The Promise and Limitations of Synthetic Data as a Strategy to Expand Access to State-level Multi-agency Longitudinal Data

Daniel Bonnéry[a], Yi Feng[b], Angela K. Henneberger[c], Tessa L. Johnson[b], Mark Lachowicz[b], Bess A. Rose[c], Terry Shaw[c], Laura M. Stapleton[b], Michael E. Woolley[c], Yating Zheng[b].
Authors listed in alphabetical order by last name.

[a]Joint Program of Survey Methodology, University of Maryland, College Park
[b]Department of Human Development and Quantitative Methodology, University of Maryland, College Park
[c] School of Social Work, University of Maryland, Baltimore

Citation:

Bonnery, D., Feng, Y., **Henneberger, A.K.**, Johnson, T., Lachowicz, M., Rose, B.A., Shaw, T., Stapleton, L.M., Woolley, M.E, & Zheng, Y. (2019). The promise and limitations of synthetic data as a strategy to expand access to state-level multi-agency longitudinal data. *Journal for Research on Educational Effectiveness, 12*(4), 616-647. https://doi.org/10.1080/19345747.2019.1631421

Author's Note: The contents of this manuscript were developed under a grant from the Department of Education. However, those contents do not necessarily represent the policy of the Department of Education, and you should not assume endorsement by the Federal Government. Additionally, this research was supported by the Maryland Longitudinal Data System (MLDS) Center. We are grateful for the assistance provided by the MLDS Center. Prior versions of this manuscript were published by the MLDS Center. We appreciate the feedback received from the MLDS Center and its stakeholder partners. All opinions are the authors' and do not represent the opinion of the MLDS Center or its partner agencies.

Abstract

There is demand among policy-makers for the use of state education longitudinal data systems, yet laws and policies regulating data disclosure limit access to such data, and security concerns and risks remain high. Well-developed synthetic datasets that statistically mimic the relations among the variables in the data from which they were derived, but which contain no records that represent actual persons, present a viable solution to these laws, policies, concerns, and risks. We present a case study in the development of a synthetic data system and highlight potential applications of synthetic data. We begin with an overview of synthetic data, what it is, how it has been utilized thus far, and the potential benefits and concerns in its application to education data systems. We then describe our federally-funded project, proposing the steps required to synthesize a statewide longitudinal data system covering high school, postsecondary, and workforce data. Last, for use as a template for other agencies considering synthetic data, we review the challenges we have confronted in the development of our synthetic data system for research and policy evaluation purposes.

Administrative data collected by governments about individuals hold great potential to advance our knowledge of key public services, policies, and programs, including those that may have an impact on education and workforce outcomes. However, confidentiality laws and procedures to protect such data typically restrict access to that data to a very limited universe of government-employed (or in some cases government-appointed) researchers and policy makers. There are a number of strategies for expanding access to government data, each having strengths and weaknesses. A common example is provision of aggregated data, which is safe but has limited research potential. Examples of sources using such a data access strategy include the State of Texas, which has a website (http://www.txhighereddata.org/) where extensive data tables about education and workforce can be reviewed by citizens, however, these tables are aggregated across units. North Carolina also has a publicly-accessible website (http://www.dpi.state.nc.us/research/data/) where datasets and variable dictionaries can be accessed, however, those datasets are also aggregated.

Disseminating granular individual-level data to a wider, more diverse, group of analysts, scholars, evaluators, and policy researchers may leverage the potential of knowledge advancement toward a broader understanding of how these systems and structures impact our population over time; nevertheless, the fundamental responsibility of data agencies remains with the protection of individual privacy. One emerging solution to this problem of restricted access is synthetic data. Synthetic data are generated based on statistical models to mimic the relational patterns among variables within and across individuals, meaning that statistical analyses with such synthetic data should yield findings substantially similar to the “real” data from which it was modeled while simultaneously reducing the risk of privacy breach.

In this manuscript, we detail the promise and limitations we have encountered in our ongoing efforts to create a synthetic version of one statewide longitudinal data system for the very purpose of increasing access to these valuable data. The core aim of this Synthetic Data Project (SDP), funded by the United States Department of Education (USDOE) through the Institute for Education Sciences, is to generate three datasets, capturing six years of data each spanning from: 1) high school to the workforce, 2) high school to postsecondary education, and 3) postsecondary education to the workforce. We begin with an overview of our ongoing project, including the current problems with access to administrative data and the potential for synthetic data to address those problems, with a brief review of the synthetic data literature. We then detail the challenges we have confronted in implementation, from constructing the simplified datasets that are the blueprints for synthesization, to selecting the synthesis models to be used, to testing the research utility and safety of the synthetic data. Throughout these sections, we provide guidance for those involved in the creation of synthetic data or interested in using synthetic data to answer substantive research and policy questions. To that end, we address several issues that must be resolved during the creation of synthetic data to ensure end-user utility, data security, and research validity, and we devote the final section to a discussion of how synthetic data might be used strategically to answer questions of relevance to policy and program evaluations.

Background

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). Since 2005, the USDOE has supported 47 states, as well as the District of Columbia, Puerto Rico, the 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). This substantial investment provides the data necessary for assessments of program and service efficacy to inform practice and policy decisions. Statewide longitudinal data systems, and administrative data in general, provide a number of advantages to researchers as compared to traditional survey measures, including larger data sets, fewer problems with attrition, lower rates of non-response bias, and more data for rare populations (Card, Chetty, Feldstein, & Saez, 2010). Furthermore, SLDSs enable a relatively cost-effective approach to answering policy questions because they obviate the need for costly and time-consuming primary data collection.

The Maryland Longitudinal Data System (MLDS) is one example of a state longitudinal data system and is the impetus for the present study. The MLDS, and the Center that houses these data, began operations in 2013 after legislation was passed in 2010 to create the data system (Md. Code, Education Article, §24.701-24.707). The State law that established this new agency also called for state agencies to share data to build the longitudinal system, matching unit record-level data of Maryland students and workers from preschool, through primary and secondary education, to postsecondary education, and ultimately to the workforce. The purpose of the MLDS Center is to generate timely and accurate information about student performance and employment outcomes that can be used to improve the State's education system and guide decision makers. To accomplish this task, the MLDS Center links individual-level student and workforce data from three State agencies: 1) the Maryland State Department of Education (MSDE); 2) the Maryland Higher Education Commission (MHEC); and 3) the Maryland Department of Labor Licensing and Regulation (DLLR). The MLDS Center has an obligation to make data accessible to researchers, policy makers, and stakeholders.

Despite the advantages of statewide administrative data, and the obligation to make data available, state longitudinal data systems are limited in their ability to share data by a myriad of federal and state confidentiality laws, including the Family Educational Rights and Privacy Act (USDOE, 2018) of 1974 and protections by the United States Department of Labor when workforce records are included (Maryland Code, § 8-625(d) of the Labor & Employment Article). To comply with federal and state regulations and protect student and worker confidentiality, states typically limit access to a small number of government officials able to access de-identified data. When research access to de-identified data is permissible, it often requires researchers to submit to a lengthy screening process including a background check and an approval process for proposed analyses or a planned research agenda. A review of state policies confirms this: Mississippi and Washington require, for example, a Memorandum of Understanding or agreement between the researcher and any institution or state agency that provides data for the research. Florida warns applicants to expect a minimum of three months from the time a completed data request proposal is submitted to the receipt of the final approval decision. Idaho requires the applicant to submit the SQL code to extract the data, a process which illustrates the burden on the state to review compliance between the submitted SQL code and the applicant's data description and data needs (see SLDS State Profiles, n.d.). North Carolina limits access to state and local government officials who must first register with the North Carolina identity management system (NCID). In Maryland, only researchers affiliated or partnering with a Maryland institute of higher education may be granted access to the MLDS data, and they must submit a detailed proposal for review by MLDS Center staff, undergo background checks, and receive extensive training (MLDS Center, n.d.).

These limitations are problematic for a number of reasons. First, policy makers often need to make decisions quickly, necessitating a quick turnaround time for analyses to inform such decisions (Hedges, 2018). Another concern is that planned analyses must go through an approval process, potentially overseen by politically-appointed individuals posing a possible conflict (Figlio, 2017). Furthermore, in states such as Maryland that require researchers who do successfully complete the extensive approval requirements to conduct all work on virtual machines housed by the MLDS Center, the costs and administrative burden to the state can be quite high. To expand access to administrative data, some agencies use statistical disclosure control methods. Such methods maintain the original information in the raw datasets but protect against the disclosure of identities (e.g., award number R305D140045 from the National Center for Educational Research; IES, 2014). Examples of disclosure control methods include 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). The majority of these methods, however, still release some elements of the raw data, and would thus not be acceptable strategies for some government agencies.

An emerging strategy, and one that would not release original raw data of any individuals, has potential to allow much greater researcher access, capacity, and latitude in statistical methods. This strategy is the development of *synthetic data sets* from the data stored in the administrative data sets. Some agencies, such as the U.S. Census Bureau, have started using such *synthetic data* (see Drechsler, 2011, and Reiter, 2002). In this approach, the raw, confidential, data are used to produce artificial data that are similar to, but distinct from, the raw data. In this way, researchers have access to microdata that closely mimic the properties of the raw data which they can then analyze to answer a variety of important research questions that

cannot be addressed from mere summaries. Importantly, with the use of synthetic data, the agencies responsible for collecting and protecting data can be assured that the true data remain confidential and that individuals from whom data were collected are exposed to minimal risk. This process, in theory, thus allows confidentiality to be maintained while also giving both researchers and policy analysts access to individual-level data. An application process and dedicated server for registered users is still necessary to track the use of the synthetic data for evaluation purposes (Abowd & Lane, 2004).

Recognizing the potential of synthetic data systems, the MLDS Center, through a federal grant, launched the Synthetic Data Project (SDP) in 2016 to test the feasibility of using synthetic data to facilitate expanded access to the MLDS data. The proposed products of the SDP would allow opportunities to undertake research and policy analyses by individuals who are not MLDS Center staff while maintaining the security of all individuals in the data. With input from an end-user group, the SDP has been evaluating the feasibility of synthetic data in the real-life setting of an actual statewide data system. Specifically, the central aims of the SDP were to answer five overarching evaluation questions: 1) What challenges arise in the process of creating synthetic data from a statewide longitudinal data system? 2) What are the best methods for assessing the quality of the synthesized data? 3) How successfully do the synthesized data fulfill the needs of the MLDS Center to provide accessible data that can inform policy while protecting individual privacy? 4) What legal and political issues arise related to the development and dissemination of synthetic data? And 5) To what extent do end users (applied researchers) find the synthetic data useful, and to what extent are the data actually used in analyses that inform policy? The SDP is currently in year 3 of 4, so this paper reports on the first phases of the project including the creation of the synthetic data and the specific issues that arise in the creation of such data with

education and workforce datasets. We also provide less detailed anticipated indications about the other phases of the project. The next sections start with an overview of synthetic data, then review successful implementations of synthetic data systems in the United States and Europe.

An Overview of Synthetic Data

As a general overview, the raw, confidential data are used to produce imputed "synthetic" data that are statistically similar but not identical to the raw data (Abowd & Woodcock, 2001; Drechsler, 2012; Rubin, 1993). In this way, researchers have access to microdata, or unit record-level data, that closely mimic the properties of the raw data. Importantly, with the use of 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 low 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 analysts access to microdata, allowing for increased data utilization toward a wide range of data analyses.

Synthetic datasets can be produced through a process in which synthesis models are fit to the original data and new, "synthetic" values are drawn from the predictive distribution from the models (Gelman, Carlin, Stern, and Rubin, 2003; Raghunathan, Reiter, and Rubin, 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 values that do not correspond to real individuals.

There are various methods that can be used to generate synthetic data, all of which require some kind of strategy for modeling relations among variables in the raw data. Synthetic data generation is traditionally accomplished with sequential regression models. Variables are arranged, and therefore synthesized, in a certain order. For each variable, a regression model is developed against a selection of predictors among the preceding variables. The models are developed in a sequential manner until a model is developed for each variable in the data (Drechsler & Reiter, 2011; Raghunathan, Lepkowski, van Hoewyk, & Solenberger, 2001; Van Buuren, 2007). Synthetic data are thus generated sequentially from the posterior predictive distribution for each variable. 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, results from analyses run on the synthetic data replicate the inference results based on the raw data (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, Harel, & Aseltine, 2010; Reiter, 2005b; 2009a; 2009b). As Matthews and Harel (2011) concisely summarize, “synthetic data sets are only as good as the models used for imputation” (p. 10). A key step in any synthetic data project is to evaluate the quality of the synthetic data as will be discussed in this article.

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 (p. 230).

Another challenge in the 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 it even though they would if the comparable raw data were available (Reiter, 2005b). Additionally, although the synthetic data mirror the raw data, the two are not equivalent and researchers might overgeneralize their conclusions. As such, Drechsler (2015) suggested that the results from synthetic data may not be appropriate for publication in academic journals. Finally, good practice in the use 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.

As an alternative to reporting results from analysis of synthetic data, the synthetic data could be used to design and develop code for statistical analysis. This code could then be passed on to agency 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, Oganian, & Karr, 2009). This process is referred to as dual-release mode or the use of a verification server and has been used in some of the applications that we discuss next.

Applications of Synthetic Data

Synthetic data have been used as a strategy for access to a few government-collected data sets in the United States and Europe. For example, the U.S. Census has generated and disseminated synthetic versions of data from two of their programs. The Survey of Income and Program Participation (SIPP) data (see Benedetto, Stinson, & Abowd, 2013) have been merged with Social Security data about retirement and disability benefits received and Internal Revenue Service data about income. SIPP synthetic data sets are referred to as the SIPP Synthetic Beta

(SSB) and currently include nine SIPP panel waves from 1984 to 2008 (U.S. Census, 2018). Substantial testing prior to the release of these synthetic datasets established a “negligible” risk of reverse identifying any actual individuals in the synthetic versions of the SSB data. In the creation and release of these synthetic versions of merged panel data, the U.S. Census, in collaboration with Cornell University faculty, took a significant step forward in the development of the methods and procedures in the use of synthetic data as a strategy to expand access to administrative data. The Census Bureau has also created synthetic data for the Longitudinal Business Database (LBD; Kinney et al. 2011). Jarmin and colleagues (2014) describe the recent work to develop synthetic data at the Census Bureau and examples are available of non-Census Bureau researchers running their code on the SIPP synthetic data before finalizing their analyses to be run on the gold standard confidential microdata (Abowd, 2016). Inspired by these examples in use by the Census Bureau, the State of Maryland decided to investigate whether synthesizing their wealth of educational data would be feasible.

A range of other efforts to undertake a synthetic data process to address data disclosure concerns involve both administrative and survey data. In Germany, Jörg Drechsler (2009; a consultant on the SDP project) led an effort to create synthetic versions of the Institute for Employment Research IAB Establishment Panel, which, along with the SSB, was an early large-scale application of synthetic data to expand access to government data (Drechsler, 2012). This panel data set was initiated in 1993 and has been collected annually since 1996 and 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. The Scottish Longitudinal Study (SLS), one of the most ambitious state-created longitudinal data bases, includes a broad range of annually collected data beginning in 1991 about a randomly selected

5.3% of the population starting from birth records, through education, health data, marriages, maternity, pollution exposure, weather, and workforce, until death (SLS, 2019). However, these rich data were only accessible to a small number of approved researchers so the SLS created three 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 (not all the variables in the SLS) that do present disclosure risk. They offer these data for public download with two stated aims: 1) so researchers can gain familiarity with the SLS data prior to applying for access, and 2) for use in university or training settings. They do not recommend disseminating analyses with the synthetic data, rather, once analyses are developed researchers should apply for access and if approved come to Edinburgh and run their analyses on secure data terminals set up for that purpose. Additional interesting examples of efforts to synthesize specific variables or sections of surveys can be found with the American Community Survey (Rodriguez, Freiman, Reiter, & Laugman, 2018) and OnTheMap (Machanavajjhala, Kifer, Abowd, Gehrke, & Vilhuber, 2008).

Gold and Synthetic Data Creation Process

As part of our project, we have identified several required steps in the process of creating a synthetic data product to share with users. This work is ongoing, as we are starting the final year of the funded project. Briefly, this process entails three steps (two of which have been completed with the third remaining to be conducted in this next year): 1) creation of gold standard datasets (GSDS), 2) synthesization, and 3) evaluation of the utility and safety of the synthetic data sets (SDS). We provide an illustrative flowchart (see Figure 1), and the following section presents a brief discussion of each step.

Creation of the GSDS (Step 1)

The creation of the SDS is built upon the structure and definition of the GSDS. Assuming that an agency does not want to synthesize its entire database, the GSDS is a simplified version of the original data that contains the variables to be synthesized; synthetic datasets then serve as a mirror image of the GSDS. The GSDS is a well-defined subset of the available data housed in the MLDS. The generation of the GSDS follows an iterative process that involves investigation of the larger data system, consultation with both end users and technical experts, and working with stakeholders to identify the cohorts, select variables, and define composite variables. Because we create the synthetic data by running predictive models on the GSDS, the definition and structure of the GSDS are of critical importance as they directly dictate the definition and the structure of the synthetic data. In this section, we discuss in depth the steps we took to create the GSDS as well as some of the challenges confronted in this process.

Data study (Step 1.1). It is critical to understand the data structure and the characteristics of the variables housed in the larger data system to make an informed decision on cohort definition and variable selection. Therefore, as the first step in creating the GSDS, we conducted a systematic study of the data housed in the MLDS. For each variable, we studied the data coding by checking the consistency between the data dictionary and the values stored in the system, examined the descriptive statistics—especially regarding outliers, missing data patterns, and in some cases the pattern of "not applicable" for some variables (if differentiable from the missing data coding), and investigated the presence of redundant or overlapping information as we have multiple data sources. For example, both the postsecondary and the K-12 education data include SAT and ACT scores. Exploring these data issues played an important role in variable selection and information aggregation (e.g., we decided to keep the highest score for each person

on the same content area of SAT or ACT in the GSDS when the scores differed across data sources).

One of the identified challenges at this step was potential record linkage errors (record linkage is the matching of individuals across data sources; see Fellegi & Sunter, 1969). The linkage of longitudinal data from multiple sources via identification information, such as the name, birthdate, student identification number and/or Social Security Number (SSN), can result in outliers such as 12th graders earning $100,000 in a quarter (in such cases the parent's SSN may have been recorded for the student in the K-12 data system). When encountering apparent errors of this type, consultation with database and content experts may help to confirm or reject evidence of record linkage errors, which then allows us to determine whether or not to reproduce or eliminate these extreme values from the GSDS. In brief, one should recognize that in a large longitudinal data system like MLDS, the issue of record linkage errors adds another layer of difficulty to the data study, as well as all the subsequent steps that we discuss below.

Evaluation of existing research questions (Step 1.2). While investigating data elements in the larger data system, we evaluated the research analyses that have used the data housed in the MLDS data system along with the current research agenda of the Center. For example, studies have evaluated the effect of dual enrollment on college attendance and performance (Henneberger, Witzen, & Preston, 2018), the impact of a state financial aid program on college persistence (Witzen, 2018), and the longitudinal impact of school-level and individual-level poverty on students' academic outcomes and employment outcomes (Henneberger, Rose, Mushonga, Nam, & Preston, 2019).

End user input about research questions and methods (Step 1.3). Having gathered detailed information about the data and the existing questions on the MLDS Center's research

agenda, we proceeded to assess the end users' needs. At this step, we convened a group of institutional researchers, scholars in the areas of education and workforce outcomes, and policy analysts from different sectors in the state of Maryland to present them with non-confidential, simplified versions of the data tables. We asked these potential synthetic data end users about their research interests and the analytic methods they would use if given access to datasets of a similar type. Some of the important feedback we received was that they were, in general, interested in conducting analyses on longitudinal panel data, or trajectories, covering a wide range of topics, including students' attendance, academic performance, financial aid, and employment conditions.

Definition of cohort and variables (Step 1.4). Based on the end user feedback, the overarching goal at this step was to generate three sets of GSDS that correspond to different trajectories, respectively: high school to postsecondary education, high school to workforce, and postsecondary to workforce. To accomplish this task, we defined the cohorts of students and selected variables to be included in the gold standard data as described in the following sections.

Cohort definition. 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 GSDS, 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 college freshmen in academic year 2010-2011 and followed this cohort for six years until academic year 2015-2016. We defined the 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 academic year 2015-2016, both in the high-school-to-postsecondary trajectory and in the high-school-to-workforce trajectory.

Variable selection. As for variable selection, the primary concern was 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 (e.g., keeping the highest score of SAT or ACT scores 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 as part of 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 and a clearly 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 or 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 earnings received. The data included in each GSDS are detailed in the Appendix.

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. As discussed in the next section, the synthesis model requires inputting a data table in wide format, with one row per individual. However, for the tables stored in the system, data are usually stored in the long format, and it is not uncommon for individuals to have multiple records on the same variable, especially over an extended period of time. One example would be the employment data, in which each row corresponds to an

employment record for a specific individual in particular quarter. 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, many populated with sparse data (i.e., 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. Similar decisions on aggregating information and reducing the level of granularity were made for other variables, such as multiple attendance records or multiple financial aid awards received by a student within the same school year, or as previously noted, multiple test score records on the same subject.

It is important to note here that all the steps we took to define and create the GSDS were completed under two anticipated constraints: 1) practicality constraints (i.e. we tried to avoid having an end product that is too complicated for users to understand and to use), and 2) legal constraints (i.e., disclosure risks; we have 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 (e.g. education, policy, psychology, sociology, economics, 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 (e.g., a well-documented data dictionary, a codebook, technical reports). Based on such practicality and user utility considerations, we decided to create a simple and straightforward data structure.

Decision point: Pass stakeholder review (Step 1.5). At this step, we presented the cohort definitions, list of variables, and simplified data structure to the major stakeholders within the MLDS Center. Although we discuss this decision in Step 1, it is important to keep in mind

that the creation of the GSDS is an iterative process, and hence this step should be repeated throughout the entire course of creating the GSDS, from cohort definition to variable selection and information aggregation.

Synthesization (Step 2)

In this section, we describe the synthesization procedure used in order to provide a synthetic version of the GSDS that satisfies a triangular trade-off between low disclosure risk, preservation of unconditional distributions, and preservation of multivariate conditional distributions. First, we justify our choice not to impute missing data and outliers. Second, we describe the technical and methodological issues germane to multi-dimensional datasets, specifically regarding the transformation of a multi-dimensional database into a single wide-format dataset. Third, we justify our choice of using the classification and regression tree (CART) procedure for synthesis, and describe how we pre-selected predictors and chose the order of variables in consideration of the triangular trade off.

Missing data and outliers. Whenever providing microdata to end users, choices have to be made about imputing missing data and correcting outliers before synthesization. Imputing all missing values and recoding outliers allows end users not to worry about imputation and data cleaning, and it decreases the number of variables to synthesize (missing data indicators do not have to be created for continuous variables). Through consultation with technical advisors for the project, we opted instead to model the missing values and outliers rather than imputing/recoding because end users may prefer to use their own imputation and outlier correction method. Furthermore, if end users wish for their code to be run on the GSDS data containing missing values and outliers, end users' code needs to have been created to address those issues.

Dealing with a multi-dimensional database. Unlike a procedure that aims to synthesize the outcome of a survey that usually consists of a "rectangular" dataset with a certain number of variables for each sampled unit such as that found with Census' SIPP or LBD, synthesizing the MLDS statewide educational longitudinal data is challenging because it is multidimensional. In database management terms, the information stored in the system corresponds to "facts" that will be determined by one or more "dimensions." For example, attendance is a Boolean variable depending on four "dimensions" (a student, a school, an academic year, and an enrollment period). In the design of the MLDS data warehouse, the adopted data model is optimal for loading the data and ease of access for dashboard developers and MLDS researchers but not necessarily outside end-users. The information in this relational database is organized with respect to a schema that describes all the tables, their relationships, and their dimensions. For example, if Student 1 attended School A in 2011, this "fact" is transcribed as a line in a specific table that contains student identifier, school identifier, year, and enrollment period variables.

To accommodate synthesization, all datasets in this multidimensional relational database needed to be transposed, or converted from "long" to "wide" format, such that the available information could be stored in a rectangular dataset with one and only one row per individual in the database. The outcome of the transposition and merging of a multi-dimensional dataset is a rectangular dataset with limited number of rows (in our case, approximately 30,000 and 60,000 for the postsecondary and high school cohorts, respectively) and a high number of variables. A back-transposing and merging procedure is then used to create the SDS with the same structure as the GSDS. For example, the transposition process completed on the postsecondary-to-workforce data, which originally contained fewer than 100 variables, led to the creation of 6,533 variables.

Addressing the longitudinal nature of the data. A particular challenge of synthesizing data from the MLDS is its longitudinal aspect. With longitudinal data, time is not simply another database dimension, but time also implies that the data structure needs to adapt to the information transmitted by the agencies over the period of data collection. Over time, this information can change in nature and in format. For instance, new dependencies between variables may arise, and "old" variables may lose their predictive power to "new" ones. Legal restrictions may appear or disappear, and these restrictions may dictate what data the center is obliged or denied to store, and for which period. Formats of the input data and data definitions change over time. These are merely a few examples of the hurdles that exist for the team in charge of loading the data for the MLDS, and these hurdles add complexity to studying and preparing the data for synthesization. One must anticipate the future of the data when defining the GSDS, which unfortunately can lead to subjective decisions when selecting predictors.

Integrated data. Most statewide longitudinal data systems involve record linkage, wherein unit-level records from disparate data systems are matched, either deterministically or probabilistically (Fellegi & Sunter, 1969). In theory, it may be possible to take into account the variability of the record linkage procedure that the MLDS Center is using and reproduce this random process in the synthesis procedure; this would involve going back to the files sent by the partner agencies. We chose instead to synthesize the MLDS Center record linkage output, which streamlines the process by not requiring us to work with original files sent by partner agencies and creates some degree of parity between the products of the SDP and the MLDS Center.

Choice of a method for synthesization. When it comes to synthesization, various methods have been developed. The scope of the project is to assess the feasibility of using existing methods in the particular setup of multidimensional, longitudinal, integrated data with a

large amount of information. For this reason, after initial testing and evaluation of the different existing methods, the decision was made to implement the CART method (described in Reiter, 2005b). According to Reiter (2005b), a CART is the outcome of a general empirical method to model a dependent variable conditionally to a set of predicting variables. It consists in a partition of the joint predictor space obtained after applying a binary partition recursively. The binary partition consists of finding the best split, e.g. identifying the predictor variable and threshold that will split the dataset in two sub-datasets (nodes) for which the within-node dispersion of the dependent variable is minimal. The process is repeated in the resulting two sub-datasets until no potential split results in a significant between-node dispersion.

This method does not require model specification for each variable other than predictor pre-selection or forcing, which streamlines our project, as we dealt with a large number of variables. Figure 2 depicts a simple example of a single regression tree, where posterior predictive distributions of individuals' term grade point average (GPA) in the 2nd term of 2015 was determined conditional on SAT math score and credits earned that term. Within each leaf, we would then sample from the empirical distribution to obtain a synthesized value of GPA. Note that two terminal nodes, or leafs, contain potentially problematic issues. Node 2 has extreme homogeneity in the GPA distribution and node 6 contains just 17 observations; these two leaves would be subject to investigation to determine potential disclosure. Another approach would have applied a parametric model; in such models it is common to "normalize" continuous variables and to fit normal linear regression models to generate data. However, each project should seek to use the model that best fits their data. In practice, most of the variables we are using are categorical, with some ordinal or count, thus complicating the parametric approach. Another practical reason for the choice of the CART approach is that it is available in the

synthpop R package (see Nowok, Raab, & Dibben, 2016). This method can be applied to rectangular datasets, where all the available information about the characteristics and experiences of one individual can be stored in one row of variables. When the number of variables is large, a consequence of this method is a potential divergence, when the variable index increases, between the distribution of a synthetic variable and the empirical distribution of the same variable in the GSDS, which can reduce the utility of the SDS. Other non-parametric synthesization methods that have been proposed are based on random forests (Caiola & Reiter 2010), or on support vector machines (Drechsler, 2010). Drechsler and Reiter (2011) compared non-parametric methods and argued that CART offers ease of application and outperforms random forests. In the online supplemental material, we present a technical description of CART, of the synthesis method, and details about potential divergence of distributions.

Predictor pre-selection and order of variables. Following the selection of an appropriate synthesis method, choices must be made regarding predictor selection and the order in which variables will be synthesized, just as in multiple imputation procedures (Little & Rubin, 1987). In general, important variables for research and variables with strong predictive power should be synthesized first to be preserve conditional relations. Ideally, predictor selection would depend only on theoretical and substantive rationale for each synthesized variable; however, automated model pre-selection is also necessary due of the high number of variables—the CART procedure fails in a reasonable amount of time when too many predictors are used. Automated model pre-selection can also incorporate researcher knowledge by forcing predictors into the model or removing predictors when it seems necessary to preserve desired conditional distributions or to eliminate attribute disclosure risks. In the following section, we describe methods used to evaluate both the final synthesized product as well as the GSDS on which it is

based. As noted previously, in wide format, the rectangular data contain more than 5,000 variables. This large number of variables creates huge challenges for CART to process properly in a reasonable time frame. Therefore, we created a set of rules to screen predictors, such that at most 300 predictors were pre-selected for each variable. The variables excluded are generally the variables with high missing rates or the variables that do not bear much additional information conditioning on other selected predictors. The predictor pre-selection is based on content knowledge as well as common sense. It reflects the choices we have made regarding which relations are more important and thus desirable to preserve in the synthetic datasets.

Evaluation (Step 3)

Evaluation of synthetic data comprises two primary steps: 1) utility assessment, and 2) disclosure risk assessment. In generating synthetic datasets, ideally researchers would strive for complete data utility (e.g., one-to-one correspondence between the synthetic data and the original data) and simultaneously zero disclosure risk (e.g., the inability of any intruder to identify individuals or sensitive attributes from the data). However, realistically, creators of synthetic data must grapple with a tradeoff between these two concepts, minimizing disclosure risk as they maximize data utility, all while maintaining fidelity to rules and regulations of any oversight agencies. Disclosure risk and data utility have been shown to be largely dependent on the synthesis models themselves (Little, 1993; Reiter, 2005a). Procedures for balancing risk and utility in the context of public use microdata have been discussed (Drechsler & Reiter, 2009; Gomatam, Reiter, Karr, & Sanil, 2005; Woo, Reiter, Oganian, & Karr, 2009), and we describe the methods employed within our SD project below.

Utility assessment (Step 3.1 & 3.2). Utility can be operationalized in different ways (Drechsler, & Reiter, 2009; Karr, Kohnen, Oganian, Reiter, & Sanil, 2006). One important

indicator of utility, which we have chosen to refer to as *inferential utility*, relates to the validity of the statistical inferences derived from the synthetic datasets; necessarily, this validity is an attribute of the synthetic data. However, given the risk-utility tradeoff, it would be impossible to perfectly reproduce all of the dependencies within the original dataset without inadvertently disclosing sensitive attributes or revealing individuals within the dataset. Considering the compromises that synthetic dataset developers must undertake to ensure security and confidentiality of the data, a second indicator of data utility, which we might term *research utility*, relates to the breadth and depth of testable hypotheses that are answerable by the synthetic data. This second component is both an attribute of the synthetic data (e.g., whether the appropriate conditional distributions have been incorporated in the synthetization model) as well as the gold standard dataset on which the synthetic data are based (e.g., whether the necessary variables, the scaling of those variables, and subpopulations are present in the dataset to conduct a particular analysis). As such, we proceed with a description of the process of assessing the utility of the data (3.1 in Figure 1) by examining first the GSDS and then the synthetic datasets.

GSDS utility assessment (Step 3.1). In assessing the research utility of the GSDS, we can examine the variables selected for the dataset in terms of: 1) scope of information, 2) quality of information, and 3) population definition. Scope of information refers to the density and diversity of the set of variables present in the dataset; in other words, the variables selected to represent each of the data system's content areas should cover the topic in enough detail to allow researchers to sufficiently address content-relevant questions without encountering redundancies or discrepancies within the dataset. Quality of information, on the other hand, refers to the completeness and consistency in measurement and recording of an individual variable or set of variables. Last, we must also consider the population represented by the data. Population

definition, which is an aspect of the *rows* rather than the *variables* of the dataset, comprises the specification of cohorts, or groups of individuals tracked over time, and directly impacts study design and internal/external validity. Evaluating these three aspects of data utility occurs iteratively throughout the GSDS creation and synthesization process. As expressed in Step 1, it entails conducting research "needs assessments" with potential internal and external end users of the dataset, identifying a comprehensive list of variables satisfying the identified research needs of target users, and consulting content and database experts on the definitions and documentation of each selected variable as well as the stability of those definitions over time. Creating a systematic process for identifying and reviewing criteria is crucial to determining the utility of the end product. This process may depend on the agency's aims, and the stakeholders may prioritize certain aspects of utility over others.

Synthetic data research utility assessment (Step 3.2). Following the satisfactory evaluation of the utility of the GSDS, the inferential and research utility of the synthetic data must be assessed as well (see Raghunathan, Reiter, & Rubin, 2003; Reiter, 2003). The bulk of the SDS evaluation involves investigating divergence, wherein, as we described above, we examine how closely the analytic results of the synthetic data align with the GS data, both in terms of relations between and among variables and in terms of univariate distributions (see Step 2 above and the online supplement for more detail on divergence). If one concludes in Step 3.1 that the GSDS possesses adequate research utility, then the SDS will also maintain the same level of research utility, assuming the two datasets do not diverge. However, divergence between the SDS and the GSDS will almost certainly occur for at least one of several reasons. First, incorporating all of the conditional distributions of the original data into the synthetic data may be infeasible or undesirable due to the dimensionality of the GSDS or due to security demands.

Furthermore, the analytic reproducibility of the SDS as compared to the GSDS may depend on the method and assumptions of the synthesis model itself as well as the order in which the variables are synthesized (Caiola & Reiter, 2010; Dandekar, Zen, Bressan, 2018; Drechsler & Reiter, 2011; Reiter, 2005a; Woo et al., 2009). Finally, while increasing the number of generated datasets from the synthesis model may improve the inferential validity of the SDS, doing so may also increase the chances of a data security breach (that is, the precision of the estimates increases with the number of generated synthetic datasets; Drechsler & Reiter, 2009; Reiter, 2005a). With these concerns in mind, we then seek to measure the distance between the two datasets to determine whether or not the SDS sufficiently reproduces the results of the GSDS without compromising feasibility or security constraints.

Several sophisticated methods for evaluation of the inferential utility of synthetic data have been proposed and empirically assessed in the literature, such as confidence interval overlap, Kullback-Leibler divergence, and use of propensity scores (Karr et al., 2006; Woo et al., 2009). Illustrations of inferential utility assessments, however, may also rely on more descriptive measures of discrepancy, such as percent differences and bivariate plots (Kinney et al., 2011). One preferred method involves taking the average distance between a list of fixed, pre-selected population parameters derived from the GSDS and their corresponding estimates from the SDS. This method has the advantage of yielding easily comprehensible values that are specific to each parameter assessed. A complementary way to measure this distance is to choose a research question, identify a statistical analysis to address this question, run the statistical model on both the SDS and the GSDS, and compare the main inferences from the two studies. Because the resulting parameter estimates will depend on both the statistical model and the subpopulation in question, the discrepancies of the SDS may be more pronounced for some analyses and

subgroups over others. One last simple, yet effective, method to assess univariate divergence involves constructing plots of the synthesized variable against the original variable; much like a quantile-quantile plot, deviations from a 45-degree angle will indicate that the synthesized variable departs from the original variable.

Although quantitative divergence measures may tempt developers to rely on hard and fast criteria for rejection or acceptance of the SDS, divergence may vary for different variables or clusters of variables as well as for different subpopulations. Developers of the SIPP Synthetic Beta and the synthetic LBD indicated in recent utility evaluations that analytic reproducibility was expected to perform well for low-dimensional analyses conducted over large groups, but that more granular analyses might require verification on the original GSDS (Benedetto, Stinson, & Abowd, 2013; Kinney et al., 2011). With that in mind, we argue that any quantitative indicators of inferential validity must be carefully evaluated in the context of the research goals of data end users. Once the inferential and research utility of the SDS has been found suitable, the dataset must be evaluated in terms of its disclosure risk, which is presented below.

To illustrate components of utility assessment, we use a subset of the postsecondary-to-workforce GSDS and three SDSs. Note that our plan is to release 30 SDSs for each GSDS; for illustration, we only show results for three SDSs here. The sample size of this cohort was 51,863 students. We illustrate specific utility by calculating the standardized difference between the estimates of interest based on the GSDS and for each SDS as $SD = \frac{\beta_{SDS} - \beta_{GSDS}}{SE_{GSDS}}$.

We also calculate the measure of confidence interval overlap for each estimate (IO; Karr, Kohnen, Organian, Reiter, & Sanil, 2006) as

$$IO = .5\left\{\frac{\min(UCL_{SDS}, UCL_{GSDS}) - \max(LCL_{SDS}, LCL_{GSDS})}{UCL_{GSDS} - LCL_{GSDS}} + \frac{\min(UCL_{SDS}, UCL_{GSDS}) - \max(LCL_{SDS}, LCL_{GSDS})}{UCL_{SDS} - LCL_{SDS}}\right\}$$

where UCL_{SDS} and LCL_{SDS} represent, respectively, the average upper and lower confidence limits for the replicated estimates based on the SDSs and where UCL_{GSDS} and LCL_{GSDS} are the confidence limits for the estimate based on the GSDS. Note that when the two confidence intervals do not overlap, the further they are away from each other the more negative the *IO* estimate will become. Ideally, standardized differences would be close to zero and confidence interval overlap would be close to one, where, ideally, intervals for SDS estimates would be slightly larger than and completely overlap the GSDS intervals (Abowd et al., 2006). Specifically, we compared the results of several multiple linear regression models in which log-transformed 2016 wage was regressed on log-transformed 2015 wage, 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 6 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 are found in Table 1 and illustrated in Figure 3. The highest utility was demonstrated by the unique association between 2016 wages and Variable 5, with an average standardized mean difference of 0.358 and an average confidence interval overlap of 0.914. The lowest utility was seen with Variable 1, with average standardized mean differences of 7.572, and average confidence interval overlap of -0.152. Our task, over this next year, is to undertake a wide range of similar analyses on a variety of models and variables, documenting specific utility, and tweaking our synthesis model to address problematic levels of divergence. Of course, given that this work is pioneering, the definition of "problematic" will also be iteratively assessed as our project continues.

Disclosure risk assessment (Step 3.3, Figure 1). Data disclosure risks are widely considered to be one of the major barriers to publication of administrative unit record-level data (Kinney et al., 2011; Weinberg, Abowd, Steel, Zayatz, & Rowland, 2007). Two primary types of disclosure risk challenge projects of this nature: 1) *identification disclosure*, and 2) *attribute disclosure*. The first type, identification disclosure, relates to the potential for an intruder to match a given record with a specific individual. The second type of risk, attribute disclosure, refers to the possibility that even aggregate data collected from these systems have the potential to disclose aspects of different subpopulations that may be sensitive in nature. Synthesization mitigates both types of risks; however, even with full synthesization, in which none of the data represent "real" individuals, the risk of identification and attribute disclosures is important to assess (McClure & Reiter, 2012; Reiter, Wang, & Zhang, 2014). As such, one of the crucial tasks facing our SD project relates to the assessment of the disclosure risks of the datasets produced. In this last year of our project, we consider two primary questions: 1) which method or methods will we use to minimize disclosure risk, and 2) both how will we quantify risk and what is an acceptable level of risk?

The first question relates to the selection and implementation of synthesis methods for minimizing disclosure risk. As noted, synthesization has the potential to mitigate risk; however, different synthesis procedures may, in fact, exacerbate disclosure risks. We thus distinguish between synthesization-based disclosure risk methods and post-hoc methods. One synthesization-based procedure to detect and prevent early instances of identity and attribute disclosure risks, associated with CART, involves the identification of all of the instances in which every record at a specific level of an explanatory variable takes the same value of the outcome variable. Consider the conditional distribution of a categorical variable on a set of one

or more categorical variables. During synthesization, we are able to examine the branch points at which individuals are so similar that all are synthesized to the same condition—for example, we could imagine that we want to synthesize graduation status (diploma vs. no diploma) based on an individual's sex (male vs. female). It is nearly unimaginable that all members of one sex across the state would carry the same graduation status. However, as we add further granularity, extreme group homogeneity becomes more probable. If we also condition on race/ethnicity, year, and a high school characteristic (e.g., size), for those attending schools categorized as small, it is not inconceivable that all members of a particular sex and race/ethnic group during a particular academic year could share the same graduation status. Thus, it follows that increasing the number of conditioning variables also increases the disclosure risk. By identifying these cases of attribute disclosure during synthesization, we can prevent them by reducing the granularity in the synthesis process. Finally, we should also note that although many disclosures are not acceptable (e.g., all Hispanic male students in a particular year from schools of a particular size failed to graduate), some disclosures may either be general enough to be acceptable or may already be public record (e.g., all diplomas obtained in a given geographic region were obtained during a particular term). These types of disclosures and the risks they present must be carefully considered and documented throughout all stages of the synthesization process.

Once we have evaluated risk prior to and iteratively throughout synthesization, we can then move on to more formal assessments of risk involving the final synthetic datasets. Several procedures for the quantification of disclosure risks for both partially synthetic and fully synthetic data have been proposed in the literature (Drechsler & Reiter, 2008; Gomatam et al., 2005; Hu, Reiter, & Wang, 2014; Reiter, 2002; Reiter & Mitra, 2009; Reiter et al., 2014). The procedures described in these papers quantify the identity disclosure risk, and are particularly

relevant in the case of partially synthetic data, where individuals in the synthetic dataset correspond to individuals in the GSDS. General descriptions of disclosure risk limitations is also given in Abowd and Woodcock (2001). Given that we use full synthesization, we cannot evaluate identity risk, only attribute disclosure risk. Thus, we apply the MLDSC reporting standards (see MLDSC Data Reporting Standards, MLDSC, 2015) to all the leaves defined by the CART procedure and drop predictors until the leaves comply with reporting policy (e.g., cell sizes cannot be lower than 10). Drechsler (2011) discusses disclosure risk for fully synthetic datasets, and describes disclosure risk assessment for the synthetic data on the German IAB Establishment panel. The discussion focuses on the risk coming from continuous variables with heavy tails: synthesis procedures based on sampling from the distributions in the CART leafs may reveal the exact value corresponding to an individual, provided that a "malicious user" has prior knowledge of the range of this variable for this individual and that he is the only one in that range. In our project, the variables with potentially heavy tails are wages and financial aid amounts. Special attention will be given to these two variables, and data coarsening may be applied to prevent this kind of disclosure. For disclosure risk measurement on fully synthetic data, see also Hu, Reiter, and Wang (2014).

Decision point: Governing Board approval (Step 3.4). Throughout this section on developing the GSDS and the SDS, we have discussed the importance of obtaining ongoing input from stakeholders. As the final stage in the evaluation process for this project, the Governing Board must approve the output of synthesization (Step 3.4 of Figure 1), both in broad terms—approval of the data release—and in specific—approval of the particular mode of release. Presenting a strong case to the Board requires the creation and evaluation processes adequately address the key concerns of the MLDS Center, and because these priorities may be

unique to a given organization, repeated consultation with the Board is necessary to identify the list of criteria to be met for approval. However, at minimum, GSDS and SDS developers must be prepared to build a reasonable argument for both the security and utility of the synthetic data, as described in the evaluation section above.

Although the Evaluation section detailed the process of balancing utility with risk, a few security issues still warrant discussion. Foremost, the release of the SDS could provide potentially useful information on the GSDS structure to an intruder, and even simple precautions may limit the ability of that intruder to compromise the security of the data. For example, currently enforced regulations for our Center forbid researchers from disclosing table names, variable names, or relations among data sources of the database. Demonstrating that review by the data team and security experts have resulted in sufficient name alteration may be a necessary first step in achieving data release approval by the Board.

Furthermore, mode of release may also pose security as well as cost concerns, which GSDS and SDS developers must carefully consider. Building and maintaining a secure platform within the MLDSC to service the preferred dual-mode data release will require ongoing resources and approval by the Governing Board. Although desirable, the dual-mode of remote access and synthetic access methods may open additional avenues for attribute disclosure: first, the synthetic data are released to the public to allow external researchers to build statistical models, and, second, internal staff are allocated to replicate these models on the GSDS and to return suppression reviewed results back to the original researchers (Abowd & Lane, 2004; Benedetto et al., 2013; Weinberg et al., 2007). Maintaining the ongoing security of the GSDS remains of utmost importance; however, at this stage in the methodological development of synthetic data, allowing researchers to verify their results on the original data is necessary to

enable the use of synthetic data with confidence in high-stakes settings, including providing recommendations to policy makers. As such, the final approved mode of data release has implications for how useful synthetic data will be, as we discuss in the final section.

Cost and benefit assessments. As an increasing number of state data repositories apply synthetic data strategies, 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 available to successfully complete this project through a grant from the Institute of Educational Sciences through the Maryland State Department of Education ($2.6 million across four years) are not an accurate reflection of the costs of applying synthetic data strategies going forward. The resources needed to continue to provide synthetic versions of MLDS data will be considerably lower going forward now that the infrastructure and procedures have been developed in the MLDS Center. Further, as more data analysts and data base engineers work on such projects and mentor students and early career practitioners (as this project has) 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; upon request, we 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 the system and to revise code for iterative data analyses to a dual-mode data release system.

Implications and Use of Synthetic Data Files for Program Evaluation

Prior sections discussed many considerations to be addressed in the construction of a synthetic data system, from creation of the GSDS, through identifying the best synthesis model, to final evaluation of synthetic data for security and research utility. Although these processes may be painstaking and lengthy, the potential of the dissemination of a synthetic dataset based on state-level longitudinal education to workforce data can be significant. Expanding access to such a unique and rich source of data will advance the ability of a range of researchers, with perhaps different perspectives, to answer important, high-impact policy and evaluation questions. In this final section, which we organize by type of evaluation analysis, we discuss several examples of how synthetic data might be used in evaluations as well as highlight the potential problems in using synthetic data to pursue evaluations of education interventions and policy.

First, while the gold standard in education intervention and policy evaluation may be the randomized trial, administrative data will rarely contain randomization information from such designs. Furthermore, natural randomized experiments often occur in educational settings, such as with a lottery to enter a charter school, yet, the information regarding the lottery assignment is rarely provided in centralized administrative data systems, making these designs difficult to study with administrative data. Similarly, data at the classroom level, such as experimental curricula, are rarely provided in centralized longitudinal data systems; classroom-level data likely contain grade level and teacher identifiers at best. However, a synthetic system could aid in some of the difficulties associated with linking data from randomized trials and natural experiments to administrative data systems. Conducting randomized trials is expensive and time consuming, and often data collection ceases when funding for the trial ends (3-5 years in a typical federally-funded trial). However, administrative data linkage is a cost effective

mechanism for examining the long-term outcomes of students in the intervention and counterfactual. Yet, many state agencies do not have the capacity to link the data and conduct the long-term evaluation. A synthetic system could serve as a solution whereby researchers interested in long-term follow up from a randomized trial leverage the system to create the code for data linkage and extract the needed variables for evaluation. That code could then be provided to the state agency to do the actual data linkage and provide the evaluation results back to the researcher. Doing so would help the education science community better understand the long-term outcomes and potential fade out associated with educational interventions.

A second consideration is whether the synthetic data can support analyses from natural experiments and quasi-experimental designs, namely regression discontinuity, interrupted time series, and observational studies with a wealth of background data that allow for conditioning on covariates (Murnane & Willett, 2010). While the data contained in the MLDS, and most statewide educational data systems, may support interrupted time series (ITS) designs, such as implementation of a new high school graduation mathematics requirement and its effects on college entrance, it is questionable whether a synthetic data system would support such analyses. In our case, for feasibility, we synthesized one cohort only. A synthetic system that would meet the requirements of an ITS design would require synthesization of multiple cohorts, before and after the policy change, along with nuanced treatment of those individuals who span multiple cohorts (such as repeaters); unfortunately, this type of synthesization has not been conducted and the methods have not been fully developed.

However, in terms of analyses for regression discontinuity designs (RDD), the potential is brighter for such studies with synthetic data systems. Per the What Works Clearinghouse (WWC) standards (Institute of Education Sciences, 2017), RDD analysis requires that studies

contain five components: 1) demonstrating the integrity of the forcing variable, 2) low sample attrition, 3) continuity of the relation between the forcing variable and the outcome, 4) an evaluation of bandwidth and model form, and 5) additional requirements for fuzzy RDDs. In theory, synthesized data can be used for such analyses, given that the original forcing variable and treatment assignment were synthesized and all variables important in the modeling of the functional form between forcing variable and outcome variable(s) were also synthesized.

Whether selection into treatment occurs at an individual level or at a cluster level (e.g., school), however, may impact the feasibility of using synthetic data. For data at the cluster level, in order to prevent data disclosure, variables likely will have been coarsened. As an example, if a school might have been eligible for additional funds based on percentage of FARMs students, it is possible that the percent FARMs variable would have been coarsened. In other words, the continuous percent measure may have been coarsened into categories of low, moderate, and high, in order to avoid attribute disclosure risk. On the other hand, when treatment is at the individual level, values on a potential forcing variable are less likely to be coarsened unless they are at outlying values. In addition to having data on the forcing variable, the RD analysis would require information regarding treatment assignment, as well as the outcome variable(s). An example RD analysis, conducted by the MLDS Center, examined the impact of the Howard P. Rawlings Educational Assistance (EA) grant in Maryland (Witzen, 2018). In this analysis, the effect of receiving EA Grant funds, which is determined using adjusted gross family income as reported on the FAFSA, on retention rate and the receipt of other financial funding (from the institution and from loans) was investigated. Although the GSDS and the SDS contain information for the forcing variable and one of the outcomes of interest (retention), information about the treatment assignment and other financial funding are not available because information

about sources of financial aid were aggregated into categories of merit-based, need-based, and loans. This aggregation was used to limit the number of variables in the table, given the dozens of types of grant awards. Had policy analysts been at the table and argued for the variable to be disaggregated for specific funding mechanisms, the policy analysis could have been accommodated. This example speaks to the necessity of having the appropriate stakeholders involved in the conversations to define the GSDS elements.

Finally, it seems that analyses that use observational data and that condition on covariates, either through multiple regressions, matching, or propensity score methods (Stuart, 2010), are the most likely evaluation analyses to be conducted on synthetic data. For such analyses, the WWC requires, among other things, demonstration of baseline equivalence on variables related to the outcome but exogenous to treatment to receive a rating of *meets standards with reservations* (Institute of Education Sciences, 2017). Given this, analyses utilizing synthetic data need access to a multitude of variables for conditioning and to show equivalence at baseline. A study conducted with the MLDS data that could be conducted on the synthetic data is a propensity score matching analysis of the effects of dual enrollment (Henneberger, Witzen, & Preston, 2018). All variables used in this dual enrollment study have been included in the GSDS and subsequently the SDS. Note that because the synthetic data will not contain student-to-school linkages, use of the synthetic data files for this analysis implies an assumption that the treatment assignment mechanism and the treatment effect are not conditional on specific school, rather they can be conditional on school characteristics that are synthesized, what Thoemmes and West (2011) would term a board inference space. Additional types of analyses, where treatment is assigned at the school level (or is a function of a school characteristic) might be effects of charter schools, effects of large vs. small schools, and so forth.

One concern voiced by Hedges regarding the use of administrative data for such quasi-experimental designs is that the available data are less nuanced than that which might be available in field trials (Figlio, 2017). However, as agencies begin to see the value of increasing access to administrative data, the elements that they are willing to share may increase.

We have discussed satisfying the data requirements of the various analysis types when using synthetic data, but have not considered the inferential validity of the use of synthetic data. All analyses using synthetic data would require the assumption that the method used to synthesize the data (in our case, CART) included the appropriate variables to predict the synthesized outcomes. This assumption is unverifiable in advance and can only be validated by applying the analysis code onto the GSDS. And thus, as Drechsler (2015) specified, it seems unwise for synthetic data on their own to be used for evaluations that lead to high-stake curricular or policy decisions. At best, they might be useful for hypothesis creation. This is not to say, however, that synthetic data would not have a role in policy evaluation. Synthetic data might best be viewed as a tool for analysis code writing, a training dataset for policy researchers to use to write their data code for merging, cleaning, imputation, and analysis that would then be applied to the GSDS via a dual-mode data release. This thoughtful process of writing code would actually lead to a needed transparency in educational evaluation settings (Hedges, 2018) as this code would be considered in the public domain.

Discussion

As interest in using administrative data in education grows with the proliferation of new sources of data, current efforts to examine disclosure risks and utility of administrative data is necessary. Our project, although not completed, suggests that the use of synthetic data may be a tool to allow for external researchers to gain easier access to this wealth of data. However, these

potential new users must have a seat at the table when defining the gold standard data files and be vocal regarding the necessary data elements, scale of those variables, and cohorts or populations to include. As Eric Hanushek noted in a recent panel hosted by the National Academy of Education, “the future of the United States depends on getting [research access to administrative data] right” (Figlio, 2017, p. 5).

As the first state agency to attempt to synthesize its longitudinal education data, we are cautiously optimistic that our pioneering work can help to shape further efforts going forward in producing data products that are accessible by many and contain the required nuance in the data. Increasing access to educational administrative data at the state level has an advantage as it would encourage comparative analyses across statewide systems, thus examining the generalizability of findings across states, a distinct benefit of the use of administrative data as considered by Hedges (2018) who argues for considering external validity just as crucial as internal validity. Furthermore, administrative data has been lauded for its ability to allow for research on subgroups that are rare (Figlio, 2017); again, these subgroups would need to be *a priori* defined for synthesization. Although Hedges makes the argument for having secured data centers for administrative data utilization (Figlio, 2017), our experience suggests that such centers may not solve the desire for fast turn-around research or broaden access to those with unique perspectives. Synthetic data represent a promising approach for increasing easy access to secure data while simultaneously protecting the confidentiality of individuals.

References

Abowd, J. (2016, September 15). *The challenge of scientific reproducibility and privacy protection for statistical agencies*. Paper presented at the meeting of the Census Scientific Advisory Committee, Suitland, MD. Retrieved from https://www2.census.gov/cac/sac/meetings/2016-09/2016-abowd.pdf.

Abowd, J. M. and Lane, J. I. (2004). New approaches to confidentiality protection: Synthetic data, remote access and research data centers. In J. Domingo-Ferrer and V. Torra, eds., *Privacy in Statistical Databases*, pp. 282–289. New York, NY: Springer-Verlag.

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. Amsterdam: North-Holland.

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. *Handbook of advanced multilevel analysis*, 313-334.

Caiola, G., & Reiter, J. P. (2010). Random forests for generating partially synthetic, categorical data. *Transactions on Data Privacy, 3,* 27-42.

Card, D., Chetty, R., Feldstein, M. & Emmanuel Saez, E. (2010). Expanding access to administrative data for research in the United States. Washington, DC: National Science Foundation White Paper No. 10–069.

Dandekar, A., Zen, R. A., & Bressan, S. (2018). *A comparative study of synthetic dataset generation techniques.* (Technical Report of The National University of Singapore, School of Computing). Retrieved from https://dl.comp.nus.edu.sg/jspui/bitstream/1900.100/7050/1/TRA6-18.pdf

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. (2010). Using support vector machines for generating synthetic datasets. In J. Domingo-Ferrer & E. Magkos (Eds.) *Privacy in Statistical Databases* (Lecture Notes in Computer Sciences 6344), Berlin: Springer, 148-161.

Drechsler, J. (2011). *Synthetic datasets for statistical disclosure control — Theory and implementation*. New York, NY: Springer.

Drechsler, J. (2012). New data dissemination approaches in old Europe — Synthetic datasets for a German establishment survey. *Journal of Applied Statistics, 39*, 243-265.

Drechsler, J. (2015). Multiple imputation of multilevel missing data — Rigor versus simplicity. *Journal of Educational and Behavioral Statistics, 40*, 69-95.

Drechsler, J., & Reiter, J. P. (2008). Accounting for intruder uncertainty due to sampling when estimating identification disclosure risks in partially synthetic data. In *International Conference on Privacy in Statistical Databases* (pp. 227-238). Springer, Berlin, Heidelberg.

Drechsler, J., & Reiter, J. P. (2009). Disclosure risk and data utility for partially synthetic data: An empirical study using the German IAB Establishment Survey. *Journal of Official Statistics, 25*, 589.

Drechsler, J., & Reiter, J. P. (2011). An empirical evaluation of easily implemented, nonparametric methods for generating synthetic datasets. *Computational Statistics & Data Analysis, 55*, 3232-3243.

Fellegi, I. P., & Sunter, A. B. (1969). A theory for record linkage. *Journal of the American Statistical Association*, *64*, 1183-1210.

Figlio, D. (2017). *Role of Administration and Survey Data in Education Research: Panel Summary*. Washington, DC: National Academy of Education.

Figlio, D., Karbownik, K., & Salvanes, K. (2017). The promise of administrative data in education research. *Education Finance and Policy*, *12*, 129-136.

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

Gomatam, S., Karr, A. F., Reiter, J. P., & Sanil, A. P. (2005). Data dissemination and disclosure limitation in a world without microdata: A risk-utility framework for remote access analysis servers. *Statistical Science*, *20*, 163-177.

Harel, O., & Zhou, X. H. (2007). Multiple imputation: Review of theory, implementation and software. *Statistics in Medicine*, *26*, 3057-3077.

Hedges, Larry V. (2018). Challenges in building usable knowledge in education. *Journal of Research on Educational Effectiveness*, *11*, 1-21.

Henneberger, A.K., Rose, B.A., Mushonga, D., Nam, B., & Preston, A. (2019). The long-term effects of school concentrated poverty on educational and career outcomes. Baltimore, MD: Maryland Longitudinal Data System Center.

Henneberger, A. K., Witzen, H., & Preston, A. (2018). What is the causal effect of dual enrollment on long-term college and workforce outcomes and do effects vary for under-represented students? Manuscript submitted for publication.

Hu, J., Reiter, J. P., Wang, Q. (2014). Disclosure risk evaluation for fully synthetic categorical data. In J. Domingo-Ferrer (Ed.), *International Conference on Privacy in Statistical Databases* (pp. 185-199). Switzerland: Springer.

Institute of Education Sciences (IES). (2014). State Longitudinal Data Systems Public-Use Project Feasibility Study. Retrieved from https://ies.ed.gov/funding/grantsearch/details.asp?ID=1479

Institute of Education Sciences. (2017). What works clearinghouse standards handbook (4th ed.). Retrieved from https://ies.ed.gov/ncee/wwc/Docs/referenceresources/wwc_standards_handbook_v4.pdf .

Jarmin, R. S., Louis, T. A., & Miranda, J. (2014). Expanding the role of synthetic data at the US Census Bureau. *Statistical Journal of the IAOS, 30*, 117-121.

Karr, A. F., Kohnen, C. N., Oganian, A., Reiter, J. P., & Sanil, A. P. (2006). A framework for evaluating the utility of data altered to protect confidentiality. *The American Statistician, 60*, 224-232.

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.

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

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

Machanavajjhala, A., Kifer, D., Abowd, J., Gehrke, J., & Vilhuber, L. (2008, April). Privacy: Theory meets practice on the map. In *Proceedings of the 2008 IEEE 24th International Conference on Data Engineering* (pp. 277-286). IEEE Computer Society.

Maryland Longitudinal Data System Center. (2015, April). Data Reporting Standards (Version 1.5). Retrieved from https://mldscenter.maryland.gov/egov/publications/DataReportingStandards_v1.5.pdf

Maryland Longitudinal Data System Center. (n.d.) Policies and Procedures for External Researcher and Grant Funded Projects. Retrieved from https://mldscenter.maryland.gov/egov/publications/ExternalResearch/MLDSCPoliciesandProceduresforExternalResearcherandGrantFundedProjects.pdf

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.

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*, 1-15.

McClure, D., & Reiter, J. P. (2012). Differential privacy and statistical disclosure risk measures: An investigation with binary synthetic data. *Transactions on Data Privacy, 5,* 535-552.

Murnane, R. J., & Willett, J. B. (2010). *Methods matter: Improving causal inference in educational and social science research.* New York, NY: Oxford University Press.

Nowok, B., Raab, G. M., & Dibben, C. (2016). synthpop: Bespoke creation of synthetic data in R. *Journal of Statistical Software, 74*(11), 1-26.

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,* 85-96.

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

Reiter, J. P. (2002). Satisfying disclosure restrictions with synthetic data sets. *Journal of Official Statistics, 18,* 531-543.

Reiter, J. P. (2003). Inference for partially synthetic, public use microdata sets. *Survey Methodology, 29,* 181-188.

Reiter, J. P. (2005a). Releasing multiply-imputed, synthetic public use microdata: An illustration and empirical study. *Journal of the Royal Statistical Society (Series A), 168,* 185-205.

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

Reiter, J. P. (2009a). Using multiple imputation to integrate and disseminate confidential microdata, *International Statistical Review*, *77,* 179 - 195.

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

Reiter, J. P., & Mitra, R. (2009). Estimating risks of identification disclosure in partially synthetic data. *Journal of Privacy and Confidentiality, 1*, 99-110.

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*, 1475-1482.

Reiter J. P., Raghunathan, E. (2007). The multiple adaptations of multiple imputation. *Journal of the American Statistical Society, 102,* 1462–1471.

Reiter, J. P., Wang, Q., & Zhang, B. E. (2014). Bayesian estimation of disclosure risks for multiply imputed, synthetic data. *Journal of Privacy and Confidentiality, 6,* 17-33.

Rodriguez, R. A., Freiman, M. H., Reiter, J. P., & Lauger, A. (2018, August). Preserving privacy in person-level data for the American Community Survey. Presentation at the *Joint Statistical Meetings*. https://www.census.gov/content/dam/Census/newsroom/press-kits/2018/jsm/jsm-presentation-person-level-acs.pdf

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

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

Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods, 7,* 147.

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

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

State Longitudinal Data Systems (SLDS) 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

State Longitudinal Data Systems (SLDS) State Profiles. (n.d.). Retrieved from http://slds.rhaskell.org/state-profiles

Thoemmes, F. J., & West, S. G. (2011). The use of propensity scores for nonrandomized designs with clustered data. *Multivariate Behavioral Research, 46,* 514-543.

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

U. S. Department of Education (USDOE). (2018). Family Educational Rights and Privacy Act (FERPA). Retrieved from https://www2.ed.gov/policy/gen/guid/fpco/ferpa/

Van Buuren, S. (2007). Multiple imputation of discrete and continuous data by fully conditional specification. *Statistical Methods in Medical Research, 16,* 219-242.

Weinberg, D. H., Abowd, J. M., Steel, P. M., Zayatz, L., & Rowland, S. K. (2007). Access Methods for United States Microdata. U. S. Census Bureau Center for Economic Studies, Paper No. CES-WP-07-25. Retrieved from https://ssrn.com/abstract=1015374. http://dx.doi.org/10.2139/ssrn.1015374

Witzen, H. (2018). The Effects of the Howard P. Rawlings Educational Assistance (EA) Grant in Maryland. Manuscript in preparation.

Woo, M-J., Reiter, J. P., Oganian, A., & Karr, A. F. (2009). Global measures of data utility for microdata masked for disclosure limitation. *Journal of Privacy and Confidentiality, 1,* 111-124.

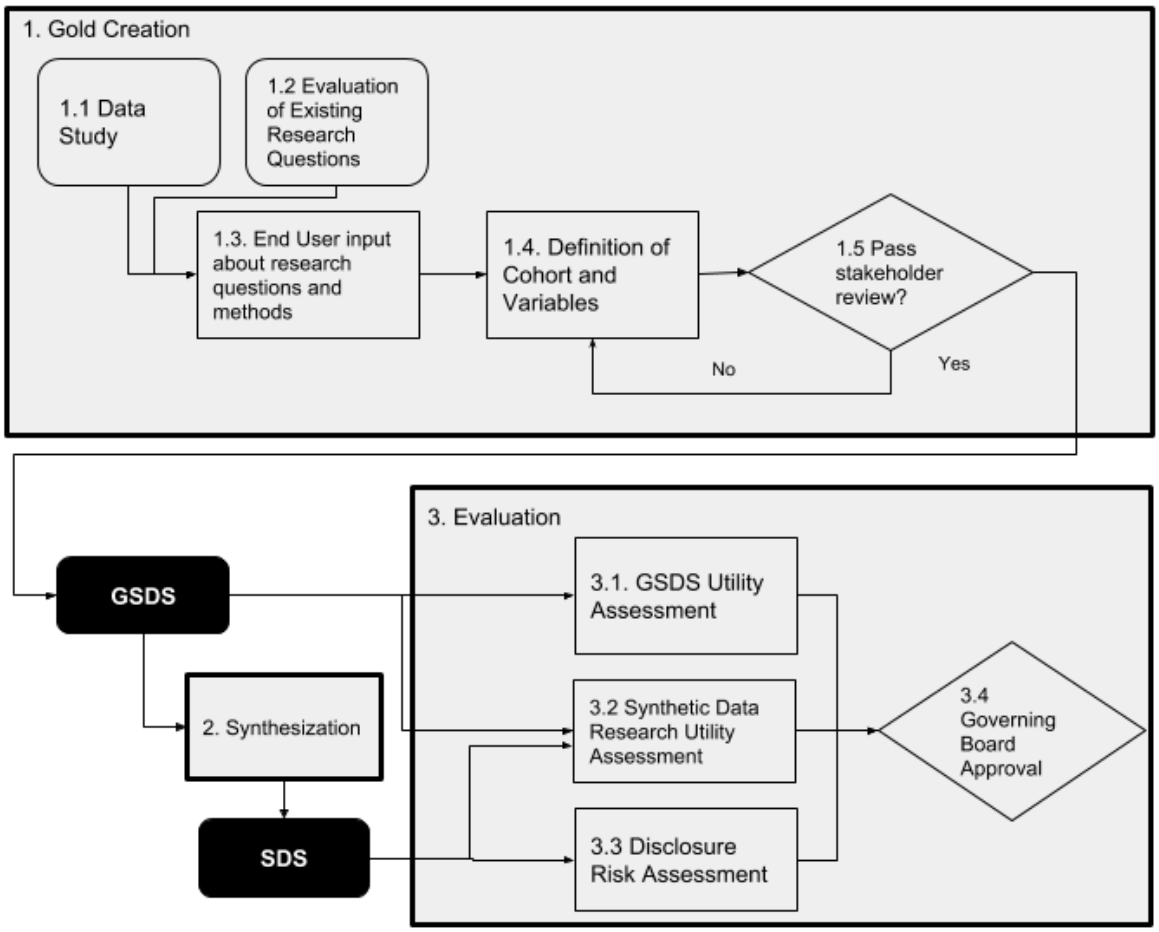

Figure 1. Gold Standard and Synthetic Dataset Creation Flowchart

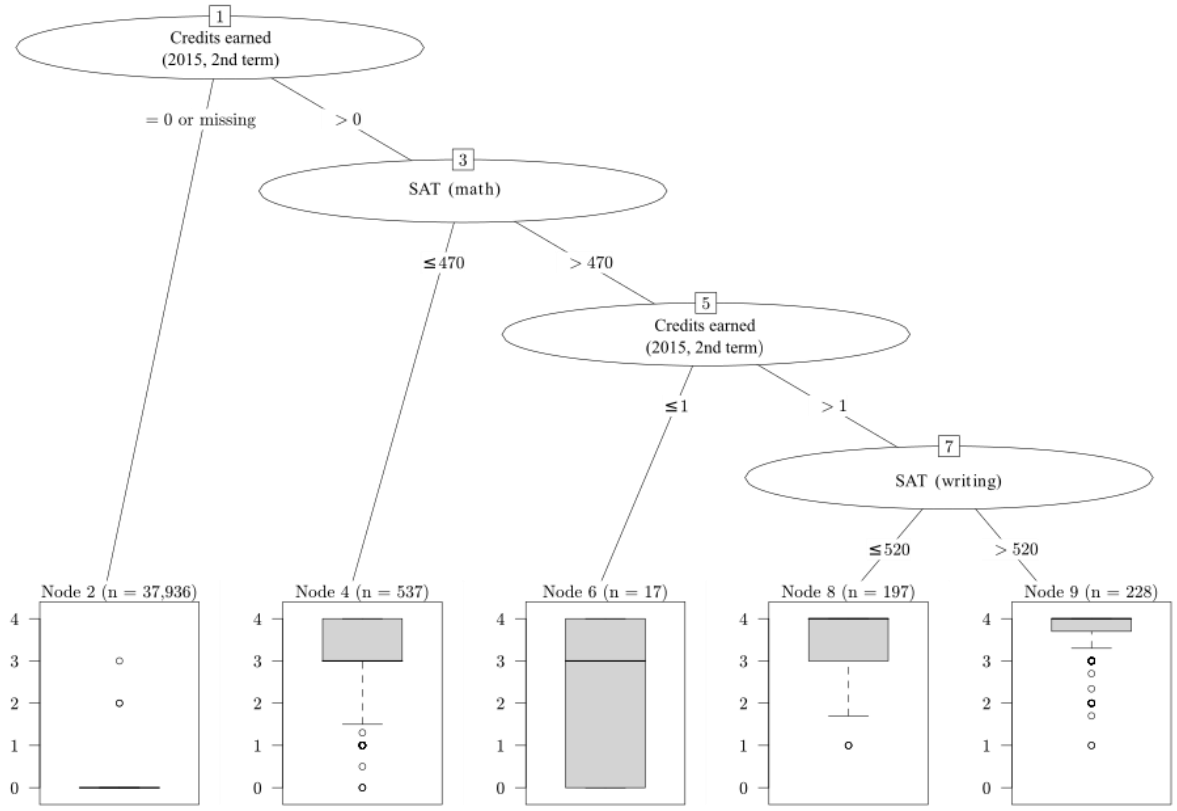

Figure 2. Simplified example of a classification tree applied to simulated term grade point average data. Model predictors are credits earned in a specific term and SAT math and writing scores.

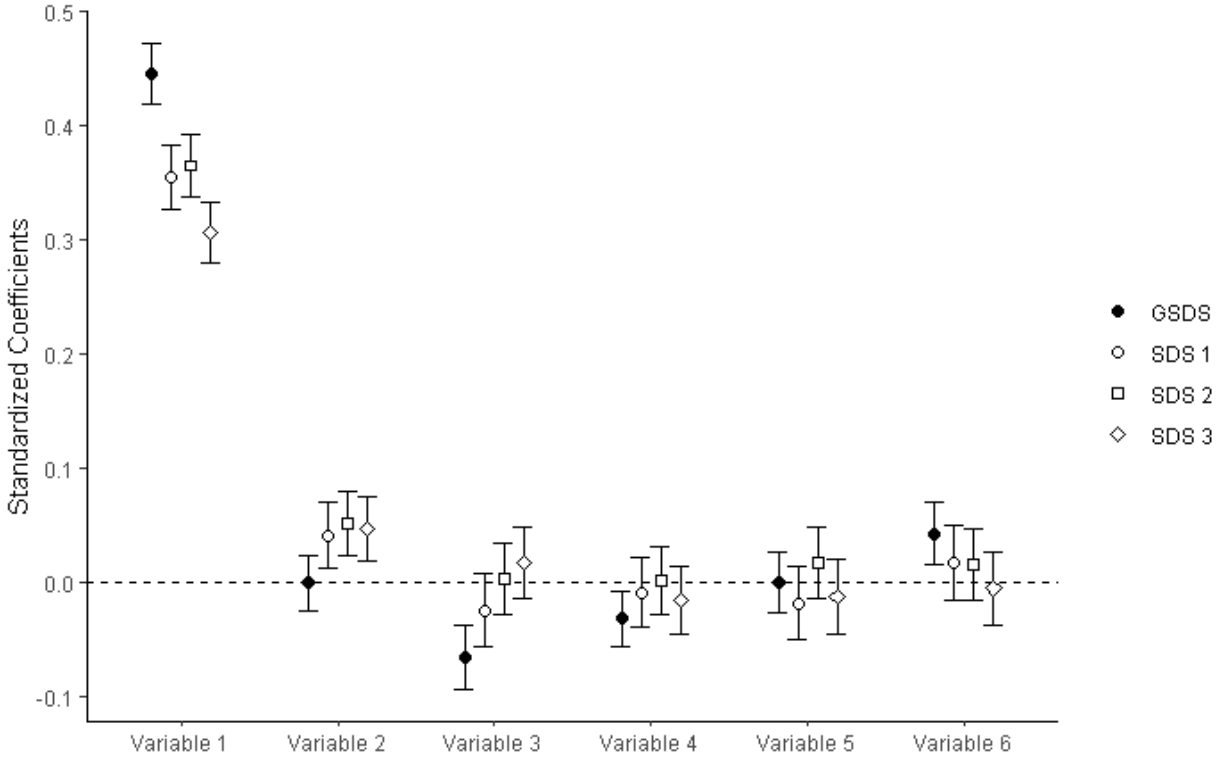

Figure 3.

Comparisons of standardized multiple regression coefficient estimates and confidence intervals from the gold standard and three synthetic datasets.

Table 1. Standardized coefficient and confidence interval comparisons of multiple regression analyses predicting 2016 wages conducted on the real and three synthetic datasets. Values of standardized coefficients, standardized differences, and confidence interval overlap presented are for the gold standard dataset and the averages across three synthetic datatsets.

Predictors	$\boldsymbol{\beta_{GSDS}}$ **(SE)**	$\boldsymbol{\overline{\beta}_{SDS}}$ **(SE)**	**SD**	**CI Overlap**
Variable 1	0.446 (0.014)	0.343 (0.033)	7.572	-0.152
Variable 2	0.001 (0.012)	0.047 (0.014)	3.823	0.107
Variable 3	-0.065 (0.014)	-0.001 (0.018)	4.526	-0.018
Variable 4	-0.031 (0.012)	-0.007 (0.015)	1.912	0.568
Variable 5	0.001(0.014)	-0.004 (0.015)	0.358	0.914
Variable 6	0.043 (0.014)	0.01 (0.016)	2.365	0.443

Note: GSDS=Gold standard dataset; SDS=Synthetic dataset; SE=Standard error; SD=Standardized difference; CI=Confidence interval

Appendix

Variables Included in the Gold Standard Datasets

Data tables	Trajectories			Descriptions	Information included
	HS → PS	HS → WF	PS → WF		
Demographic Information	✓	✓	✓	This data table contains the demographic information for each cohort member.	• Race • Gender • Ethnicity • Birth year and birth month
Assessments	✓	✓	✓	This table contains the standardized assessment scores for each cohort member. Scores on the same assessment reported by MSDE and MHEC are both included in the PSWF and HSPS data tables. HSWF data table only includes the assessment scores reported by MSDE. When	• College admissions exams (SAT and ACT) • Remedial assessment scores at college entrance

			there are multiple records for each person on the same assessment component reported by the same agency, only the maximum score for that specific component is kept.	
High School Achievements	✓	✓	This data table contains high school graduation information for the cohort members.	● Academic year ● Students' graduation status in that academic year (certificate of completion, HS diploma, or early college admission)
High School Completion Status	✓	✓	This data table contains high school completion status information for the cohort members. It provides information on the ways in which a high school student met a graduation or completion requirement by a Maryland public school.	● Academic year and grade level ● Students' high school completion status in that academic year (completed the requirement for The University System of Maryland (USM), completed the requirement for approved occupational program, completed the requirement for both USM and approved occupational program, other HS completions, non-completers)
High School Attendance	✓	✓	This data table contains the High school attendance records for the cohort members. There	● Academic year and grade level ● Numbers of days of attendance and absence within the academic year ● Entry and exit status

			can be multiple attendance record entries for the same person in the same academic year and at the same school. For each student, a limited number of attendance record entries per year are kept, prioritizing records associated with the greatest number of days attended.	● The length of attendance for each attendance record ● Indicator of promotion status ● Indicator of participation in the reduced meal program ● Indicator of homelessness ● Indicator of English language ● Indicator of receiving special education services
Postsecondary Attendance	✓	✓	This data table contains the enrollment information at public postsecondary institutions for cohort members. There can be multiple enrollment records for the same person within the same academic year in the same term. In that case, we only keep the first 2 attendance records for each person in the same term in the same year (they can be in different colleges), prioritizing the attendance records with most credit hours registered and completed, as well as in the academic terms with earliest starting date.	● Academic year and academic term ● The level of degree being sought ● The group name for the instructional program defined by the CIP code ● The total number of credit hours completed at the reporting institution as of the current term ● The number of credit hours the student registered during the current term that can be applied towards the degree completion ● The permanent legal residency for the student at the time of admission ● Student's GPA as of the current term earned in courses with credits applicable towards the degree

Postsecondary Achievements	✓	✓	This data table contains the achievement records for students earning 1-2 year certificates, associate's degrees, bachelor's degrees, or master's degrees since academic year 2010-2011. When there are multiple records of a student for the same type of degree within the same academic year, we only keep the first two records associated with the largest values of the number of credit hours required to complete the degree, the total number of credit hours, and the total number of native credit hours the student has earned for this degree.	• Academic year • The postsecondary degree the student earned • The group name for the instructional program defined by the CIP code • Cumulative GPA • Number of credit hours required to complete the degree • The total number of credit hours the student earned for this degree • The total number of native credit hours the student earned for this degree
Financial Aid	✓	✓	This data table contains the grants information for students those who have applied for and received PS funding from reported sources. When there are multiple records for the same person within the same academic year with the same type of	• Type of the grant/award (undergraduate grant, undergraduate loan, undergraduate scholarship, and undergraduate work-study) • The academic year the grant/award was received • The total award amount received by the student within the academic year for the same type of reward

			grant/award, we aggregate the award amount across records and only keep one entry in the table. These data are only to be used to evaluate effectiveness of financial aid programs.	
Employment/Earnings	✓	✓	This data table contains employment and wages information for cohort members who were employed by a non-federal Maryland employer starting from academic year 2010-2011 through 2015-2016. When there are multiple employment records for the same person within the same calendar year in the same quarter term in the same industry group as defined by the North American Industry Classification (NAIC) 2-digit code, we aggregate the wage amount across records and only keep one entry in the data table.	• Wage amount • Calendar year • Quarter term number • Two-digit NAIC group code

Note: HS = High School, PS = Postsecondary, WF = Workforce.