Authors: Christopher Jack[1*], Craig Parker[2*], Yao Etienne Kouakou[3], Bonnie R. Joubert[4], Kimberly McAllister[4], Matthew Francis Chersich[2], Gloria Maimela[2] ,Sibusisiwe Makhanya[5], Stanley Luchters[6], Etienne Vos[5] Kristie Ebi[6], Gueladio Cisse[3] on behalf of the HE2AT Center
*Equal first authors

HE[2]AT Center Group (alphabetical): Abdoulaye Tall, Adja Ferdinand Vanga, Brama Kone, Christopher Jack, Craig Mahlasi, Craig Parker, Iba Dieudonné Dely, James Mashiyane, Kimberly McAllister, Lisa van Aardenne, Madina Doumbia, Maliha Ilias, Pierre Kloppers, Piotr Wolski, Sibusisiwe Makhanya, Tamara Govindasamy, Tatenda Makanga, Toby Kurien, Yao Etienne Kouakou

Author Affiliations:

- Climate System Analysis Group, University of Cape Town
- Wits Reproductive Health and HIV Institute (Wits RHI), University of the Witwatersrand, Johannesburg
- University Peleforo Gon Coulibaly, Abidjan
- National Institute of Environmental Health Sciences (NIEHS), Durham NC, USA
- IBM Research Africa, Johannesburg
- Centre for Sexual Health and HIV & AIDS Research, Harare
- The University of Washington, Seattle

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

LEVERAGING DATA SCIENCE AND MACHINE LEARNING FOR URBAN CLIMATE ADAPTATION IN AFRICA: A PROTOCOL FOR THE HE[2]AT CENTER'S SECOND RESEARCH PROJECT

Abstract
Rapid urban growth, significant levels of informality, and increasingly stretched health services, intersecting with observed past and projected future temperature increases, have resulted in a critical intersection between development patterns and climate change in African cities. The HE[2]AT Center's second research project aims to use data science and machine learning, including natural language processing and geospatial analysis, to combine and explore multiple data sources to understand the spatial and demographic complexity of heat-related health impacts in Abidjan, Côte d'Ivoire and Johannesburg, South Africa. The study will acquire existing health datasets from clinical research, as well as socioeconomic and geospatial climate datasets and satellite imagery, to map heat hazards at an urban scale and quantify heat-health vulnerability and impact on morbidity-specific health outcomes. Statistical, machine learning, and deep-learning techniques will be used to develop heat-health models and optimize an app-based Heat-Health Early Warning System. The results of this project will inform efforts to find innovative solutions for African cities to adapt to their changing climate.

Commented [MK([1]: Not sure what is meant by informality?

The study, approved by the Wits Human Research Ethics Committee in 2022 (reference no: 220606), adheres to all relevant guidelines and legislation, including the Declaration of Helsinki, ICH Good Clinical Practice, Ethics in Health Research, and the Protection of Personal Information Act (POPIA) 2013. Data will be managed according to ethics committee approved procedures, and results will be disseminated through workshops, policy and research forums, scientific conferences, and journal publications.

Keywords: urban, heat, health, early warning systems, intra-urban vulnerability, socio-economics and environment, exposure mapping, hazard mapping, Heat-related health impacts, African cities, Data science and machine learning

THE STRENGTHS AND LIMITATIONS OF THIS STUDY

Commented [BJ2]: Should this section be moved to later in the paper?

Commented [MK([3R2]: Yes this seems odd to start paper with?

The strengths and limitations of this study can be presented as follows:

Strengths:

- Large sample size and generalizability: The study will collect data from multiple sources, including randomized controlled trials, cohort studies, socioeconomic data, census data, remote sensing data, and ~~weather records~~both observed and simulated climate data, in two large African cities, allowing for greater generalizability of the findings.
- Use of novel machine learning techniques: The study will apply advanced data analytics techniques. This includes utilizing pattern machine learning algorithms, Quantile Regression Forests, Gated Recurrent Unit models, and natural language processing. In addition, it~~,~~ will also leverage the power of pre-trained large learning models to further enhance the performance and efficiency of our analysis.
- Multi-disciplinary team and approach: The study is supported by experts from various fields, including climate science, data science, public health, epidemiology, and environmental epidemiology, ensuring a comprehensive and well-rounded approach.

Limitations:

- Challenges in managing variation and bias when using multi-source data: The study may face challenges in managing variation and bias when using data from multiple sources. These challenges will be addressed through dimensionality reduction and controlling for confounding bias.
- Difficulty in gathering data from various sources: The process of locating and acquiring data from various sources may be challenging. The study will prioritize the management of data transfer to ensure the smooth collection of data.

INTRODUCTION

BACKGROUND/RATIONALE

High ambient temperatures above long-term averages during summer months and discrete heat extremes (e.g. heat waves) are associated with excess mortality and considerable morbidity[1-4]. The World Health Organization predicts that by 2030, there will be almost 92,000 deaths per year from heat waves, with sub-Saharan Africa among the worst affected regions[5].

Anthropogenic climate change has already resulted in a more than 1°C rise in temperature globally since the pre-industrial times (1850-1900) ~~Industrial Revolution~~[6]. However, this increase is not evenly distributed across the planet, or even within local areas[7]. Regional differences and the effect of urban development and land use change mean that many parts of Africa are experiencing higher than average temperature increases, and more frequent, intensive, and longer-lasting heat waves[8].

The Urban Heat Island (UHI) effect is a phenomenon in which the presence of concrete, non-reflective surfaces~~,~~ and low levels of greenery and wind result in temperatures considerably higher than in surrounding areas, leading to increased morbidity and mortality during heatwaves[7]. This is a particular concern in Africa, as it is the most rapidly urbanizing continent in the world, with an estimated 59% of its population living in cities by 2050[9]. In many African cities, a large proportion of the population lives in informal dwellings in unplanned settlements ~~or "slums,"~~ which are often located in hot, low-lying areas of the city and lack vegetation, shade, and natural ventilation[10]. Housing materials in these informal settlements, such as iron metal sheeting, can also exacerbate heat exposure, with temperatures inside these dwellings commonly 3-4°C warmer than outdoors[10-13].

The UHI effect is particularly concerning in African cities, where high heat exposure and limited insulation is common. The urban poor are particularly vulnerable to heat exposure due to their heightened sensitivity and lowered adaptive capacity[14]. Elderly individuals, those with pre-existing respiratory conditions, and those with HIV, malnutrition, or non-communicable diseases, are more sensitive to heat exposure, and those without access to cool water, air-conditioned spaces, health services, or occupational protections may have lower adaptive capacity and be unable to protect themselves from heat ~~stress~~related morbidity and mortality [15-19]. Occupational settings, such as manual labor in factories, construction sites, or other outdoor activities, can also ~~reach~~ result in dangerous levels of heat exposure[20].

Commented [BJ4]: Do you need to define heat stress at first instance? And for context of this sentence, is it only heat stress or more broadly, deleterious health effects of heat, or mortality? The references look specific to mortality.

To address these challenges, ~~it is important to use~~ data science innovations ~~solutions~~ offer an important opportunity to assess the impact of urban heat on health in African cities and improve preparedness to avoid heat-related morbidity and mortality[21]. This study aims to lay the foundation for an African urban heat health early warning system by ~~providing forecast modeling~~integrating advances in short (days to 1 week) and seasonal (weeks to months) weather forecasting with ~~and~~ identified~~ying~~ demographic and socioeconomic factors that increase susceptibility to heat stress[22]. By collecting and analyzing data from two large African cities, this project will provide valuable insights into the dangers of heat in urban environments in sub-Saharan Africa.

In addition to improving health outcomes, this research project aims to contribute to the broader goal of building more climate resilient cities in Africa. By understanding the complex interplay

between climate change, urbanization, and health, we hope to develop strategies that can help African cities better adapt to the challenges of a changing climate. This includes finding innovative solutions for managing heat hazards and protecting vulnerable populations from the impacts of extreme heat. The results of this study will be used to inform the development of an app-based Heat-Health Early Warning System, which can be used by city planners, public health officials, and community leaders to better prepare for and respond to heatwaves in African cities.

AIMS AND OBJECTIVES

1. Map intra-urban heat vulnerability and exposure across urban areas in large African cities: This aim involves using data science and machine learning techniques, including geospatial analysis and natural language processing, to combine and explore multiple data sources to understand the spatial and demographic complexity of heat-related health impacts in African cities. The study will acquire existing health datasets from clinical research, as well as socioeconomic and geospatial climate datasets and satellite imagery to map heat hazards at an urban scale and quantify heat-health vulnerability and impact on morbidity-specific health outcomes.
2. Develop a spatially and demographically stratified heat-health outcome forecast model: The second aim of this project is to use statistical, machine learning, and deep learning techniques to develop a heat-health outcome forecast model that is capable of predicting the probability of adverse health outcomes at different temperature thresholds. This model will be stratified by geography and demographics, allowing for more precise and targeted forecasts that are tailored to the specific needs of different populations and neighborhoods.
3. Develop an Early Warning System reflective of geospatial and individualized risk patterns: The final aim of this project is to develop an app-based Heat-Health Early Warning System that is reflective of the unique risk patterns identified through the mapping and forecasting activities described above. This system will be designed to provide timely and accurate warnings to city planners, public health officials, and community leaders, helping them to better prepare for and respond to heatwaves in African cities. The goal is to use the results of this study to inform the development of an Early Warning System that is tailored to the specific needs of African cities and capable of helping to mitigate the risks of heat-related health impacts in these regions

METHODS AND ANALYSIS

STUDY SETTING

Abidjan and Johannesburg are two large cities located in Côte d'Ivoire and South Africa, respectively. Both cities are experiencing rapid urban growth, significant levels of informality, and increasingly stretched health services, intersecting with observed past and projected future temperature increases[23]. This has resulted in a critical intersection between development patterns and climate change in these cities.

Johannesburg is the largest city in South Africa and is located in the Highveld region of the eastern plateau. It has a diverse and rapidly growing economy and faces significant health challenges,

Commented [BJ5]: Suggest converting this to paragraph text, and condense, for peer reviewed publication. And follow format of target journal (check instructions for authors on formatting requirements)

Commented [MK([6R5]: I agree-generally don't see lists in paper unless in appendix or in table?

Commented [BJ7]: Some text in this section seems more suitable for the background/introduction of a manuscript. For methods section of a peer reviewed manuscript, may be good to focus on what was done for this project, and write in first person. E.g., We collected data from Abijan and Johannesburg. Both cities are experiencing...etc.

Commented [BJ8]: This is unclear to me but if considered known by target audience just ignore

including high rates of HIV and tuberculosis, as well as non-communicable diseases[24]. Abidjan is the largest city in Côte d'Ivoire and is located on the southeastern coast of the country. It is a major economic hub and faces significant health challenges, including high rates of malaria and other infectious diseases, as well as non-communicable diseases[25, 26].

Both cities are characterized as urban heat islands because they exhibit higher temperatures than their surrounding rural areas. Johannesburg, famed for its enormous urban forest of over 10 million trees, suffers from the urban heat island effect, in which temperatures in metropolitan regions are greater than in adjacent rural areas[27]. The density of people and structures, as well as the amount of vegetative cover in the city, can all have an impact on this effect. The intra-urban spatial heterogeneity of vegetation levels across residential areas in Johannesburg contributes to the heat island effect[28]. Residential areas with high levels of vegetation may have weaker heat island effects due to evapo-transpirative cooling, whereas places with low levels of vegetation may be more sensitive to the heat island effect[29].

The district of Cocody in Abidjan is experiencing the urban heat island effect due to rapid urbanization and associated changes in land use and land cover that are occurring in the area[30]. The concentration of buildings and lack of green spaces in Cocody may be contributing to higher temperatures in the district compared to surrounding rural areas[31].

The comparison of Johannesburg to other cities, such as the tropical coastal city of Abidjan and the high-elevation inland subtropical city of Johannesburg, allows for the evaluation of the generalizability of our models and techniques in different contexts.

DATA SOURCES/MEASUREMENT

HEALTH VARIABLES OF INTEREST:

- Clinical data: This includes vital signs (e.g., body temperature, blood pressure, heart rate), symptoms and signs of heat-related illness (e.g., headache, dizziness, fatigue, nausea), and information on pre-existing medical conditions (e.g., hypertension, diabetes, cardiovascular disease) that may increase the risk of heat-related illness.
- Laboratory data: This includes blood tests (e.g., electrolyte levels, liver function tests, kidney function tests), markers of inflammation and oxidative stress, and tests for infectious diseases (e.g., malaria, dengue fever, leptospirosis) that may be exacerbated by heat. It also includes HIV tests, including viral load and CD4 count.
- Demographic data: This includes basic demographic information (e.g., age, sex, race, ethnicity), socioeconomic factors (e.g., education, income, occupation), and information on housing and urban infrastructure (e.g., availability of air conditioning, ventilation, shading) that may affect heat exposure and vulnerability.

The primary health data for this study will be collected from HIV clinical trial and cohort studies. These types of studies typically involve a large number of participants and are conducted over an extended period of time, allowing for the collection of detailed health data that can be used to identify trends and patterns. Possible outcomes of interest include heat stroke, heat exhaustion, and heat-related deaths.

Commented [BJ9]: May want to start this section with this text, rather than bulleted list of health variables.

Health variable	Description
Vital signs	Body temperature, blood pressure, and heart rate
Symptoms and signs of heat-related illness	Headache, dizziness, fatigue, and nausea
Pre-existing medical conditions	Hypertension, diabetes, and cardiovascular disease
Laboratory tests	Electrolyte levels, liver function tests, kidney function tests, markers of inflammation and oxidative stress
Tests for infectious diseases	Malaria, dengue fever, and leptospirosis
HIV tests	HIV viral load test and CD4 count test
Demographic data	Age, sex, race, ethnicity, education, income, occupation
Housing and urban infrastructure	Availability of air conditioning, ventilation, and shading
Other possible outcomes of interest	Heat stroke, heat exhaustion, heat-related deaths

Table:xx

Commented [MK([10]: Nice table-needs title

OTHER DATA TYPES(CLIMATE AND SOCIO-ECONOMIC)

This study will use a range of data sources to understand the impacts of heat on health in African cities. Climate-related data will be obtained from open data repositories, such as the Copernicus Climate Data Store (CDS) and Earth System Grid Federation (ESGF), which provide observational-based datasets, historical re-analyses, and climate simulations[32]. The IBM-PAIRS platform will also be used as a comprehensive and reliable source of climate data, including data from climate models, weather stations, and satellite observations[33]. This will provide a detailed picture of the historical and future climate conditions overin Africa, including the frequency, duration, and intensity of heatwaves.

Commented [MK([11]: Might want to define this if not done previously

In addition to climate data, the study will also use remote sensing data obtained from satellite sensors, including optical imagery and indicators of physical measures such as land surface temperature, soil moisture, vegetation condition, and land use and cover[34]. In cities where this information is available, researchers will combine data from existing sensor networks with information on urban land use and building density to create ~~transferable urban temperature models that can be used in other cities~~[35].

Commented [CJ12]: I'm not really sure about the transferability...

The study will also analyze geospatial socio-economic data, including household economic ~~status~~indicators, access to services, and dwelling type[36]. This data will be obtained from sources such as national census data and focused household and demographic surveys, and will include information on individuals' and households' income, education, employment, living conditions, and access to healthcare, education, and transportation[37].

By combining climate, remote sensing, and socio-economic data, the study aims to create vulnerability maps of cities that show areas where individuals and households are most vulnerable to the effects of heat on health. These maps will be useful for public health officials and policymakers in identifying areas of need and developing targeted interventions and policies to address these risks.

Study Identification

In our study, we will examine the relationship between heat and health in Johannesburg and Abidjan. To identify relevant clinical trials and cohort studies, we will conduct a systematic review of the literature. This will involve searching relevant databases and search engines using a list of predetermined keywords and inclusion/exclusion criteria. Two reviewers will independently screen the titles and abstracts of the identified studies, and full-text articles will be obtained for those that meet the inclusion criteria. The quality of the included studies will be assessed using the Newcastle-Ottawa Scale, and data will be extracted and synthesized according to the study design and focus[38]. Any discrepancies will be resolved through discussion and consensus. This systematic review will provide a comprehensive overview of the current evidence on the topic and inform the direction of our research.

To be considered for this study, a research project must meet the following criteria:

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

Commented [MK([13]: Not clear -needs at least one of identified clinical or lab variables?

Table: xx

METHODS

Commented [BJ14]: You already have a methods and analysis section earlier. Modify this to a different subheader?

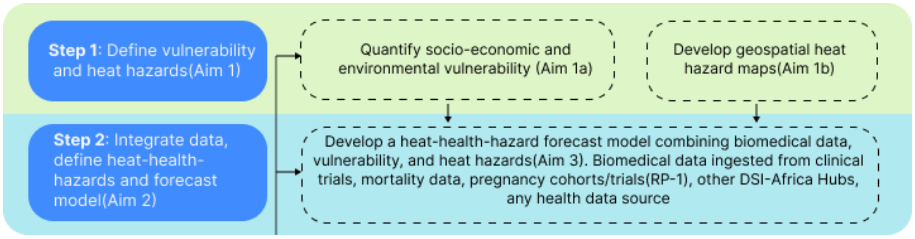

Figure: xx

Commented [MK([15]: Add title

Measure the socioeconomic and environmental vulnerability within cities

The study's first goal is to map and index intra-urban socioeconomic and environmental vulnerability and heat hazard exposure in African cities. Vulnerability refers to the interconnected set of factors

that determine whether a hazard (such as high temperatures) causes a health problem[39]. To measure vulnerability, the study will use a range of data sources, including OpenStreetMaps, sentinel satellite imagery, and socioeconomic data from censuses and household surveys[40]. These data sources provide detailed information on the physical and social characteristics of cities, including the location and density of buildings, the availability of green space and other heat-mitigating features, and the socio-economic status of residents.

Model transferability is a key consideration in this study, as the goal is to develop models that can be used in various cities across Africa[41]. To enable model transferability, the study will use dimensionality reduction methods such as Principal Component Analysis (PCA) to identify dominant correlation structures across variables and to combine the components into a single indicator that confers combined socioeconomic and environmental vulnerability[42]. Spatial techniques such as spatial principal component analysis and geodemographic clustering methods will also be used to account for spatial variability in the data[43]. The resulting output map will be used to identify regions that are vulnerable in the same way, and to develop targeted interventions that address these risk factors.

Develop high-resolution urban temperature hazard maps

Developing high-resolution urban temperature hazard maps is an important aspect of this study, as they serve as the foundation for subsequent machine learning training. To create these maps, the study will use a range of techniques, including downscaling and imputation methods to fill in gaps caused by cloud cover[44]. Land surface temperatures will be derived from multi-spectral band data from Landsat 8, with a focus on using the thermal infrared band and bands 5 and 7 to produce 100 m resolution grids[45, 46]. However, to better understand the correlation between temperature and health outcomes, the study will also investigate methods for producing higher spatial resolution temperature grids at 30 m resolution. This will involve the development and testing of statistical downscaling models using machine learning approaches such as random forest regression kriging and quantile random forest regression kriging, as well as the use of physics-based models to derive near-surface air temperature from LST data[47, 48].

The temperature patterns in each city will be influenced by the city's unique climate and physical characteristics, including altitude, land cover, and building density[49]. These models will incorporate digital elevation models and land surface maps to account for these factors and to create detailed, high-resolution maps of temperature hazards in each city[50]. The results of these models will be compared to existing field observations to ensure their accuracy and reliability. These maps will provide valuable information on the dangers of heat in African cities and will be used to inform efforts to find innovative solutions for adaptation to changing climates.

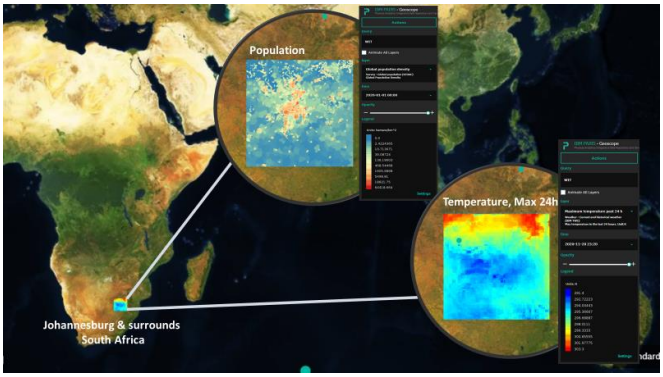

Figure: xx

Commented [MK([16]: Add title

CREATE A MODEL THAT STRATIFIES HEAT-HEALTH OUTCOMES BASED ON GEOGRAPHY AND DEMOGRAPHICS

The study will develop a model to predict the likelihood of adverse health outcomes at different temperatures based on geographic and demographic factors. The model will be trained using data on weather hazards at a high resolution, vulnerability data stratified by socioeconomic status and location, and information about individual biomedical outcomes. Machine-learning models, such as quantile regression forests, will be used to determine the strongest predictor variables from the suite of socioeconomic variables at each geographic location[51]. These models must be geographically coincident and include downscaled near-surface air temperature estimates, estimates of socioeconomic conditions, and an indicator of socioeconomic vulnerability to heat[52]. Initial research will focus on exploring different manual aggregations and exploratory data analysis with guidance from biomedical or epidemiological researchers.

Fine-tuning the machine-learning models will capture the associative relationships between high temperatures and negative health outcomes[53]. The significance of these health predictors in the models will be estimated within different populations within the investigated cities, indicating different susceptibility levels to heat-induced health conditions based on patient demographics and risk factors[54]. Some potential health co-morbidities that could be explored using data from clinical trials and machine-learning models include cardiovascular disease, respiratory disease, renal disease, and HIV.

The database's participants will be divided into subgroups based on factors such as age and socioeconomic status. The applicability of automatic sub-group discovery methods, including the use of open-source tools such as Pysubgroup and IBM-developed auto-stratification tools, will be explored[55, 56]. The outcomes of machine-learning models will be validated using clinical trial data to ensure their accuracy and reliability in predicting adverse health outcomes.

COMPARING MACHINE LEARNING AND DEEP LEARNING APPROACHES

The relationship between heat and human health is complex and non-linear, with clear "tipping points" at which the response to heat changes dramatically[7]. To accurately model this relationship, we will use a Transformer architecture as our primary model for analysis. Transformer models have recently demonstrated state-of-the-art performance on a wide range of natural language processing tasks and have the potential to capture complex dependencies within our dataset[57].

To further improve the performance of our model, we will also explore the use of transfer learning, which involves fine-tuning pre-trained models on a new dataset[58]. This approach can be particularly useful when the available data is limited or noisy, as is often the case with health data[59]. We will experiment with various pre-trained models and fine-tuning strategies to identify the most effective approach for our specific dataset and research question.

As a secondary analysis, we will also evaluate the efficacy and accuracy/sensitivity/specificity of other machine-learning techniques, including recurrent neural networks (RNNs), long short-term memories (LSTMs)[60], and gated recurrent units (GRUs), as well as traditional machine-learning techniques, such as the Multi-Layer Perceptron (MLP), Bayesian Neural Network (BNN), Radial Basis Functions (RBF)[61].

By using a Transformer architecture as our primary model and exploring the use of transfer learning, we aim to build the best possible model for predicting the health effects of extreme heat[62]. Our hypothesis is that this approach will produce the most reliable forecasts due to the ability of Transformer models to capture complex dependencies within our data, potentially enhanced through the use of transfer learning. However, we will rigorously test this hypothesis through our analysis and compare the performance of the Transformer model to other machine-learning techniques.

Analysis	**Description**
Primary model	Use of Transformer architecture to predict health effects of extreme heat
Transfer learning	Fine-tuning of pre-trained models on new dataset to improve model performance
Secondary analysis	Evaluation of efficacy and accuracy/sensitivity/specificity of RNNs, LSTMs, GRUs, MLP, BNN, RBF, KNN, and GRU for predicting health effects of extreme heat
Statistical analysis	Comparison of model performance using statistical measures such as mean absolute error, mean squared error, and Pearson's correlation coefficient
Sensitivity analysis	Evaluation of model performance under different assumptions and scenarios
Interpretability	Use of techniques such as feature importance and partial dependence plots to understand the factors driving model predictions

Table xx

Commented [MK([17]: Add title and make sure you refer in text for each table and figure in paper

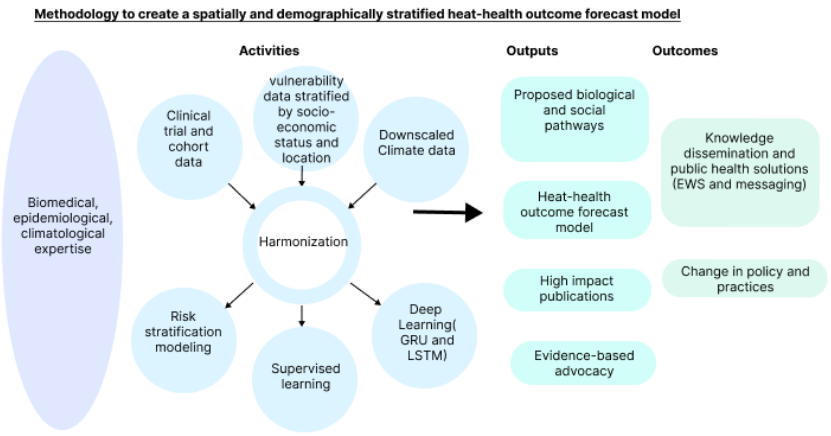

Figure:xx

Commented [MK([18]: Add title

Develop an Early Warning System reflective of geospatial and individualized risk patterns

The goal of our study is to develop an early warning system, including a digital app-based system developed using Flutter, that reflects the geospatial and individualized risk patterns of heat-related health impacts in Abidjan and Johannesburg[63]. This app, available on both Android and iOS, will allow users to set their own thresholds for triggering warnings based on the forecasted health effects of extreme heat. An interactive system that invites users to submit data could facilitate continuous recalibration and learning through crowd interaction and knowledge sharing. This information is critical for verifying the accuracy of the thresholds and gauging the App's overall performance in preventing adverse health outcomes during heat waves.

Commented [MK([19]: I was wondering if you wanted to say more about how applicable or not the findings may be to other African cities or cities across the world?

The early warning system developed in this study will be tested for its effectiveness in preventing adverse health outcomes during heat waves[64]. It will be integrated into the department of health's existing processes and protocols for managing heat waves and other weather-related hazards at the district level. Health workers at the district level will use the system to monitor weather conditions and the predicted likelihood of adverse health outcomes in real-time, and use this information to plan and implement interventions to prevent or mitigate the effects of heat waves on vulnerable populations in the district. The interactive nature of the app will also enable health workers at the district level to collect and share information with the broader community, facilitating continuous learning and improving the accuracy and effectiveness of the early warning system.

Commented [20]: There are several strategies that can be used to manage bias in big data projects that involve multiple sources of varying quality. One approach is to carefully examine the sources of data and assess their quality and reliability before incorporating them into the project. This can involve checking the sources for completeness, accuracy, and bias, and excluding any sources that do not meet certain standards. Another approach is to use multiple sources of data and combine them in a way that reduces the impact of any biases in the individual sources. For example, data from multiple sources can be weighted according to their quality and reliability, and the resulting data can be used to produce more accurate and unbiased results. Additionally, data from multiple sources can be combined using statistical methods, such as meta-analysis, to produce more robust and reliable results.

Managing Bias

To manage bias in the data, we will adopt a multi-pronged approach that involves careful selection of data sources, statistical adjustment of potential biases, and using a diverse range of data sources[65]. Specifically, we will carefully select clinical trials and cohort studies to ensure that they represent the modeled population. We will also adjust for potential biases in the data using statistical techniques such as weighting and stratification[66]. In addition, we will use a variety of

Commented [BJ21]: This could potentially be included in the strengths and limitations section depending on journal formatting guidelines

Commented [BJ22R21]: Do these techniques address potential selection bias? Suggest listing the specific methods rather than "statistical techniques" broadly.

high-quality data sources, including clinical trials and cohort studies, to provide a more comprehensive and balanced view of the data. By using these strategies, we aim to reduce the impact of bias and improve the accuracy and reliability of the modeling and early warning system.

ETHICAL CONSIDERATIONS

ETHICAL CONSIDERATIONS

The University of Witwatersrand's Health Research Ethics Committee has approved the study protocol for the use of secondary data in research, with the approval dated June 24, 2022 (reference no. 220606). The HE2AT Center's second research project involves the use of multiple data sources to understand the impact of heat on health in two African cities, Abidjan and Johannesburg. In conducting this research, we are committed to upholding the highest ethical standards and following all relevant guidelines and regulations.

One of the key ethical considerations in this study is the need to ensure that informed consent was obtained from participants for the primary studies from which we will be using data. We will carefully evaluate whether participants provided broad consent for data sharing, narrow consent for specific purposes, or whether it may be necessary to obtain re-consent or a waiver for informed consent**[67]**. We recognize that the rights and dignity of participants must be respected at all times, and we will take all necessary measures to protect these rights throughout the research process.

To protect the confidentiality of participant data, we will require data providers to provide assurances in the data sharing agreement that informed consent was obtained and that they have individual participant consent to share the data for this study. In addition, we will take steps to prevent privacy breaches, such as storing the data on a password-protected server and employing minimization principles to keep only essential study data. We will also ensure that all data is handled and stored in accordance with relevant data protection legislation.

To further protect personally identifiable information, including location data, we will follow US Department of Health and Human Services guidelines and may aggregate street addresses into regions and add random values to latitude/longitude coordinates. Only a limited number of named individuals will have access to the encryption keys for the 256-bit AES-encrypted data. We will also adhere to the US government's requirement for the use of NIST FIPS 140-2 verified cryptography modules for all sensitive unclassified data**[68]**.

We will also adhere to the principles of the Declaration of Helsinki and the ICH Good Clinical Practice guidelines in conducting this research**[69, 70]**. Any potential conflicts of interest will be disclosed to the appropriate parties.

Overall, our goal is to conduct this research in an ethical and responsible manner, while also protecting the privacy and confidentiality of the participants. We believe that this research has the potential to inform efforts to find innovative solutions for African cities to adapt to their changing climate, and we are committed to conducting it in a way that upholds the highest ethical standards.

END OF STUDY

The project is funded to run until 2026.

STUDY OVERSIGHT

Prof. Chersich, Prof. Luchters, and the Hub Administrator direct the study. Steering Committee members represent six institutes from South, East, and West Africa. This study is led by Prof. Cisse of Ivory Coast's Peleforo Gon Coulibaly University.

DISSEMINATION

To maximize the effectiveness of the HE2AT Center, it is essential that we promptly disseminate our research findings. We have developed a publication strategy that outlines the types of publications, authors, and release dates for our research. We will share our findings with our research partners and other relevant stakeholders to inform local, state, federal, and international activities and revise recommendations as needed. Effective and timely dissemination is critical to the success of the HE2AT Center and its mission.

STUDY STATUS

Ongoing.

Contributors: GC, MC, CJ, and SL were involved in the conception and design of the research. CP, MC, and DL obtained ethics approval. CP, MC and SL prepared the figures. CP, CJ drafted the manuscript. All authors edited and revised the manuscript. All authors approved the final version of the manuscript.

FUNDING:

Research reported in this publication was supported by the Fogarty International Center and 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 responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

COMPETING INTERESTS:

The authors declare the following financial interests and personal relationships as potential competing interests: Through their pension funds, MF, DL, GM, and CP have investments in the fossil fuel industry. The University of the Witwatersrand has endowments and other financial reserves that are invested in the fossil fuel industry.

Data sharing statement: Data from the HEAT002 study can be made available upon request. Interested researchers should contact Chris Jack on cjack@csag.uct.ac.za

REFERENCES

1. Analitis, A., et al., *Synergistic Effects of Ambient Temperature and Air Pollution on Health in Europe: Results from the PHASE Project.* Int J Environ Res Public Health, 2018. **15**(9).
2. Chersich, M.F., et al., *Associations between high temperatures in pregnancy and risk of preterm birth, low birth weight, and stillbirths: systematic review and meta-analysis.* BMJ, 2020. **371**: p. m3811.
3. Gasparrini, A., et al., *Mortality risk attributable to high and low ambient temperature: a multicountry observational study.* Lancet, 2015. **386**(9991): p. 369-75.

4. Manyuchi A, et al., *Extreme heat events, high ambient temperatures and human morbidity and mortality in Africa: A systematic review.* Environmental Research (under review), 2020.
5. WHO, *Quantitative risk assessment of the effects of climate change on selected causes of death, 2030s and 2050s.* https://www.who.int/globalchange/publications/quantitative-risk-assessment/en/, 2014.
6. Watts, N., et al., *The 2020 report of The Lancet Countdown on health and climate change: responding to converging crises.* Lancet, 2020.
7. *IPCC, 2018: Global Warming of 1.5°C. An IPCC Special Report on the impacts of global warming of 1.5°C above pre-industrial levels and related global greenhouse gas emission pathways, in the context of strengthening the global response to the threat of climate change, sustainable development, and efforts to eradicate poverty [Masson-Delmotte, V., P. Zhai, H.-O. Pörtner, D. Roberts, J. Skea, P.R. Shukla, A. Pirani, W. Moufouma-Okia, C. Péan, R. Pidcock, S. Connors, J.B.R. Matthews, Y. Chen, X. Zhou, M.I. Gomis, E. Lonnoy, T. Maycock, M. Tignor, and T. Waterfield (eds.)].*
8. Engelbrecht F, et al., *Projections of rapidly rising surface temperatures over Africa under low mitigation.* 10, 2015. **8**.
9. United Nations Department of Economic and Social Affairs Population Division, *World Urbanization Prospects: The 2018 Revision (ST/ESA/SER.A/420). New York: United Nations.* https://population.un.org/wup/, 2019.
10. Scott, A.A., et al., *Temperature and heat in informal settlements in Nairobi.* PLoS One, 2017. **12**(11): p. e0187300.
11. Bidassey-Manilal, S., et al., *Students' perceived heat-health symptoms increased with warmer classroom temperatures.* International journal of environmental research and public health, 2016. **13**(6): p. 566.
12. Naicker, N., et al., *Indoor temperatures in low cost housing in Johannesburg, South Africa.* International journal of environmental research and public health, 2017. **14**(11): p. 1410.
13. Wright, C.Y., et al., *Indoor temperatures in patient waiting rooms in eight rural primary health care centers in northern South Africa and the related potential risks to human health and wellbeing.* International journal of environmental research and public health, 2017. **14**(1): p. 43.
14. Lee, Y.Y., et al., *Overview of Urban Heat Island (UHI) phenomenon towards human thermal comfort.* Environmental Engineering and Management Journal, 2017. **16**: p. 2097-2111.
15. Azongo, D.K., et al., *A time series analysis of weather variability and all-cause mortality in the Kasena-Nankana Districts of Northern Ghana, 1995-2010.* Glob Health Action, 2012. **5**: p. 14-22.
16. Kovats RS, Wilkinson P, and Mohamed H, *Weather and cause-specific mortality in Cape Town, South Africa.* Epidemiology, 2005. **16**(5): p. S47-48. Program and Abstracts: The Seventeenth Conference of the International Society for Environmental Epidemiology (ISEE).
17. Kynast-Wolf, G., et al., *Seasonal patterns of cardiovascular disease mortality of adults in Burkina Faso, West Africa.* Trop Med Int Health, 2010. **15**(9): p. 1082-9.
18. Leone, M., et al., *A time series study on the effects of heat on mortality and evaluation of heterogeneity into European and Eastern-Southern Mediterranean cities: results of EU CIRCE project.* Environ Health, 2013. **12**: p. 55.
19. Wichmann, J., *Heat effects of ambient apparent temperature on all-cause mortality in Cape Town, Durban and Johannesburg, South Africa: 2006-2010.* Sci Total Environ, 2017. **587-588**: p. 266-272.
20. International Labour Organization, *Working on a warmer planet: The effect of heat stress on productivity and decent work.* 2019. **https://www.ilo.org/global/publications/books/WCMS_711919/lang--en/index.htm**.
21. Beyene, J., et al., *A Roadmap for Building Data Science Capacity for Health Discovery and Innovation in Africa.* Frontiers in Public Health, 2021. **9**.

22. Thiaw, W.M., et al., *Toward Experimental Heat-Health Early Warning in Africa.* Bulletin of the American Meteorological Society, 2022.
23. Tischler, J., et al., *Environmental Change and African Societies*. Chapter 11 Increasing Urbanisation and the Role of Green Spaces in Urban Climate Resilience in Africa. 2019: Brill. 265-295.
24. Rees, H.V., et al., *At the Heart of the Problem: Health in Johannesburg's Inner-City.* BMC Public Health, 2017. **17**.
25. Koné, A.B., et al., *[The impact of urbanization on malaria infection rate and parasite density in children in the municipality of Yopougon, Abidjan (Côte d'Ivoire)].* Medecine et sante tropicales, 2015. **25 1**: p. 69-74.
26. Dongo, K., et al., *Analysing Environmental Risks and Perceptions of Risks to Assess Health and Well-being in Poor Areas of Abidjan.* World Academy of Science, Engineering and Technology, International Journal of Environmental, Chemical, Ecological, Geological and Geophysical Engineering, 2010. **4**: p. 31-37.
27. Li, X., L.C. Stringer, and M. Dallimer, *The Impacts of Urbanisation and Climate Change on the Urban Thermal Environment in Africa.* Climate, 2022. **10**(11): p. 164.
28. Hardy, C. and A. Nel, *Data and techniques for studying the urban heat island effect in Johannesburg.* ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, 2015. **XL-7/W3**: p. 203-206.
29. Hardy, C.H. and A.L. Nel, *Data and techniques for studying the urban heat island effect in Johannesburg.* ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, 2015: p. 203-206.
30. 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.
31. Dongo, K., A.K.M. Kablan, and F.K. Kouame, *Mapping urban residents' vulnerability to heat in Abidjan, Côte d'Ivoire.* Climate and Development, 2018. **10**: p. 600 - 613.
32. 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.
33. 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.
34. 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.
35. Zumwald, M., et al., *Mapping urban temperature using crowd-sensing data and machine learning.* Urban Climate, 2021. **35**: p. 100739.
36. 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.
37. *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/.*
38. *Newcastle-Ottawa Scale (NOS) for assessing the quality of non-randomized studies. (n.d.). Retrieved from https://www.ohri.ca/programs/clinical_epidemiology/oxford.asp.*
39. Noy, I. and R. Yonson, *Economic Vulnerability and Resilience to Natural Hazards: A Survey of Concepts and Measurements.* Sustainability, 2018. **10**(8): p. 2850.
40. Feldmeyer, D., et al., *Using OpenStreetMap Data and Machine Learning to Generate Socio-Economic Indicators.* ISPRS International Journal of Geo-Information, 2020. **9**(9): p. 498.
41. Farahani, A., et al., *A Concise Review of Transfer Learning.* 2020 International Conference on Computational Science and Computational Intelligence (CSCI), 2020: p. 344-351.

42. 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.
43. 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.
44. Mokari, E., et al., *Spatiotemporal imputation of MODIS land surface temperature using machine learning techniques (Case study: New Mexico's Lower Rio Grande Valley).* Remote Sensing Applications: Society and Environment, 2021. **24**: p. 100651.
45. Roy, D.P., et al., *Landsat-8: Science and product vision for terrestrial global change research.* Remote Sensing of Environment, 2014. **145**: p. 154-172.
46. Rozenstein, O., et al., *Derivation of land surface temperature for Landsat-8 TIRS using a split window algorithm.* Sensors (Basel), 2014. **14**(4): p. 5768-80.
47. Sekulić, A., et al., *Random Forest Spatial Interpolation.* Remote Sensing, 2020. **12**(10): p. 1687.
48. Xu, Y., et al., *Spatially Explicit Model for Statistical Downscaling of Satellite Passive Microwave Soil Moisture.* IEEE Transactions on Geoscience and Remote Sensing, 2019. **PP**: p. 1-10.
49. Coseo, P. and L. Larsen, *How factors of land use/land cover, building configuration, and adjacent heat sources and sinks explain Urban Heat Islands in Chicago.* Landscape and Urban Planning, 2014. **125**: p. 117-129.
50. Guth, P.L., et al., *Digital Elevation Models: Terminology and Definitions.* Remote Sensing, 2021. **13**(18): p. 3581.
51. Hu, L., et al., *Quantile Regression Forests to Identify Determinants of Neighborhood Stroke Prevalence in 500 Cities in the USA: Implications for Neighborhoods with High Prevalence.* J Urban Health, 2021. **98**(2): p. 259-270.
52. Crosson, W.L., M.Z. Al-Hamdan, and T.Z. Insaf, *Downscaling NLDAS-2 daily maximum air temperatures using MODIS land surface temperatures.* PLoS One, 2020. **15**(1): p. e0227480.
53. Deng, C., et al., *Integrating Machine Learning with Human Knowledge.* iScience, 2020. **23**(11): p. 101656.
54. Gronlund, C.J., *Racial and socioeconomic disparities in heat-related health effects and their mechanisms: a review.* Curr Epidemiol Rep, 2014. **1**(3): p. 165-173.
55. *Introduction." PySubgroup, readthedocs.io, pysubgroup.readthedocs.io/en/latest/tutorials/introduction.html.*
56. Tadesse, G.A., et al., *Model-free feature selection to facilitate automatic discovery of divergent subgroups in tabular data*. 2022.
57. Khurana, D., et al., *Natural language processing: state of the art, current trends and challenges.* Multimedia Tools and Applications, 2023. **82**(3): p. 3713-3744.
58. Han, X., et al., *Pre-trained models: Past, present and future.* AI Open, 2021. **2**: p. 225-250.
59. Kim, H.E., et al., *Transfer learning for medical image classification: a literature review.* BMC Med Imaging, 2022. **22**(1): p. 69.
60. Mirzaei, S., J.-L. Kang, and K.-Y. Chu, *A comparative study on long short-term memory and gated recurrent unit neural networks in fault diagnosis for chemical processes using visualization.* Journal of the Taiwan Institute of Chemical Engineers, 2021. **130**.
61. Cuomo, S., et al., *Scientific Machine Learning Through Physics–Informed Neural Networks: Where we are and What's Next.* Journal of Scientific Computing, 2022. **92**(3): p. 88.
62. Hommel, B.E., et al., *Transformer-Based Deep Neural Language Modeling for Construct-Specific Automatic Item Generation.* Psychometrika, 2022. **87**(2): p. 749-772.
63. Kalkstein, L.S., S.C. Sheridan, and A.J. Kalkstein. *Heat/Health Warning Systems: Development, Implementation, and Intervention Activities*. 2009.

64. Pascal, M., et al., *Evolving heat waves characteristics challenge heat warning systems and prevention plans.* International Journal of Biometeorology, 2021. **65**(10): p. 1683-1694.
65. Schwartz, R., et al., *Towards a Standard for Identifying and Managing Bias in Artificial Intelligence.* 2022.
66. Howe, C.J., et al., *Selection Bias Due to Loss to Follow Up in Cohort Studies.* Epidemiology, 2016. **27**(1): p. 91-7.
67. Rothstein, M.A., *Informed consent for secondary research under the new NIH data sharing policy.* Journal of Law, Medicine & Ethics, 2021. **49**(3): p. 489-494.
68. Schaffer, K., *FIPS 140-3 Derived Test Requirements (DTR): CMVP Validation Authority Updates to ISO/IEC 24759*. 2019, National Institute of Standards and Technology.
69. Goodyear, M.D., K. Krleza-Jeric, and T. Lemmens, *The Declaration of Helsinki.* Bmj, 2007. **335**(7621): p. 624-5.
70. Vijayananthan, A. and O. Nawawi, *The importance of Good Clinical Practice guidelines and its role in clinical trials.* Biomed Imaging Interv J, 2008. **4**(1): p. e5.