Datasheet for climate-health mortality collected at Iganga -Mayuge HDSS in Uganda under Lacuna Project

We present the climate and health datasheet created by INSPIRE Network at APHRC, Kenya. We followed the dataset datasheet framework developed by Gebru et al. (2021).

Motivation	
Who created the dataset?	The dataset was created by a team from Iganga-Mayuge Health and Demographic Surveillance Site (IMHDSS) in Uganda.
For what purpose was the data set created?	The dataset was created to understand the causes of deaths in Iganga Mayuge district in Uganda
Was there a specific task in mind?	Yes, our goal was to integrate mortality data with remote sensing climate data, estimate the impact of climate change on mortality, and apply data science methods to predict its effects on health outcomes
Who funded the creation of the dataset?	This work was conducted with the support of the Lacuna Fund
Composition	
What do the instances that comprise the dataset represent?	The instances contain information on deaths that occurred among residents within the Health and Demographic Surveillance System (HDSS) study area and climatic factors.
How many instances are there in total (of each type, if appropriate)?	The dataset comprises 5,054 observations detailing the causes of death in the Iganga-Mayuge district and climatic factors such as average, maximum, and minimum temperature, and precipitation
Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set?	Yes, the dataset represents a sample of instances from a larger population.
What data does each instance consist of? “Raw” data or features?	The data consists of verbal autopsy data and climate data. Each data point is accompanied by attributes; gender, birthdate, deathdate, age, deathyear, birthyear,

	codmethod, ICD10causeofdeath, ICD10codes, General CoD, causeofdeath, climate record date, average temperature, maximum temperature. Minimum temperature, and precipitation.
Is there a label or target associated with each instance? If so, please provide a description.	No
Is any information missing from individual instances?	No
Are relationships between individual instances made explicit?	No
Are there recommended data splits (for example, training, development/validation, testing)?	We do not specify any data splits.
Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.	None
Is the dataset self-contained, or does it link to or otherwise rely on external resources?	The verbal autopsy data has been linked to climate data.
Does the dataset contain data that might be considered confidential?	No
Does the dataset contain data that, if viewed directly, might be of- offensive, insulting, threatening, or might otherwise cause anxiety?	No
Collection Process	

How was the data associated with each instance acquired?	The dataset was collected through Verbal Autopsy procedures, which involve interviewing surviving relatives or caregivers of the deceased to gather detailed information about the circumstances and events leading to the death. The remote sensing climate data was sourced from the Copernicus Climate Data Store
What mechanisms or procedures were used to collect the data?	Verbal Autopsy procedures, which involve interviewing surviving relatives or caregivers of the deceased to gather detailed information about the circumstances and events leading to the death. The remote sensing climate data was sourced from the Copernicus Climate Data Store
If the dataset is a sample from a larger set, what was the sampling strategy?	The dataset is not from a larger set.
Who was involved in the data collection process?	The Iganga Mayuge Data Team
Over what timeframe was the data collected?	January 2007 to December 2022
Were there any ethical review processes conducted (for example, by an institutional review board)?	The study was approved by the Makerere University School of Public Health Internal Review Board and the Uganda National Council of Science and Technology.
Preprocessing, Cleaning, and Labelling	
Was any pre-processing/cleaning/labelling of the data done	Yes. We worked with two datasets: a daily remote sensing climate dataset and a mortality dataset, both spanning from January 1, 2007 to December 31, 2022. The goal was to integrate mortality data with remote sensing climate data. The following pre-processing steps were done. **Step 1:** Preparation of Mortality Dataset

	The mortality dataset contained individual-level information including birthdate, deathdate, age, and birth year. Since deaths were recorded only on the specific date they occurred, a daily time series was created spanning the full study period (2007–2022). A new “date” column was generated representing every single day in this period. The date column was then populated such that: • If a death occurred on that specific day, the row was filled with the corresponding death record. • If no death occurred on that day, the row remained with NA values for all columns, except for the date. **Step 2:** Preparation of Climate Data The climate dataset provided daily weather information for the same period and location, including: date, average, maximum, and minimum temperature, and precipitation. **Step 3:** Linking the two datasets Using the date column as the common key, the mortality and climate datasets were merged. This left a combined dataset where: • For dates with recorded deaths, both mortality and climate data were present. • For dates without deaths, only climate data were available, while the mortality-related fields remained NA.

	Step 4: Creating a mortality status variable To facilitate analysis and filtering, a new column 'Status' was created using the 'ICD10codes' column from the mortality data: We categorize each day as either a day when someone died ("*dead*") or a day when no death occurred ("*alive*"), based on the presence or absence of ICD-10 cause of death codes. **Step 5:** Filtering for Relevant Records Since the study focused on understanding the impact of climate variables on days when deaths occurred, the dataset was filtered to retain only records where *status = "dead"*. This produced a final subset consisting of: • Dates with at least one recorded death • Complete climate data corresponding to those dates **Step 6:** Handling Missing Data Following the filtering step, the dataset was checked for missing values. No missing values were found in the filtered dataset, as: • Only days with death records and corresponding climate observations were retained • All climate variables had complete data for those days This final, cleaned dataset was then used for subsequent analysis.
Was the "raw" data saved in addition to the pre-processed/	None

cleaned/ labelled data (for example, to support unanticipated future uses)? If so, please provide a link or other access point to the "raw" data.	
Is the software that was used to pre-process/clean/label the data available? If so, please provide a link or other access point.	We used R programming language for data preprocessing.
Uses	
Has the dataset been used for any tasks already? If so, please provide a description.	Yes, we have used the dataset.
Is there a repository that links to any or all papers or systems that use the dataset?	No
What (other) tasks could the dataset be used for?	The dataset can be utilized for building time-series models and conducting trend analysis.
Is there anything about the composition of the dataset or the way it was collected and pre-processed/cleaned/labelled that might impact future uses?	No
Distribution	
Will the dataset be distributed to third parties outside of the entity (for example, company, institution, organization) on behalf of which the dataset was created? If so, please provide a description.	Yes. The data provided in the Harvard dataverse represents a random 10% of the entire dataset. If you require access to the complete dataset, please follow this link for more information: https://microdataportal.aphrc.org/index.php/catalog/165

How will the dataset be distributed (for example, tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?	The dataset and its associated metadata are hosted on the Harvard Dataverse, an open-source data repository. However, the version available there represents only 10% of the full dataset. For access to the complete dataset, please follow this link for more information: https://microdataportal.aphrc.org/index.php/catalog/165 The dataset has been assigned a Digital Object Identifier (DOI) for easy referencing: https://doi.org/10.7910/DVN/DHDNIC
When will be the dataset be distributed?	Upon approval, you will receive a link to download a compressed file containing the dataset in Stata format.
Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)?	The dataset is licensed under the CC BY license
Have any third parties imposed IP-based or other restrictions on the data associated with the instances?	No
Do any export controls or other regulatory restrictions apply to the dataset or to individual instances?	Yes. The data privacy and protection regulations in the host country.
Maintenance	
Who will be supporting/hosting/maintaining the dataset?	The dataset will be maintained by the INSPIRE Network. The team will support, host, and maintain the dataset.
How can the owner/curator/manager of the dataset be contacted (for example, email address)?	The dataset owner can be contacted via email (INSPIRE Network).

Is there an erratum?	No
Will the dataset be updated (for example, to correct labelling errors, add new instances, delete instances)?	All updates to the dataset will be documented and communicated via the INSPIRE Network.
Will older versions of the data-set continue to be supported/hosted/ maintained? If so, please describe how.	Yes, previous versions will be stored on the APHRC microdata portal.
If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so?	Researchers interested in contributing to the dataset can reach out to INSPIRE Network for further discussions.

References

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Iii, H. D., & Crawford, K. (2021). Datasheets for datasets. *Communications of the ACM*, *64*(12), 86-92.