DATA MANAGEMENT AND SHARING PLAN

If any of the proposed research in the application involves the generation of scientific data, this application is subject to the NIH Policy for Data Management and Sharing and requires submission of a Data Management and Sharing Plan. If the proposed research in the application will generate large-scale genomic data, the Genomic Data Sharing Policy also applies and should be addressed in this Plan. Refer to the detailed instructions in the application guide for developing this plan as well as to additional guidance on sharing.nih.gov. The Plan is recommended not to exceed two pages. Text in italics should be deleted. There is no “form page” for the Data Management and Sharing Plan. The DMS Plan may be provided in the *format* shown below.

Public reporting burden for this collection of information is estimated to average 2 hours per response, including the time for reviewing instructions, searching existing data sources, gathering, and maintaining the data needed, and completing and reviewing the collection of information. An agency may not conduct or sponsor, and a person is not required to respond to, a collection of information unless it displays a currently valid OMB control number. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to: NIH, Project Clearance Branch, 6705 Rockledge Drive, MSC 7974, Bethesda, MD 20892-7974, ATTN: PRA (0925-0001 and 0925-0002). Do not return the completed form to this address.

**Element 1: Data Type**

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

This research uses elements from six different publicly available datasets: Nielsen Consumer Panel (HomeScan), Nielsen Retail Scanner (RetailScan), American Community Survey (ACS) 5-year estimates, the National Longitudinal Study of Adolescent Health (Add Health), the Hispanic Community Health Study/Study of Latinos (HCHS/SOL), and the Multi-Ethnic Study of Atherosclerosis (MESA). We will include references to each of these datasets in any published material.

Using HomeScan, RetailScan, and ACS data, we will generate a model for estimating food price indices and assign food price indices to US census tracts for each calendar quarter from 2018 to mid-2023. Will combine Add Health, HCHS/SOL, and MESA datasets into one dataset and assign food price indices.

We will also generate a synthetic population for Aim 3, based on data from multiple sources, including the ACS, NHANES, BRFSS, and effect sizes from aims 1 and 2 (see Table 4 in the proposal for more detail)

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

We will make a copy of our price index, a data file containing stratum-specific price indices, and a data file containing census-tract specific price indices, and instructions and code to be used with RetailScan, Add Health, HCHS/SOL, and MESA to add price indices. A similar approach will be employed with the synthetic population data for aim 3.

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. We do not plan to make these datasets public as they can be directly downloaded from the data originators and turned into copies of our analytic datasets using data preparation code that we will make publicly available.

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

We will write detailed documentation for all published data and code. Data documentation will include codebooks for analytic datasets, instructions for using price index datasets and models, and detailed in-document annotation on published code files. Some information that will be found within these documents includes which originating datasets were used in our analysis, where to find them, and descriptions of any data or models were derived. We will crosslink these resources with any published work using doi’s and other metadata.

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

**Data preparation and analytic software includes R. R is open-source statistical software that runs on most UNIX platforms, Windows, and MacOS (https://www.r-project.org). The R environment will be managed using the *renv* package and we will make our environment publicly available so that future researchers can easily rerun our code. All code will be published on GitHub. We will assign the GitHub page a persistent identifier and reference that identifier in any published work.**

**Code to produce the synthetic populations and aim 3 is based on the IMPACTNCD model. This is a microsimulation modelling framework based on open-source code (GPL v3 licence) already available on GitHub (https://github.com/ChristK/workHORSE). The license allows anyone to adapt and use it without any restrictions as long as it remains open-source. We will release the adapted code under a similar license.**

**Element 3: Standards:**

**All originating data will be kept as its original file type. Working analytic datasets will be saved as rds files for all analyses conducted in R. Published datasets will be saved as comma separated value (csv) files, which can be easily read by any major statistical software. We will write codebooks for all analytic datasets that contain each variable name, variable category keys, the name and year of the originating dataset, and a brief description of the variable. We will also include links in our codebooks to the codebooks for the originating datasets.**

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

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

All previously unshared data will be deposited to OPENICPSR (https://www.openicpsr.org/). Code, models, and other documentation will also be archived as a GitHub project.

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

Data will be findable through the OPENICPSR collection or by following a persistent identifier linking to our GitHub project that we will include in all our published work.

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

The research community will have access to all previously unshared data and models at the end of the grant award. We will also make data and models available to the community alongside the release any publications (including ones that occur prior to the end of the award).

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

1. **Factors affecting subsequent access, distribution, or reuse of scientific data:**  
   All originating data are publicly available, but may be subject to data use agreements specific to each of those datasets. Our price index model, census tract price indices, and associated documentation will be freely available for non-commercial use.
2. **Whether access to scientific data will be controlled:** **Researchers can directly request the originating data from the holding organizations. Access to models, data, and code we publish will be directly available on OPENICPSR or GitHub.**
3. **Protections for privacy, rights, and confidentiality of human research participants:**

All data used in this research are deidentified and 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:**

**The two co-Principal Investigators (Drs. Langellier and Bilal), along with the Biostatistics Scientific Collaboration Center will be responsible for compliance and monitoring of this plan, and any adaptations that may result from changes in the project, through regularly scheduled meetings at least twice a year.**