ATTACHMENT I – PROJECT TOPIC

Synthetic Data Generation with Large, Real-World Data

Key Objective

The processing and ingestion of large, real-world data can be challenging because of the volume and complexity of the source data. In addition, there may be constraints due to privacy risk that make accessing these data difficult. Tiered access mechanisms, such as synthetic data creation, have been explored to increase access to data while protecting privacy and maintaining utility. Developing synthetic data involves using statistical or machine learning techniques to create a dataset that contains new records with similar aggregate statistical properties as the original dataset. As a privacy protection technique, synthetic data may limit privacy concerns because the records within the dataset are not the records in the real, underlying data.

The objective of this project is to improve understanding of how synthetic data generators work with large real-world data (RWD) (e.g., datasets with over 30 billion rows of data) to inform a synthetic data generator toolkit. For this project, open-source synthetic data generators that utilize AI techniques will be assessed for use with large RWD on a super computing platform (e.g., National AI Research Resource (NAIRR) pilot secure compute environment). Typically, the synthetic generation of extremely large files is completed by blocking or stratifying data to run in parallel, however advancements in computing platforms enable the processing of large data files for synthetic data generation.

This project would inform the National Secure Data Service (NSDS) Demonstration Project and the NAIRR pilot by providing tools and techniques that support tiered access through the development and assessment of different tools for synthetic data generation.

The intent is to build on previous synthetic data generation methods and test the methods using large RWD files. This unique approach will provide insight into approaches that can then be compiled to inform the creation of a synthetic data toolkit that will be accessible within a future NSDS.

Background

America's DataHub Consortium brings together capabilities and infrastructure to securely fill information gaps and to take on key analytic questions and evidence building challenges. As demand for access to confidential federal data assets increases alongside novel analytical approaches, privacy protections must be in place to ensure the protection of privacy and the confidentiality of the data. Use of synthetic data can reduce disclosure risk while allowing data users to access microdata for research and other statistical purposes. The production and use of synthetic data is being explored to increase the utility of the data while ensuring strong privacy protections.

The proposed availability of a synthetic dataset would provide an additional tiered access option that would allow researchers access to previously inaccessible data. This option could prove valuable to researchers in making maximum use of these data while enabling the government to ensure privacy. The production of synthetic data, though, can prove challenging and resource intensive especially with large datasets. Mitigation strategies may include the use of a super compute platform that could ingest and process large RWD. There are different methods for the production of synthetic data and,

regardless of the methodology, all approaches must be scientifically valid, and the risk for re-identification will need to be determined prior to using the data for research. This project will address this using a large RWD file (e.g., a case study with data from the National Institutes of Health, N3C data and linked N3C data). This project will also inform the infrastructure and governance needed to expand the use of large-scale compute environments for synthetic data generation for other types of confidential federal data (e.g., CIPSEA data).

The following steps will be completed for synthetic data generation methods with a large RWD file:

1. Conduct an initial assessment to determine what variables for a synthetic dataset will be used. This will include outreach to stakeholders and subject matter experts to identify critical variables for inclusion.
2. Decide on an open-source synthetic data generator to be utilized in the creation of the large RWD. The tool will be selected based on its ability to process large RWD and ability to be deployed in a super computing environment.
3. Once produced, evaluate the dataset to assess quality and disclosure risk. Evaluation criteria will include but not be limited to an assessment of the alignment of the synthetic data with the underlying restricted data. In addition, assessments will be made of the quality of estimates produced from the synthetic data (bias, fidelity to true data, and disclosure risk). In addition, verification metrics will need to be developed so that researchers, data users, and other stakeholders can request them on an as needed basis. Note: no estimates based on the restricted data will be shared but rather a metric indicating alignment of the estimates (e.g., distribution comparisons, correlation heat maps, metrics on distance between synthetic and true data).
4. Identify use cases for this synthetic data to assist in determining how the synthetic data could be optimally utilized by researchers, data users and other stakeholders.
5. Develop a plan for accessing the synthetic data, verification metrics, and messaging regarding this new synthetic data product.
6. Write a report outlining the considerations that are needed for synthetic data generation with large RWD and a framework of the governance considerations to create and release synthetic data. In addition, the report should address the ethical considerations around data privacy and the limitations of synthetic data.

Information Gaps

This project will identify:

- bias in synthetic data estimates when compared to the truth data
- disclosure metrics and assessments for synthetic data created
- a framework to inform a synthetic data toolkit that will include guidance and governance for synthetic data generation with a large RWDF file

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:

- Which novel techniques for data, privacy, and confidentiality protections can be used while maintaining utility for large RWD?

- Are the resulting synthetic data fit for purpose to support research, evidence building, and/or policy making?
- What mechanisms are needed to access the resulting data that uphold privacy requirements?
- How will the results of this work inform a synthetic data toolkit that will benefit the research community and guide future synthetic data generators?

Deliverables

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

- Monthly status reports on progress towards project objectives.
- Monthly or bi-weekly status update meetings with project team.
- Quarterly lessons learned based on the previous quarter to inform a future NSDS and the NAIRR pilot.
- All code (clearly documented), documentation of synthetic data methodology, documentation of data quality assessment, and any other documentation created under this award. Data should be made available in an agency designated repository.
- Documentation of verification metrics and alignment with true estimates.
- A report that outlines the framework to inform a synthetic data toolkit with guidance for synthetic data generation with a large RWD file. The report will describe the lessons learned through this project including but not limited to the creation of synthetic data and whether the resulting data and models fit are able to inform research questions. In addition, the report will describe how this approach could inform a tiered access model and contribute to a potential NSDS and inform the NAIRR pilot.