data, which includes information on household economic conditions, service availability and residential characteristics—referring to factors such as housing type, construction materials used and the quality and condition of living spaces.[26] 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.[27] For Johannesburg, the Gauteng City-Region Observatory datasets will provide key variables for the study. In the case of Abidjan, equivalent data will be sourced from the National Institute of Statistics of Côte d'Ivoire, which provides comprehensive socioeconomic and demographic data.[27 28]

Table 2 Summary of data sources for each objective

Objective	Data sources
1. Mapping intraurban heat risk and exposure	► Socioeconomic data (census, surveys, GCRO datasets) ► Geospatial data (land use, building density, OpenStreetMaps) ► Climate data (WRF, UrbClim models, downscaled CDS and ESGF data, IBM-PAIRS platform)
2. Creating a stratified heat–health outcome forecast model	► Health data with clinical variables (eg, 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) ► Socioeconomic and environmental data (household economic conditions, service availability, residential characteristics)
3. Establishing an early warning system	► Integrated health and socioeconomic 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, socioeconomic statuses)

CDS, Copernicus Climate Data Store; ESGF, Earth System Grid Federation; GCRO, Gauteng City-Region Observatory; MODIS, Moderate Resolution Imaging Spectroradiometer; PAIRS, Physical Analytics Integrated Data Repository and Services; WRF, Weather Research and Forecasting.

Remote sensing data will be retrieved from satellite sensors, including optical images and indicators of physical aspects such as land surface temperature (LST), soil moisture, vegetation condition and land use and coverage.[29] 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.[26 27] Although Landsat and Moderate Resolution Imaging Spectroradiometer (MODIS) data primarily measure 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.[30]

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 reanalyses and climate simulations. While the CDS and ESGF provide valuable climate data, their spatial resolution may not be sufficient to distinguish different parts within the city.[31] 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 and UrbClim urban climate models. These models offer detailed results on heat stress for cities, allowing for a more precise analysis of intraurban heat variations and can improve the accuracy of our heat risk assessments for Johannesburg and Abidjan.[32 33]

Additionally, the IBM Physical Analytics Integrated Data Repository and Services (PAIRS) platform will be employed as a source of climate data, including data from climate models, weather stations and satellite observations.[34] To further enhance our analysis, we will integrate datasets from the European Space Agency's WorldCover portal and the Global Human Settlement Layer, which provide detailed land cover and human settlement data, respectively.[35 36] This will provide a comprehensive snapshot of Africa's past and future climate conditions, including the frequency, duration and intensity of heat waves.

Health trials and cohort data

In this study, we use cohort data due to the limited availability and generally poor quality of administrative health data in Abidjan and Johannesburg. These data also commonly contain limited variables on characteristics and health outcomes. Clinical trial data offer a robust alternative, providing detailed health outcomes and covariates, essential for minimising biases in heat–health studies. These studies (primarily HIV prevention and COVID-19) 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. This approach aligns with findings from Gasparrini *et al* and others, who used diverse data sources to analyse heat mortality associations.[37 38] 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:

1. Clinical variables: including vital signs (eg, body temperature, blood pressure, and heart rate), indicators of heat-related illness (eg, headache, dizziness, fatigue, and nausea), and details on pre-existing medical conditions (eg, hypertension, diabetes, and cardiovascular disease) that could increase the risk of heat-related illness, and documentation of adverse events potentially related to heat exposure.
2. Laboratory variables: including blood tests (eg, 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.
3. Demographic and SDOH variables: involving basic demographic information (eg, age, sex, race and ethnicity) (We acknowledge the complex interplay between race, ethnicity and health outcomes, recognising them as social constructs rather than biological determinants. We explicitly consider systemic racism and socioeconomic factors in our analysis, informed by Chokshi *et al* (2022), O'Reilly and Jones, to ensure a nuanced interpretation of demographic data), socioeconomic factors (eg, education, income and occupation) and data on housing and urban infrastructure (eg, air conditioning availability, ventilation and shading) that could influence heat exposure and the degree to which individuals and households are at an increased risk.[40–42]

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

Integration of datasets

Our study relies on integrating socioeconomic, clinical, environmental and geospatial data to understand heat's impact on health in African cities. We will cross-reference health trial participant geolocations with socioeconomic and environmental data, applying spatial jittering to protect privacy while retaining spatial trends. Additionally, we will 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 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.

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

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 predefined 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.

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, adjusting the study's scope and using synthetic data if necessary.

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.[43]

We will adjust the analysis phase when potential biases are identified. Specific statistical methods such as 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.[44]

Objective 1: assessing the degree of increased risk within cities

The methodology for quantifying intraurban vulnerability to heat in Johannesburg employs dimension reduction techniques such as principal component analysis to identify critical variables impacting heat vulnerability.[45] These identified components are aggregated using a scientifically derived weighting system, which reflects their relative importance and contribution to heat vulnerability. Aggregating these weighted components forms a composite vulnerability index, effectively quantifying socioeconomic and environmental susceptibility to heat.[46]

The creation of this index serves as a crucial step towards synthesising a unified 'heat risk index' that consolidates multiple vulnerability factors into a single, actionable metric. This index underpins our spatial multicriteria analysis, which uses a weighted overlay approach to produce a vulnerability map. This map critically informs policy interventions and resource allocation, guiding targeted measures to mitigate heat risks in the most vulnerable urban zones.[45–49]

Objective 2: 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] 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.[30]

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.

Sample size considerations are integral at this stage to ensure precision of study findings, with acceptable uncertainty ranges. The selection of adequate sample sizes is based on the statistical power required to detect significant differences in heat-related health outcomes, including across the different geographical and demographic strata, where possible. This involves detailed calculations to ensure that the study has sufficient power to validate the predictions made by our heat–health models.[51]

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.[52] 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.[53 54] 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 and long short-term memory 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 the literature on deep learning and ensemble tree-based machine learning models.[55–66] However, recognising that simpler statistical models may be effective, we plan to build on the work by Boudreault *et al* to compare the performance of deep learning models with tree-based approaches and nonlinear statistical models in our analysis.[57]

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 comorbidities to be explored include cardiovascular disease, respiratory disease, renal disease and HIV status.[67]

We will use k-fold cross-validation to assess model performance and generalisability, 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 such as accuracy, precision, recall, F1 score, Mean Squared Error (MSE) and Mean Absolute Error (MAE). Special attention will be paid to model performance during heatwave periods to ensure effectiveness in predicting heat-related health outcomes.[56]

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.[68]

Objective 3: develop an EWS reflective of geospatial and individualised risk profiles

The third objective is to develop an 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.

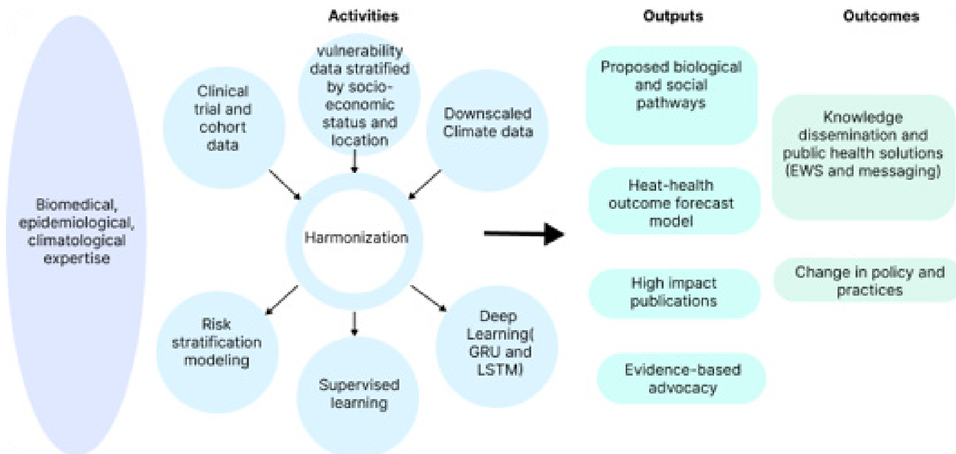

Figure 2 Methodological framework for the stratified heat–health outcome forecast model. This illustrates the methodology for developing a forecast model that predicts heat-related health outcomes, stratified by demographic and geographic variables. It involves harmonising clinical and cohort data with socioeconomic and climatic factors, using machine learning methods such as gated recurrent unit (GRU) and long short-term memory (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.

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 interagency coordination and community outreach for effective heat risk mitigation.[69]

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.

Patient and public involvement

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

Project timeline

The project is funded to run from 2022 to 2026.

ETHICS AND DISSEMINATION

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 220606) on 30 June 2022 and the National Ethics Committee for Life and Health Sciences, Côte d'Ivoire, on 25 November 2022 (reference number 176-22/MSHPCMU/CNESVS-kp) and will follow the US 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

MC, SL and the Hub Administrator direct the HE^{2}AT Center. Steering committee members represent six South, East and West African institutes. This study is led by GC of Ivory Coast's Peleforo Gon Coulibaly University and co-led by CJ of the University of Cape Town.

Dissemination

Prompt dissemination of research findings is crucial to the HE^{2}AT 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 HE^{2}AT Center's success and mission.

Study status

Ongoing.

Author affiliations
[1]Climate System Analysis Group, University of Cape Town, Rondebosch, Western Cape, South Africa
[2]Wits Planetary Health Research, Faculty of Health Sciences, University of the Witwatersrand, Johannesburg, South Africa
[3]University Peleforo Gon Coulibaly, Korhogo, Côte d'Ivoire
[4]Centre Suisse de Recherches Scientifiques, Abidjan, Côte d'Ivoire
[5]National Institute of Environmental Health Sciences, Durham, North Carolina, USA
[6]National Heart Lung and Blood Institute, Bethesda, Maryland, USA
[7]Climate and Health Directorate, Wits Reproductive Health and HIV Institute, Hillbrow, Gauteng, South Africa
[8]Public Health and Primary Care, School of Medicine, Trinity College Dublin, Dublin, UK
[9]IBM Research—Africa, Johannesburg, South Africa
[10]Centre for Sexual Health and HIV & AIDS Research (CeSHHAR), Harare, Zimbabwe
[11]Liverpool School of Tropical Medicine, Liverpool, UK
[12]Surveying and Geomatics Department, Midlands State University, Gweru, Zimbabwe
[13]University of Washington, Seattle, Washington, USA
[14]Gastroenterology, University of Michigan, Ann Arbor, Michigan, USA
[15]Ann Arbor VA Medical Center, VA Center for Clinical Management Research, Ann Arbor, Michigan, USA

X Craig Parker @logicSA06182, Yao Etienne Kouakou @KOUAKOUYAOETIE1 and Gloria Maimela @masebotja

Collaborators HE^{2}AT Center Group (alphabetical): Abdoulaye Tall, Adja Ferdinand Vanga, Craig Mahlasi, Darshnika Lakhoo, Iba Dieudonné Dely, James Mashiyane, Lisa van Aardenne, Madina Doumbia, Nicholas Brink, Pierre Kloppers, Piotr Wolski, Sibusisiwe Makhanya, Tamara Govindasamy, Toby Kurien.

Contributors CJ, CP, SL, MC and GC were involved in the research's conception and design. CP, GM and MC obtained ethics approval. CP, MC and YEK were involved in data acquisition. CP, MC and SL prepared the figures, and CP and CJ drafted the manuscript. AKW was involved in the conception and design, reviewing the structure of the paper. All authors were involved in the planning, conduct and reporting of the work, editing and revising the manuscript and approving the final version for submission.

Funding 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 MC, 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.

Patient and public involvement Patients and/or the public were involved in the design, or conduct, or reporting, or dissemination plans of this research. Refer to the Methods section for further details.

Patient consent for publication Not applicable.

Provenance and peer review Not commissioned; externally peer reviewed.

ORCID iDs
Christopher Jack http://orcid.org/0000-0002-0936-7277
Craig Parker http://orcid.org/0000-0003-2517-929X
Yao Etienne Kouakou http://orcid.org/0000-0002-4150-1499
Gloria Maimela http://orcid.org/0000-0002-7704-6803

REFERENCES

1 Harnessing data acience for health discovery and innovation in Africa (DS-I Africa), Available: https://commonfund.nih.gov/AfricaData
2 Johnson DP, Wilson JS, Luber GC. Socioeconomic indicators of heat-related health risk supplemented with remotely sensed data. *Int J Health Geogr* 2009;8:57.
3 Jung J, Uejio CK, Kintziger KW, *et al*. Heat illness data strengthens vulnerability maps. *BMC Public Health* 2021;21.
4 Xu R, Zhao Q, Coelho MSZS, *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 Med* 2020;17:e1003369.
5 Ncongwane KP, Botai JO, Sivakumar V, *et al*. A literature review of the impacts of heat stress on human health across Africa. *Sustainability* 2021;13:5312.
6 Pasquini L, van Aardenne L, Godsmark CN, *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;747:141355.
7 Wright CY, Dominick F, Kapwata T, *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;34:40.
8 Igun E, Xu X, Shi Z, *et al*. Enhanced nighttime heatwaves over African urban clusters. *Environ Res Lett* 2023;18:014001.
9 Enete I. Assessment of health related impacts of urban heat island (UHI) in Douala Metropolis, Cameroon. *IJEPP* 2014;2:35.
10 Rohat G, Flacke J, Dosio A, *et al*. Projections of human exposure to dangerous heat in African cities under multiple socioeconomic and climate scenarios. *Earth's Future* 2019;7:528–46.
11 Thiaw WM, Bekele E, Diouf SN, *et al*. n.d. Toward experimental heat–health early warning in Africa. *B am Meteorol Soc*103:E1843–60.
12 Chapman S, Birch CE, Marsham JH, *et al*. Past and projected climate change impacts on heat-related child mortality in Africa. *Environ Res Lett* 2022;17:074028.
13 Lwasa S. Managing African urbanization in the context of environmental change. *Id* 2014;2.
14 Wang YP, Kintrea K. Urban expansion and land use changes in Asia and Africa. *Environ Urban Asia* 2021;12:S13–7.
15 Abrahams C, Everatt D. City profile: Johannesburg, South Africa. *Environment and Urbanization ASIA* 2019;10:255–70.
16 Rees H, Delany-Moretlwe S, Scorgie F, *et al*. At the heart of the problem: health in johannesburg's inner-city. *BMC Public Health* 2017;17:17.
17 Macrotrends. Johannesburg, South Africa Metro area population 1950-2023. 2023. Available: https://www.macrotrends.net/cities/22486/johannesburg/population
18 Venter ZS, Shackleton CM, Van Staden F, *et al*. Green apartheid: urban green infrastructure remains unequally distributed across income and race geographies in South Africa. *Landscape and Urban Planning* 2020;203:103889.
19 Granado S, Manderson L, Obrist B, *et al*. Appropriating 'malaria': local responses to malaria treatment and prevention in Abidjan, Cote D'Ivoire. *Med Anthropol* 2011;30:102–21.
20 Djomand G, Roels T, Ellerbrock T, *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:S5–15.

21 World Population Review. Abidjan population 2023. 2023. Available: https://worldpopulationreview.com/world-cities/abidjan-population
22 Dongo K, Kablan M, Kouamé F. Mapping urban residents' vulnerability to heat in Abidjan, Côte D'Ivoire. *Clim Dev* 2018;10:1–14.
23 Wolf ST, Vecellio DJ, Kenney WL. *Adverse Heat-Health Outcomes and Critical Environmental Limits*. PSU HEAT Project, 2022.
24 Schubert S. An update on experimental climate prediction and analysis products being developed at NASA's global modeling and assimilation office. 2011.
25 Riedel M, Dosso SE, Beran L. Uncertainty estimation for amplitude variation with offset (AVO) inversion. *GEOPHYSICS* 2003;68:1485–96.
26 Alonso L, Renard F. 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. *Int J Environ Res Public Health* 2020;17:1004.
27 Observatory GC-R. Quality of life in the Gauteng city-region: A report on key indicators, 2019. Available: https://www.gcro.ac.za/about/annual-reports
28 Anderson TM, Shammami MA, Taddei SM, *et al*. How to use a mutant library to identify genes required for biofilm formation in the pathogenic *fungus candida albicans* *UJEMI* 2020;2:1–13.
29 Hofierka J, Gallay M, Onačillová K, *et al*. Physically-based land surface temperature modeling in urban areas using a 3-D city model and multispectral satellite data. *Urban Climate* 2020;31:100566.
30 Hooker J, Duveiller G, Cescatti A. A global Dataset of air temperature derived from satellite remote sensing and weather stations. *Sci Data* 2018;5:180246.
31 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.
32 Copernicus climate data store (CDS). Copernicus Climate Change Service; 2024.
33 Earth system grid Federation (ESGF). 2024.
34 Albrecht CM, *et al*. Pairs (Re)loaded: system design & benchmarking for scalable geospatial applications. IEEE Latin American GRSS & ISPRS Remote Sensing Conference (LAGIRS); 2020:488–93.
35 10 m Worldcover 2020 V100. European space Agency (ESA). 2021
36 European Comission. *The Global Human Settlement Layer 2019 (GHSL 2019) Public Release*. Luxembourg: Publications office of the European Union, 2021.
37 Gasparrini A, Guo Y, Hashizume M, *et al*. Temporal variation in heat–mortality associations: a multicountry study. *Environ Health Perspect* 2015;123:1200–7.
38 Sahani J, Kumar P, Debele S, *et al*. Heat risk of mortality in two different regions of the United Kingdom. *Sustainable Cities and Society* 2022;80:103758.
39 Arifwidodo SD, Ratanawichit P, Chandrasiri O. 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 Chokshi DA, Foote MMK, Morse ME. How to act upon racism—not race—as a risk factor. *JAMA Health Forum* 2022;3:e220548.
41 Jones CP. Invited Commentary:'race,' racism, and the practice of epidemiology. *Am J Epidemiol* 2001;154:299–304.
42 O'Reilly K. *AMA: Racism Is a Threat to Public Health*. American Medical Association, 2020.
43 Narod SA. Countercurrents: the bias of choice. *Curr Oncol* 2019;26:e712–3.
44 Schwartz R, *et al*. Towards a standard for identifying and managing bias in artificial intelligence. 2022.
45 Friesen CE, Seliske P, Papadopoulos A. Using principal component analysis to identify priority Neighbourhoods for health services delivery by ranking socioeconomic status. *Online J Public Health Inform* 2016;8:e192.
46 Abson DJ, Dougill AJ, Stringer LC. Using principal component analysis for information-rich socio-ecological vulnerability mapping in Southern Africa. *Applied Geography* 2012;35:515–24.
47 Liu Y, Singleton A, Arribas-Bel D. A principal component analysis (PCA)-based framework for automated variable selection in geodemographic classification. *Geo-Spatial Information Science* 2019;22:251–64.
48 Sera F, Armstrong B, Tobias A, *et al*. How urban characteristics affect vulnerability to heat and cold: a multi-country analysis. *Int J Epidemiol* 2019;48:1101–12.
49 Yao F, Coquery J, Lê Cao K-A. Independent principal component analysis for biologically meaningful dimension reduction of large biological data SETS. *BMC Bioinformatics* 2012;13:24.
50 Janatian N, Sadeghi M, Sanaeinejad SH, *et al*. A statistical framework for estimating air temperature using MODIS land surface temperature data. *Intl Journal of Climatology* 2017;37:1181–94.
51 Riley RD, Ensor J, Snell KIE, *et al*. Calculating the sample size required for developing a clinical prediction model. *BMJ* 2020;368:m441.
52 Kelleher JD, Tierney B. *Data Science*. MIT Press, 2018. Available: https://direct.mit.edu/books/book/3667/data-science
53 Hastie T, Tibshirani R, Friedman J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer Series in Statistics*. New York, NY: Springer, 2009. Available: http://link.springer.com/10.1007/978-0-387-84858-7
54 Xu J, Zhang F, Jiang H, *et al*. Downscaling Aster land surface temperature over urban areas with machine learning-based area-to-point regression kriging. *Remote Sensing* 2020;12:1082.
55 Usmani RSA, Pillai TR, Hashem IAT, *et al*. Air pollution and cardiorespiratory hospitalization, predictive modeling, and analysis using artificial intelligence techniques. *Environ Sci Pollut Res Int* 2021;28:56759–71.
56 Boudreault J, Campagna C, Chebana F. Machine and deep learning for modelling heat-health relationships. *Sci Total Environ* 2023;892:164660.
57 Boudreault J, Campagna C, Chebana F. Revisiting the importance of temperature, weather and air pollution variables in heat-mortality relationships with machine learning. *Environ Sci Pollut Res Int* 2024;31:14059–70.
58 Wang C, Feng L, Qi Y. Explainable deep learning predictions for illness risk of mental disorders in Nanjing, China. *Environmental Research* 2021;202:111740.
59 Wang C, Qi Y, Chen Z. 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;171:107689.
60 Lee W, Lim Y-H, Ha E, *et al*. Forecasting of non-accidental, cardiovascular, and respiratory mortality with environmental exposures adopting machine learning approaches. *Environ Sci Pollut Res* 2022;29:88318–29.
61 Nishimura T, Rashed EA, Kodera S, *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;74:103203.
62 Ke D, Takahashi K, Takakura J, *et al*. Effects of Heatwave features on machine-learning-based heat-related ambulance calls prediction models in Japan. *Sci Total Environ* 2023;873:162283.
63 Kim Y, Kim Y. Explainable heat-related mortality with random forest and shapley additive exPlanations (SHAP) models. *Sustainable Cities and Society* 2022;79:103677.
64 Ogata S, Takegami M, Ozaki T, *et al*. Heatstroke predictions by machine learning, weather information, and an all-population registry for 12-hour heatstroke alerts. *Nat Commun* 2021;12:4575.
65 Park M, Jung D, Lee S, *et al*. Heatwave damage prediction using random forest model in Korea. *Applied Sciences* 2020;10:8237.
66 Zhang K, Li Y, Schwartz JD, *et al*. What weather variables are important in predicting heat-related mortality? a new application of statistical learning methods. *Environ Res* 2014;132:350–9.
67 Arsad FS, Hod R, Ahmad N, *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;19:16356:23:.
68 Lin C-Y, Su C-J, Kusaka H, *et al*. Impact of an improved WRF urban canopy model on diurnal air temperature simulation over northern Taiwan. *Atmos Chem Phys* 2015;16:1809–22.
69 Jaiswal A, Sarkar S. Climate leadership: Ahmedabad's 6th heat action plan. NRDC; 2018.