MLDS Center
Maryland Longitudinal Data System

Address 550 West Baltimore Street
Baltimore, MD 21201
Phone 410-706-2085
Email mlds.center@maryland.gov
Website www.MLDSCenter.org

Memorandum

To: The MLDSC Governing Board

From: Synthetic Data Project (SDP) Team

Date: September 13, 2019

Subject: Assessment of Synthetic Data and Beta Testing

The SDP Team will be presenting an update at the September 13, 2019 Governing Board meeting. There are three central topics to address:

1. Findings from our assessment of the Research Utility of the synthetic data;
2. Findings from our assessment of the Disclosure Risk of the synthetic data, and
3. An overview of our planning and anticipated benefits from Beta Testing, which we will then be requesting permission from the Governing Board to implement.

Below please find a brief description of our assessment of Research Utility and Disclosure Risk, and our plan for Beta Testing. On page two, please find brief definitions of key terms. After our Governing Board presentation next week, we look forward to responding to your questions.

Research Utility

Research utility refers to how well the properties of the original data are represented in the synthetic data. There are two types of research utility:

1. General Utility, which is the extent to which individual variables in the synthetic data look like the variables in the original data (for example counts, means, and distributions); and
2. Specific Utility, which is the extent to which relationships *between* variables in the synthetic data are similar to those relationships in the original data (for example, bivariate correlations and multivariate analyses).

Our initial assessment of Research Utility did identify aspects of the synthetic data that needed improvement. As a result, we made changes to the models used to generate the synthetic data and our current synthetic data sets are robust research tools that provide substantially similar findings as the data from which they were generated (the GSDS, definition below).

Disclosure Risk.

There are two types of disclosure risk:

1. Identification Disclosure, which is the potential for an intruder to match a given record with a specific individual; and
2. Attribute Disclosure, which is the potential for an intruder to identify sensitive characteristics about small subpopulations in the data.

We have assessed both types of Disclosure Risk and consulted with experienced experts at the US Census Bureau, who have been generating and releasing synthetic data for 9 years, and concluded that the fully synthetic data sets (see definition below) we have created are robust, very safe and pose no discernible identification or attribute disclosure risk.

Beta Testing

While we have determined the synthetic data sets have both strong research utility and security, we are still proposing an additional Beta Testing step in the project. Beta Testing will add to the rigor of our assessment of the synthetic data sets. We want to recruit six to eight Research Utility and two Disclosure Risk Beta Testers who will develop analyses with the synthetic data (which we will then compare to the same analyses with the original data) or try and identify individuals or values of sensitive variables for small subgroups in the data. We anticipate these outside Beta Testers will lead to analyses and attempts to penetrate the security of the data we might not have anticipated and therefore offer a *real world* field test of the synthetic data and further our assessment of both Research Utility and Disclosure Risk. These Beta Testers will be known trusted professionals who will be given access only to the synthetic data and be 0vetted through the standard procedures as affiliate researchers of the MLDSC.

Definitions

Gold Standard Data Set – The gold standard data sets (GSDS) are simplified versions of the data housed in the Maryland Longitudinal Data System and includes the cohorts of students, variables, and values that will be synthesized. GSDS variables represent a comprehensive set of academic and workforce indicators. The GSDS for the Synthetic Data Project consists of two cohorts: 1) students beginning postsecondary programs in the 2010-2011 academic year and followed into the workforce, and 2) students beginning high school in the 2010-2011 academic year and followed into postsecondary education and the workforce. Eight years of data are available for both cohorts.

Synthetic Data Set – A synthetic data set (SDS) is comprised of variables that are created from a computational model that is based on the characteristics and relationships among variables in the GSDS. In short, a synthetic data when analyzed will act like the original data it was based upon. The SDS created for this project has the same variables as the GSDS, but what is critically important is that the distributions of synthesized variables will differ from those in the GSDS, which provides a strong layer of privacy protection.

Fully Synthetic Data Sets – There are two types of synthetic data sets: partially synthetic and fully synthetic. Fully synthetic data maximizes data security and privacy protection in relation to partially synthetic data. For example, the US Census Bureau, an early adopter of synthetic data as a data access strategy, started using a partial synthetic data strategy and have switched to a fully synthetic data strategy. For the MLDSC Synthetic Data Project, we made the decision that all our data sets would be fully synthetic data sets. Fully synthetic data sets are comprised completely of synthesized variables, all variables and values are synthesized in creating such data sets. One of the reasons fully synthetic data are more secure is that having no original values means an individual record in the synthetic data cannot be identified as belonging to a real individual.

Partially Synthetic Data Sets – Partially synthetic data sets are comprised of a combination of synthesized and non-synthesized variables (which have the original "real" values). Variables considered sensitive (confidential) are synthesized, while non-sensitive variables (age, gender, race) are the same as in the GSDS. Given a few real values in each case record in a partially synthetic data set means there is a small potential to identify an individual in those data based on those few real values with the synthesized sensitive values then offering a general indication of what that person's data may look like (although not the true values).