Synthetic Data Generation with Large, Real-World Data (DG-RWD-24) FAQ

As of July 3, 2024

	Question	Answer
1	Does the contractor need to identify the dataset, or will the dataset be provided?	The data set will be defined by the government team. It will likely be the N3C data from the National Institutes of Health
2	What specific linkages are currently provided to help researchers by the NCSES?	Linkages are not part of this project. Other linkage projects can be found on the ADC website. This project is looking at creating synthetic data with a very large real world data file.
3	Is the synthetic data required for 'un-structured' data like images, videos, blobs, or hand-written notes as well?	The project will start with codified data and may possibly expand to other multi-modal data types.
4	Is there a link or reference to the NIH data?	It's the N3C data. More information can be found at these websites: https://ncats.nih.gov/research/research-activities/n3c/data-overview https://ncats.nih.gov/sites/default/files/OMOP_CDM_COVID.pdf https://ncats.nih.gov/research/research-activities/n3c/faqs
5	What tools are available in the secure compute environment that can be used for synthetic data generation? Like in terms of CPUs/GPUs, memory available, etc.	Several open-source packages are available in the enclave. In addition, there are tools currently available through N3C and other tools being developed as part of the National Secure Data Service Demonstration project that can be imported into the NAIRR secure compute enclave.
6	Are all analyses and data expected to stay within the NAIRR or are other computing environments (i.e., AWS) open for consideration?	The workspace will be within the NAIRR secure compute environment. Detailed discussions of workspace will occur upon award.
7	Would you be open to use Commercial Cloud based products to generate synth data - instead of building something from scratch? This is especially when the type of data changes with changing dimensions.	A requirement of the request for solution is that it is an open source, synthetic data generator, so that would take out the commercial aspect.
8	Who is the intended audience for the deliverable reports, code, project outputs, etc.? Will these also require "tiered access" for disclosure protection, especially for metrics that depend on confidential data (like statistical distances to confidential data)	It is the Government and the project team, but it is possible that the reports will be made public, In terms of the code and the project outputs, it would be for the government project team. A key aspect of the project is figuring out how to release the verification metrics to enhance utility while protecting privacy.,

	Question	Answer
9	Does the variable selection described in Step 1 imply that the dataset(s) to be synthesized could be represented in a single table vs. a relational database structure?	The idea is that the variable selection would be based on the true data, and we would hope to create a Microdata file that could be analyzed where researchers or data users would not have access to the true data, but they would be able to use the synthetic data in similar ways that they would use a restricted data source.
10	When are expected award dates and award amounts?	Please refer to the Request for Solutions for the expected award date and award amount.
11	Where would the data be generated? Just wanted to make sure we are clear on whether we will work on infrastructure or if that is established, whether it must be processed strictly "on site" within NAIRR	It will be within the NAIRR secure compute environment.
12	Is the data a single dataset or interrelated datasets with foreign keys?	There are many datasets in N3C that are connected through an inter-relational database.
13	Would we get access via an API?	The NAIRR secure compute environment is a secure enclave and is not accessed through an API.
14	How is the cost of the compute environment handled - is that something that needs to be budgeted	There will be no cost for the vendor to work within the NAIRR secure compute environment.
15	Will the dataset(s) have unique identifiers that correspond to natural persons? The RFS mentions that the RWD could be >30billion total rows	The true dataset will have unique identifiers, but the synthetic data should not include any identifiers of patients.
16	Are there any expectations on performance/ runtime? Generate synthetic data in hours? days? weeks?	This is part of the project goals since we plan to test capabilities.
17	Can all anticipated NAIRR secure computing resources be accessed remotely, or will any require an in-person presence?	This can all be done remotely.