ATTACHMENT I – PROJECT TOPIC

Utilizing Privacy Preserving Record Linkage to Link Data from Two Federal Statistical Agencies

Background

The Advisory Committee on Data for Evidence Building (ACDEB) was charged with making recommendations to the OMB Director on how to promote the use of federal data for evidence building. This includes utilizing the CIPSEA 2018 requirements to enhance how the federal statistical system facilitates evidence building and provide necessary frameworks to inform the development of the National Secure Data Service (NSDS).

The ACDEB Year two report includes several recommendations, notably,

- Recommendation 1.5, OMB, in coordination with the ICSP and other relevant federal councils, should identify mechanisms for streamlining data-sharing agreements across federal agencies,
- Recommendation 1.7, OMB, in coordination with the ICSP, should promote the use of privacy-preserving technologies in the tiered access framework required under Title III of the Evidence Act by identifying an initial set of promising tools over the next 1 to 3 years, and
- Recommendation 3.12. The NSDS should promote the use of privacy preserving technologies that can support working with data in situ, coordinating with the research community to develop efficient, scalable tools for users from all levels of government (including through open competitions).

To build on these recommendations, this project will serve as a proof of concept to develop a data sharing agreement between two federal statistical agencies that have not previously developed data sharing relationships, deploy a privacy preserving record linkage (PPRL) tool to link data from two federal statistical agencies and utilize a secure environment to analyze the resulting linked data file. In order to conduct the linkage, a privacy preserving record linkage (PPRL) tool will be selected. PPRL is a method that can be used to link de-identified data, using encrypted tokens and a trusted third party. PPRL will need to be performed in a secure environment that meets CIPSEA and FEDRAMP standards. Once the data are linked using a PPRL tool and stripped of the encrypted tokens, they will be available for analysis in a secure environment that meets the standards outlined above.

This work will link data from the National Center for Health Statistics (NCHS) and the National Center for Science and Engineering Statistics (NCSES). The two data sources to be linked are the NCHS National Hospital Care Survey (NHCS) and the NCSES Survey of Earned Doctorate (SED). The NCHS has a wealth of information on inpatient and Emergency Department encounters through the course of a calendar year. However, NCHS is missing a key demographic variable, education. NCHS is developing an imputation model for patient-level educational attainment. The SED is an annual census conducted since 1957 of all individuals receiving a research doctorate from an accredited U.S. institution in a given academic year. Linking these two sources can provide vital information to assess the validity of the imputation model for education, a key variable addressing social determinants of health (SDOH). Increasing the ability to

collect, link, and analyze health equity and SDOH-related data from within and outside the Department of Health and Human Services is critical to answer key research questions for evidence building.

America's DataHub Consortium is unique in bringing together capabilities and infrastructure to securely fill information gaps and to take on key analytic questions and evidence building challenges using innovative and novel approaches like PPRL.

Objectives

The project objectives are:

1) To develop a data sharing agreement between two federal statistical agencies.
2) To utilize a PPRL tool to link NHCS data to the SED through a trusted third party hosted by NCSES. This linkage would serve as an example of two federal agencies utilizing PPRL for interagency data sharing and linking which will inform future interagency initiatives.
3) To demonstrate the ability to analyze the linked data in a secure environment. The resulting linked file will be used to provide validation statistics back to NCHS to provide additional insight on the validity of the education imputation methodology.

PPRL would be used to link NHCS data to the SED. This work will highlight the ability to deploy a PPRL tool in a secure environment with a trusted third party, utilize the tool to link NCHS survey data to NCSES data, and then analyze the resulting linked file to inform the imputation methodology. The work must be completed by a CIPSEA designated agent, and the secure environment needs to meet CIPSEA and FEDRAMP standards. This work will inform linkages across the federal government, using the development of agreements and deployment of PPRL as a model to improve the availability, quality, accessibility, and interoperability of data sharing. This work will support the technical infrastructure that would be needed to allow linkages to occur quickly and efficiently for specific projects across federal agencies.

Information Gaps

This project will identify key components necessary to inform a future/potential NSDS including:

- What is needed to create a standardized data sharing agreement between two federal statistical agencies, one of which would serve as a trusted agent who would host the compute environment to conduct the linkage.
- What is needed to conduct PPRL to link these sources without ever exchanging direct personally identifiable information.
- What is needed to build an infrastructure to inform linkages across the federal government, using the development of agreements, utilization of a central, shared compute environment, and deployment of PPRL as a model to improve the availability, quality, accessibility, and interoperability of data sharing and linking.
- What is needed to assess the feasibility of using linked data to inform/validate imputation models that will support evidence-based policymaking.

Key Evidence Building Considerations

- Key focus questions (address one or more) to assess innovation in the following areas: data acquisition, data security, data linking, privacy, and engagement:

- What are key challenges with locating, acquiring, accessing, linking, and using disparate data and information?
- What are the lessons learned to using the presumption of accessibility in the Evidence Act?
- Which novel techniques for data, privacy, and confidentiality protections can be used?
- What models and estimation methods are best suited to fill information gaps?
- Are the resulting data and models fit to inform policy discussions and to make data available more equitability?
- What types of collaboration and stakeholder engagement are needed to help inform these questions, data, and analysis?
- What mechanisms are needed to access the resulting data that uphold privacy requirements?

Deliverables

At a minimum, offerors will provide the following if selected for an award. Additional deliverables may be required.

- Selection of a PPRL tool that meets certain minimum requirements, including but not limited to successful past performances with documented high precision and recall. The tool should demonstrate performance linking with identifiers such as partial social security number, name, and address.
- All code (clearly documented), linked data file, documentation of linkage process, documentation of linked data file (including codebook), and any other data/documentation created under this award.
- A report describing the lessons learned through this project including but not limited to the selection of the PPRL tool, and the key challenges/opportunities with locating, acquiring, accessing, and linking data using PPRL.