Data Sharing Plan

Types of data/information created

This project will create research artifacts consisting of three distinct types of information: (1) the sourcecode for the open-source software implementing PRL algorithms that achieve high-velocity record linkage for the Census Bureau's large volume data; (2) the documentation and tutorials demonstrating and detailing how to use and extend this open-source software; and (3) a testbed of databases that are publicly available and sharable for testing alternative approaches to PRL and aiding development and application of PRL methods.

Tentative dates by which data will be shared

As an academic organization, the University of Washington's Institute for Health Metrics and Evaluation (IHME) is committed to sharing study results as quickly and widely as possible. Because of the iterative and participatory nature of this project, we will be able to share preliminary versions of sourcecode, documentation, and databases in real-time as we develop and refine them.

This means we would be able to share our first validation databases by June, 2022 (and likely sooner); our documentation on large volume PRL by Dec, 2022; our documentation on high velocity PRL by Dec, 2024; and full documentation, sourcecode, and testbed databases by the conclusion of the project in Dec, 2026.

Standards to be used for data/metadata format and content

We plan to use an open-source license such as the 3-clause BSD license for the sourcecode developed in this project, to ensure that the work is software can readily be used and extended by knowledgeable personnel of large-scale statistical organizations within the academia, private sector, state, and local government, as well as the federal government.

For documentation, we will use a copyleft license such as the Creative Commons Attributional 4.0 International License (CC BY 4.0), that ensures we receive recognition for our work (and blame!) but also permits others to share and adapt this material for their own purposes as long as we receive appropriate credit.

For testbed databases, we will use an appropriate format that balances file size, access speed, and usability. We will determine precisely what format to use during the course of the project through our participatory process. We will use a public domain license for this data, such as CC Zero, and document clearly why analysis with this data does not constitute human subjects research under the Common Rule and may therefore be exempt from institutional review board approval at US universities.

Policies and procedures for data stewardship and preservation, and for providing access and security (including prior experience in publishing such data)

IHME's information dissemination strategy emphasizes transparency, with a commitment to share as much data as we can through the Global Health Data Exchange (GHDx), publication in peer-reviewed journals, our website, presentation at scientific and policy-based conferences, as

well as training and outreach opportunities to help others understand our research methods and output. The Global Burden of Disease study is compliant with the Guidelines for Accurate and Transparent Health Estimates Reporting (GATHER) recommendations on documentation of data sources, estimation methods, and statistical analysis (*Lancet* 2016 Vol 388 (10062): PE19-E23). IHME seeks to publish all relevant information while complying with any data use agreements under which the source data were obtained, and while safeguarding the privacy of human subjects as appropriate.

To ensure preservation of artifacts created during this project, we will use an open-access research repository, such as Zenodo, to store sourcecode, documentation, and testbed databases. For each artifact, we will create a Document Object Identifier (DOI), which is automatic in many of the systems such as Zenodo, and will make each artifact citable and trackable.

To ensure easy access to the sourcecode, we will use an online repository, such as GitHub, that makes it easy to keep track of changes to code through a concurrent versioning system. We may also use a system like GitHub Large File Storage (LFS) to make distribution of large testbed databases easier.

Our security considerations will include particular attention to the privacy risks of large-scale testbed databases, especially those that rely on real-world data in some way (e.g. to create realistic noise). As part of this project, we will develop ethical principles specifically relevant to PRL algorithms and databses.

This team has years of prior experience sharing data/information of all of the three distinct types that will be produced in this work. Our open-source software includes statistical modeling software for the bednets supply chain, for generic disease modeling, for computer coding of verbal autopsies, and for public health microsimulation [4,14,16,32]. The documentation for these projects ranges from a technical appendix in published literature to an edited volume of theory and applications to automatically updated websites. We have also developed and distributed databases from scientific research, including replication archives to allow other researchers to reproduce our published results, as well as new databases to facilitate researchers and practitioners work in developing and selecting appropriate methods for verbal autopsy.[17–25]