**DATA MANAGEMENT AND SHARING PLAN**

**Element 1: Data Type**

1. **Types and amount of scientific data expected to be generated in the project:**

The research project will not conduct primary data collection. The project will use existing data from a variety of sources: (a) vital registration data, including counts of deaths and characteristics of decedents; (b) temperature data; and (c) neighborhood-level characteristics. Vital registration data for Philadelphia comes from the PA department of health; Vital registration data for Brazil comes from IBGE/DATASUS; Vital registration data for Panama comes from INEC; Vital registration data for Guatemala comes from RENAP. Temperature data at the city level was generated by the SALURBAL study (see Research Strategy of Research Project). Temperature data at the city-level for Philadelphia comes from PRISM (see Research Strategy of Research Project); and for Sao Paulo from Aina Roca-Barcelo (LSTHM), via modeling based on weather stations. Neighborhood level modifiers data comes from the ACS, Census Bureau (US), census of Brazil, IBGE (Brazil), census of Guatemala, INE (Guatemala), and census of Panama, INEC (Panama).

Pilot projects will likely generate data, either of quantitative of qualitative nature.  Pilot projects will be required to conform with NIH’s Data Management and Sharing plans.

1. **Scientific data that will be preserved and shared, and the rationale for doing so:**

We will make a copy of all data generated in the research project, administrative core, and community engagement core. We will preserve copies of all original and analytic datasets on a secure network drive for 3 years at minimum. All data on the drive automatically receive daily backups and twice daily snapshots that may be used for file recovery. Data sharing for the research project will be guided by the data use agreements of all data collected from other sources. Pilot projects that generate newly collected data will use a similar process to preserve and share data.

1. **Metadata, other relevant data, and associated documentation:**

Documentation of metadata (procedural and descriptive information) happens at multiple levels: project, dataset, variable level. Firstly, at the variable level, we will propose a baseline set of metadata fields including variable definition, value type, categorical coding, derived vs origin indicator, licensing information, public access indicator, publication linkages, geographic identifiers, longitudinal identifiers. Secondly, at the dataset level we will leverage a GitHub based workflow (see Element 2) to document as datasets are being generated. Each dataset will have its own repository which includes dataset specific code, issues, discussions, and change logs. Lastly, project level documentation will be available via a project web interface which can be used to navigate across the datasets and teams of a project; the content of which is designed to increase the findability, accessibility, interoperability, and reusability of assets within our project both internally and externally.

Pilot projects will follow their own protocols, but will need to comply with this data management and sharing plan. However, we will provide these pilot projects access to providing their data for our data infrastructure, which will also incorporate associated metadata.

**Element 2: Related Tools, Software and/or Code:**

We propose to use GitHub to manage our codebases. This is a cloud-based version control platform, essential for research center cooperation. Its project management features, such as Issues, Discussions, Projects, Pages, facilitate effective collaboration and real-time documentation throughout the project lifecycle. Importantly, its version control features (git, branches, pull requests) allow us to leverage software best practices to build in a reproducible product that not just quickly but also safely.

For building data infrastructure, we will leverage open-source programming languages (R/Python) that integrate well with modern big data tools such as columnar storage (Parquet), cross-language big data structures (Apache Arrow), big data infrastructure (Apache Spark, DuckDB, Azure Synapse). To effectively orchestrate the transformations done by these tools, we will utilize DBT (Data Build Tools) a industry convergent framework for building efficient and maintainable project level data warehouses.

Analyses will be performed by several data analysts and co-investigators using both open source (R, python) and propriety software (SAS, STATA); to accommodate this we will utilize .parquet as our fundamental storage format but be sure to make analytic datasets available in a variety of formats (.csv, .sas7bday, .dta) to increase interoperability for our researchers. 

**Element 3: Standards:**

In terms of interoperability standards with the larger/global research community we will adapt DDI (Data Documentation Initiative) standards which is commonly used in the Social Sciences. During the ‘archiving’ stage of the project all of our metadata will be converted into DDI formatted .xml which is compatible with our long-term data preservation plan. Moreover, within the project we implement additional standards to improve reproducibility and machine actionability. Firstly, to ensure reproducibility all raw data will be in historical storage and transformations into downstream (base, intermediate, analytic) datasets are under version control.  Secondly, establishing a baseline set of metadata fields for all variables of this project as detailed in `Element 1.C`. Other relevant metadata may arise during the process of actual data management, and we will append as needed; but these default metadata will be implemented consistently to ensure baseline machine actionability of data/metadata throughout project lifecycle. Thirdly, dataset and manuscript documentation/codebases will be available and developed through GitHub.

**Element 4: Data Preservation, Access, and Associated Timelines**

1. **Repository where scientific data and metadata will be archived:**

All shareable data will be available in our integrated web repository, including data and metadata sharing, based on a previously developed repository for the SALURBAL study. Code, models, and other documentation will also be archived as a GitHub project. Novel/high-value code that contribute new features to the open-source statistical ecosystem will be published as R packages. This includes packages that wrap the methods developed in the Research Capacity Building Core and the Research Project. For analytic datasets and code that require archival storage, we will leverage open data repositories such as Zenodo. 

1. **How scientific data will be findable and identifiable:**

To comply with FAIR principles, we will ensure within the project uniqueness and persistence of data identifiers and plan on making data/metadata publicly available throughout the project life cycle on our integrated web repository. This repository will serve to improve findability, accessibility and reusability both internally and externally. At project end, we will archive our data in ICPSR and utilize their repository for findability and minting of globally persistent and unique data identifiers.

1. **When and how long the scientific data will be made available:**

In the short term (during the project life cycle) data will be available as they become operationalized then released on our integrated web repository. In the long term (after project end), data will be available for perpetuity via ICPSR. 

**Element 5: Access, Distribution, or Reuse Considerations**

1. **Factors affecting subsequent access, distribution, or reuse of scientific data:**   
      
   For the Research Project, two kinds of sources of data are used: fully publicly available data, and data that we obtained through data use agreements (DUA). We anticipate that our data on neighborhood modifiers will be available to be unconditionally shared. However, data obtained through DUAs (e.g., mortality and temperature) will be subject to their specific DUAs.

Factors affecting subsequent access and reuse of data from pilot projects will be determined prior to awarding the pilot projects to ensure they comply with NIH guidance.

1. **Whether access to scientific data will be controlled:**   
      
   Access to scientific data will aim to be as open as possible. At minimum, metadata will be publicly available via our integrated web repository. Derived datasets that we produce based on publicly available sources will be made available without restriction. Scientific data if allowable to be public by originating data DUA or legal restrictions will be made publicly available. If originating data DUA is restrictive, those interested can reach out to our admin core and request authorization, access and mechanisms will be in place to distribute data accordingly.
2. **Protections for privacy, rights, and confidentiality of human research participants:**

All data used in this research are deidentified or publicly available. All research our team conducts with these data – along with our plan for dissemination of code and limited data – will be evaluated by our institution’s IRB.

**Element 6: Oversight of Data Management and Sharing:**

Usama Bilal (co-Lead of Administrative Core) will be responsible for the oversight of this data management and sharing plan, and will report to the executive committee on the performance of this plan quarterly.