MLDS Center
Maryland Longitudinal Data System

Better Data • Informed Choices • Improved Results

September 2020

Expanding MLDS Data Access and Research Capacity with Synthetic Data Sets

Submitted by:

Maryland Longitudinal Data System Center

Ross Goldstein, Executive Director
Angela K. Henneberger, Ph.D., Director of Research

Authored by:

Michael E. Woolley, Ph.D.[1]
Laura M. Stapleton, Ph.D.[2]
Daniel Bonnéry, Ph.D.[2]
Mark Lachowicz, Ph.D.[2]
Terry V. Shaw, Ph.D.[1]
Angela K. Henneberger, Ph.D.[1]
Bess A. Rose, Ed.D.[1]
Tessa L. Johnson, M.S.[2]
Yi Feng, M.A.[2]

[1] University of Maryland School of Social Work
[2] University of Maryland, College Park

Maryland Longitudinal Data System Center
550 West Baltimore Street
Baltimore, MD 21201
410-706-2085
mlds.center@maryland.gov
http://mldscenter.maryland.gov/

Suggested Citation

Woolley, M. E., Stapleton, L. M., Bonnéry, D., Lachowicz, M., Shaw, T. V., Henneberger, A. K., Rose, B.A., Johnson, T. L., & Feng, Y. (2020). *Expanding MLDS data access and research capacity with synthetic data sets.* Maryland Longitudinal Data System Center: Baltimore, MD.

Acknowledgement

This report was prepared by the Research Branch of the Maryland Longitudinal Data System Center (MLDSC). The contents of this report were developed under a grant from the Department of Education. However, these contents do not necessarily represent the policy of the Department of Education, and you should not assume endorsement by the Federal Government. We are grateful for the data, technical, and research support provided by the MLDS Center and its agency partners. The views and opinions expressed are those of the authors and do not necessarily represent the views of the MLDS Center or its agency partners.

If you have questions regarding this publication, please contact mlds.center@maryland.gov.

Table of Contents

This page intentionally left blank

Executive Summary

State education and longitudinal data systems are advancing and growing in number, and the use of these data systems for education and workforce research holds great promise (Figlio, Karbownik, & Salvanes, 2017). However, access to such data sets is typically restricted, and synthetic data, which statistically mimic the statewide longitudinal data system, are a promising solution to providing access while maintaining confidentiality. Yet, there have been no efforts to-date to synthesize data from statewide longitudinal data systems. As such, the MLDS Center synthetic data project (SDP) was funded by the Institute of Education Sciences (IES; $2.6 million) and had five core project aims: **Aim 1)** Design and create three gold standard data sets (GSDS) from the longitudinal data housed in the data system; **Aim 2**) Create innovative synthetic versions of those GSDS, by the application of well-established data imputation algorithms used to replace missing values in data to create entirely imputed versions of real data sets; **Aim 3**) Assess those synthetic data sets for research utility to determine whether those data substantially reflect analyses with the raw data they were imputed from; **Aim 4**) Assess the disclosure risk of those synthetic data sets to determine if they sufficiently protect the confidentiality of protected information about the real individuals whose data were synthesized; and **Aim 5**) Explore the feasibility of creating Cluster-specific synthetic versions of the MLDS data. This report provides an overview of the completion of tasks for each aim of the project.

This page intentionally left blank

Introduction

State education and longitudinal data systems are advancing and growing in number, and the use of these data systems for education and workforce research holds great promise (Figlio et al., 2017). Since 2005, the USDOE has supported 47 states, as well as the District of Columbia, Puerto Rico, the U.S. Virgin Islands, and American Samoa in their development of statewide education data systems (SLDS Grant Program, 2018b), representing an overall investment of $721 million in federal funding as of May 2018 (SLDS Grant Program, 2018a). State longitudinal data systems provide several advantages to researchers as compared to traditional survey measures, including larger sample sizes, fewer problems with attrition, and low rates of non-response bias. Administrative data systems are a relatively cost-effective approach to answering policy questions as there is no need for costly and time-consuming primary data collection. Access to such data sets is typically restricted; however, synthetic data, which statistically mimic the statewide longitudinal data system, are a promising solution to providing access while maintaining confidentiality. Yet, there have been no efforts to date to synthesize data from statewide longitudinal data systems[1]. A few large state datasets have been synthesized in the past. For example, the U.S. Census SIPP survey, the Institute for Employment Research IAB Establishment Panel in Germany, and the Scottish Longitudinal Study have all created and released synthetic versions of large longitudinal data sets (see below for more details). This project is the first to create and assess synthetic versions of longitudinal education to workforce data.

The MLDS Center synthetic data project (SDP) was funded by the Institute of Education Sciences (IES; $2.6 million) and had five core project aims: **Aim 1)** Design and create three gold standard data sets (GSDS) from the longitudinal data housed in the data system; **Aim 2**) Create innovative synthetic versions of those GSDS, by the application of well-established data imputation algorithms used to replace missing values in data to create entirely imputed versions of real data sets; **Aim 3**) Assess those synthetic data sets for research utility to determine whether those data substantially reflect analyses with the raw data they were imputed from; **Aim 4**) Assess the disclosure risk of those synthetic data sets to determine if they sufficiently protect the confidentiality of protected information about the real individuals whose data were synthesized; and **Aim 5**) Explore the feasibility of creating cluster-specific synthetic versions of the MLDS data. The SDP has been a collaborative research endeavor across two schools on two different campuses of the University of Maryland, the School of Social Work in Baltimore and the School of Education in College Park, as well as the Maryland Longitudinal Data System (MLDS) Center, and the Maryland State Department of Education (MSDE).

---

[1] One emerging strategy currently being investigated is the use of statistical disclosure control methods, which keep the original information in the raw datasets, but protect against the disclosure of identities (Award number R305D140045 from the National Center for Educational Research; PI: Hedges). A range of disclosure control methods exist, from data swapping across individuals, perturbing observations with random error, categorizing sensitive continuous measures into discrete categories, and suppressing sensitive variables and records altogether (see Little, 1993). These methods differ from synthetic data strategies as the majority of these methods still release some elements of the original data.

Background

The overarching goal of the SDP was to investigate the application of synthetic data as a data sharing strategy for the MLDS Center that could serve two purposes. The first purpose was that the Center would be able to effectively and safely fulfill data requests by providing synthetic versions of the data housed in the MLDS. The second potential purpose was that synthetic data sets could be strategically made available to policy analysts, institutional researchers, or academic scholars, while maintaining the security and confidentiality of the data in the MLDS. Such strategically released synthetic data sets could leverage the potential benefits of the data housed in the MLDS to inform education and workforce policy and programming in Maryland and advance knowledge about education and workforce activities to Maryland and beyond. Such advancements could come from the various potential end users of such synthetic data sets; for example, state-level policy makers and analysts, institutional researchers in school districts, colleges, universities, and employers; as well as scholars and researchers across disciplines interested in education and workforce programs, practice, and outcomes.

We begin this report with an overview of synthetic data: what it is, what it is for, and what has previously been accomplished with synthetic versions of state-collected data. We then discuss the design and construction of three GSDS, the data sets derived from the data in the MLDS in order to create synthesizable data sets. Next, we detail the processes to create synthetic versions of those gold standard data sets. We then describe our efforts assessing two key characteristics of synthetic data: one, the robustness of those synthetic data sets as a research tool – do they give similar findings to the raw data?, and two, the security of those synthetic data – do the synthetic data present any risk of exposure of the protected and sensitive values about the individuals, Maryland students and employees, represented in the raw data?

Brief Overview of Synthetic Data

The creation of synthetic data from the MLDS involves starting with a set of raw data from the MLDS, and through the application of an imputation model selected based on the nature of that data, imputing multiple "synthetic" versions of that data set that are statistically similar but not identical to the raw data (see for example, Abowd & Woodcock, 2001; Drechsler, 2012; or Rubin, 1993). In this way, researchers have access to microdata, or unit record-level data, that closely mimic the properties of the raw data. Synthetic data sets can be partially synthetic or fully synthetic. In a partially synthetic data set some variables will maintain the raw (original) values, usually variables that are not protected or sensitive, while protected or sensitive variables are synthesized. In a fully synthetic data set, all variables, categorical or numerical, sensitive or not, are synthesized. Importantly, with the use of fully synthetic data, those who collect and are ultimately responsible for the data can be assured that the risk of disclosure of the true data is virtually nonexistent and that individuals about whom the data were collected are not exposed (Drechsler, 2011). In theory, this process allows confidentiality to be strongly maintained, while also giving policy analysts and researchers access to microdata (individual-level granular data), allowing for increased data utilization toward a wide range of

data analyses. Such data can be used to answer a much wider range of research and policy questions than other strategies for sharing restricted confidential data.

There are various methods that can be used to generate synthetic data, all of which require the application of a statistical model to capture relations among variables and across the "cases" (in the raw data, a *case* is a set of values for an individual, in the instance of a synthetic data set, a *case* is a set of imputed values that do not reflect a real individual) in the raw data. Synthetic data generation is traditionally accomplished with sequential regression models. Variables are arranged, and therefore synthesized, in a certain order, which is chosen to best capture the relations between variables and cases. For each variable in turn, a regression model is developed against a selected set of predictors from among the available variables. Such models are developed in a sequential manner until a model is developed for each variable in the data (Drechsler & Reiter, 2011; Raghunathan et al., 2001; Van Buuren, 2007). Synthetic data are thus generated from the posterior predictive distribution for each variable. Synthetic datasets can be produced through a Bayesian posterior predictive distribution as first detailed by Gelman et al. (2003) and Raghunathan et al. (2003). Values are randomly drawn to create the synthetic data in a process reminiscent of multiple imputation, except instead of imputing select missing values, entire data records for "individuals" are imputed (Drechsler, 2011; Harel & Zhou, 2007; Rubin, 1987; Schafer & Graham, 2002). The synthetic data will thus have similar statistical properties to the raw data (because they come from the same multivariate distributions provided that the statistical model is adequately specified) but will be comprised of *cases* each with a set of variable values that do not correspond to real individuals.

Although the idea of synthesization seems fairly straightforward conceptually, it can be difficult to create an appropriate probability distribution such that, across various statistical analyses of the synthetic data set, the results replicate the results and therefore inferences that would be derived from the same analyses on the raw data from which the synthetic data were based (Drechsler, 2011; Reiter, 2009a). The quality and usefulness of synthetic data therefore are highly reliant on the modeling process used to capture the relevant nuances of the raw data (Matthews et al., 2010; Reiter, 2005; 2009a; 2009b). As Matthews and Harel (2011) concisely summarize, "synthetic data sets are only as good as the models used for imputation" (p. 10). Therefore, when generating synthetic data, a critical determinant of success is the adequacy of the model used to capture the relations between variables. Further, a key step in any synthetic data project is to systematically evaluate the quality, or *research utility*, of the synthetic data. This evaluation is conducted by first examining the characteristics (distributions, descriptive statistics) of the variables in the raw data as compared to the synthetic data, then running analyses on the SDS and the raw data to see how closely the results mimic each other.

A number of challenges exist when creating synthetic data from complex education-based data systems. First, a particular challenge of educational data is the complex hierarchical structure where students are often cross-classified or have multiple memberships (Beretvas, 2011). For instance, students who move during a school year could belong to multiple school districts and students who attend the same middle school may not all attend the same high school. Currently, statistical theory has yet to devise a method for creating synthetic data with such a complex hierarchical structure, and Reiter (2009b) argues that this is a key area of future research. Another challenge in the creation and use of a synthetic data system is whether end-

user researchers have sufficient confidence in the data. Some may not trust the synthetic data and choose not to use them even though they would use the comparable raw data if they were available (Reiter, 2005). Additionally, although the synthetic data mirror the raw data, the two are not equivalent and researchers might overgeneralize their conclusions.

Drechsler (2015) suggested that the results from synthetic data may not be appropriate for publication in academic journals. As an alternative to reporting results from analysis of synthetic data, the synthetic data could be used to design and develop code for statistical analyses. This code could then be passed on to MLDS Center staff, who could run the code using the raw data and pass the results along to the end user without ever having to disclose any raw data (Reiter et al., 2009); this process has been used by other governmental agencies who have created and released synthetic data (detailed below). Finally, good practice in the application of a synthetic data system is to create several different synthetic data sets from a single multivariate probability distribution, as is done in multiple imputation. Such replication allows for proper estimation of variance (Raghunathan et al., 2003); however, properly utilizing such a set of synthetic data replicates can be complicated.

Overview of Prior Use of Synthetic Data by Governmental Agencies

Synthetic data have been used as a strategy for public release of a few government collected data sets in the United States and Europe. For example, the U.S. Census Bureau has generated and disseminated synthetic versions of data from two of their programs. The Survey of Income and Program Participation (SIPP) data (see Benedetto et al., 2013) have been merged with Social Security data about retirement and disability benefits received and Internal Revenue Service data about income. These SIPP Synthetic Beta (SSB) data sets currently include nine SIPP panel waves from 1984 to 2008 (U.S. Census, 2018). The Census, through the creation and release of these synthetic data sets in collaboration with Cornell University faculty, greatly advanced the development of the methods and procedures in the use of synthetic data as a strategy to expand access to administrative data. The U.S. Census Bureau has also created synthetic versions of the Longitudinal Business Database (LBD; Kinney et al., 2011). This synthetic data set covers the LBD from 1976 to 2000, representing 21 million records including 3-digit Standard Industrial Classification codes, annual employment, and payroll information.

Jörg Drechsler (2009; a consultant on the current project) led an effort to create synthetic versions of the Institute for Employment Research IAB Establishment Panel in Germany, which was the first large scale application of synthetic data to expand access to government data. This panel data set was initiated in 1993 and has been collected annually since 1996. The data includes a stratified sample of German employment data from the German Social Security Data and is integrated with health, pension, unemployment insurance, and employer data.

A fourth example comes from the Scottish Longitudinal Study (SLS), one of the most ambitious state-created longitudinal databases in the world, which contains data collected annually beginning in 1991 from a randomly selected 5.3% of the population, including data starting from birth records, through education, health data, marriages, maternity, pollution exposure, weather, and workforce, until death (SLS, 2019). These rich data were only accessible

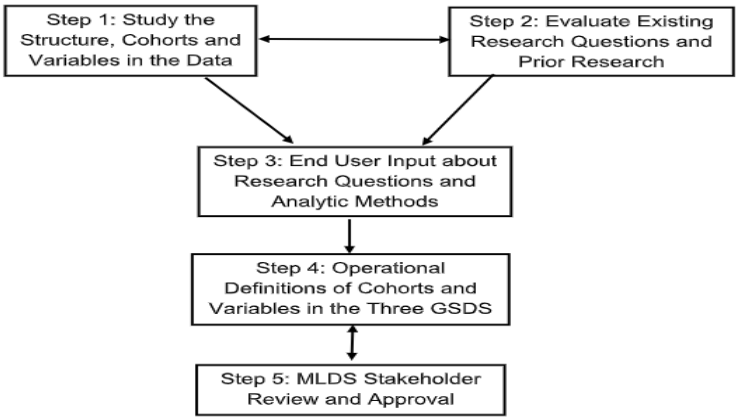
to a small number of government researchers, so they created three synthetic data sets for public access. These data sets each include a limited number of variables with real values that do not present disclosure risk supplemented by synthetic versions of an additional number of variables (i.e., partially synthetic data sets). These data are available for public download with two stated aims: (1) researchers can gain familiarity with the SLS data prior to applying for access, and (2) for use in university or training settings. Disseminating analyses with the synthetic data is not recommended. Instead, it is suggested that once analyses are developed, researchers may apply for access and, if approved, come to Edinburgh, Scotland and run their analyses on the raw SLS data on secure data terminals set up for that purpose.

Project Specific Aims

From this brief overview of how synthetic data are generated, what they can be used for, and who has previously created synthetic versions of state collected and maintained data sets, we now turn to describe our successful completion of the five central aims of the SDP over the four years of the project. Figure 1 details the flow process of the aims of the SDP, including creating GSDS, synthesis, assessment of the synthetic data, and presentation of results to the MLDS Governing Board. We further detail the completion of each aim below.

Figure 1.

Gold Standard Data Set Development Flow Chart



Note: This Figure was informed by a Synthetic Data Project flowchart in Bonnéry et al., 2019.

Aim 1: Creating the Gold Standard Data Sets (GSDS)

A GSDS is a well-defined subset of the available data housed in a data system. Synthetic data are only as good as the raw data sets from which they are imputed. The first project aim, the creation of the GSDS, was therefore critical to the success of the project. Given that the MLDS is very large and complex, creating GSDS necessitated designing smaller less complex "synthesizable" data sets that maintain the most essential information contained in the raw data in the system. This was the first time this process was done on a data system as large and

as complex as the MLDS, and it was a lengthy, detailed, and arduous task. The GSDS should be refreshed periodically when identities are updated and variables and years of data are added to the system. Each time the GSDS are refreshed, the process should be faster and easier, as the road map for this critical procedure is already created and updated.

The generation of the three GSDS for this project followed an iterative process that involved: (1) investigation of the larger data system and the variables included in the data, (2) assessing the purposes of the data elements for research and policy analysis, and (3) consultation with potential end users of the data as to the types of research questions and analyses they would seek to pursue with the synthetic data (see Figure 1).

Operational Definitions of Cohorts and Variables in Three GSDS. Based on our study of the data and variables, prior research with the data, and end user feedback, we decided to design and create three GSDS that correspond to different trajectories, respectively: *Data Set One,* high school to workforce; *Data Set Two,* high school to postsecondary education; and *Data Set Three,* postsecondary to workforce. Once these three data sets were chosen, we then defined the cohorts of students and years of data included, then engaged in a variable-by-variable process to select the variables and decide on recoding of values to be included in these three GSDS as described in the following sections.

Cohort Definitions. When defining the cohorts, we were primarily concerned with the availability of data in the system, as complete data were not available for all cohorts over the entire span of the designated longitudinal trajectories (at the time of creating GSDSs, 2008 was the earliest year of data availability and 2016 was the latest). We defined the first cohort for the GSDS as students who first registered as Maryland public college freshmen in academic year 2010-2011 and followed this cohort for six years until academic year 2015-2016. We defined a second cohort as students who attended their first year of high school, 9th grade, in academic year 2010-2011. We also followed this high school cohort for six years until the academic year 2015-2016, both in the high school to postsecondary data set and in the high school to workforce data set.

Variable Selection. As for variable selection, the primary concern was about the trade-off between data utility, confidentiality, and practical feasibility in the process of synthesization. For some variables, decisions on aggregating information or even creating new variables are needed (for example, keeping the highest score of SAT over multiple records or combining several sources of financial aid under a category of “need-based” aid). In such cases, it is also important to consider the need for a straightforward definition and clear documentation, which lays the foundational work for creating an end-user data dictionary for the final product. Based on the information gathered at previous steps, we first made a broad selection of variables from those with an acceptable missing rate (i.e. we wanted to avoid variables that were not applicable or not available for large portions of individuals) and a clear, documented, definition. Next, we prioritized variables according to their level of research utility. Based on the data study and our knowledge about end users’ needs, we structured the GSDS to capture multiple aspects of performance in high school and postsecondary environments, including attendance, standardized assessment scores, completion status, and financial aid information. Non-identifying attributes of high schools and postsecondary institutions were also retained. Workforce data were simplified to capture the organization’s industry sector, quarter/year of employment, and wages received.

Granularity. The last step of variable selection and definition was to reduce the level of granularity in the tables within the GSDS, due to the constraints imposed by data security and utility, as well as practical concerns regarding synthesization. Aggregations reduce the level of granularity, but help with practical considerations for synthesizing data. Examples of the data aggregations included aggregating multiple attendance records for a student within the same school year, multiple financial aid awards received by a student within the same school year, or multiple test score records on the same test and subject area.

All steps taken to define and create the GSDS were completed under two anticipated constraints: (1) practicality constraints—we try to avoid having an end product that is too complicated for users to understand and to use, and (2) legal constraints—disclosure risks in that it was imperative to protect data confidentiality. As stated previously, the overarching goal of the SDP project was to better meet the needs of external researchers, who may come from a wide array of backgrounds (for example, education, policy, psychology, sociology, economics, or public health) and thus, have very different research interests. Therefore, it is desirable to provide them with clearly defined variables in data tables that are not overly complicated, along with good documentation (for example, a well-documented data dictionary, a codebook, and technical reports). Based on such practicality and user utility considerations, we decided to create a simple and straightforward data structure.

MLDS Center Stakeholder Review. As a final step, we presented the cohort definitions, list of variables, and simplified data structure to the major stakeholders within the MLDS Center, including individuals who are intimately familiar with the data in the MLDS as well as the source data from the state agencies. We gained feedback on the development and design of the GSDS, and we incorporated that feedback in an iterative process in order to ensure that the data are useful to a broad range of stakeholders.

Aim 2: Synthesization of GSDS

Below, we describe the synthesization procedure used to create multiple (30) implicates of each of the three GSDS. These synthetic data sets (SDS) needed to balance the trade-off between strong research utility and disclosure risk. We would conclude that the SDS met high research utility standards if the data preserved the unconditional distributions and multivariate conditional distributions of the GSDS. At the same time, we would only conclude that the data upheld disclosure risk standards if the data exhibited minimal, near-zero risk that the synthetic data could be used to ascertain sensitive protected values about Maryland students or employees whose data were included in the raw data that were synthesized.

Managing a Multi-dimensional Database. Synthesizing the MLDS statewide educational longitudinal data was particularly challenging in direct proportion to the multidimensional nature of the data system. The data in the MLDS are housed with multiple rows per person (i.e., long format). To accommodate synthesization, all datasets in this multidimensional relational database that correspond to the GSDS needed to be transposed so that all the important available information could be stored in a *rectangular two-dimensional* dataset with one and only one row per individual in the database (i.e., wide format). The outcome of the transposition and merging of a multi-dimensional dataset was a dataset with limited number of rows (in our case, approximately 30,000 and 60,000 for the respective postsecondary or high school cohort) and a very high number of variables. Once synthesis was completed, each SDS

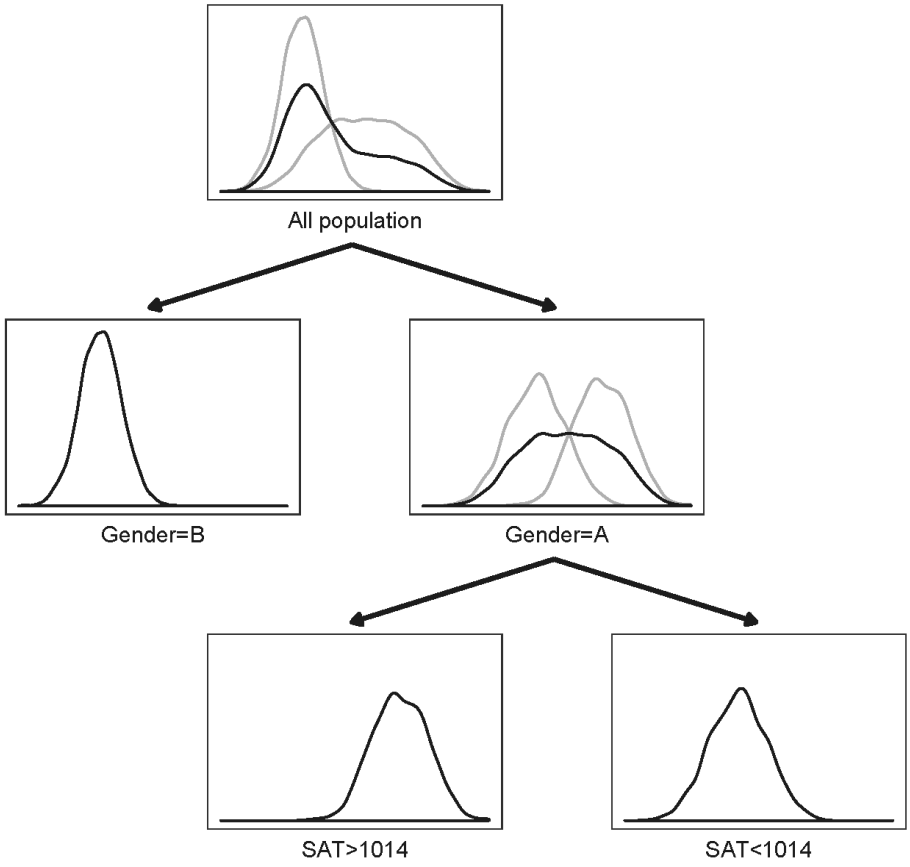
needed to then be back-transposed and merged to create a SDS with the same structure and dimensions as the GSDS from which it was imputed. The choice of variables and strategy that define the transposition must balance between having a manageable number of variables and having meaningful variables. One example would be the employment data, in which each row corresponds to an employment record for a specific individual. As individuals can hold multiple jobs and change jobs over time, keeping the same level of granularity in the GSDS as in the MLDS would result in an excessive number of columns in the transposed wide-format data tables, and many columns would be populated with sparse data (essentially, many zeros). Therefore, we decided to retain only the total wage amount an individual earns each quarter per year in each industry sector, by aggregating information over multiple job records.

Choice of a Method for Synthesization. When it comes to synthesization, various methods have been developed. The scope of the project was to assess the feasibility of using existing methods in the particular setup of integrated longitudinal multidimensional data with a very large amount of information. For this reason, after initial testing and evaluation of the different existing methods, and after convening and consulting with the synthetic data experts who had been recruited for this project, we made the decision to implement the Classification and Regression Trees (CART) method (Breiman et al., 1984; described in Reiter, 2005). CART offers ease of application and outperforms other methods (Drechsler & Reiter, 2011). Essentially, the CART approach is an algorithm for generating or predicting an outcome (e.g., wages) based on an individual's values of other variables. Figure 2 below depicts a simple example of a single regression tree, where posterior predictive distributions of individuals' wages were determined conditional on gender and overall SAT score.

Predictor Pre-selection and Order of Variables. Once the datasets were transposed and merged into a single rectangular file, the variables that will be used as predictors as well as the order in which variables will be synthesized needed to be chosen, as would be done in multiple imputation procedures (Little & Rubin, 1987). The chosen order has an impact on the utility of the final SDS because of the large number of variables. In general, important variables for research and variables with strong predictive power need to be synthesized first. However, arbitrary model pre-selection is necessary because of the high number of variables: the CART procedure fails in a reasonable amount of time when too many predictors are used. This procedure should be as inclusive as possible, but within the limit of the number of predictors that the synthesis procedure allows. The predictor pre-selection is based on content knowledge as well as common sense.

Creating the data-generating model and synthetic data. Using the CART method, we created a data-generating model that captured the relations among all of the variables in the GSDS. Once the data-generating model was identified, we were ready to create the synthesized data. This was accomplished by creating a set of 30 datasets, each with a similar number of cases to our actual GSDS with similar characteristics, and then recursively predicting, using the CART approach, another variable, until all of the desired variables were generated.

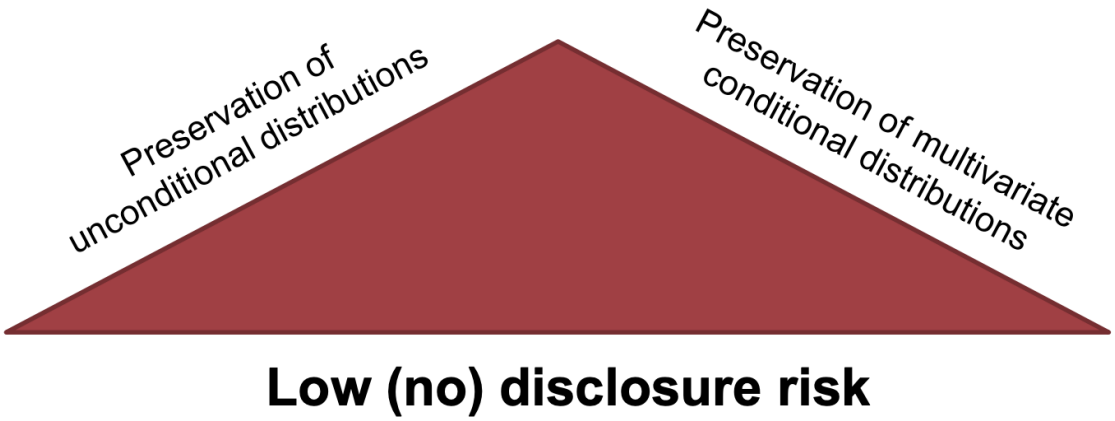
Figure 2.
CART Model Decision Tree with Gender and SAT Scores



Aim 3: Evaluation of Research Utility of Synthetic Data

In the assessment of synthetic data, research utility (RU) can be construed as the range of research questions that can be investigated by users and the quality of inferences yielded by analyses to address those questions. Further, the range of research questions is a characteristic of both the synthetic data and the gold standard data from which the synthetic data were generated. Thus, our assessment of RU of the SDS created for this project had two primary tasks: (1) evaluation of the RU of the GSDS, and (2) evaluation of the RU of the SDS, which is an iterative process to both assess the performance of the synthesis process to inform refinements and then ultimately assess the RU of the final SDS. In general, the goal is to satisfy a triangular trade-off (see Figure 3) between preservation of unconditional model distributions, preservation of multivariate conditional distributions, and no disclosure risk.

Figure 3.
Triangular Trade-off in Evaluating Synthetic Data



In general, the RU of the GSDS was assessed in terms of the scope of information—density and diversity of variables in the dataset, information quality—completeness and consistency of measured variables, and population definition—specification of cohorts (see Bonnéry et al., 2019 for detailed description of our assessment of the RU of the GSDS). Model specification is important for RU, and due to the large number of variables and observations in the current project, misspecifications by model were expected. Therefore, the RU of the SDS was carefully evaluated by examining discrepancies between the SDS and GSDS. Of particular interest was the degree of fidelity in analytic inferences drawn from the SDS, or inferential utility, essentially comparing analytic results across the GSDS and SDS.

We assessed the RU of the SDS across three dimensions. The first dimension was the performance of the synthetic data-generating CART model, including confirmation of the statistical relationships present in the GSDS and the CART modeling approach and determination of the number of predictors in each CART model. The second dimension consisted of exploratory comparisons of specific variables—for example distributions and descriptive statistics, termed *general utility*. We conducted this utility assessment by numerically investigating descriptive statistics—frequencies, means, standard deviations, skewness, kurtosis—of synthetic and their companion gold standard variables, and graphically with plots of variable distributions. An example of summary statistics comparing SDS to GSDS variables can be found in Table 1. The values for the statistics from the GSDS closely mimic the average values across the synthesized datasets. In general, categorical variables with relatively large cell counts and continuous variables with low rates of missing data tended to produce synthetic data that closely resembled the gold standard on a univariate basis. This was expected because typically CART-based generation is better able to reproduce idiosyncrasies of variable distributions compared to other synthesization methods (Raab, 2016).

Table 1.

Comparison of Summary Statistics in the Gold Standard Data Set to Averages Across 30 Synthetic Datasets to Assess General Utility for Cumulative GPA

Statistic	GSDS	AVG SDS M (SD)
N	26864	27898.79 (184.60)
Missing	2267	2217.24 (44.52)
Mean	3.20	3.19 (0.00)
SD	0.47	0.47 (0.00)
Skew	-0.27	-0.25 (0.02)
Kurtosis	2.51	2.50 (0.11)
Min	0	0.16 (0.47)
Q10	2.56	2.56 (0.00)
Q25	2.87	2.86 (0.01)
Median	3.22	3.21 (0.01)
Q75	3.57	3.56 (0.01)
Q90	3.83	3.82 (0.01)
Max	4.26	4.25 (0.01)

Notes. GSDS = gold standard data set; SDS = synthesized data set; M = Mean; SD = standard deviation; Q = quintile.

The third dimension of RU is a comparison of results of model-specific analyses across the GSDS and SDS, referred to as *specific utility*. A set of gold standard analyses were proposed to test several types of variable associations. These included analysis of mean differences, contingency tables, bivariate correlations, longitudinal analyses, and multiple linear and logistic regressions. One of the challenges of creating a synthetic version of the MLDS was modeling the longitudinal structure of the data, so we present the results of two linear regression analyses to illustrate how the generating model was modified based on early findings. We compared the results of several multiple linear regression models in which log-transformed 2016 wage was regressed on log-transformed 2015 wages, SAT math, gender, race (using two dummy codes of Black and other race with a referent of White), and a Hispanic ethnicity indicator. The predictor variables were renamed to be Variables 1 to 7 in random order, given that this example analysis should not be construed as a careful examination of the relation of wages to SAT and personal characteristics.

Results for specific utility for the early synthesis model are found in Figure 4 below. A visual inspection shows that the results for variables 2-7 were very close to their corresponding GSDS estimates (Average *SD* = 0.09, average *IO* = 0.81). However, the discrepancy between the estimates for Variable 1 revealed that the generating model was not well tuned for this particular variable (there was a large discrepancy). These results led us to refine the synthesis model, and a comparison of Figure 4 and Figure 5 shows that the effects from the refined

model were substantially closer to the effect in the GSDS for Variable 1, with Variables 2 – 7 remaining close to GSDS estimates.

Figure 4.

Assessing Specific Utility: Comparison of Pooled SDS Estimates Compared to GSDS from Early Synthesis Model

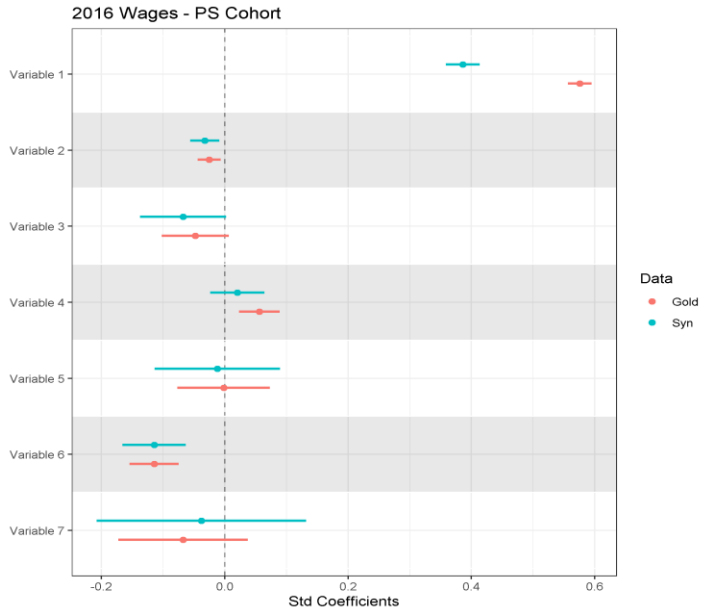
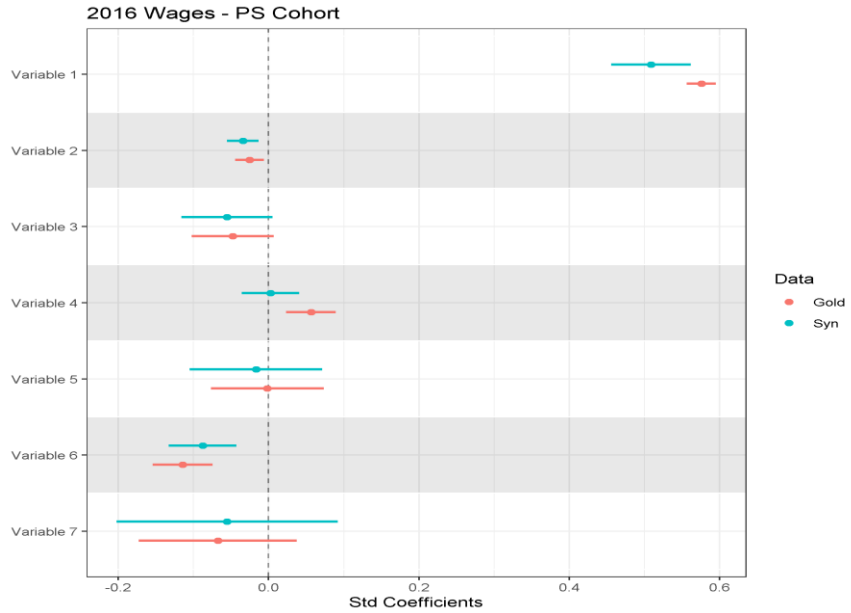


Figure 5.

Assessing Specific Utility: Comparison of Pooled SDS Estimates Compared to GSDS from Refined Synthesis Model



Beta Testers of the Research Utility of the SDS. To further test SDS usability and expand the range of RU analyses that were conducted, external researchers were recruited to serve as *beta testers* and were invited to implement analyses on the SDS. These beta testers were asked to develop four research questions based on the variables available in the SDS, translate these questions into statistical models and conduct analyses using various software packages, then pool results across the 30 SDS implicates. When these analyses were compared to those analyses run on the GSDS by the SDP research team results of these analyses yielded findings similar to those in our RU assessment.

Aim 4: Assessing Disclosure Risk of the Synthetic Data Sets

One barrier to the access and availability of unit level records from administrative state data is the risk of inadvertently disclosing sensitive or legally protected information. Disclosure risk is a challenge that continues to vex synthetic data projects as concerns about the safety of releasing synthetic data continue. There are two main types of disclosure risk: identification disclosure and attribute disclosure. Identification disclosure risk relates to the potential for an intruder to learn the original values of sensitive information about an individual through the accurate identification of a specific person in the synthetic dataset. In cases using fully-synthesized data, such as the synthetic data created for the SDP, the values for each variable in each record are randomly selected from an array of possibilities based on the statistical imputation of the relationship between the variable and all prior variables. We used a unique CART procedure that allowed for missing data while utilizing a method for trimming unique values for data synthesis (details provided in Bonnéry et al., 2019). Additionally, the process entails making available multiple random samples (replicates) of the synthesized data to users rather than the entirety of the data. In this scenario *identification disclosure* is determined as impossible as there are no "real" records—sets of data for real individuals—in the data set. While no real individuals exist in the synthetic data, there remains the possibility that aggregate sensitive information might be obtained for "rare" sub-groups of individuals who share a number of variable values in the data. Attribute disclosure risk falls into two categories – attribute disclosure through identity matching and attribute disclosure through correlations (Andreou et al., 2017). The threat of attribute disclosure through identification is a function of the risk of identification disclosure. If identification disclosure is established to be near zero, as it is in the fully synthesized datasets created in this project, then this type of attribute disclosure (through identification) will be near zero as well. Attribute disclosure through correlation does not rely on pre-identifying a record, but instead uses correlations across a series of variables to infer which records in the synthetic data have the highest probability of having the same values as a set of target records, which are associated with a rare subgroup in the original data. Attribute disclosure relies on some foreknowledge of such rare sub-populations or the presence of an external dataset that has some information on such rare sub-populations that are used as a comparison. Through our examination of disclosure risk with the SDSs, we found that the fully synthetic data structure did an exceptional job of protecting both individual disclosure risk and attribute disclosure risk (see Table 2 for sample statistics on disclosure risk).

Table 2.

Sample Disclosure Risk Statistics for the High School and Postsecondary Cohorts

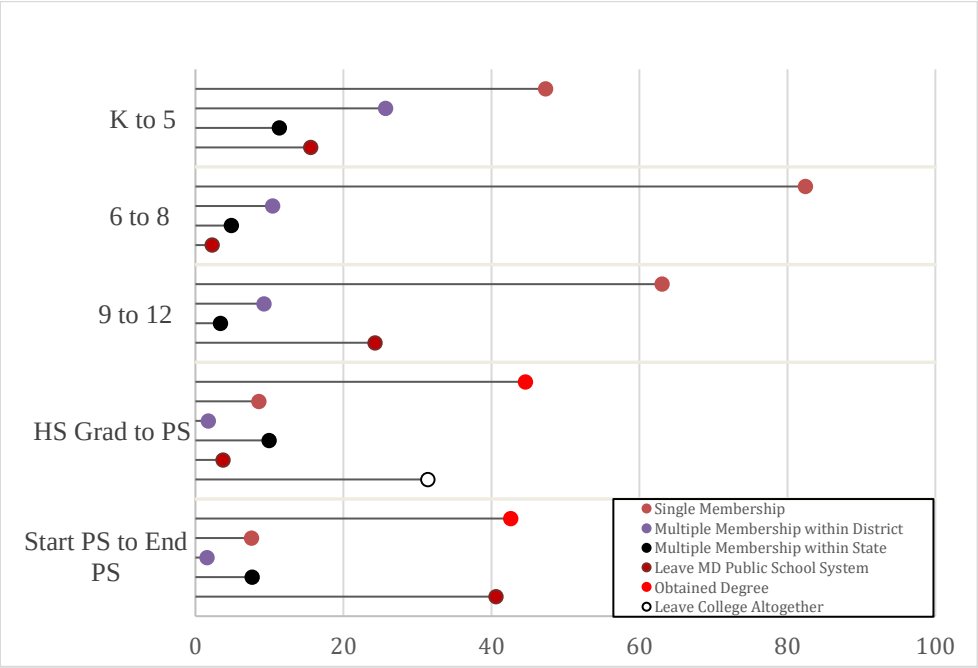
	High School Cohort	Postsecondary Cohort
Overall Disclosure Risk	0.000002	0.000006
Disclosure Risk for Average Person (records near the median across categories examined)	0.000029	0.000114

Aim 5: Assessing the Feasibility of Creating Cluster-Specific Synthetic Data

The SDP did not synthesize meaningful school, district, and other organizational clustering, which is commonly used in educational research when students are nested within schools and schools are nested within districts. As such, the cross-sectional hierarchies like school and district that exist in the original data are not meaningful in the synthetic data. Instead, all school- or district-level identifiers were completely randomly assigned to each student. For example, the School ID variable present in the SDS is purely a stand-in and bears no relation to any school identifiers in Maryland. As a result, any multilevel analyses conducted with the current SDS would spuriously find zero variance at the school-level precisely because the organizations have been fully ignored during synthesis.

The first step in assessing the feasibility of synthesizing clustered data was to fully understand the clustering in the MLDS. A common example of complex clustering is when students attend multiple schools (e.g., multiple membership). Using the MLDS data, we found that 53% of elementary school students, 18% of middle school students, and 37% of high school students changed schools at least once during the elementary, middle, and high school periods, respectively (Henneberger et al., 2019). We also examined the multiple membership rates in a high school cohort moving into postsecondary and in a postsecondary cohort (see Figure 6). Multiple membership rates varied by student characteristics (for example, English language learners and students receiving special education had higher multiple membership rates) and school characteristics (for example, students attending schools with higher percentages of students eligible for free and reduced price meals and urban schools had higher multiple membership rates).

Figure 6.
Multiple Membership Rates Across Maryland in the MLDS Data



In Figure 7 below, we present a set of local school system-level network plots that further display the movement of students (arrows) across high schools (dots) within a single academic year. In examining these plots, we see the intricate ways in which students change schools in Maryland public schools.

These examinations are particularly relevant to synthesizing clustered data because multiple memberships violate statistical assumptions of the traditional multilevel models (hierarchical linear modeling; Raudenbush & Bryk, 2002). Multilevel models assume that each student is nested within only one school, an assumption that is not upheld with the multiple membership structure of the high school data across Maryland. Henneberger and colleagues (2019) compared the results of a longitudinal analysis using a single membership multilevel modeling approach and a multiple membership modeling (MMREM; Beretvas, 2011) approach and found that the estimated effects of the predictors remained fairly stable across modeling approaches.

A few promising methods have emerged recently for implementing random effects CART models, which would be needed to synthesize clustered data: REEMtree (Sela & Simonoff, 2011) and Multilevel CART (Lin & Luo, 2019). For an exploratory attempt at creating clustered synthetic data, we went through a similar process to the one detailed in a previous section in order to produce a new set of synthetic data that maintained the clustered structure using a smaller subset of the data from the GSDS to reduce the computational burden of this method. In this subset, we retained high school attendance records and postsecondary enrollment

records for one cohort of students. The unique school identifier was retained for each student prior to synthesizing the rest of the data. Most person-level covariates were then synthesized using the original CART procedure. The focal outcome, individual 4-year college enrollment, was synthesized at the last step.

Figure 7.

Graphic Representation of Student Mobility across High Schools in Four Maryland School Systems

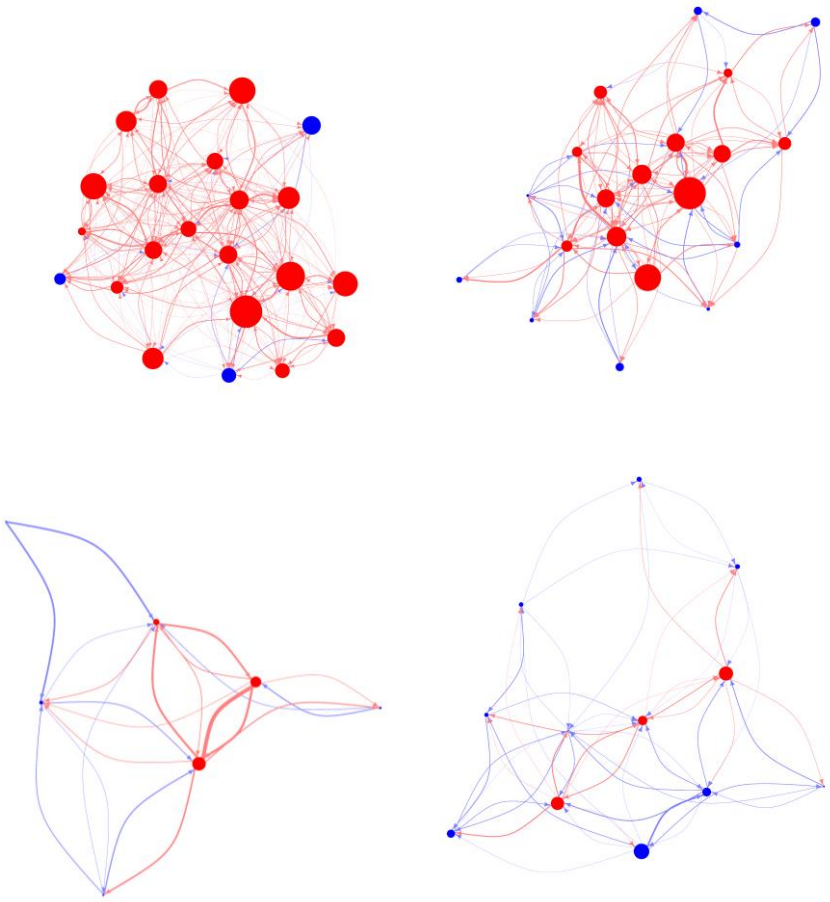


Figure Note: Network graphs of student mobility across schools within four local school systems in Maryland. The nodes, or circles, represent schools, and curved arrows represent rate of students moving from one school to another (thinner lines = lower rates). Larger node size is related to a higher proportion of students receiving free or reduced-price lunch and the color of the node is related to the school's pass rate for the Maryland High School Assessment (blue nodes had pass rates above the county median while red nodes had pass rates below the county median). The color of the arrow is derived from its origin.

We conclude that producing such clustered data is feasible as proposed above. However, the process would be met with two main challenges. First, the computational burden of synthesizing the entire GSDS is great. Future work should investigate different algorithms and computational advances that would facilitate synthesizing a large amount of clustered

data. Second, the synthesized data should be thoroughly assessed, both from the standpoint of research utility and disclosure risk, as was previously done with the current iteration of the SDS. These advancements and assessments are beyond the scope of the current research project.

Costs and Benefits Considerations of a Synthetic Data Strategy for the MLDS

As an increasing number of state data repositories apply synthetic data strategies, we predict the marginal costs to develop additional GSDS and SDS will decline. Just as we have benefitted from those who have created SDS previously, those who venture to create synthetic data sets going forward will benefit from the detailed descriptions of our efforts and lessons learned. We therefore assert the resources needed to successfully complete this project over the past 4 years through the SLDS grant from the IES through MSDE ($2.6 million) are substantially higher than the needed resource costs of applying synthetic data strategies going forward for the MLDS Center. The resources needed to continue to provide synthetic versions of MLDS data will be considerably lower now that the infrastructure and procedures have been developed in the MLDS Center. Further, as more data analysts and database engineers work on such projects and mentor students and early career practitioners, the knowledge and skills needed will become more widely available. As others have predicted previously (Drechsler, 2012; Rubin, 1993) we expect that synthetic data as a data access strategy will continue to expand. Maryland researchers have already consulted with other emerging state-level efforts to apply synthetic data. That said, we do not yet know what the costs may be for the MLDS Center to provide continued support, nor do we yet know the costs for researchers in terms of time invested in learning how to analyze synthetic data.

Conclusion

In closing, we assert synthetic data as a strategy to advance access to, and therefore the application and use of, state integrated data systems. Synthetic data is a promising way to leverage the expense of creating those data systems. That leverage comes from those with expanded access using those synthetic data sets to inform program delivery, policy development, and knowledge about education and workforce services. As the methods to create synthetic data advance, two critical dynamics of creating synthetic data will progress. First, what can be done analytically with synthetic data will continue to expand as we gain the models to create synthetic data that capture complexities, such as clustering in education data. Second, the costs to benefits formula will increasingly tilt toward the benefits as the costs go down both in terms of expended resources, for example as the availability of individuals who have developed the expertise in creating synthetic data increases and the efficacy and availability of the needed methods and related software programs advance.

References

Abowd, J. M., & Woodcock, S. D. (2001). Disclosure limitation in longitudinal linked data. In P. Doyle, J. Lane, L. Zayatz and J. Theeuwes, (Eds.), *Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies,* pp. 215–277. North-Holland.

Andreou, A., Goga, O., & Loiseau, P. (2017, July). Identity vs. attribute disclosure risks for users with multiple social profiles. In *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017* (pp. 163-170).

Benedetto, G., Stinson, M., & Abowd, J. M. (2013). *The creation and use of the SIPP Synthetic Beta.* U.S. Census technical report. Retrieved from https://www.census.gov/content/dam/Census/programs-surveys/sipp/methodology/SSBdescribe_nontechnical.pdf

Beretvas, S. N. (2011). Cross-classified and multiple-membership models. In J. J. Hox and J. K. Roberts (Eds.) *Handbook of Advanced Multilevel Analysis*, pp. 313-334. European Association of Methodology.

Bonnéry, D., Feng, Y., Henneberger, A. K., Johnson, T. L., Rose, B., Shaw, T. V., Stapleton, L. M., Woolley, M. E., & Zheng, Y. (2019, online). The Promise and Limitations of Synthetic Data as a Strategy to Expand Access to State-level Multi-agency Longitudinal Data. *Journal of Research on Educational Effectiveness, 12*(4), 616-647. https://doi.org/10.1080/19345747.2019.1631421

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification and Regression Trees (CART).* Wadsworth.

Drechsler, J. (2009, December). *Synthetic datasets for the German IAB Establishment Panel*. Conference of European Statistics, United Nations Commission and Economic Commission for Europe, Bilbao, Spain.

Drechsler, J. (2011). *Synthetic Datasets for Statistical Disclosure Control — Theory and Implementation*. Springer.

Drechsler, J. (2012). New data dissemination approaches in old Europe — Synthetic datasets for a German establishment survey. *Journal of Applied Statistics*, *39*(2), 243-265. https://doi.org/10.1080/02664763.2011.584523

Drechsler, J. (2015). Multiple imputation of multilevel missing data — Rigor versus simplicity. *Journal of Educational and Behavioral Statistics*, *40*(1), 69-95. https://doi.org/10.3102/1076998614563393

Drechsler, J., & Reiter, J. P. (2011). An empirical evaluation of easily implemented, nonparametric methods for generating synthetic datasets. *Computational Statistics & Data Analysis, 55*(12), 3232-3243. https://doi.org/10.1016/j.csda.2011.06.006

Figlio, D., Karbownik, K., & Salvanes, K. (2017). The promise of administrative data in education research. *Education Finance and Policy*, *12*(2), 129-136. https://doi.org/10.1162/EDFP_a_00229

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2003). Hierarchical models. In *Bayesian data analysis* (pp. 120-160). Chapman Hall/CRC.

Harel, O., & Zhou, X. H. (2007). Multiple imputation: Review of theory, implementation and software. *Statistics in Medicine*, *26*(16), 3057-3077. https://doi.org/10.1002/sim.2787

Henneberger, A.K., Feng, Y., Johnson, T., Zheng, Y., Rose, B., Stapleton, L.M., Sweet, T., & Woolley, M.E. (2019). *Prevalence of and Statistical Approaches for Handling Multiple Membership in the Maryland Longitudinal Data System: A Technical Report.* Baltimore, MD: Maryland Longitudinal Data System Center.

Kinney, S. K., Reiter, J. P., Reznek, A. P., Miranda, J., Jarmin, R. S., & Abowd, J. M. (2011). Toward unrestricted public use business microdata: The synthetic Longitudinal Business Database. *International Statistical Review, 79*, 362-384. https://doi.org/10.1111/j.1751-5823.2011.00153.x

Lin, S., & Luo, W. (2019). A new multilevel CART algorithm for multilevel data with binary outcomes. *Multivariate Behavioral Research, 54*(4), 578-592. https://doi.org/10.1080/00273171.2018.1552555

Little, R. J. (1993). Statistical analysis of masked data. *Journal of Official Statistics*, *9*(2), 407-426.

Little R. J., & Rubin D. B. (1987). *Statistical Analysis with Missing Data*. Wiley.

Matthews, G. J., & Harel, O. (2011). Data confidentiality: A review of methods for statistical disclosure limitation and methods for assessing privacy. *Statistics Surveys*, *5*, 1-29. https://doi.org/ 10.1214/11-SS074

Matthews, G. J., Harel, O., & Aseltine, R. H. (2010). Assessing database privacy using the area under the receiver-operator characteristic curve. *Health Services and Outcomes Research Methodology*, *10*(1-2), 1-15. https://doi.org/10.1007/s10742-010-0061-3

Raab, G. M., Nowok, B. and Dibben, C. (2016) Practical data synthesis for large samples. *Journal of Privacy and Confidentiality*, *7*(3), 67–97. https://doi.org/10.29012/jpc.v7i3.407

Raghunathan, T. E., Lepkowski, J. M., Van Hoewyk, J., & Solenberger, P. (2001). A multivariate technique for multiply imputing missing values using a sequence of regression models. *Survey Methodology*, *27*(1), 85-96.

Raghunathan, T. E., Reiter, J. P., & Rubin, D. B. (2003). Multiple imputation for statistical disclosure limitation. *Journal of Official Statistics, 19*(1), 1-16.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods.* Sage Publications.

Reiter, J. P. (2005). Using CART to generate partially synthetic public use microdata. *Journal of Official Statistics, 21*(3), 441–462.

Reiter, J. P. (2009a). Using multiple imputation to integrate and disseminate confidential microdata, *International Statistical Review*, *77*(2), 179 - 195. https://doi.org/10.1111/j.1751-5823.2009.00083.x

Reiter, J. P. (2009b). Multiple imputation for disclosure limitation: Future research challenges. *Journal of Privacy and Confidentiality, 1*(2), 223-233.

Reiter, J. P., Oganian, A., & Karr, A. F. (2009). Verification servers: Enabling analysts to assess the quality of inferences from public use data. *Computational Statistics & Data Analysis*, *53*(4), 1475-1482. https://doi.org/10.1016/j.csda.2008.10.006

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

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. John Wiley & Sons.

Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods*, *7*(2), 147–177. https://doi.org/10.1037/1082-989X.7.2.147

Scottish Longitudinal Study. (SLS; 2019, February 14). *Scottish Longitudinal Study. Development and Support Unit*. Retrieved from https://sls.lscs.ac.uk/

Sela, R. J. & Simonoff, J. S. (2011). REEMtree: Regression Trees with Random Effects. R package version 0.90.3.

State Longitudinal Data Systems Grant Program. National Center for Education Statistics, Institute for Education Statistics. (2018a). *History of the SLDS grant program: Expanding states' capacity for data-driven decision making.* Retrieved from https://nces.ed.gov/programs/slds/pdf/History_of_the_SLDS_Grant_Program_May2018.pdf

State Longitudinal Data Systems Grant Program. National Center for Education Statistics, Institute for Education Statistics. (2018b). *Grant information*. Retrieved from https://nces.ed.gov/programs/slds/grant_information.asp

U. S. Census Bureau (2018). *Survey of Income and Program Participation: Synthetic SIPP Data. Retrieved from*: https://www.census.gov/programs-surveys/sipp/guidance/sipp-synthetic-beta-data-product.html

Van Buuren, S. (2007). Multiple imputation of discrete and continuous data by fully conditional specification. *Statistical Methods in Medical Research*, *16*(3), 219-242. https://doi.org/10.1177/0962280206074463