SLDS Issue Brief

Maryland's Synthetic Data Project

The Maryland Longitudinal Data System Center (MLDS Center) is investigating the use of a synthetic data method to increase the amount of rigorous policy research conducted with MLDS data while protecting confidential individual data.

The method would allow policy analysts and researchers to use synthetic data without going through the lengthy approval process required to use confidential data. In addition to increasing access to MLDS data and the data's impact on policy and practice, the project could be a model for states seeking to protect confidential data while encouraging statewide longitudinal data system (SLDS) use for research, training, and evaluation.

Maryland's synthetic data project is the work of the MLDS Center and the Maryland State Department of Education (MSDE) as part of a 2015 SLDS grant awarded by the U.S. Department of Education. The MLDS Center partners with Maryland Higher Education Commission; Maryland Department of Labor, Licensing, and Regulation; MSDE, the University of Maryland, Baltimore; and the University of Maryland, College Park.

What Are Synthetic Data?

The concept of synthetic data was first proposed by Harvard University Professor Donald Rubin in 2012 in response to the access constraints of sensitive individual-level data.[1] The goal of developing synthetic data is to provide publicly available datasets that can be used for valid research analyses in place of the confidential data.

Producing synthetic data requires identifying variables of interest and creating "gold-standard" files that contain the original confidential information. The gold-standard files serve as the basis for creating and evaluating synthetic datasets. Borrowing from imputation methods, or the process of replacing missing data with substituted values, MLDS Center staff members would construct joint distributions of the original variables. Then, they would randomly select values from the joint distributions to create multiple sets of new, or synthetic, data that mimic the actual data.

The synthetic datasets would then be evaluated to verify that their statistical characteristics were sufficiently similar to those of the original data. Before the synthetic datasets would be released, a disclosure risk assessment would be conducted. That assessment would ensure negligible risk of linking synthetic data records to the students, workers, schools, or employers represented in the gold-standard files.

To help verify results, external researchers completing analyses with synthetic datasets could request that the Center replicate the analysis with actual data. Currently, the Survey of Income and Program Participation (SIPP) synthetic data project of the U.S. Census Bureau provides such an option for external researchers.

Maryland's Synthetic Data Project

The MLDS Center serves as a central repository of data from all levels of the state's education and workforce programs. Because such data could be linked to individual students, workers, schools, and employers, the Center treats the data as confidential.

[1] Donald Rubin curriculum vitae, *https://statistics.fas.harvard.edu/files/statistics/files/rubin-cv-june2017.pdf*

This product of the Institute of Education Sciences (IES) SLDS Grant Program was developed with the help of knowledgeable staff from state education agencies and partner organizations. The information presented does not necessarily represent the opinions of the IES SLDS Grant Program. We thank the following people for their valuable contributions:

Chandra Haislet
Maryland State Department of Education

Laura Stapleton
Maryland Longitudinal Data System Center and the University of Maryland, College Park

Michael Woolley
Maryland Longitudinal Data System Center and the University of Maryland, Baltimore

Xiaying Zheng
Maryland Longitudinal Data System Center and the University of Maryland, College Park

Corey Chatis
SLDS Grant Program, State Support Team

Carla Howe, Ph.D.
SLDS Grant Program, State Support Team

For more information on the IES SLDS Grant Program or for support with system development, please visit http://nces.ed.gov/programs/SLDS.

Notes on Data Privacy

The federal Family Educational Rights and Privacy Act (FERPA) allows schools to share student data with the SLDS even if families have "opted out" of releasing directory information for other uses. Properly de-identified data can be released from the SLDS without violating FERPA and should be the preferred option for fulfilling research requests.

Review applicable state, federal, and local privacy laws, as well as current privacy policies in each agency involved with data integration. Check with the state attorney to determine state and local requirements for using or releasing student data.

The Privacy Technical Assistance Center (PTAC) team is available to answer questions about data privacy. Contact PTAC at PrivacyTA@ed.gov.

State law[2] limits direct access to data in the MLDS Center to authorized MLDS Center staff members. Additionally, the MLDS Governing Board's policy and MLDS Center regulations strictly limit access to MLDS data, which in turn limits researcher access to unit-record-level data.

The synthetic data method would expand access to such data while protecting confidential data.

The MLDS Center's synthetic data project has four goals:

1. *Create three gold-standard files, which cover K12 to postsecondary education, postsecondary education to workforce, and K12 education to workforce.* The specific variables to be included in the gold-standard files are being determined, in part, by an end-user panel convened in April 2017 to define the needs of interested education and workforce researchers from across Maryland.
2. *Generate multiple sets of synthetic data based on the gold-standard files.* The utility of the multiple synthetic datasets and the potential disclosure risk will be extensively examined before the decision to release the data is brought to the MLDS Governing Board. MLDS Center staff are examining the possibility of taking non-parametric approaches to the synthesis process—specifically, a classification and regression tree (CART) approach—and using information about the research questions of interest to end users.
3. *Disseminate information about the Center's synthetic data via a summit for education and workforce researchers.* An online access portal and training materials will be developed to ensure the utility of synthetic files for interested users.
4. *Examine the feasibility of using synthetic datasets for cluster-level inference analysis.* Education data typically have a hierarchical or multi-level structure in which students are clustered within classrooms and schools, and schools are clustered within local education agencies. The question of whether synthetic data can be developed and used to capture cluster-specific deviations (i.e., random effects) has not been explored previously, but the synthetic data project provides the opportunity to further understand this possibility.

What are the benefits of using synthetic data?

There are several benefits of using synthetic data.

1. Synthetic data allow external researchers to access data at a granular, individual level to allow for more nuanced analyses. Traditionally, a common strategy for protecting sensitive or confidential data was to provide aggregated data, which cannot be used to answer detailed questions and do not allow for many types of analysis.
2. When properly generated, synthetic data can yield comparable results to those from the original data without violating confidentiality.
3. Creating gold-standard datasets and re-running analyses on the original data could help the data-hosting agency better understand its own data and improve the analytic validity of the synthetic datasets.
4. Synthetic data greatly expand the number of researchers—each bringing different backgrounds, expertise, and orientations—to important education and workforce issues in Maryland.

The synthetic data method has great potential to help SLDS agencies efficiently use limited resources. Data stored in the SLDS are expected to become exponentially more useful with the size of the database; the demand from researchers would predictably also increase. Allowing external researchers direct access to individual-level synthetic data through the SLDS could alleviate the burden of handling research requests.

[2] Education Article, §24-707, Annotated Code of Maryland, *http://mgaleg.maryland.gov/webmga/frmStatutesText.aspx?article=ged&section=24-707&ext=html&session=2017RS&tab=subject5*

Additional Resources

Maryland Longitudinal Data System Center
https://mldscenter.maryland.gov

Maryland Longitudinal Data System Center Policy Reports
https://mldscenter.maryland.gov/PolicyReports.html

Maryland State Department of Education
https://www.marylandpublicschools.org/Pages/default.aspx

Nowok B., Raab, G.M., Dibben, C. (2016). synthpop: Bespoke Creation of Synthetic Data in R. *Journal of Statistical Software, 74(11)*, 1-26.
https://www.jstatsoft.org/article/view/v074i11

Raghunathan, T. E., Reiter, J. P., & Rubin, D. B. (2003). Multiple imputation for statistical disclosure limitation. *Journal of Official Statistics, 19(1)*, 1-17.

Reiter, J. P. (2004). Inference for partially synthetic, public use microdata sets. *Survey Methodology, 29(2)*, 181-188.

Reiter, J. P., & Raghunathan, T. E. (2007). The multiple adaptations of multiple imputation. *Journal of the American Statistical Association, 102(480)*, 1462-1471.

Rubin D. B. (1993). Discussion: statistical disclosure limitation. *Journal of Official Statistics, 9(2)*, 461-468.

U.S. Census Bureau's Survey of Income and Program Participation (SIPP) synthetic data project
https://www.census.gov/programs-surveys/sipp/guidance/sipp-synthetic-beta-data-product.html