DATA GOVERNANCE AND PRIVACY

Expanding Data Access through Synthetic Datasets and Verification Metrics

The SEDSyn-23 Pilot for the Survey of Earned Doctorates, Final Report (ADC-SEDSyn-23-N04)

Claire McKay Bowen
Urban Institute

Jingchen Hu
Vassar College

Saki Kinney
RTI International

Jeremy Seeman
Urban Institute

Joshua Snoke
RAND

Aaron R. Williams
Urban Institute

URBAN INSTITUTE

RESEARCH REPORT

January 2026

ABOUT THE URBAN INSTITUTE

The Urban Institute is a nonprofit research organization founded on one simple idea: To improve lives and strengthen communities, we need practices and policies that work. For more than 50 years, that has been our charge. By equipping changemakers with evidence and solutions, together we can create a future where every person and community has the opportunity and power to thrive.

America's DataHub Consortium (ADC), a public-private partnership, implements research opportunities that support the strategic objectives of the National Center for Science and Engineering Statistics (NCSES) within the U.S. National Science Foundation (NSF). These results document research funded through ADC and are being shared to inform interested parties of ongoing activities and to encourage further discussion. Any opinions, findings, conclusions, or recommendations expressed above do not necessarily reflect the views of NCSES or NSF. Please send questions to ncsesweb@nsf.gov. NCSES has reviewed this product for unauthorized disclosure of confidential information and approved its release (NCSES-DRN26-015).

Contents

Acknowledgments v

Executive Summary vi

Expanding Data Access through Synthetic Datasets and Verification Metrics 7

The National Secure Data Service 7

Why Synthesize the Survey of Earned Doctorates? 9

Synthetic Data 11

Why Use Synthetic Datasets? 11

What Are Synthetic Datasets? 13

How Are Synthetic Datasets Generated? 14

What Are the Modeling Considerations? 15

How Are Synthetic Data Evaluated? 16

Utility 17

Disclosure Risks 18

Survey of Earned Doctorates Synthetic Data 20

Defining Use Cases through User Engagement 20

Gold Standard Dataset Creation 21

Synthetic Data Processing 22

Start Data 23

Synthesis Order 23

Models and Samplers 24

Software 24

Description of Utility Metrics for Evaluation 25

Missing Data Patterns 25

General Distributional Measures 25

Specific Analysis Measures 27

Assessing the Results of the Utility Evaluation 29

Description of the Disclosure Risk Metrics for Evaluation 30

Disclosure Risk Parameters for the Synthetic SED 30

Empirical Attack Models 31

Metric Configuration for Empirical Disclosure Risk 33

Verification Metrics 35

Definition of Verification Metrics and Verification Server 35

Candidate Verification Metrics for a Synthetic SED 36

Adaptations of Candidate Measures for Analyses of a Census 37
SED User Engagement and Feedback on Verification Metrics 38
Existing Models for Verification and Validation Servers 40
National Center for Health Statistic (NCHS) Public-Use Synthetic Linked Data 40
Cornell Synthetic Data Server 41
Statistics of Income (SOI) Division of the Internal Revenue Service (IRS) Prototype Server 41
Implementation Considerations When Building a Verification Server 42
Consider the Three-Way Trade-Off between Automation, Availability, and Privacy 42
Make Joint Decisions for Synthetic Data and Verification Servers 43
Future Research to Inform the Implementation of Verification Servers for the National Secure Data Service 44
Mitigating Privacy Loss in Releasing Verification Metrics 44
Verification Metrics for Small Subsets 44
Further Considering Potential Users of a Verification Server 44
Assessing Whether Verification Servers Provide the Right Tier of Access 45

Conclusion: Impact and Implications of Synthetic Data Generation 46
Time and Resources Investment 46
Empirical Disclosure Risk Evaluation Limitations and Common Practice Improvements 47
Synthesis Code Disclosure Review Requirements 48
Version Control in Secure Environments 48
Utility Metrics in the Broader Context 48
Shared Taxonomy for Improvement of Communication across Teams 49
Consideration of Balancing the Three-Way Trade-Off among Automation, Availability, and Privacy 49

References 52

About the Authors 55

Statement of Independence 57

Acknowledgments

This project was funded through the National Secure Data Service Demonstration (NSDS-D) project, which is operated by the National Center for Science and Engineering Statistics (NCSES) within the US National Science Foundation. We are grateful to them and to all our funders, who make it possible for Urban to advance its mission.

The views expressed are those of the authors and should not be attributed to the Urban Institute, its trustees, or its funders. Funders do not determine research findings or the insights and recommendations of Urban experts. Further information on the Urban Institute's funding principles is available at urban.org/fundingprinciples.

The authors thank the following individuals for their contributions to this project:

- Erin Doyle, Urban Institute
- Y. Patrick Hsieh, formerly at RTI International
- Kimberly Janda, RTI International

Executive Summary

Access to high-quality administrative and survey data is critical for enabling evidence-based policymaking across the government, academia, and the broader research community. However, significant barriers—such as restrictive eligibility criteria, geographic limitations, and lengthy approval processes—continue to hinder access for many potential users. These challenges disproportionately affect researchers and practitioners who lack proximity to secure data facilities or the resources to navigate complex application procedures.

To address these limitations, policymakers and researchers have advocated for a tiered data access model that bridges the gap between public-use files and secure data enclaves. One promising approach involves the creation of synthetic datasets that replicate key statistical properties of confidential data while minimizing disclosure risks. These synthetic files, when paired with verification metrics, can offer a viable alternative for preliminary analysis, educational use, and broader data exploration.

In alignment with this vision, the National Secure Data Service (NSDS) Demonstration, operated by the National Center for Science and Engineering Statistics (NCSES) within the US National Science Foundation, is a federal initiative authorized under Section 10375 of the CHIPS and Science Act of 2022 to inform a governmentwide effort on strengthening data linkage and data access infrastructure. Through a public–private partnership with America's DataHub Consortium (ADC), NSDS Demonstration projects facilitate research collaboration across the government, industry, and academia in topics including but not limited to data linkage, access, security, and privacy to inform the establishment of a future NSDS. This report details one such project with NCSES, ADC, and a collaborative team from the Urban Institute, RTI International, RAND, and Vassar College to develop a synthetic version of the Survey of Earned Doctorates (SED).

This pilot project explores the feasibility of generating microlevel synthetic data that supports a tiered access model and improves usability while maintaining privacy protections for the SED. Through continuous stakeholder engagement and rigorous evaluation of utility, disclosure risk, and verification metrics, this project (SEDSyn-23) aims to inform future data dissemination strategies and expand the value of SED data for a diverse user base.

Expanding Data Access through Synthetic Datasets and Verification Metrics

Data access is essential to informing evidence-based policymaking for government officials, policymakers, social science researchers, and data practitioners. Despite the clear value of administrative and survey data, several barriers limit access for many potential users. These include eligibility requirements (e.g., US-based researchers), geographic proximity to a federal statistical research data center (often concentrated in large cities with major universities or institutions), and lengthy approval processes for certain datasets (e.g., processing of legal agreements). As a result, a substantial portion of potential data users may never gain access to the data needed to support data-informed decision-making.

Recognizing these challenges, many public policymakers and researchers have proposed creating multiple tiers of access between the two extremes of public-use files and secure data enclaves (Advisory Committee on Data for Evidence Building 2022). One proposed tier involves developing public-use files through synthetic data generation that preserve key statistical properties of the underlying data (e.g., means, variances, and covariances). These synthetic files could be accompanied by validation or verification metrics to help data users assess the accuracy and reliability of their analyses (see Verification Metrics section on page 34). However, these metrics should be carefully balanced against privacy or disclosure risk metrics as greater usability or utility of the data tends to increase the risk of disclosure.

The National Secure Data Service

The National Secure Data Service (NSDS) aims to expand and streamline access to administrative data across all levels of government to support evidence-based policymaking. In other words, a "future NSDS is envisioned as a government-wide set of shared services for data and evidence building."[1] The NSDS Demonstration (NSDS-D) projects "will inform whether an NSDS will be established in the future. The goal is to pilot potential services, technologies, techniques, and shared services that could

[1] "Vision for a Future NSDS,". National Center for Science and Engineering Statistics, accessed September 17, 2025, https://ncses.nsf.gov/initiatives/national-secure-data-service-demo/vision-for-a-future-nsds.

be used within a potential NSDS. The focus is on novel research collaborations, data linkage methodologies, and privacy preserving technologies and techniques."[2] Many NSDS-D activities are currently being implemented and tested through America's Data Hub Consortium (ADC). Established in 2021 by NCSES as a public-private partnership, the ADC aims to address key research questions through innovative solutions and a streamlined acquisition process. Given the alignment with NSDS-D goals, such as better understanding data linkage and privacy preserving technologies, the ADC is being leveraged to inform requirements for the demonstration project.[3]

As part of the NSDS-D, the project, "Creation of Synthetic Data and Development and Use of Verification Metrics (Survey of Earned Doctorates)," is supported under ADC.[4] In addition to producing synthetic datasets for potential tiered access to SED, the broader goals of the project are to expand the available open-source toolkit for developing and evaluating synthetic datasets, explore the use of synthetic data for evidence building, and test the use of verification metrics to validate estimates produced from synthetic data. As synthetic datasets may be released to the public with fewer restrictions if the synthesized records reduce the disclosure risk compared with the confidential records, they have potential to provide tiered access that would allow researchers to access previously inaccessible data. This supports the NSDS-D project goal of informing efforts to develop a shared services model in support of increasing evidence-building capacity.

This NSDS-D activity brought together a collaborative team from the Urban Institute, RTI International, RAND, and Vassar College to generate a synthetic version of the SED. The SEDSyn team developed and executed a pilot approach that involved (1) running stakeholder engagement to understand the needs of a potential synthetic SED userbase, (2) creating prototype synthetic data files using reproducible code architecture, and (3) developing appropriate utility, disclosure, and verification metrics that aligned with the specific context of the SED and user needs..

[2] "The National Secure Data Service Demonstration," National Center for Science and Engineering Statistics, accessed October 16, 2025, https://ncses.nsf.gov/initiatives/national-secure-data-service-demo

[3] "The NSDS-D Project and America's Data Hub," National Center for Science and Engineering Statistics, accessed September 17, 2025, https://ncses.nsf.gov/initiatives/national-secure-data-service-demo/nsds-adc.

[4] "Creation of Synthetic Data and Development and Use of Verification Metrics (Survey of Earned Doctorates)" America's Datahub Consortium, accessed October 16, 2025, https://www.americasdatahub.org/award-sedsyn/

Why Synthesize the Survey of Earned Doctorates?

The SED is an annual census of all individuals receiving a research doctorate from an accredited US institution in a given academic year. SED is sponsored by NCSES and by three other federal agencies: the National Institutes of Health, the US Department of Education, and the National Endowment for the Humanities. SED collects information on educational history, demographic characteristics, graduate funding sources and educational debt, and postgraduation plans. SED data are used to assess trends in doctoral education and degrees, characteristics, and career trajectories of the doctoral population. The SED is a census with a consistently high response rate (over 90 percent annually).[5] This situation creates unique challenges for synthesizing the SED compared with other sample surveys, such as the Annual Business Survey. Given that the probability of inclusion in the SED is almost 1, this impacts the way users may choose to analyze the data, which impacts which utility metrics are used to evaluate the synthesis. Likewise, careful consideration of disclosure risk is required because standard assumptions of sample surveys may not apply. Specifically, standard measures of identification risk may not apply to a census, and the data do not benefit from any privacy amplifications due to sampling. Synthetic data generation is a widely used method for enabling data access while protecting respondent privacy. While there are a variety of synthetic data generation methods with different privacy and utility properties, generated synthetic microdata ultimately aim to reproduce key statistical properties of the original data while minimizing disclosure risk. A large amount of literature of statistical approaches exists on generating synthetic data, starting with Rubin (1993) and including seminal work such as Raghunathan, Reiter, and Rubin (2003), Reiter (2003, 2005), National Center for Health Statistics (2025), and others. Several federal statistical agencies have implemented synthetic datasets in practice (e.g., Benedetto, Stanley, and Totty 2018; Bowen et al. 2022; Kinney, Reiter, and Miranda 2014), and the field has expanded into domains like health care, where machine learning and computational advances have enabled large-scale synthesis.

The high response rate, small and easily identifiable population, and sensitive questions make it risky to share SED data without proper data governance and statistical disclosure control. Due to privacy concerns, SED data dissemination has, for much of its history, consisted primarily of published tables. In more recent years, the number of tables available to researchers has expanded through online table generators, which currently include the NCSES Table Builder[6] and SED Restricted Data

[5] "Is the Study Participation Voluntary?" SED FAQ, Survey of Earned Doctorates, accessed August 29, 2025, https://sedsurvey.org/Home/FAQS.

[6] "Survey of Earned Doctorates Table Builder," National Center for Science and Engineering Statistics, accessed September 17, 2025, https://ncsesdata.nsf.gov/builder/sed.

Analysis System.[7] Despite this expanded data access, these systems come with limitations, such as restrictions on the granularity of crosstabulations produced by the Table Builder. Also, microdata is only available to eligible researchers (i.e., individuals employed at US based institution or organization) who submit a proposal through the Standard Application Portal[8] that meet with access criteria (e.g., applicant justification that their research or policy questions cannot be informed using publicly available data products).[9] This application process can take several months and once granted access, all output must undergo NCSES disclosure review.

Currently, researchers cannot develop analyses while waiting for access to the confidential microdata nor can students learn about using the data. Synthetic data offer a promising solution by approximating the statistical properties of a confidential dataset while preserving privacy and making the data more accessible and usable. This NSDS-D project SEDSyn-23 seeks to provide a roadmap for how agencies can strike a balance between access and disclosure risk to expand the uses and value of datasets, using SED as a case study.

[7] "Restricted Data Analysis Tool for the Survey of Earned Doctorates," National Center for Science and Engineering Statistics, accessed September 17, 2025, https://ncsesdata.nsf.gov/rdas/.

[8] "The Standard Application Process," National Center for Science and Engineering Statistics, accessed December 12, 2025, https://ncses.nsf.gov/initiatives/standard-application-process.

[9] "Restricted Use Data Licensing," National Center for Science and Engineering Statistics, accessed September 17, 2025, https://ncses.nsf.gov/licensing.

Synthetic Data

The goal of SEDSyn-23 is to "explore the creation of a synthetic data file, demonstrate examples of uses of synthetic data for evidence-building, and test the use of verification metrics in validating estimates produced from synthetic data."[10] This section outlines the background of synthetic data, the process of synthetic data generation, and the key considerations for evaluating the synthetic data quality, utility, and disclosure risk.

Why Use Synthetic Datasets?

Data privacy and confidentiality are often managed through de-identification techniques, which assess whether specific statistics can be safely released (Hundepool et al. 2012). These methods typically involve procedural safeguards, such as removing personally identifiable information (PII; e.g., names, addresses) or suppressing statistics that describe small population groups (e.g., rural communities). Many privacy laws, such as the Family Educational Rights and Privacy Act (FERPA),[11] explicitly reference these traditional techniques, such as suppressing small cell counts. While useful, these methods have two major limitations:

1. **Traditional de-identification techniques can still leave confidential data vulnerable to privacy risks.** Even after applying de-identification techniques such as suppression, a data intruder or malicious actor may infer information about data subjects. In other words, removing PII fields is no longer sufficient to guarantee privacy protection (Ohm 2009).
2. **Traditional de-identification techniques are often not sufficient to expand access to critical datasets.** As stated previously, more sensitive or granular data that might support data users in making data-informed decisions are often inaccessible or otherwise difficult to access due to privacy concerns. Because of these concerns, de-identification techniques can be either overly restrictive in accessing useful data or provide insufficient disclosure risk protections (Dwork and Naor 2010).

[10] "Demonstration Projects," National Center for Science and Engineering Statistics, accessed September 19, 2025, https://ncses.nsf.gov/initiatives/national-secure-data-service-demo/projects.

[11] "Family Educational Rights and Privacy," Protecting Student Privacy, accessed December 29, 2025, https://studentprivacy.ed.gov/ferpa

These limitations highlight the need for alternative approaches to data access, ones that offer greater flexibility while maintaining strong privacy protections. Synthetic data address this need by offering more flexibility in addressing privacy concerns.

BOX 1

Data Access Ecosystem Terminology

- **Data subjects or participants:** persons described by the data
- **Data collectors:** persons participating in the creation of data sources
- **Data curators, maintainers, or stewards:** persons tasked with governing and disseminating data and data byproducts
- **Data privacy experts or researchers:** persons who specialize in developing data privacy and confidentiality methods
- **Data utility experts:** persons providing assistance in evaluating the utility of data pipelines
- **Data users and practitioners:** persons who use data, both for intended and unintended purposes

Data Privacy and Confidentiality Terminology

- **Confidential dataset:** a dataset that contains personal or sensitive information that is not publicly accessible
- **Synthetic dataset:** a dataset designed to imitate a confidential dataset while limiting information about individual records in the confidential dataset.
- **Privacy-enhancing technology (PET):** a computational or algorithmic process or method used to limit the unintended leakage of personal information in data processing tasks
- **Gold standard dataset (GSDS):** a dataset based on the confidential dataset, containing all relevant information in a format that meets the needs of data users for conducting analyses. It serves as the target dataset that PETs, such as synthetic data generation, aim to protect.
- **Disclosure risk:** the risks of unintentionally disclosing personal information about contributors to a confidential dataset by inferring information about them from published statistics or data
- **Data utility:** the ability for synthetic data to mimic the properties of a confidential dataset, either in general or in specific data processing tasks

It is important to note that all data access poses some privacy risks, regardless of how the data are released. A perfectly secure dataset is one that is not accessible, which cannot be used. Conversely, a perfectly useful dataset is one that is released without any privacy protection.[12]

What Are Synthetic Datasets?

A synthetic dataset is created to imitate a confidential dataset while limiting information about individual records. Synthetic data can refer to either the methodology used to produce synthetic datasets or the synthetic data output itself (Raghunathan 2021). In this context, imitating a confidential dataset means that analyses conducted on the synthetic version should yield results similar to those from the confidential data. The closer these outputs align, the higher the utility or usability of the synthetic data. At the same time, limiting information about individual records prevents the data adversary from inferring information about the data subjects in the confidential data.

Synthetic data specifically aim to mitigate disclosure risks or the possibility that sensitive information about data subjects could be inferred from published statistics or datasets derived from the confidential data. Using synthetic data offers two key advantages:

- **Greater disclosure risk protections.** As mentioned previously, even when suppressing or removing data fields containing PII, providing access to confidential data without further modification can still enable users to infer personal information about individuals in the confidential dataset (Dwork et al. 2017). Synthetic data involve randomization based on a statistical model to provide greater protection, making inferences about individuals much harder, thus limiting disclosure risks.
- **Expanded data accessibility.** Processes for getting access to confidential datasets can be burdensome. Synthetic data can be more easily shared than confidential data, often requiring only a modest data-use agreement instead of more extensive processes. This reduces the administrative burden for both data curators and data users.

[12] In this case, perfect usefulness is defined by the confidential data, acknowledging that there may be measurement or other errors in the confidential data not accounted for in this definition.

How Are Synthetic Datasets Generated?

Effective synthetic data deployment involves multiple steps, outlined in figure 1 on the following page.

- **Step 0. Defining use cases and thresholds.** Data curators work with data users to define use cases for synthetic data. This critical step ensures that synthetic data meets the needs of prospective data users, which may vary significantly. Some data users may only need a few key variables, such as demographic information or postgraduation outcomes. Other users may need more granular information, such as detailed course enrollment information for individual students. As part of this process, data curators can identify thresholds, or goals for the success of a particular synthetic data task. For example, program evaluators may want to produce statistics that differ by a relative error of no more than 10 percent from those based on the confidential data. Additional considerations include identifying key variables, subpopulations, and analyses of interest along with establishing accuracy targets. It is also important to clarify the intended use of the synthetic data. Will synthetic data serve as a substitute for confidential data in certain cases, function primarily as exploratory data for researchers preparing access applications, or be used mainly in educational settings?
- **Step 1. Preprocessing and creating gold standard data.** Data curators and privacy experts work together to preprocess the data. The goal of this stage is to design and create a gold-standard dataset (GSDS), or a dataset that contains information data users need to perform their analyses. Synthetic data aim to imitate one specific GSDS, but there are many possible ways to design and produce GSDSs.

In the next two steps, data privacy experts generate and evaluate the synthetic data.

- **Step 2. Privacy step: Synthesis and testing.** The data curators and privacy experts use statistical or algorithmic modeling techniques to generate synthetic data based on the GSDS (this is described in more detail in the next subsection).
- **Step 3. Postprocessing and evaluating.** The data curators and privacy experts evaluate the synthetic data to ensure it meets privacy protection and utility goals.

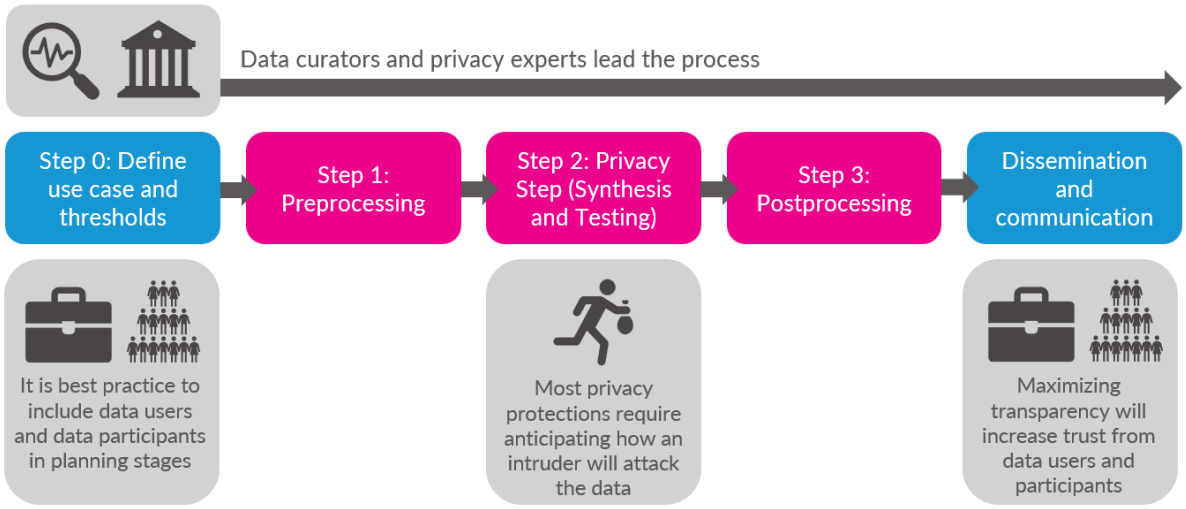
Steps 2 and 3 are often performed iteratively; data curators and privacy experts might evaluate multiple synthetic data configurations or try new synthetic data configurations motivated by evaluation results.

- **Step 4. Disseminating and communicating.** Synthetic data are disseminated to potential data users. In this step, it is critical to make sure the data users are well-informed about how the

synthetic data were generated, how the synthetic data were evaluated, and how the synthetic data should be used in practice. For example, if a synthetic dataset is sufficiently accurate for producing two-way contingency tables but not for complex modeling, data users need to know this information to responsibly use the synthetic data.

FIGURE 1

Data Curators and Privacy Experts Lead the Process



URBAN **INSTITUTE**

Source: Created by authors.

What Are the Modeling Considerations?

In general, synthetic data are generated in two stages: first, the synthetic data experts (which could be data curators, subject matter experts, data privacy experts, or a combination) build a statistical model to capture relationships between variables that appear in the confidential dataset. Next, those same experts randomly sample synthetic records or values from these models.

As stated previously, these two stages occur more than once. Many synthetic data methods, including the ones used by Urban Institute's open-source software (*tidysynthesis* and *syntheval*), use what's known as fully conditional synthesis. This type of synthesis means new variables (i.e., tabular data columns) are iteratively added one by one to construct synthetic data.

The effectiveness of synthetic data hinges on successful model building. The following considerations have the biggest impact on both disclosure risks and utility; see Drechsler and colleagues (2024) for more methodological details.

- **Synthesis inputs and ordering.** When using fully conditional synthesis, variables are synthesized sequentially, and the order in which they are synthesized can impact the quality of the synthetic data. For example, the synthetic data will yield more accurate results for models where the outcome variable has been synthesized after the explanatory variables. It is important to note that not all variables can be effectively modeled using the same predictor variables. For example, if a model for field of study relies on predictors that fail to capture statistically meaningful differences in field of study, the resulting synthetic data may be lower quality or have less utility. This situation is why it is important to synthesize variables in an order that captures increasingly complex relationships. Doing so improves the quality of variables synthesized later in the process.

Model complexity. There are countless different methods for modeling the relationships between variables. Generic model fitting methods (sometimes known as nonparametric, empirical or "black box" methods such as classification and regression tree [or CART] models, a kind of machine learning model) allow for the construction of more complex, flexible models than simpler methods (such as linear regression models). Although more complex models can produce higher utility synthetic data, they can also produce models that overfit to the confidential data. Overfitting occurs when model parameters reflect individual data points instead of broader statistical trends, which poses privacy concerns for producing tabular data. Tuning the underlying model complexity allows synthetic data experts to modify the degree of overfitting that may occur when generating synthetic data.

How Are Synthetic Data Evaluated?

Synthetic data should be evaluated to ensure they provide both sufficient utility and disclosure risk protection prior to being released to data users. Evaluating synthetic data helps data curators like NCSES communicate about how synthetic data should (or should not) be used and the privacy risks involved with making synthetic data more widely available.

Utility

Utility metrics, broadly speaking, are used to measure statistical aspects of synthetic data, with higher values corresponding to greater amounts of "utility" for the users of the synthetic data. There is no single measure of utility, since utility can be defined in many ways, depending on the intended use case. Utility metrics form one of the main sources of information when determining whether to release a synthetic dataset, For the sake of decisions about when to release synthetic data, it is imperative to be clear about how utility metrics are computed and what they measure.

This project assumes that the baseline case for the highest utility is a user who can work directly with the restricted, confidential GSDS (i.e., using the GSDS represents the maximal utility). It is important to note that the GSDS is a modified version of the restricted confidential file, which contains all the available confidential data collected in the SED; to produce the GSDS, some variables have been aggregated or suppressed from the original survey data. The results using the synthetic data are compared with *the GSDS*. This ensures that the confidential and synthetic datasets we are comparing have the same structure.

Utility metrics are commonly divided into two main classes: (1) general measures describe the structural or distributional similarity between the GSDS and the synthetic data, and (2) specific measures describe the similarity in statistics computed for specific analyses using the GSDS and the synthetic data (Snoke et al. 2018). The distinction between these two categories is not strict, since a specific analysis of interest might be a comparison of distributions (Snoke et al. 2018). Generally, though, it is a helpful distinction between defining utility based on whether the GSDS and synthetic data "generally look like each other" and defining utility based on whether a user would "obtain similar results" for an analysis using either dataset.

Specific utility measures can also be subdivided, depending on whether the metric (1) computes the distance between the estimated statistic (e.g., the difference in linear regression coefficients estimated using the GSDS versus the synthetic data) and (2) computes the difference in the resulting inference, (e.g., whether a hypothesis test for the significance of a linear regression coefficient has the same outcome on both datasets). Again, this distinction is primarily helpful for thinking about the meaning of the utility metric, whether we define utility as "how much the coefficients differ" or "whether or not they have the same inference."

Later in this report is a list of a wide range of utility metrics from each of these two categories. Data curators should determine whether a synthetic dataset meets their data quality criteria and

should decide what type of utility is most relevant for their intended users (e.g., by conducting focus groups) and then computing the corresponding utility metrics.

Disclosure Risks

Disclosure risk refers to the ability to infer information about data subjects' confidential datasets based on the synthetic data. Measures of disclosure risk require making assumptions about the type of inferences about individuals someone might attempt, the amount of additional information available about the target individuals and the population, and the computational power available for generating the statistics used to make inferences (sometimes called the attack model). We focus on the accuracy of these inferences as our measure of disclosure risk, but it is also important data curators and privacy experts to consider the severity and implications of disclosures when determining whether it is safe to release synthetic data. It may be possible to infer certain information, but not all disclosures will be perceived as equally consequential. Sensitivity to disclosures will also vary by data subject, as subjects with less frequently occurring features may be more prone to disclosure risk. The goal when releasing synthetic data should be to reasonably minimize the risk of disclosures, but determining the acceptable level of risk is a decision to be made by those responsible for releasing the data that should consider the potential severity of specific types of disclosures.

Disclosure risk metrics can be broadly classified based on two different kinds of inference about data subjects. Identification inference measures how well someone can infer an individuals' uniqueness in the population, which has been connected to the legal concept of "singling out" (Cohen and Nissim 2020). Attribute inference measures how well someone can infer sensitive information about data subjects within the confidential dataset. Note that given sufficient background information, successful identification or attribute inferences can also reveal data utility. In other words, some synthetic data models that produce higher utility will capture relationships between variables that exist within the broader population, not just the specific confidential sample. Because of this, it is important to compare disclosure risk metrics with a baseline measure that reflects what might be inferred simply from known characteristics of the population.

An empirical disclosure risk evaluation approach is used in the SEDSyn-23 project, which means measures of disclosure risk are generated for observed synthetic and confidential datasets under a set of specific assumptions, which will account for both identification and attribute inference. This differs from approaches that seek to control the theoretical maximum disclosure risk (or privacy loss) under a formal definition, such as differential privacy (Dwork and Roth 2014). The advantage of formal approaches like differential privacy is that they are less sensitive to the assumptions we make and

allow for the release of information concerning the disclosure risk assessment without potentially leaking additional information. However, satisfying formal definitions, such as differential privacy,[13] requires additional randomization in the model fitting stage that often fails to provide sufficient data utility.

This SEDSyn-23 project did not explore satisfying formal definitions of privacy, so all the disclosure risk metrics presented in the evaluation are sensitive to the assumptions made about the information and computational power available to a potential malicious party. The estimates of the accuracy of the inferences should be understood as reasonable approximations, not upper bounds, of the disclosure risk.

[13] Formal definitions refer to a set of mathematical concepts or frameworks that can mathematically prove the worst-case, or maximum, privacy loss that will result from releasing information based on the confidential data.

Survey of Earned Doctorates Synthetic Data

This section outlines the synthetic data generation and evaluation strategies used to produce the pilot synthetic SED data file.

Defining Use Cases through User Engagement

A focus group study (approved through OMB 3145-0215) was conducted to better understand the needs of SED data users and how they might use synthetic SED microdata in their research. Potential participants were identified from lists of restricted-use license holders, logs of information requests, registered Restricted Data Access System users, and literature review of papers utilizing SED data or Survey of Doctorate Recipients data, which is a subset of the SED data.[14] Of the 27 total participants, 21 were from academia, with the remaining coming from government and industry. Eleven had previously accessed restricted-use SED microdata. Participants were provided background information about synthetic data and verification metrics and then asked to provide feedback on the potential use of synthetic SED data in their research. The feedback from these focus groups was used to inform other SEDSyn activities, including the synthetic data generation, verification metric roadmap, and disclosure risk assessment.

The information obtained in these focus groups was supplemented by feedback from internal SED experts when the project was closer to completion. These experts provided feedback on SED variables in the GSDS, analytic models of broad interest that users might create using synthetic SED data, and the impact of tiered access models across user groups.

The major thematic findings that emerged from these activities regarding the perception and opinion of synthetic SED data include the following:

1. Potential benefits of synthetic SED data include increasing data accessibility and analytical capabilities for researchers, and for teaching and training purposes for professors.

[14] "Survey of Doctorate Recipients" National Center for Science and Engineering Statistics, accessed December 12, 2025, https://ncses.nsf.gov/surveys/doctorate-recipients/2023.

2. Concerns regarding synthetic SED data included incompatibility with some research and the difficulty in assessing the representativeness and potential bias of the synthetic data.
3. Researchers, particularly those interested in modelling and potentially applying for a restricted-use license, prefer synthetic data that more closely resembles the restricted-use data file (i.e., with more variables and minimal coarsening).
4. Journal reviewers are unlikely to accept research outcomes from analyzing synthetic data. The perception of synthetic data analysis from the broader research community remains a key obstacle that prevents researchers from using synthetic data.
5. Verification metrics—particularly those tailored to users' analytic models—and detailed documentation about the synthesizing methodology will instill confidence about the quality of synthetic data; however, there was skepticism that synthetic data could replace restricted-use microdata as the gold standard for their research. On the other hand, users primarily interested in learning about the SED population, such as for institutional benchmarking, may be satisfied with synthetic data combined with appropriate verification metrics.

Gold Standard Dataset Creation

Recall that a GSDS refers to a specific representation of the confidential data that a synthetic dataset aims to imitate. Synthetic data generation algorithms take a GSDS as input; as a result, the synthetic data have the same format, data definitions, and structure as the GSDS. Creating the GSDS involves many design decisions, including but not limited to deciding which variables should be included in the synthetic data and how they should be represented. Releasing synthetic datasets with more variables and more detailed representations of said variables increase disclosure risks, regardless of the method used to produce the synthetic data. Therefore, GSDS specification is an important step that influences the quality of the eventual synthetic data and should be given adequate time and attention prior to beginning the synthesis process.

To determine the appropriate GSDS, the SEDSyn team used researcher input, focus groups of SED data users, and a scan of research products using SED data to identify a collection of candidate variables. Once this subset of variables was identified, baseline levels of disclosure risk were assessed to determine whether different data representations would provide more favorable privacy-utility trade-offs. The SED data include many categorical variables with distinct and sparse levels (for

example, 'First baccalaureate institution location'). This creates two issues for synthesis: first, it is difficult to generate quality predictive models for categorical variables with many sparse levels. Second, sparse levels create tangible disclosure risks that cannot always be sufficiently mitigated by synthetic data generation. To address this challenge, four different coarsening schemes were developed for categorical variables, ordered from the most to least coarse. Whenever possible, multiple forms of nonresponse (like logical skip, questions refused, questions unanswered, etc.) were combined and categories were collapsed to increase the number of observations per category.

In addition to considering different coarsening schemes, different populations of interest for the GSDS were also considered. SED is collected one year at a time, and the first draft of the synthetic SED was based on one year of data. However, to allow for greater detail about fields of study, which was commonly requested by SED data users, a pooled file that combines doctorate recipients over a five-year period, without identifying the specific graduation year, was considered. The pooled five-year dataset benefits from larger sample size, allowing more flexibility to build higher-quality synthetic data models and thus synthetic data. The combined five-year dataset also allows for more favorable trade-offs with coarsening by limiting year-specific information in the dataset and reducing the number of sparse categories in categorical variables. However, the five-year dataset comes at the expense of disabling users from looking at temporal trends within the five-year period. This trade-off is faced by other data products that release pooled data.[15] In the case of SED, longer-term trends are generally of greater interest than single-year changes because most demographic and outcome variables are relatively stable year to year.

Based on utility analyses and feedback from NCSES and user engagement, the five-year pooled dataset and the second-most coarse variable specification were chosen for the GSDS and corresponding synthetic data.

Synthetic Data Processing

The following is the process and considerations of each step of the synthesis process along with details of the software used.

[15] "Using 1-Year or 5-Year American Community Survey Data," US Census Bureau, accessed August 29, 2025, https://www.census.gov/programs-surveys/acs/guidance/estimates.html.

Start Data

The starting variables in the synthesis process serve two important purposes. First, the start data variables are the only predictors for the first predictive model in the synthesis. Second, the start data variables determine the number of observations in the synthetic dataset. Sometimes, it is desirable to add noise to the number of observations. The number of observations in SED is public information, so noise is not added. SED synthesis started with synthesizing one variable, using a bootstrap sample from the confidential variable.

Modeling categorical variables with many categories can be challenging when there are many predictors. Hence, it is often desirable to choose a variable with high cardinality as the start variable.

Synthesis Order

The synthesis order, or visit sequence, is the order in which variables are synthesized. As noted earlier, this order can affect the quality of the synthetic data. Unfortunately, even modest datasets can have an overwhelming number of potential visit sequences, making brute-force search for the optimal ordering computationally infeasible. For example, there are more than 3.6 million ways to order a synthesis with only 10 variables.

Accordingly, the following generic guidelines are used to construct the visit sequence. First, low-variance variables are generally prioritized over high-variance variables. Next, variables with no or minimal missing data are prioritized over those with more missing values. Finally, variables with smaller and less complex supports are prioritized over those with larger or more complex supports (for example, categorical variables with many levels).

For the synthetic SED implementation, the visit sequence started with categorical variables that had no missing values, in order from lowest to highest empirical distribution entropy. This approach allowed the synthesis of “easier” variables first before moving to more complicated distributions. For categorical variables with missing values, missingness was treated as variable level to model. Categorical variables with missing values were synthesized next, according to the same ordering conventional as categorical variables without missing values.

Numeric variables were synthesized last. For numeric variables with missing data, each variable was modeled as a mixture of a “missing data” distribution and an “observed data” distribution. First, intermediate binary missingness indicators were created. These indicators for missingness were synthesized in order from the lowest proportion of missing values to the highest. This maximized the

amount of data available for synthesizing missingness indicators where they have the greatest impact on synthetic data quality. Finally, the numeric variables were synthesized in increasing order from lowest to highest sample standard deviation excluding missing values. The synthesized variables are combined with the synthesized indicators to produce numeric synthetic variables with missingness.

Models and Samplers

The general strategy for model specification relied on nonparametric tree-based modes. Tree-based models are one of the most common and well-established forms of models for data synthesis (Bonnéry et al. 2019; Drechsler and Hu 2021) and have been used for many federal statistical agency data products (Bowen et al. 2022). Numeric variables were synthesized using decision tree regression models, and categorical variables were synthesized using decision tree classification models. For more information about how CART models, a kind of machine learning model, are used to generate synthetic data, see Reiter (2005).

Decision tree hyperparameters play an important role in the privacy-utility trade-off. Parameters like maximum tree depth, minimum number of observations per terminal node, and cost complexity determine the degree to which these models fit to (or potentially overfit to) the confidential data. Choosing parameters that allow more complex models increases both data utility and disclosure risks, and vice versa.

Software

The synthetic SED data was generated using *tidysynthesis*,[16] an open-source R package for generating synthetic data created and maintained by the Urban Institute. *tidysynthesis* focuses on using sequential parametric and nonparametric models to generate synthetic data that matches the joint distribution of confidential data while reducing disclosure risks for the entities behind the synthetic data.

tidysynthesis is built on top of the *tidymodels* framework.[17] *tidymodels* presents a full suite of tools for predictive modeling including common interfaces to many predictive models available through R

[16] "Synthetic Data with *tidysynthesis*." Urban Institute, May 23, 2025, https://ui-research.github.io/tidysynthesis-documentation/.

[17] Max Kuhn and Hadley Wickham, "Tidymodels: A Collection of Packages for Modeling and Machine Learning Using Tidyverse Principles," 2020, http://tidymodels.org/.

packages. *tidysynthesis* has access to most of this functionality, which is essential for developing good predictive models for synthetic data generation.

syntheval[18] is an open-source R package created and maintained by the Urban Institute for evaluating utility and disclosure risks in synthetic data. The package contains functions to evaluate univariate utility with summary statistics, bivariate utility, multivariate utility, discriminant metrics, and specific utility metrics. The package also contains implementations of empirical disclosure risk metrics like membership inference tests and attribute inference tests. This package was used to generate metrics described in the next section.

Description of Utility Metrics for Evaluation

This section provides a brief description of the utility metrics that are computed to evaluate the synthetic SED dataset. More detailed descriptions can be found on GitHub.[19]

Missing Data Patterns

- **Missingness frequency** compares the rates of missingness for each variable in the GSDS and synthetic data. Higher utility is demonstrated by plotting points on or near the 45-degree line, where the rates of missingness are equal in the GSDS and synthetic data.
- **Missingness co-occurrence** looks at the proportion of observations that are missing in both variables for pairs of variables in the GSDS and synthetic data. Higher utility is demonstrated by values close to 0, which indicates that the co-occurrence structure is similar in the GSDS and synthetic data.

General Distributional Measures

Full Joint Distribution Comparisons

- **Nonparametric discriminators** combine the GSDS and synthetic data and then develop models to predict if a given observation is from the GSDS or synthetic data. Nonparametric decision

[18] "syntheval," Urban Institute, accessed on August 29, 2025, https://github.com/UrbanInstitute/syntheval.

[19] "syntheval," Urban Institute.

tree models are used for our discriminator's model. Higher synthetic utility is demonstrated by a discriminator that struggles to distinguish between the sources of observations.

- The discriminator is used to evaluate utility in several ways. The discriminator model's predicted values are evaluated on the training and testing data. Using the predicted probabilities, the **propensity mean-squared error (pMSE)** are computed from random chance, and the **Area Under the Receiving Operating Characteristic Curve (ROC AUC)** of the discriminator model. ROC AUC quantifies how far the predicted probabilities are from random chance, and the pMSE score represents the average distance squared from the predicted probabilities to the marginal probabilities (i.e., the proportion of records in the combined data that stem from the synthetic data). In this case, where the GSDS and synthetic data are the same size, that means probabilities predicted far from 0.5 indicate the model strongly suspects an observation is or is not synthetic. High utility is demonstrated by ROC AUC near 0.5 and high utility is demonstrated by pMSE near 0.
- The variables that are most helpful in discriminating between the GSDS and the synthetic data are also assessed. These variables may indicate where synthesis models could be improved.

Single Variable Distribution Comparisons for Categorical Variables

- **Proportions overview** compares the frequencies for all categories in all variables for the GSDS and synthetic data. Higher utility is demonstrated by points on or near the 45-degree line, where the frequencies of categories in the GSDS and synthetic data are equal.
- **Proportions overview (top categories)** compares the frequencies for highly ranked categories for all variables for the GSDS and synthetic data. The first panel compares the largest categories in each variable, the second panel compares the second largest categories in each variable, etc. Higher utility is demonstrated by points on or near the 45-degree line.
- **Proportions detailed** shows the detailed frequencies for categories in each categorical variable. Higher utility is demonstrated when the frequencies in the synthetic data closely match the frequencies in the GSDS.

Single Variable Distribution Comparisons for Numeric Variables

- **Histograms** compare empirical distributions on numeric variables from the GSDS and synthetic data. Higher utility is demonstrated by histograms that closely overlap.

- **Moments** compare the proportional difference in means, standard deviations, skewness, and kurtoses between the GSDS and synthetic data. Higher utility is demonstrated by moments in the synthetic data and GSDS that closely match.
- **Kolmogorov-Smirnov distance** is the maximum vertical distance (D) between empirical cumulative distance functions calculated on the synthetic data and GSDS. Higher utility is demonstrated by D close to zero.

Bivariate Distribution Comparisons for Categorical Variables

- **Relative mutual information (RMI)** measures the reduction in entropy (i.e., uncertainty) in one variable when observing another variable. It quantifies the relationship between two variables, generalizes to categorical variables, and is in the interval [0,1].
- **RMI difference** is the difference between an RMI matrix from the synthetic data and an RMI matrix from the GSDS. Similar to a correlation matrix for pairwise correlation, an RMI matrix calculates the RMI for all pairs of variables in a dataset. The element-wise differences are evaluated, and the numeric summaries of the differences are calculated. Higher utility is demonstrated by the difference close to 0.
- **RMI side-by-side plots** compares RMI matrices from the synthetic SED and GSDS. Higher utility is demonstrated by visually similar matrices.

Bivariate Distribution Comparisons for Numeric Variables

- **Side-by-side correlation plots** are a simple side-by-side comparison of the correlation matrices from the GSDS and synthetic dataset. Higher utility is demonstrated by correlation matrices that visually closely match.
- **Correlation differences** are heat maps of the differences between the GSDS and synthetic data correlation matrices. Higher utility is demonstrated by correlation matrices that closely match. The element-wise differences are evaluated, and the numeric summaries of the differences are calculated.

Specific Analysis Measures

Models of general interest to SED data users were identified as part of the user engagement described in the “Defining Use Cases” section. Three regression models that could be constructed using the

GSDS and synthetic data were selected to demonstrate the use of utility measures tailored to specific analyses. These models were based on the following research questions:

1. Within field of study, what variables are associated with time to degree completion?
2. What individual-level characteristics are associated with time to degree completion?
3. What are the career and earning outcomes of doctorate earners with disabilities?

These models are fitted on both the GSDS and the synthetic data and compare the estimated models using the following metrics.

- **Number of observations** used to estimate the model. This value can differ because models are only fit on the complete cases.
- **Confidence interval overlap** is the proportion of overlap between the confidence intervals for a coefficient in the GSDS and synthetic SED models. 1 indicates perfect overlap, 0 indicates no overlap, and values grow increasingly negative as intervals move further apart. Confidence intervals are calculated based on commonly used formula for confidence intervals, which can be obtained from built-in R functions. When it comes to complex survey data where weights and design variables are involved, such confidence interval calculation is more complicated, which may or may not be readily available from existing R functions such as those from the survey R package.
- **Coef_diff** is the difference of coefficient values between the GSDS and the synthetic SED.
- **Std_coef_diff** is the standardized difference of coefficient values between the GSDS and the synthetic SED.
- **Ci_widths_ratio** is the ratio of the confidence interval widths between the GSDS and the synthetic SED.
- **Sign_match** is 1if the sign of the coefficients is the same between the GSDS and the synthetic SED and 0 if it is not.
- **Significance_match** is 1 if the statistical significance of a t-test for non-zero coefficients against the null hypothesis of a coefficient equal to zero, for a given alpha level of confidence, is the same between the GSDS and synthetic SED models and 0 if not.
- **SSO_match** is 1 if the signs match, the statistical significance match, and the confidence intervals overlap at least some between the GSDS and the synthetic SED and 0 if any of these

three are not true. Consider this a combination of the sign_match, significance_match, and overlap metrics described above.

Assessing the Results of the Utility Evaluation

Given the numerous utility metrics provided, the SEDSyn project team utilized the following steps for reviewing the results of the utility evaluation and recommend that others reviewing a similar set of metrics use a similar process.

1. Start by reviewing the schema and missing data pattern comparison to ensure that the synthetic data correctly captures the structure of the GSDS. The synthetic data should contain the same dimensions, variable types, and ranges. While the amount of missing data can vary slightly, it should not dramatically differ between the synthetic data and the GSDS.
2. Review the results of the full joint distribution comparison based on the various evaluation metrics listed earlier. First, this provides simple statistics to compare across candidate syntheses to summarize their relative utility. Second, looking at the variable importance plot helps inform which variables may contribute to lower utility and deserve further investigation.
3. Based on the assessment of the discriminant model, review the single variable and bivariate distribution comparisons for the variables that may contribute to lower utility. For categorical variables, this includes reviewing the proportions and the relative mutual information. For numeric variables, this includes reviewing the moments, empirical distributions, and correlation plots. Variables, or pairs of variables, may be targeted for improvement if the differences between the synthetic data and the GSDS are considered too large.
4. Finally, review the results of the specific regression analyses. While the distributional comparisons give a general idea of how the synthetic data compares with the GSDS, the specific comparisons help understand how the results of a user's model changes by using the synthetic data in place of the GSDS. For each mode, review how the overall fit (i.e., adjusted r-squared) changes, how much the point estimates and confidence intervals differ, and what percentage of the coefficients retain the same sign and statistical significance.

Description of the Disclosure Risk Metrics for Evaluation

Our goal for disclosure risk metrics is to assess the inferences that could be made about participants in the SED from 2019 to 2023 based on (1) a candidate synthetic dataset and (2) a candidate synthesis model. The SEDSyn team started with the GSDS and synthesized data to capture the same variables, formats, and coding schemes that were selected in creating the GSDS. The GSDS preprocessing decisions, such as how variable levels were collapsed, are assumed to be released as public information in accompanying documentation. Another assumption is that the specification of the synthesis model (i.e., the model types and hyperparameters) are released, but the specific parameters of the model are not released.

Disclosure Risk Parameters for the Synthetic SED

Two types of variables are defined which form the basis of the disclosure risk evaluation.

- **Quasi-identifying keys (QIKs):** the characteristics assumed to be gathered externally with reasonable accuracy
- **Sensitive attributes (SAs):** the characteristics assumed not to be accurately gathered externally and those individuals would be more likely to want to keep confidential

An attacker attempting to make inferences under the assumption that participation in the SED is *known* is considered, and that QIK values for individuals known to participate have already been collected externally prior to making inferences. This is a common assumption in disclosure risk assessments. It reflects the case where the attacker has collected information about individuals, and they are seeking to learn more information using the released synthetic data and accompanying preprocessing and model information. Under this assumption, the more QIKs that are included, the stronger the attack because it assumes more information was collected about individuals a priori.

Although the SED is a census, about 8 percent of new doctorates do not respond each year. This small amount of nonresponse uncertainty could provide some privacy protection because it may not be known with certainty that an individual participated in the SED, so our method may slightly overstate the accuracy of the inferences. Because there is evidence that nonresponse is not completely random, for example some institutions require SED completion, any additional protection from nonresponse uncertainty is not expected to be equally distributed across the population.

Given the population of interest, the working definition of QIKs is information that could be accurately gathered through professional websites, such as LinkedIn or organization employee pages.

To test the sensitivity of the measures to the amount of assumed QIKs, disclosure risk metrics are produced under four different sets of QIKs.

SAs are identified as variables that individuals in the SED would likely want to keep confidential and that would not be generally available through professional websites. These can include variables like citizenship or post-graduation salary.

The QIKs and SAs form the basis for our disclosure risk evaluation. Two commonly used statistics for empirical risk assessments are defined, which are used to inform our inferences:

- **k-frequency (k) (Sweeney 2002):** for a given set of QIKs, the number of records sharing the same QIK values
- **l-diversity (l) (Machanavajjhala et al. 2007):** for a given set of QIKs and a selected SA, the number of distinct values of the SA for records sharing the same QIK values

Empirical Attack Models

Our method for measuring empirical disclosure risk is to generate predictions of information an attacker may want to learn, and to use the GSDS to evaluate the accuracy of these predictions. The predictions are compared with commonly used empirical measures based on k-frequency or l-diversity uniques in the GSDS (Raab, Nowok, and Dibben 2024), with an l-diversity of 1 in the synthetic data, and have the same QIK and SA values.

These measures also implicitly try to measure the accuracy of an attack. In this case, the assumed attack is that someone would look for synthetic records with a k of 1 or and l of 1 and then assume the values match in the GSDS. The assumed inference of the attacker in the first case is that the individual they are seeking is unique in the GSDS, and in the second case the assumed inference is that they have learned the individual's SA value.

These approaches have a few shortcomings. First, they do not leverage any information about the synthetic data model, which can be significant. Second, they only produce predictions for a small set of synthetic records, while an attacker might reasonably try to make inferences about many more records. In other words, the attacker might infer with some probability that a record has a k of 1 in the GSDS, even if the corresponding QIKs in the synthetic data show up more (or less) than 1 time. Third, because these attacks rely on precisely replicated QIK values, they get weaker as the number of QIKs increases, since the likelihood of producing any matches decreases. This is the opposite of the truth,

which is that the attacks should become more powerful as the assumption is the attacker has more information a priori.

To produce stronger empirical estimates of identification risk, "resynthesis" replicates are generated (i.e., new synthetic datasets that use the same synthesis model) *with the synthetic SED as the input* in place of the GSDS. This allows us to model the relationship between the frequency of QIKs in the input data used in a synthesis model and the synthetic data it generates. This model is then used to make predictions about the k-frequency in the GSDS based only on information in the synthetic SED.

For attribute risk, the relationship is modeled between the QIKs and the selected SA in the synthetic data. For variables where the synthetic data have encoded the relationships, an attacker might reasonably use this model to make predictions of the SA (assuming the QIKs are predictive of it). To avoid measuring utility, the evaluation of the accuracy is restricted of these predictions to cases in the GSDS where the l-diversity is equal to 1. The resynthesis replicates can also be used to increase the strength of the attribute inference. These datasets capture some of the variability in the relationship between the QIKs and the SAs, so they can be used to tune the hyperparameters of the predictive models. Due to computational limitations, we were unable to tune our attack models in this way, but we recommend doing so if possible.

One common type of empirical risk measure that we do not utilize is the membership inference attack (Stadler, Oprisanu, and Troncoso 2022). The goal of this attack is to measure an attacker's ability to determine whether a particular record (e.g., containing certain QIK values) was part of the confidential data used to create the synthetic data. This is accomplished by simulating synthetic datasets under the case where the single record is included or not included in the confidential dataset. We then compare the outputs to understand the extent to which the synthetic data model encodes that particular record and determine how well we can predict membership of that record. For the SED, we chose not to evaluate this risk metric because it is a near census and membership in the dataset can be assumed with high certainty without running any attack. We chose not to run the analogous attribute inference attacks, which attempt to ascertain the presence of a particular record containing certain QIK values *and* SA values, because these attacks were computationally infeasible for this project. The empirical attacks we used are designed to produce similar inferences but are easier to run with limited resources.

Metric Configuration for Empirical Disclosure Risk

The measures of identification and attribute risk are defined by the accuracy with which an attacker might make the following inferences.

Identification Risk

$$P(k_{GSDS} = 1 \,|k_{SEDSyn} = k, k_{GSDS} > 0)$$

In other words, given the observed number of records (k) sharing QIK values in the synthetic SED, what is the probability that only one record in the GSDS has these characteristics (i.e., k equal to 1)? Predictions are made for all records observed in the GSDS and compare the predictions to the true GSDS k values.

The accuracy of these predicted probabilities are compared with the accuracy of a prediction model that is 1 if k_{SEDSyn} is 1 and 0 otherwise. This corresponds to the implicit attack model underlying the measurement of replicated uniques.

Attribute Risk

$$P(SA_{GSDS} = a \,|QIK_{GSDS}, f_{SEDSyn}, l_{GSDS} = 1)$$

where f is the model is estimated using the synthetic SED for predicting the SA based on the QIKs.

In other words, given the observed synthetic QIK and selected SA values in the synthetic SED, what is the probability that a record in the GSDS with certain QIK values has a certain value of the SA? Predictions are made for all records observed in the GSDS which have an l-diversity of 1 and compare the predictions to the true GSDS SA values.

The accuracy is compared of these predicted probabilities with the accuracy of a prediction model that is the observed value of SA_{SEDSyn} if $l_{SEDSyn} = 1$ and 0 otherwise. This corresponds to the implicit attack model underlying the measurement of the TCAP and related measures.

To measure accuracy of the disclosure risk predictions, two statistics are used, precision (also known as the positive predictive value) and recall (also known as the true positive rate), to define the accuracy of the predictions.

In the context of identification and attribute risk, precision is the proportion of synthetic SED records that the attacker predicts have a k of 1, or SA value of $\boldsymbol{a}$, in the GSDS that does. Recall in the

proportion of GSDS records that have a k of 1 (or SA value of $\boldsymbol{a}$) that the attacker predicts have a k of 1 (or SA value of $\boldsymbol{a}$).

Simply put, precision is the rate at which an attacker would be right about their predictions, and recall is the fraction of all the possible correct predictions that the attacker is able to make. Generally, precision will decrease as recall increases because the attacker needs to make less confident predictions (and be wrong more often) to increase the percentage of correct inferences out of all possible ones. When the predictions are probabilities, this trade-off is summarized using a precision-recall curve that shows all the possible values of precision and recall for different target cutoffs (e.g., only assuming those with probability greater than 0.9 are true predictions is one cutoff choice). When the predictions are binary values, as in the case of our comparisons with replicated uniques or TCAP, only one set of values of precision and recall are obtained.

Verification Metrics

Although the SEDSyn team recommends conducting extensive utility and disclosure risk evaluations, verification metrics are necessary to provide vital information to potential users about the accuracy of analyses conducted on synthetic data. However, releasing results from verification metrics also introduces new risks to the disclosure of information about individuals in the data. Verification requests must be controlled to ensure that a malicious entity does not disclose information through a series of well-constructed requests.

Definition of Verification Metrics and Verification Server

Verification metrics are statistical measures that describe the similarity of inferences made using the synthetic data compared with the same analyses made using the confidential data. The goal of these metrics is to provide users with information about the accuracy of the results obtained through the synthetic data, with respect to the results they would obtain using the confidential, or gold-standard data. Verification metrics are calculated by conducting the same analysis on both the synthetic and gold standard data, and then computing a measure of similarity between the two results that is returned to the user. Verification metrics are traditionally considered a subset of a broader class of *utility metrics*, which seek to describe different aspects of the synthetic data, with higher values corresponding to great amounts of "utility" for the users of the synthetic data.

Consider three categories of utility metrics: (1) distributional distance measures, (2) specific statistics distance measures, and (3) agreement of hypothesis test results. All of these categories define maximal utility by statistical estimates produced using the confidential data. Most prior work on verification metrics focused on category 3, but depending on what the goal of the verification is, any of these buckets could contain verification metrics. Verification metrics do not report on the validity of the inferences made using the synthetic data.

For this case study, using SED verification metrics could help inform users as to whether the inferences they draw using the synthetic data are similar to what they would have obtained with the GSDS. We note that these measures do not necessarily verify what may have been obtained using the Doctorate Records File,[20] the restricted-use file containing SED microdata going back to 1920, since

[20] "Survey of Earned Doctorates," National Center for Education Statistics accessed on December 12, 2025, https://nces.ed.gov/statprog/handbook/sed.asp.

the Doctorate Records File includes additional variables and detail that are removed when creating the gold-standard data (see Gold Standard Dataset Creation section).

A *verification server*, first proposed by Reiter, Oganian, and Karr (2009), is a system for providing users with verification metrics for a specified synthetic dataset and the corresponding confidential data. A verification server could provide preselected metrics, so that users can generate results similar to a table builder. Alternatively, a server could allow users to request semi-arbitrary metrics by submitting code from a list of approved statistics (e.g., linear regression models) that will be run on both the synthetic and confidential data to produce the verification metrics. Privacy for these servers can be maintained through either manual disclosure review: by predetermined metrics or statistics as they are requested, or by automatic processes that systematically adding noise through a formal privacy approach, such as differential privacy.

Finally, we note that *validation servers* have also been proposed in the literature (e.g., Tyagi et al. 2025; US Census Bureau 2024; Vilhuber 2025). They share much in common with verification servers, and research on them can provide useful lessons for verification servers, but they provide a different type of output for the users. Rather than providing metrics on the similarity of results between those run on the synthetic and confidential data, validation servers directly provide the statistics run on the confidential data (albeit with noise added in many cases). In the validation server paradigm, users use the synthetic data to explore their models and then receive *validation* estimates of the same models run on the confidential data. The verification server paradigm, while similar, assumes that users will be satisfied with *verification* that their synthetic results are close to what they would have received using the confidential data, without receiving the estimates derived from confidential data.

Candidate Verification Metrics for a Synthetic SED

An initial list of verification metrics was developed based on a review of the relevant prior literature (Barrientos et al. 2018, 2019; Lin and Reiter 2024; McClure and Reiter, 2012; Reiter 2009; Yu and Reiter 2018). Given the focus on regression-based methods reported by users of the SED in the focus groups conducted, the following metrics are used:

1. Confidence interval overlap
2. Confidential estimate above or below a user-provided threshold
3. Point estimate sign alignment

4. Statistical significance alignment

Confidence interval overlap (Karr et al. 2006; Snoke et al. 2018) measures the percentage of overlap between the synthetic and confidential interval for an estimated statistic with a corresponding confidence interval at a given level of significance. The optimal value of this metric is 1, when the intervals overlap entirely; it is 0 when they do not overlap and negative the further they move away from each other (using the Snoke et al. 2018 formulation of the metric).

Confidential estimate above or below a user-provided threshold (Barrientos et al. 2018) is a binary or probabilistic metric which measures whether the confidential statistic, typically a regression coefficient, is above or below a certain value. This value is commonly informed by the estimates obtained using the synthetic data, and the verification metric is used to determine whether the magnitude obtained in the confidential data surpasses a threshold that may be meaningful for the analyses.

Sign and significance overlaps (Barrientos et al. 2019; Reiter 2009) are binary or probabilistic measures that measure whether the synthetic estimate shares the same sign and significance, for a given level, with the corresponding confidential estimate. These metrics are useful for determining whether the inferences drawn from the synthetic data about the direction and significance of the estimate are the same as would have been found in the confidential data.

All metrics can be computed using the R software and are publicly available in the *syntheval*[2] package. These metrics could be easily automated and reproduced in a future verification server. As we will discuss further below, there does not exist current code to automate the disclosure control protections for releasing these metrics, so additional work would be required to ensure the verification server maintains privacy.

Adaptations of Candidate Measures for Analyses of a Census

As previously mentioned, the SED comprises a census of individuals who earn research doctorates each year in the United States from an accredited university. Many researchers analyze the dataset as a finite population, with no sampling uncertainty, while others generalize inferences outside of the observed data and use confidence intervals and hypothesis testing, in effect treating the data as a sample from a superpopulation. Some of the verification metrics discussed in the previous section

assume that analyses will incorporate uncertainty estimates. If users of the synthetic SED want to analyze it as a census of a finite population, with no sampling uncertainty, the following measures may still be utilized:

1. Confidential estimate above or below a user-provided threshold
2. Point estimate sign alignment

In addition, the following measures not traditionally considered verification metrics may be considered as types of verification metrics:

1. Absolute difference between estimates
2. Relative difference between estimates

SED User Engagement and Feedback on Verification Metrics

As described earlier, the project team conducted a focus group study to better understand the needs of SED data users and how they might use synthetic SED data and verification metrics in their research. The focus group participants were provided information about three categories of utility and verification metrics: (1) measures of distributional similarity, (2) metrics summarizing closeness of estimates to the corresponding estimates based on the confidential data, and (3) verification measures for population inferences. Generally, participants noted that the availability of evaluation metrics and detailed documentation about the synthesizing methodology would instill confidence in the synthetic data quality; however, some participants, particularly those who had previously used the restricted-use microdata, remained skeptical about whether verification metrics would be sufficient for research publications.

Overwhelmingly, participants believed that the metrics should be tailored to a researcher's analytic model, and that generic prepublished metrics would be less helpful. Some participants also expressed a preference for metrics that were easy to understand and verification or validation processes that did not require substantial effort on their part; for example, a general stamp of approval that did not require researchers to sift through the details of verification for every analysis. Even with metrics tailored to their analyses, some researchers believed they might still need to access the confidential data. Factors including complexity of verification information, complexity of analytic

models, and verification metrics indicating lack of agreement with the confidential data might make the synthetic data plus verification metrics a less than adequate substitute for the confidential data.

The keys points around discussions of specific evaluation metrics were as follows:

- Measures of distributional similarity that provide general information about the quality of synthetic data might be useful for determining whether to use the synthetic data at all; however, they will be less useful for instilling trust in specific analyses, which are often multivariate models. The measures of distributional similarity discussed were side-by-side histograms, correlation heatmaps, and distance metrics between confidential and synthetic data.
- More researchers believed the metrics summarizing closeness of estimates to the corresponding estimates based on confidential data would be useful for multivariate analyses and for helping analysts decide whether to use synthetic data. The closeness metrics discussed include confidence interval overlap, standardized or absolute difference, and ratio of confidence interval widths.
- Similarly, the verification measures for population inference were preferred for their ability to inform multivariate analyses. The measures presented include confidential estimate above or below a threshold, point estimate sign alignment, and statistical significance alignment.

Based on the feedback from focus group participants, most users are interested in measures for population inference in a verification setting. These metrics can be used to support researchers conducting multivariate analyses with synthetic data and are easily interpreted. However, some users are primarily interested in learning about the SED population and analyze tables rather than build models (e.g., institutional users using SED data for benchmarking purposes), and measures of distributional similarity might be better suited to their needs.

Second, the focus groups indicated that users want an interactive verification server where they can request specific models. They have responded this way because they were not presented with a specific plan for what models the verification server would include, but regardless it is true that users would not find a server useful if it did not contain the ability to verify their specific models of interest. Overall, users seemed more concerned that the verification metrics could be estimated for their specific analyses than they were about which metrics were provided.

In conclusion, it seems clear that any implementation decisions regarding a verification server should utilize feedback from additional focus groups. The focus groups we conducted provide a useful

starting place, but participants were only presented with a theoretical synthetic data set and verification server. In particular, these additional focus groups should discuss the variables which would be made available in the synthetic dataset and what analyses could be run through the verification server. Additional focus groups can also help determine whether more potential users would benefit from metrics for verifying population inferences or measures of distributional similarity. In the event that resources are limited, findings from the focus groups and the accompanying literature review conducted for this project could inform next steps.

Existing Models for Verification and Validation Servers

The SEDSyn team reviewed three existing models for verification and validation in the literature to help provide implementation considerations for a future NSDS. Note that a related concept to verification is validation. With validation, users submit code that has been successfully run on a synthetic dataset to be run on the confidential data and receive output based on the confidential data. The output must meet agency criteria for disclosure, which may require manual screening and applications of disclosure limitation procedures such as rounding or noise infusion.

The three existing models are (1) NCHS public-use synthetic linked data (National Center for Health Statistics 2025), (2) Cornell Synthetic Data Server, and (3) a prototype Statistics of Income server (Tyagi et al. 2024). The first model is a verification server while the last two are validation servers.

National Center for Health Statistic (NCHS) Public-Use Synthetic Linked Data

The NCHS produced a public-use synthetic linked data file, along with a predetermine and noninteractive verification server (NCHS 2025). The verification server is run by NCHS staff who can provide metrics for binomial logistic regression models in response to emailed requests by users. Clients must complete a request form selecting outcome and predictor variables for their regressions from a set of predetermined variables, and models can be estimated on a preselected list of sub-populations. Users receive metrics including side-by-side histograms, confidence interval overlap (in discrete bins), sign match, significance match, and whether the synthetic coefficient is contained in the confidential confidence interval. This verification server provides privacy by only providing statistics on models that have undergone disclosure review, so no further review is required once users are given approval to review the results for their model. The NCHS server is the only existing verification server that could be identified in the literature review.

Cornell Synthetic Data Server

An early example of synthetic data validation is the Cornell Synthetic Data Server, which between 2014 and 2022 provided access to two synthetic versions of Census Bureau datasets. Users were able to request validation of results from the Census Bureau. Despite providing users with a virtual desktop environment very similar, though not containerized, to the restricted environment, code often failed to replicate, creating an excessive time burden to validate requests (Vilhuber 2025). In 2024, the Census Bureau established a new validation procedure with some limitations (Benedetto et al. 2024; US Census Bureau 2023). Specifically, users who submitted validation requests were responsible for fixing replication errors and following guidelines to comply with disclosure avoidance rules, reducing the required staff level of effort.

Drawing on experience with the Cornell Synthetic Data Server, Vilhuber (2025) proposed a framework for validation and verification using a containerized computing environment to streamline processes and facilitate code replicability in a scalable manner. This framework could solve many of the issues faced by the previously established servers, but using containers and a streamlined process does not provide disclosure control. To deploy these methods data providers would need to determine methods for protecting the output, either through review or automated noise addition.

Statistics of Income (SOI) Division of the Internal Revenue Service (IRS) Prototype Server

A more recent example of an automated validation server has been developed by the SOI division of the IRS, in collaboration with a team from the Urban Institute, RAND Corporation, and Florida State University (Taylor et al. 2021; Tyagi et al. 2024, forthcoming). The goal of the project is to combine a synthetic dataset, also in development by the Urban Institute (Bowen et al. 2020, 2022), with an automated server where researchers can request a limited set of statistics from the confidential data by submitting code written in R. Specifically, users can query univariate statistics, such as means, standard deviations, or quantiles, or they can query linear or logistic regression estimates, including coefficients, confidence intervals, and model goodness of fit statistics. The results of these queries are protected through noise addition to preserve privacy and reduce the need for human administrative review. While verification is not currently a goal of the IRS server, the infrastructure they have built provides some lessons that may be useful for others that are interested in establishing a verification server. We detail the combined lessons from these three prior case studies in the following section.

Implementation Considerations When Building a Verification Server

Considering the Three-Way Trade-Off among Automation, Availability, and Privacy

From the three existing models of verification and validation servers, future efforts should consider balancing the three-way trade-off between automation, availability, and privacy. We describe each pair of the three-way trade-off below.

First, there exists a trade-off between automation and complexity of the available statistics, in addition to the typical trade-off of accuracy and privacy. To automate systems (i.e., provide users with verification results without human review), the statistics that can be verified must be limited or risk serious privacy leakages. To achieve full automation while allowing users to request arbitrary statistics, the system would need to use a theoretically rigorous approach such as differential privacy, but this would come at the cost of accuracy in the statistics (Barrientos et al. 2024).

Among the three existing models of verification and validation servers, the NCHS verification server limits users the most. In exchange for not adding any noise to the verification results, users may only request results from a range of preselected models. The idea of supplying users with verification metrics from only predetermined models was raised during the focus groups conducted for this project, and some users expressed concern that this would limit the value of the server. Specifically, they were worried that the server would not include the types of models or the specifications they needed for their research, in which case the server would not provide any value.

Second, we see a trade-off between automation and privacy. The greater the automation, the greater the need for rigorous privacy protections, while less automated systems can allow for more flexibility in the privacy evaluations. One the one hand, the prototype SOI validation server seeks to provide statistics to users while minimizing the need for human review, but the server adds noise to every released statistic to ensure privacy loss is managed. On the other hand, the Cornell Synthetic Data Server and Census Bureau validation processes allow users to request statistics without adding noise, but the results must be approved through the standard disclosure review process, which requires manual review.

Finally, a trade-off exists between privacy and availability. The prototype SOI validation server limits the types of queries users can make to regression models and univariate statistics, on a limited

set of variables, to reduce the amount of noise that must be added to protect privacy. Alternatively, the NCHS server only offers a limited range of statistics, but they are provided with less noise added.

For the case study of SED, if a verification server is developed alongside a synthetic SED, or if this approach is used with other data for the NSDS, it will need to determine the balance between automation, availability, and privacy. Users will need to understand that a bespoke system that allows them to verify any arbitrary result will only be possible with either a lengthy human review (low automation) or potentially low accuracy statistics (high levels of noise).

Prior to creating a verification system, it is important to establish a clear goal as to whether users should be able to specify models versus select from predetermined models. It is inevitable that the latter approach results in variable utility for researchers, so it is necessary to assess whether it is more important to have a system with less administrative management required or one that allows for a more bespoke set of models.

Decisions will also need to be made determining the appropriate levels of acceptable risk to SED respondents and respondents from other NCSES datasets and selecting the space of possible statistics that can be requested. Because a technical trade-off exists, as described above, this implies the lack of any one "correct" answer to a verification process to address these trade-offs. Any implementation will need to determine how much noise to add to sufficiently protect privacy, depending on the level of automation versus human review. Sufficient protections will also need to consider access questions, such as how many queries users can make and what types of queries are allowed. In general, allowing more queries or a broader allowable range introduces increased privacy risks that must be mitigated by adding more noise to the statistics. If adding more noise is not acceptable, the query ability will need to be limited. Regardless, the existing models make it clear that there is value to verification, even if determining the specific decisions for each dataset will need further work. This situation stresses the importance of setting clear goals early in the process for releasing synthetic data and developing an accompanying verification or validation server to ensure proper implementation and alignment within a future NSDS.

Making Joint Decisions for Synthetic Data and Verification Servers

Finally, the existing models show that decisions concerning the development of synthetic data should be made in tandem with decisions about the verification servers. If the verification server will be populated with a limited set of preselected models, it may make sense to ensure that the synthetic data accurately capture these statistics. If users find that the synthetic data perform poorly for the

provided verification metrics, that may diminish the value of the synthetic data. Conversely, if users understand that the synthetic data are intended for broad exploration, though not guaranteed to replicate results from the confidential data, then the verification server would be more valuable if it provides a wider range of possible queries.

Future Research to Inform the Implementation of Verification Servers for the National Secure Data Service

Mitigating Privacy Loss in Releasing Verification Metrics

The most significant area for further research is the potential for privacy loss from releasing verification metrics. Some metrics, such as absolute or relative error, cannot be released directly without revealing the confidential estimate. Other measures, such as confidence interval overlap or confidential estimate thresholds, have the potential to leak information, particularly in combination with other metrics. Prior work has utilized additional privacy protections, such as noise addition (Barrientos et al. 2018, 2019) to release verification metrics. These approaches have the drawback of adding uncertainty to the verification itself, which is meant to increase confidence in the synthetic data estimates.

Verification Metrics for Small Subsets

Multiple users in our engagement study discussed conducting analyses on small subsets of SED respondents and wanting to ensure that they could verify the results, since the synthetic data may not be designed to specifically capture these groups. In general, small subsets present additional difficulties for privacy preserving methods (Bowen and Snoke 2023). Similarly to common practice in surveys or other privacy preserving implementations, future work should investigate the need for suppressing output for estimates that are based on subsets that do not meet a sample size threshold.

Further Considering Potential Users of a Verification Server

While we found that focus group participants wanted a verification server, we acknowledge the caveat that these participants were largely those who already have access to the restricted data or expect to publish based on the synthetic SED. It may be that other potential users of the synthetic SED have different goals or needs, which do not involve the same level of verification for inference.

Ongoing research should consider what needs could be met if we are only able to share a limited amount of verification metrics and how this may relate to alternative user profiles. For example, as noted under the Defining Use Cases through User Engagement section, there are some SED users who primarily use tabular data to learn about the doctorate population. For these users, verification measures might be sufficient for them to trust the synthetic data.

Assessing Whether Verification Servers Provide the Right Tier of Access

The NSDS will contain a variety of tiers of access. Determining whether verification servers should play a role in the future service will depend on whether they provide value to researchers with an acceptable level of administrative burden. This challenge highlights a three-way trade-off among automation, availability of statistics, and privacy, which is further discussed in the next section. Given the relatively little amount of prior research on verification servers, it is too soon to say whether this can be a general tool for use in the NSDS.

Conclusion: Impact and Implications of Synthetic Data Generation

This section summarizes the lessons learned throughout the project along with a potential roadmap on how to release a synthetic SED.

Time and Resources Investment

While tools such as *tidysynthesis* have greatly reduced the time and effort to produce synthetic data, producing a high-quality synthetic version of a complex dataset still requires significant investment. Using synthetic data generation tools such as black-box synthetic data generators is unlikely to produce synthetic data that provides both good data utility and good privacy protection. In addition, any data supported on a future NSDS will require specifying the appropriate GSDS, which is a time-consuming and labor-intensive process that should be given adequate time and attention prior to beginning synthetic data generation. Including data curators, stakeholders, and users in the process will result in a GSDS and a corresponding synthetic dataset that best meets the desired use cases.

Simply put, generating synthetic data is a multipart process that includes defining the objectives of the synthetic data, defining a GSDS, generating synthetic data, evaluating synthetic data, and working with decisionmakers to release synthetic data. Ideally this process should be scalable, so each synthesis is lower cost than the previous synthesis. The generation of a synthetic dataset will scale, but not all parts of the process will scale the same. The actual generation of the synthetic data may scale very well with the addition and refinement of the synthesis process. The time-consuming process of working with stakeholders to define and refine the data will likely scale less well.

The resources and time required to create synthetic data should be compared with the process of collecting and processing high-quality survey data. There are underlying methods that can be replicated between different synthetic datasets, but the process as a whole cannot be automated. Similar to surveys, there are design choices and unique elements to consider in each different context of generating synthetic data.

For repeated cross-sectional surveys like SED, the effort required to produce synthetic data for additional years of data will be significantly lower than the initial investment. As with the collection of the SED, we can more easily repeat methods for future years once a standard approach has been

developed. Further, the lessons learned from the earlier efforts may reduce the time to produce synthetic data for related datasets with overlapping data users.

Empirical Disclosure Risk Evaluation Limitations and Common Practice Improvements

Common empirical disclosure risk measures tend to be too simple and do not provide a meaningful measure of the actual risk, such as the percentage of replicated uniques as a measure of identification risk or the total correct attribute percentage as a measure of attribute risk. The evaluations in this project show that these measures significantly underestimate the ability of an attacker to use the synthetic data to draw inferences about individuals in the confidential data. This project showed that empirical evaluations for synthetic data can be strengthened compared with commonly used approaches. We also note that the SED is a census, and therefore there is little uncertainty in the composition of the GSDS.

However, empirical evaluations depend directly on our assumptions about what information is available to the attacker and how much computational power and time they can utilize in their attack. The measures proposed in this work give a better assessment of the risk, as compared with simplistic measures commonly used, but they are still limited by the computational resources that were available for this project. In particular, predictive models could not be tuned, which would likely have strengthened the attack. The SEDSyn team could also only measure single implicates of the synthetic data because of computational limitations and could not make statements about the variation in the disclosure risk estimates.

More work is required to develop best practices for conducting empirical disclosure risk evaluations for synthetic datasets generation within a future NSDS. This project has shown that measures such as the percentage of replicated uniques are insufficient, but a broader consensus is required concerning which empirical measures are considered sufficient. The strength of the empirical measures that are available will depend directly on what computational resources the NSDS can provide. The strongest possible empirical disclosure risk metrics typically require fitting multiple models on datasets, generating multiple synthetic data replicates, and building adaptive tests for membership inference to compare disclosure risk across records while accounting for sampling uncertainty (if applicable). Establishing sufficient computing resources, as well as the ability to pre-

allocate either time and/or resources in off hours, would allow future NSDS users to perform more complex tasks like those needed for stronger empirical disclosure risk metrics.

Synthesis Code Disclosure Review Requirements

Any code used to generate and evaluate synthetic data within the NSDS should undergo separate review focused on disclosure avoidance prior to removal from the NSDS secure computing environment. For example, pseudo-random number generation details should not be included in shared code to prevent exact duplication of randomness. Other implementation code specific to confidential data should also be reviewed for potential risks.

Version Control in Secure Environments

Working in secure environments like the NSDS Secure Computing Environment Testbed[21] presents challenges for collaborative code development. Some secure environments do not support distributed versioning tools like Git, requiring teams to manually version control code and documentation. Future support for compliant tools like GitLab would significantly improve collaboration across NSDS projects. Furthermore, many secure computing environments lack the computational power to support specific empirical disclosure risk metrics.

Utility Metrics in the Broader Context

Decisions concerning the development of synthetic data should be made in tandem with decisions about the verification servers or other access models that may be utilized in the NSDS. Utility metrics should be considered in light of how users expect to interact with and use data products. For instance, existing examples of verification or validation servers show that decisions concerning the development of synthetic data should be made in tandem with decisions about the verification servers. If a verification server will be provided with a limited set of preselected models, it would make sense to ensure that the synthetic data accurately capture these statistics. If users find that the synthetic data perform poorly for the provided verification metrics, that may diminish the value of the synthetic data. Conversely, if users understand that the synthetic data are intended for broad exploration, though not

[21] "Secure Computing Environment Testbed (SCET)," NORC, accessed September 17, 2025, https://www.norc.org/research/projects/secure-compute-environment-testbed-scet.html.

guaranteed to replicate results from the confidential data, then the verification server would be more valuable if it provides a wider range of possible queries.

Understanding the broader context requires engagement with the data user community, both those currently using related data products and those who want to access data but have been unable to due to current barriers. In particular, potential users who do not have access to restricted data products may provide the most information regarding what would make a useful product. Prior to evaluating the utility of synthetic data, a specific plan for how the synthetic data will be used and combined with other data products should be developed. This will help guide the selection of appropriate utility metrics.

Shared Taxonomy for Improvement of Communication across Teams

The various NSDS projects involve researchers and methodologists with diverse training—statisticians, data scientists, survey methodologists, and social scientists—with varying technical backgrounds. Establishing operational definitions for terms like “key variables” (e.g., for this case study information from professional websites or organizational papers) and “sensitive variables” (e.g., information that individuals would want kept private) helps align disclosure and utility risk analyses, even if the definitions are imperfect.

Consideration of Balancing the Three-Way Trade-Off among Automation, Availability, and Privacy

From the three existing models of verification and validation servers, future efforts should consider balancing the three-way trade-off among automation, availability, and privacy. We describe pairs of the three-way trade-off below in detail.

First, a trade-off exists between automation and complexity of the available statistics, in addition to the typical trade-off of accuracy and privacy. To automate systems (i.e., provide users with verification results without human review), the statistics that can be verified must be limited or risk serious privacy leakages. To achieve full automation while allowing users to request arbitrary statistics, the system would need to use a theoretically rigorous approach, such as differential privacy, but this would come at the cost of accuracy in the statistics (Barrientos et al. 2024).

Among the three existing models of verification and validation servers, the NCHS verification server limits users. In exchange for not adding any noise to the verification results, users may only request results from a range of preselected models. The idea of supplying users with verification metrics from only predetermined models was raised during the focus groups conducted for this project, and some users expressed concern that this would limit the value of the server. Specifically, they were worried that the server would not include the types of models or the specifications they needed for their research, in which case the server would not provide any value. Nevertheless, the NCHS verification is an important use case and has important contribution to the field of setting up a verification and/or validation server. Second, we see a trade-off between automation and privacy. The greater the automation the more need there is for rigorous privacy protections, while less automated systems can allow for more flexibility in the privacy evaluations. For example, the prototype SOI validation server seeks to provide statistics to users while minimizing the need for human review, but the server adds noise to every released statistic to ensure privacy loss is managed. The Cornell Synthetic Data Server and Census Bureau validation processes allow users to request statistics without adding noise, but the results must be approved through the standard disclosure review process, which requires manual review.

Finally, there is a trade-off between privacy and availability. The prototype SOI validation server limits the types of queries users can make to regression models and univariate statistics on a limited set of variables to reduce the amount of noise that must be added to protect privacy. Alternatively, the NCHS server only offers a limited range of statistics on models that have undergone disclosure review, but they are provided with less noise added.

If a verification server is developed alongside a synthetic SED, or if this approach is used with other data within the NSDS, it will need to determine the balance among automation, availability, and privacy. Users will need to understand that a bespoke system that allows them to verify any arbitrary result will only be possible with either a lengthy human review (low automation) or potentially low accuracy statistics (high levels of noise). Such clear and open communication with the users would benefit the users and NSDS in future endeavors of developing and improving a verification server.

Prior to creating a verification system, it is important to establish a clear goal as to whether users should be able to specify models versus select from predetermined models. It is inevitable that the latter approach would be useful for different researchers, so it is necessary to assess whether it is more important to have a system with less administrative management required or one that allows for a more bespoke set of models.

Determining the framework for each dataset will require further discussions and work. This situation stresses the importance of setting clear goals early in the process for releasing synthetic data and developing an accompanying verification or validation server to ensure proper implementation and alignment within a future NSDS.

References

Advisory Committee on Data for Evidence Building. 2022. *Advisory Committee on Data for Evidence Building: Year 2 Report.* Washington, DC: U.S. Bureau of Economic Analysis.

Barrientos, Andrés, Alexander Bolton, Tom Balmat, Jerome P. Reiter, John M. de Figueiredo, Ashwin Machanavajjhala, Yan Chen, Charley Kneifel, and Mark DeLong. 2018. "Providing Access to Confidential Research Data through Synthesis and Verification: An Application to Data on Employees of the US Federal Government." *Annals of Applied Statistics* 12: 1124–56. https://doi.org/10.1214/18-AOAS1194.

Barrientos, Andrés, Jerome Reiter, Ashwin Machanavajjhala, and Yan Chen. 2019. "Differentially Private Significance Tests for Regression Coefficients." *Journal of Computational and Graphical Statistics* 28 (2): 440–53. https://doi.org/10.1080/10618600.2018.1538881.

Barrientos, Andrés F., Aaron R. Williams, Joshua Snoke, and Claire McKay Bowen. "A Feasibility Study of Differentially Private Summary Statistics and Regression Analyses with Evaluations on Administrative and Survey Data." Journal of the American Statistical Association 119, no. 545 (2024): 52-65.

Benedetto, Gary, Jordan Stanley, and Evan Totty. 2018. *The Creation and Use of the SIPP Synthetic Beta v7.0*. Washington DC, US Census Bureau.

Benedetto, Gary, Rolando Rodríguez, Jordan Stanley, and Evan Totty. 2024. "Validation Services for Confidential Data." In Handbook of Sharing Confidential Data, 194–210. New York: Chapman & Hall and CRC Press.

Benschop, Thijs, and Matthew Welch. 2019. "Statistical Disclosure Control for Microdata: A Practice Guide." SDC Practice Guide. International Household Survey Network Revision.

Bonnéry, Daniel, Yi Feng, Angela K. Henneberger, Tessa L. Johnson, Mark Lachowicz, Bess A. Rose, Terry Shaw, Laura M. Stapleton, Michael E. Woolley, and Yating Zheng. 2019. "The Promise and Limitations of Synthetic Data as a Strategy to Expand Access to State-Level Multi-agency Longitudinal Data." *Journal of Research on Educational Effectiveness* 12 (4): 616–47. https://doi.org/10.1080/19345747.2019.1631421.

Bowen, Claire, and Joshua Snoke. 2023. "Do No Harm Guide: Applying Equity Awareness in Data Privacy Methods." Washington, DC: Urban Institute.

Bowen, Claire, Victoria Briant, Leonard Burman, John Czajka, Surachai Khitatrakun, Graham MacDonald, Robert McClelland, Livia Mucciolo, Madeline Pickens, Kyle Ueyama, Aaron Williams, Doug Wissoker, and Noah Zwiefel. 2022. "Synthetic Individual Income Tax Data: Methodology, Utility, and Privacy Implications." Paper presented at International Conference on Privacy in Statistical Databases, Paris, France, September 21–23.

Bowen, Claire, Victoria Bryant, Leonard Burman, Surachai Khitatrakun, Robert McClelland, Philip Stallworth, Kyle Ueyama, and Aaron Williams. 2020. "A Synthetic Supplemental Public Use File of Low-Income Information Return Data: Methodology, Utility, and Privacy Implications." Paper presented at International Conference on Privacy in Statistical Databases, Taragona, Spain, September 23–25.

Census Bureau. 2024. "Guidelines for Requesting Results of Analysis on the Confidential Data" Washington, DC: US Census Bureau.

Census Bureau. 2023. "Validating Results," Washington, DC: US Census Bureau.

Cohen, Aloni, and Kobbi Nissim. 2020. "Towards Formalizing the GDPR's Notion of Singling Out." *Proceedings of the National Academy of Sciences* 117 (15): 8344–52. https://doi.org/10.1073/pnas.1914598117.

Drechsler, Jörg, Daniel Kifer, Jerome Reiter, and Aleksandra Slavković. 2024. *Handbook of Sharing Confidential Data: Differential Privacy, Secure Multiparty Computation, and Synthetic Data*. New York: Chapman & Hall and CRC Press.

Drechsler, Jörg, and Jingchen Hu. 2021 "Synthesizing Geocodes to Facilitate Access to Detailed Geographical Information in Large-Scale Administrative Data." *Journal of Survey Statistics and Methodology* 9: 523–48. https://doi.org/10.1093/jssam/smaa035.

Dwork, Cynthia, Adam Smith, Thomas Steinke, and Jonathan Ullman. 2017. "Exposed! A Survey of Attacks on Private Data." *Annual Review of Statistics and Its Application* 4: 61–84. https://doi.org/10.1146/annurev-statistics-060116-054123.

Dwork, Cynthia, and Aaron Roth. 2014. "The Algorithmic Foundations of Differential Privacy." Foundations and Trends Theoretical Comput. Science 9 (3–4): 211–407. https://doi.org/10.1561/0400000042.

Dwork, Cynthia, and Moni Naor. 2010. "On the Difficulties of Disclosure Prevention in Statistical Databases or the Case for Differential Privacy." *Journal of Privacy and Confidentiality* 2 (1): 93–107. https://doi.org/10.29012/jpc.v2i1.585.

Hundepool, Anco, Josep Domingo-Ferrer, Luisa Franconi, Sarah Giessing, Eric Schulte Nordholt, Keith Spicer, and Peter-Paul De Wolf. 2012. *Statistical Disclosure Control*. Vol. 2. New York: John Wiley & Sons, Ltd.

Karr, Alan, Christine Kohnen, Anna Oganian, Jerome Reiter, and Ashish Sanil. 2006. "A Framework for Evaluating the Utility of Data Altered to Protect Confidentiality." *The American Statistician* 60 (3): 224–32.

Kinney, Satkartar, Jerome Reiter, and Javier Miranda. 2014. "SynLBD 2.0: Improving the Synthetic Longitudinal Business Database." *Statistical Journal of the IAOS* 30 (2):129–35. https://doi.org/10.3233/SJI-140808.

Lin, Tong, and Jerome Reiter. 2024. "Differentially Private Verification of Survey-Weighted Estimates." arXiv preprint arXiv:2404.02519. https://doi.org/10.48550/arXiv.2404.02519.

Machanavajjhala, Ashwin, Daniel Kifer, Johannes Gehrke, and Muthuramakrishnan Venkitasubramaniam. 2007. "l-diversity: Privacy Beyond k-Anonymity." *ACM Transactions on Knowledge Discovery from Data (TKDD)* 1 (1): 3-es. https://doi.org/10.1109/ICDE.2006.1.

McClure, David, and Jerome Reiter. 2012. "Towards Providing Automated Feedback on the Quality of Inferences from Synthetic Datasets." *Journal of Privacy and Confidentiality* 4 (1). https://orcid.org/0000-0002-8374-3832.

National Center for Health Statistics, Division of Analysis and Epidemiology. 2025. *Synthetic Linked 2018 NHISHUD-CMS Data: User Guide*. Washington, DC: Centers for Disease Control and Prevention.

Ohm, Paul. 2009. "Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization." *UCLA Law Review* 57: 1701. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1450006#.

Raab, Gillian, Beata Nowok, and Chris Dibben. 2024. "Practical Privacy Metrics for Synthetic Data." arXiv preprint arXiv:2406.16826.https://doi.org/10.48550/arXiv.2406.16826.

Raghunathan, Trivellore, Jerome Reiter, and Donald Rubin. 2003. "Multiple Imputation for Statistical Disclosure Limitation." *Journal of Official Statistics* 19: 1–16. https://www2.stat.duke.edu/~jerry/Papers/jos03.pdf.

Raghunathan, Trivellore. 2021. "Synthetic Data." *Annual Review of Statistics and Its Application* 8 (1): 129–40. https://doi.org/10.1146/annurev-statistics-040720-031848.

Reiter, Jerome, Anna Oganian, and Alan Karr. 2009. "Verification Servers: Enabling Analysts to Assess the Quality of Inferences from Public Use Data." *Computational Statistics and Data Analysis* 53: 1475–82. https://doi.org/10.1016/j.csda.2008.10.006.

Reiter, Jerome. 2003. "Inference for Partially Synthetic, Public Use Microdata Sets." *Survey Methodology* 29: 181–89. https://scholars.duke.edu/publication/689864.

Reiter, Jerome. 2005. "Using CART to Generate Partially Synthetic, Public Use Microdata." *Journal of Official Statistics* 21: 441–62. https://api.semanticscholar.org/CorpusID:2977015.

Rubin, Donald B. 1993. "Statistical Disclosure Limitation." *Journal of official Statistics* 9 (2): 461–68. https://ecommons.cornell.edu/server/api/core/bitstreams/dd0b63ff-4494-4491-96ba-a69811563dee/content.

Snoke, Joshua, Gillian Raab, Beata Nowok, Chris Dibben, and Aleksandra Slavkovic. 2018. "General and Specific Utility Measures for Synthetic Data." *Journal of the Royal Statistical Society Series A: Statistics in Society* 181 (3): 663–88. https://doi.org/10.48550/arXiv.1604.06651.

Stadler, Theresa, Bristena Oprisanu, and Carmela Troncoso. 2022. "Synthetic Data–Anonymisation Groundhog Day." Presentation given at the 31st USENIX Security Symposium, Boston, MA, August 10–12.

Sweeney, Latanya. 2002 "K-Anonymity: A Model for Protecting Privacy." *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 10 (5): 557–70. https://doi.org/10.1142/S0218488502001648.

Taylor, Silke, Graham MacDonald, Kyle Ueyama, and Claire Bowen. 2021. "A Privacy-Preserving Validation Server Prototype." Washington, DC: Urban Institute.

Tyagi, Erika, Silke Taylor, Graham MacDonald, Deena Tamaroff, Claire McKay Bowen, and Joshua Snoke, forthcoming, "A Privacy-Preserving Validation Server Version 3.0." Washington, DC: Urban Institute.

Tyagi, Erika, Silke Taylor, Graham MacDonald, Deena Tamaroff, Josh Miller, Aaron Williams, and Claire Bowen. 2024. "A Privacy-Preserving Validation Server Version 2.0." Washington, DC: Urban Institute.

Vilhuber, Lars. 2025. "Using Containers to Validate Research on Confidential Data at Scale." *Harvard Data Science Review*. https://doi.org/10.1162/99608f92.4d1853ce.

Yu, Haoyang, and Jerome Reiter. 2018. "Differentially Private Verification of Regression Predictions from Synthetic Data." *Transactions on Data Privacy* 11 (3): 279–97. https://hdl.handle.net/10161/16424.

About the Authors

Claire McKay Bowen is a senior fellow and leads the Data Governance and Privacy Practice Area at the Urban Institute. Her research focuses on developing technical and policy solutions to safely expand access to confidential data for advancing evidence-based policymaking. She is also interested in improving science communication and ensuring everyone is responsibly represented in data. In 2024, she became an American Statistical Association Fellow "for her significant contributions in the field of statistical data privacy, leadership activities in support of the profession, and commitment to mentoring the next generation of statisticians and data scientists." Further, she is also a council member of the Inter-university Consortium for Political and Social Research, an advisory board member of the Future of Privacy Forums, and she is an adjunct professor at Stonehill College.

Jingchen (Monika) Hu is an associate professor of statistics at Binghamton University. Her main research interests are statistical data privacy, where she works on methodology advancement in confidential data dissemination and applied projects on dissemination strategies. Monika also works on building connections between the academia, government, and private sectors through collaborations, society leadership, conferences, and education training. She designs and provides data privacy training to staff members at several government agencies and consulting services on these topics. This work was completed while Monika was an associate professor at Vassar College.

Saki Kinney is a Senior Research Statistician at RTI International with 20 years of experience in applied research, with emphasis on statistical methods in privacy and confidentiality. Her expertise in synthetic data is complemented by extensive practical experience with data dissemination, tiered access, disclosure review processes, and designing and implementing strategies for safely expanding access to data for different federal and state agencies. She also serves as the statistical lead on the SED data production team.

Jeremy Seeman is a research associate in the Data Governance and Privacy Practice Area at the Urban Institute. His research focuses on technical and policy solutions for responsible data sharing, with an emphasis on deploying privacy-enhancing technologies (PETs) to expand access to sensitive data. Jeremy has worked on PETs methodology and deployments for the US Census Bureau, the National Science Foundation, and the Internal Revenue Service. Outside of the Urban Institute, Jeremy is on the steering committees for the National Institute of Statistical Sciences New Researchers Network and the American Statistical Association Privacy and Confidentiality Interest Group. He is an adjunct

professor at the Institute for Social Research and a faculty affiliate of the Center for Ethics, Society, and Computing at the University of Michigan.

Joshua Snoke is a senior fellow in the Massive Data Institute at Georgetown University McCourt School of Public Policy. He has expertise broadly in statistical data privacy and has published technical papers and policy reports on topics such as synthetic datasets, model estimation across multiple databases, and differentially private algorithms. He focuses mainly on evaluating and developing practical applications of privacy-preserving methodologies to identify technologies which increase access to administrative or survey data while maintaining privacy and confidentiality. This work was completed while he was a Statistician at the RAND Corporation.

Aaron R. Williams is a lead data scientist for statistical computing in the Data Governance and Privacy Practice Area at the Urban Institute. He works on data privacy, data imputation, microsimulation modeling, and survey analysis with a focus on income, wealth, tax, and retirement policies. Williams has developed systems for safely releasing administrative data for research, including the Urban-Brookings Tax Policy Center's synthesis of individual tax records and the US Department of Labor's Safe Transfer, Restricted-Use Data Lake. He is a member of the Bureau of Labor Statistics Technical Advisory Committee and an adjunct professor in the McCourt School of Public Policy at Georgetown University.

Statement of Independence

The Urban Institute strives to meet the highest standards of integrity and quality in its research and analyses and in the evidence-based policy recommendations offered by its researchers and experts. We believe that operating consistent with the values of independence, rigor, and transparency is essential to maintaining those standards. As an organization, the Urban Institute does not take positions on issues, but it does empower and support its experts in sharing their own evidence-based views and policy recommendations that have been shaped by scholarship. Funders do not determine our research findings or the insights and recommendations of our experts. Urban scholars and experts are expected to be objective and follow the evidence wherever it may lead.

500 L’Enfant Plaza SW
Washington, DC 20024

www.urban.org