**PROTECTING PRIVACY AND EXPANDING ACCESS IN A MODERN ADMINISTRATIVE TAX DATA SYSTEM**

Leonard Burman, Barry Johnson, Victoria Bryant, Graham MacDonald, Robert McClelland

July 9, 2024

Leonard Burman: Urban Institute, Washington, DC, USA (lburman@urban.org); Barry Johnson: Statistics of Income Division, IRS, Washington, DC, USA (Barry.W.Johnson@irs.gov); Victoria L. Bryant: Statistics of Income Division, IRS, Washington, DC, USA (Victoria.L.Bryant@irs.gov); Graham MacDonald, Urban Institute, Washington, DC, USA (rmcclelland@urban.org); Robert McClelland: Urban Institute, Washington, DC, USA (rmcclelland@urban.org).

ABSTRACT

Providing access to confidential tax data for research while protecting taxpayers’ privacy has grown increasingly difficult using traditional disclosure control methods. For that reason, we have been working to implement two new tools to protect taxpayer privacy and expand access to tax data for research. This paper reports on progress in producing a high-quality synthetic public use file of individual income tax information and a safe way for researchers to perform statistical analysis on confidential tax data without seeing individual records. We discuss modern privacy methods, the advantages and limitations inherent in applying them to tax data, and implementation challenges.

JEL Codes: C15, C18, H24

# Introduction

For several years, the Urban Institute and Tax Policy Center have been working with the Statistics of Income (SOI) Division of the Internal Revenue Service (IRS) to create a modern data access system that expands the use of tax data for research purposes while also protecting those data from growing threats of attacks on public data releases.

An exciting feature of the new system will be two new tools: a fully synthetic public use file (PUF) designed to reflect the distribution of actual individual income tax information while containing no actual individual’s income tax return information, and a validation server that would allow selected researchers to submit statistical programs to execute on confidential tax data with noise added to estimates to protect privacy. The synthetic PUF will replace the traditional PUF, a sample of tax returns altered using legacy statistical disclosure control (SDC) methods that SOI has produced for decades. Unfortunately, growing threats to privacy have made such methods inadequate to protect privacy without completely undermining statistical validity. A fully synthetic PUF will be superior to the traditional PUF in several ways: for example, useful policy variables such as state of residence and business can be safely included in the synthetic PUF, but not the traditional PUF. In addition, researchers will be able to test and debug statistical programs using synthetic data that could be submitted to a validation server that would produce privacy-protected statistical estimates derived directly from confidential data.

The new system will include multiple tiers of access. The first tier is the extensive tabular data and reports that SOI has produced and published for many decades. To supplement these publicly available reports, SOI is introducing a Data Services team that will leverage existing legal authority to produce custom tabulations in support of evidence building. The second tier is the synthetic PUF, designed to represent the statistical properties of income tax microdata without containing actual tax return records. While the synthetic PUF will produce reliable estimates for simple statistics and most microsimulation analyses, its estimates will be less reliable for more complex statistical queries. This is where the third tier – a validation server— comes into play. Authorized users will be able to test and debug programs using the synthetic data, and then submit them to a remote access system to execute on confidential tax data with random noise added to parameter estimates as needed to protect privacy. The fourth tier will allow a select few trusted researchers to access the confidential data directly for purposes that the validation server cannot fulfill. This capability currently exists under SOI’s Joint Statistical Research Program (JSRP), which allows a very small number of researchers to access confidential data for research under the supervision of SOI staff.[[1]](#footnote-1) We anticipate that many researchers will be able to complete projects that would currently be done under the JSRP using the validation server, which will involve a more streamlined and widely accessible process. SOI will grant access to the higher tiers based on the qualifications and needs of researchers and in compliance with the Internal Revenue Code restrictions.

This paper provides an update on progress to produce a useful synthetic PUF and an update on our prototype validation server, capable of producing tabular statistics as well as ordinary least squares and logistic regression analyses.

# Robust privacy protection for tax data

The IRS possesses invaluable administrative data from tax and information returns that could vastly expand our understanding of how tax policies affect behavior and how those policies could be made more effective. However, access to and use of these data is precisely described in Title 26 of the United State Code, better known as the Internal Revenue Code, to prevent the disclosure of information that could be associated with specific taxpayers. SOI’s mission is to produce products that help realize the potential benefits of tax data while protecting taxpayer privacy.

Many users access tax data through tabulations—for example, of income and deductions by filing status or geographic location. SOI publishes voluminous statistics online at [www.irs.gov./statistics](http://www.irs.gov./statistics) and in publications such as the quarterly *SOI Bulletin*.

For decades, SOI also released an annual traditional PUF, based on a privacy-protected, representative sample of about 200,000 individual income tax returns. Each file was statistically weighted to approximate the population of tax returns filed for a specified tax year, which is generally the calendar year during which income is earned. For example, income earned during 2012 would be classified as Tax Year 2012 income and would be reported on tax returns filed with the IRS during the calendar year 2013 tax filing season. The traditional PUF included about 200 data items, a fraction of the total information reported on Form 1040 and attached schedules. Although based on actual tax return information, data in the traditional PUF were protected using a range of SDC techniques, including rounding, data coarsening, aggregating, data swapping, and data suppression. Among the suppressed items were personal identifiers and most demographic data, including address information, although these would be of keen interest for studying important tax policy questions. Data for tax returns containing extreme positive or negative values for any amount variable are aggregated into one of four composite records. In the 2015 PUF, 1,124 returns were aggregated (Bryant 2022).

In the last decade, how institutions collect, store, and disseminate information has drastically changed due to increased computational power and more sophisticated technologies. These technological advances make legacy SDC methods like de-identification, sampling, coarsening, and swapping less effective. In 2019, computer scientists from the Imperial College London and Université Catholique de Louvain examined simple de-identification and sampling. They estimated that 99.98% of Americans in an anonymized health dataset could be uniquely identified with 15 detailed attributes that included ZIP Code, date of birth, gender, and number of children (Rocher et al. 2019). Concerned about computational advances, the United States Census Bureau conducted a simulated attack on the 2010 Decennial Census, which used coarsening, sampling, and swapping to protect against disclosure. They discovered they could re-identify one in six of the U.S. population using publicly available data from sources such as Facebook (Leclerc, 2019). These examples show that re-identification attacks are a serious risk for datasets with both detailed geographic information and detailed categorical variables.

Awareness of the growing threats to public-use microdata, and a general concern for protecting individual taxpayer’s data privacy, led SOI to progressively restrict and distort more and more information being released in the traditional PUF. This made traditional PUF data much less comprehensive—and thus less useful for microsimulation modeling and policy analysis. For example, recent PUFs have excluded most detail about small business income, making it difficult to model the effects of business tax reforms. In addition, the data were modified in ways, such as grouping many extreme records into four aggregate records and “blurring” certain sensitive variable values (Bryant 2022), that can undermine statistical validity. Continuing to remove or distort more variables in response to growing threats would further undermine the traditional PUF’s already limited usefulness.

But several organizations, including the American Enterprise Institute, the Urban-Brookings Tax Policy Center, the Tax Foundation, the Institute on Taxation and Economic Policy, and the National Bureau of Economic Research have developed PUF-based microsimulation models that help inform the public on potential impacts of policy proposals. SOI intends the current PUF for tax year 2015 to be the last traditional PUF so it is imperative that a new tool be developed to continue to make such analyses possible.

##  A modern approach to privacy

SOI’s modernized approach to privacy follows the recommendations of the Advisory Committee on Data for Evidence Building (ACDEB, 2023).[[2]](#footnote-2) ACDEB recommended a tiered access approach where analysts have access to the data they need in a way that protects privacy.

For many users, that access will continue to be through extensive tabulations that are annually updated. As noted, SOI is also working to improve its capacity to produce custom aggregations, on a reimbursable basis, to support evidence building, with an initial focus on meeting the needs of federal, state, local and tribal government agencies. This approach will ideally focus on a set of standard outputs for which major components of the production process can be automated to reduce SOI resources required to generate the needed statistics and speed delivery (O’Hara et al. 2024).

Recognizing that the legacy approach to producing public-use data has become obsolete, both in terms of privacy protection and the utility of the data provided to users, SOI is working with the Urban Institute to create fully synthetic public-use files. These, while informed by the characteristics of the confidential data and applicable tax laws, do not contain any actual tax return records. SOI plans to distribute these synthetic files free of charge making them widely available. The traditional PUF came with a fee and had very limited distribution.

For researchers who need access to the confidential data to perform statistical analyses that cannot be performed on public-use files, there will be a validation server—a mechanism to derive statistical estimates from the confidential data, with noise added to estimates to protect privacy. Finally, for analyses not possible using the validation server, some vetted and trusted researchers might gain access to the confidential microdata through a process similar to the current JSRP. Table 1 summarizes the 4 target access tiers.

When fully implemented, we anticipate that accessing the confidential tiers will be accomplished as part of the [Standard Application Process](https://ncses.nsf.gov/about/standard-application-process) (SAP) established by the Office of Management and Budget’s Interagency Committee on Statistical Policy and currently hosted by the National Center for Science and Engineering Statistics at the National Science Foundation (NSF). The SAP is a way for researchers to easily gain access to confidential data products, potentially from multiple agencies, to support a research project. From the NSF web portal, “Applications with the SAP must be for a statistical purpose and will be reviewed by the statistical agency with ownership of the data. Applicants can use the SAP to apply for access to data from multiple agencies for the same project and track the application as it moves through the review process.”[[3]](#footnote-3)

**Table 1. Tiers of Access to Tax Data**

|  |  |  |
| --- | --- | --- |
| Tier | Access | To Whom |
| 1 | Tabular data and reports | Anybody – via website and published reports |
| 2 | Synthetic individual income tax return data | Anybody who needs it – upon request to SOI  |
| 3 | Validation server, an automated system that allows researchers to access confidential tax return information in an environment that protects against disclosure | Researchers vetted by SOI with a research plan that could not be completed using tier 1 or tier 2 access. |
| 4 | Access to confidential microdata | Researchers vetted and trusted by SOI who must access microdata; for example, to merge datasets, perform data cleaning on raw data files. or conduct more complex analyses not supported by the validation server.  |
|  |  |  |

##  Privacy concepts

While not a discourse on formal privacy methods, this paper necessarily draws on some of the concepts and terminology from that literature. We provide a very high-level overview of some of them here.[[4]](#footnote-4)

The revolutionary innovation in data privacy is the development of **formal privacy** methods, which provide a provable guarantee of privacy risk for a released statistic that is not dependent on assumptions about the data or technology an intruder might possess now or in the future. (An **intruder**is another bit of jargon, referring to someone who would like to infer confidential information from published data or statistics.)

**Differential privacy** is one formal privacy criterion. Simply put, it says that statistics should change very little when any record is added or subtracted from the database used to derive the statistic. **Pure differential privacy** requires that, after application of a “sanitization algorithm,” the ratio of probabilities of any particular value for a statistic generated with and without the inclusion of any record must be less than a constant, exp(ε), where ε is chosen by the data custodian. Put more simply, released statistics cannot change much even if the most extreme records are added or subtracted from a dataset. This is sometimes called **ε-differential privacy**. To meet the differential privacy criterion, random noise is added to released statistics. The parameter, ε, encapsulates the trade-off between **utility** and privacy. As ε approaches zero, the released statistic has no connection to the underlying data (e.g., it is all noise). As ε approaches infinity, statistics may be released with no privacy protection (i.e., the unaltered statistics derived from the confidential data).[[5]](#footnote-5)

A key feature of any formally private definition is that it is possible to calculate the cumulative privacy loss of statistics derived from a particular database. For example, if *k* differentially private statistics are generated with privacy loss ε on the same part of the data, the cumulative privacy loss of all the statistics is *k*ε. If *k* statistics are generated with privacy loss ε on disjoint parts of the data, the cumulative privacy loss of those statistics is just ε.[[6]](#footnote-6) This accounting allows data stewards to set and enforce **privacy budgets**, both for individual users and for all the statistics released from a particular dataset. When the individual’s privacy budget is exhausted, they may not receive additional statistics. When the total privacy budget for a dataset is reached, no additional statistics are produced.[[7]](#footnote-7)

In many cases, the original formal privacy definition, differential privacy, makes unrealistic assumptions about the severity of the privacy threat, which may be too restrictive and thus requires too much noise be added to statistics. For that reason, relaxations to differential privacy have been developed, such as (ε,δ)-differential privacy, that are formally private algorithms but require less noise.

Implementing formal privacy methods often requires an assessment of the **global sensitivity** of statistics with respect to the presence or absence of the most extreme outliers—not only the most extreme values in the current data, but the most extreme possible values that might arise in the universe of all possible versions of data. Global sensitivity is often hard to assess *a priori*. For example, what is the largest income or loss that might be reported on an income tax return in a given year?

For that reason, privacy researchers may apply a more limited range to the possible values, called **local or observed sensitivity**. They are based on sensitivity of statistics to the presence or absence of any actual value in the underlying data. (Chetty et al. 2019) These methods require less noise infusion and are often easier to implement and more flexible in practice than methods based on global sensitivity. But because they depend on the actual range of the data, they are by definition not formally private.

The synthetic datasets and validation server discussed below are also not formally private. The methodology used to create synthetic tax data are more able to reflect the constraints and complexities of the US income tax system and include less noise (i.e., are more reliable) than formally private synthetic data would be. The methodology for creating a validation server is informed by the logic of formal privacy, but it is based on observed sensitivity and thus does not come with a formal privacy guarantee.

For that reason, the privacy budget enforced for statistics released by the validation server also does not have a formal privacy guarantee. Rather, it is an approximation designed to limit disclosure risk (as well as p-hacking and data mining).

#  Synthetic data

Work continues to produce the full synthetic PUF, the first public release of which will be for Tax Year 2016, although development has focused on earlier tax years to facilitate comparison with legacy PUF outcomes and refine and improve the synthesis methodology. Bowen et al. (2022) and Bowen et al. (2024) reported on early efforts to create synthetic data for tax year 2012.

The goal in data synthesis is to produce an empirical model of a statistical distribution function that looks like the universe of tax returns. The challenge is that the distribution is highly skewed and irregular. Synthetic data are drawn from an empirical representation of a joint distribution function that represents the multivariate relationships among variables to be included in the PUF. To generate this empirical distribution, we start with a subsample of the confidential Individual and Sole Proprietor (INSOLE) dataset, a cleaned stratified sample of individual income tax returns, created by SOI for policy analysis. Unlike the INSOLE data, but like the traditional PUF, our source dataset samples all income strata at rates less than 100 percent.

We then use a nonparametric machine-learning model called classification and regression trees (CART) to generate pseudo-tax-records that look like tax return information but do not exactly reproduce the information on any particular tax return. CART uses predictor variables (X) to sort observations of an outcome variable (Y) into relatively homogeneous groups and then randomly draws a value from the relevant group. This provides a nonparametric estimate of the value of Y given X. Because this data-driven method is more flexible than parametric approaches, such as regression-based models, it can account for unusual variable distributions and nonlinear relationships that can be hard to identify and model explicitly. CART tends to outperform regression-based parametric methods (Goldstein et al. 2020; Drechsler and Hu 2020). CART is also computationally tractable.

Variables are synthesized sequentially, and each synthesized variable is a function of the variable synthesized before it. For additional privacy protection, noise is added so that the synthesized values do not exactly match the confidential data values. More noise is added in sparse parts of the empirical distribution (for example, the largest reported capital gains). Tax data create a host of challenges. The Urban Institute team has created new synthesis tools that enhance and extend the standard CART algorithm to incorporate the constraints and anomalies in tax data in the synthesis process.[[8]](#footnote-8)

While the full synthetic PUF is still in the development and testing process, SOI has already produced and released a nonfiler synthetic public-use data set or Supplemental Synthetic PUF (Bowen et al. 2020). This file uses a sample of information returns provided to IRS from third party payers, including Forms W-2 documenting wages and salaries, and Forms 1099 documenting payments from banks, brokerages, and retirement accounts, to construct a synthetic file representing individuals who did not file an income tax return and whose income is estimated to be below the income tax filing threshold in effect for each tax year. SOI has released the Supplemental Synthetic PUF for tax year 2012 and is in the process of developing files for more recent years. The 2012 Supplemental Synthetic PUF was the first public release of nonfiler data in any form. It provides information critical for representing the lower tail of the income distribution.

##  Advantages of synthetic data

The process of synthesis provides a safer and more systematic privacy protection than other SDC methods. The combination of the CART method, sampling, and noise-infusion provides systematic protection against disclosure (Bowen et al. 2024). As a result, there are several advantages of a synthetic PUF over the traditional PUF.

First, some variables can be more accurately represented in the synthetic PUF than in the traditional PUF. Bowen et al. (2022) showed that a number of variables in a Tax Year 2012 synthetic PUF were closer to the confidential data distribution than in the traditional PUF. Importantly, the synthetic PUF does not need to contain aggregate records to pool information from tax returns with very large positive or negative incomes. Values in the tails of the distribution must have a lot of noise added because of the heightened risk of disclosure, but the synthetic data more accurately represent the relationships between variables on very high-income returns than aggregate records in the traditional PUF.

Second, more data may be included in a synthetic PUF. For example, the traditional PUF has omitted information about state of residence for many years because of disclosure concerns, but the synthetic PUF can safely include geographic information. The synthetic PUF can also include details about business income and deductions that are suppressed on the traditional PUF. Some information that is limited in the traditional PUF can be safely included in the synthetic PUF with no or fewer restrictions. For example, capital losses are only reported subject to the $3,000 limit on immediate deduction in the traditional PUF but capital loss before limitation may be included in the synthetic PUF. And some of the voluminous data reported to the IRS on information returns can be safely included in the synthetic PUF. SOI will determine the exact list of variables to be included in the synthetic PUF in consultation with the user community.

Third, as SOI builds its capacity to create synthetic datasets, the lag time between completion of the public-use files and creation of the confidential data file that is used by SOI, the Joint Committee on Taxation (JCT), and Treasury can be significantly reduced. The most recent traditional PUF is for tax year 2015, six years behind the most recent confidential data for tax year 2021. More timely release would make the synthetic PUF more useful for researchers and analysts to examine recent policy developments. It will also ease the process of extrapolating tax data forward in time to simulate future effects of tax changes, which is necessary to produce revenue and distributional analyses of policy proposals.

Fourth, over time, more synthetic datasets could be safely produced and released. It would be possible to produce a synthetic dataset of select information from tax returns with more detailed geographic information. It might even be possible to produce a synthetic file that includes imputed indicators for race and ethnicity.

In addition, SOI staff create valuable datasets derived from unpublished data based on information returns and other sources as part of special studies. Those datasets are unavailable to researchers outside IRS and a few government agencies, but synthetic versions of the data might be safely produced and disseminated. Outside analysts could use the synthetic data to calculate simple statistics such as means, sums, and counts. Perhaps more importantly, the synthetic data could be used as the basis for developing econometric analysis to be conducted on the confidential data using the validation server.

## Limitations of synthetic data

Like the traditional PUF, statistical analyses performed on synthetic data may produce statistically invalid results, especially for analyses that are outside the purposes for which the data were designed. We continue to improve the synthetic PUF. Our goal is for it to produce accurate tabular estimates, simple correlations between most pairs of variables, and to generally produce reliable results from microsimulation analysis (except for those that are heavily dependent on outlier observations). They will not produce reliable estimates for models that rely on kinks, notches, and discontinuities in the underlying data because the synthesis process effectively smooths out those anomalies (Abowd and Schmutte, 2015).

Note that no single summary measure of the quality of a synthetic dataset exists, in part, because its usefulness depends on the purpose to which the data are applied. Bowen, et al. (2022) address this by reporting a set of utility measures for the 2012 synthetic PUF focusing on specific applications. For example, comparing means, counts, and correlations is useful to assess the value of the dataset for producing tabular data. The confidence interval overlap test indicates how likely it is that the results from specific ordinary least squares regressions generated using the synthetic data would be statistically equivalent to those based on the confidential data. Since the PUF is often used for microsimulation modeling, the synthetic data must produce estimates close to those generated using the administrative data. An advantage of this criterion is that there is a finite set of actual enacted policies that may be used to assess synthesis quality. We reported tests for some of these in Bowen (2022) and plan to explore many more, some with the help of beta testers (see below), before completion of the 2016 synthetic PUF.

Another issue is the tradeoff between privacy and utility. The theoretical literature on privacy often shows a privacy-utility frontier that is analogous to the production-possibility frontier in economics. The end points are zero privacy and maximal utility (original unaltered data) and maximum privacy and zero utility (perfectly safe, but completely uninformative). An advantage of using differentially private methods is that the single privacy parameter, ε, makes the tradeoff between privacy and utility relatively simple to visualize. (See II.B.) But when releasing multiple sanitized statistics, there is no unique tradeoff between privacy and utility because the utility from releasing any one depends on the utility of the statistics already released. When creating synthetic data, the number of potential tradeoffs is much larger. Moreover, the non-formally private methods we use involve many separate steps designed to protect privacy. We are currently exploring how the decisions made at each step affect the tradeoff between privacy and utility and will report on this in future research. Some kinds of data are less amenable to producing reliable synthetic data. For example, a synthetic file with information at the zip code level would have to be very noisy to prevent disclosure of information about residents of areas with small populations. The challenges of protecting privacy are magnified in panel data that follows individuals over time (Mitra et al. 2020).[[9]](#footnote-9) That is why SOI decided decades ago to stop creating public use tax panels. A panel of individual income tax returns would require much more noise than a cross-section file. A synthetic dataset of corporate income tax data would also have to be extremely noisy to protect privacy in highly concentrated industries.

However, even very noisy synthetic datasets could be useful as dummy datasets designed to allow researchers to program and test statistical queries that could be inputs to a validation server and for other training purposes.

## Timeline for completion of synthetic datasets

SOI and the Urban Institute plan to complete a synthetic PUF of tax year 2015 data by fall of 2024. SOI will share that file with some current PUF users and interested government agencies to assess the quality of the data and their usefulness for particular tasks such as microsimulation modeling. Based on that feedback, we may refine the synthesis algorithm and produce a second version of the synthetic PUF for another round of testing. The beta testers will then delete the synthetic datasets from their computers because allowing both a synthetic and traditional PUF to coexist would present an unnecessary disclosure risk.

Starting with tax year 2016 data, SOI will produce and release only a synthetic PUF. That file will reflect everything we have learned from internal development and testing and the experience and feedback of beta testers. We expect that synthetic PUFs for later tax years will be completed and released on a much faster schedule than the traditional PUF. Supplement PUF files will be released concurrently with new synthetic PUFs.

# A validation server for administrative tax data

An automated validation server is a digital tool that creates an intermediate layer between a researcher and the confidential data.[[10]](#footnote-10) With this intermediate layer, a researcher can analyze the confidential data without actually seeing them. To use a validation server, a researcher would first develop analysis code using synthetic data, then apply for access to the server. Once accepted, the researcher would submit their analysis code to the validation server, which would return a table of results from the analysis with noise added to protect privacy.[[11]](#footnote-11)

##  Current capabilities

The Urban Institute has built a prototype validation server that can perform a wide range of analyses, including ordinary least square and logistic regressions, tabular counts, and summary statistics. The prototype validation server is not yet publicly available. Statistics have noise added to protect privacy based on the maximum observed sensitivity of parameter estimates to the addition or subtraction of observations from underlying data. (Chetty at al. 2019) A researcher could submit a logistic regression analysis to run on the validation server and receive noise-infused, statistically valid, coefficients that would be suitable to publish in a refereed journal. A researcher might also query the confidential data to tabulate noise-infused means to use as targets to calibrate a microsimulation model, or to produce a table for a report.

The noisy results produced by the validation server are statistically valid if the underlying analysis is, but somewhat less precise than the estimates derived from directly using the confidential data. The validation server provides more robust privacy protection than manual review of statistical releases at less cost to federal agencies and in a way that could ultimately be completely automated. And they are an ideal complement to synthetic data since they allow researchers to test hypotheses generated using the synthetic data.

##  Advantages of a validation server

Traditionally, access to government administrative data is granted either as publicly available data via various websites and data portals, as part of an agreement with a contractor for a specific piece of work, or via employment or contract that requires a lengthy security clearance process. In contrast, in time, an automated validation server that does not permit a user to directly access confidential data could significantly reduce barriers to analyses on confidential government data by introducing a simpler, streamlined application and approval process. In addition, the validation server could in principle be fully automated. This would reduce the role of senior SOI staff in manually reviewing every statistic for possible disclosure concerns. Manual reviews are time-intensive and costly and may fail to detect some privacy risks that would be avoided by the validation server algorithm.

Another advantage of the validation server is that many more researchers will be able to gain access to tax data with relatively modest demands on SOI staff. Researchers will not need to travel potentially long distances to an IRS office or secure data enclaves like the Federal Statistical Research Data Center (FSRDC). This means users who live in remote areas or don’t have access to FSRDCs will have the same access to this invaluable data resource as researchers at more remote or less well-endowed institutions.[[12]](#footnote-12)

## Limitations

The prototype validation server allows for a wide range of analyses, but not every analysis is possible. While researchers may preprocess the data as much as they wish using a wide range of program libraries before running a regression or tabular analysis, they are restricted to a set of simple statistics and regression and machine learning models that do not violate the constraints of the privacy-preserving algorithm that is implemented on the validation server.

For now, that means that certain weighted tabulations are not possible, though we hope to enable their use in future releases. This also means that certain multi-stage analysis techniques, such as two-stage least squares and instrumental variables regressions, may only reveal to the researcher the result of the final stage. The addition of noise to results also means that results that may have otherwise summed neatly to a total may not add up as expected.

The validation server also restricts the accuracy and amount of data researchers can release through a privacy budget, so unlimited queries of highly accurate statistics are not permitted.

A privacy budget is necessary to prevent excessive leakage of information from statistical queries. Each user is allocated a privacy budget, ε (see discussion in II.B) representing the total information loss allowed for all the analyses that may be published by that user.[[13]](#footnote-13) The interface shows the user how the amount of noise added varies with the portion of their privacy budget they choose to allocate to a particular analysis so analysts can choose between producing a small number of more precise statistics versus a larger number of noisier estimates. Researchers may also be permitted a larger privacy budget for analyses that will not be released, for example, as part of sensitivity testing subject to the proviso that the additional estimates not be published.

Although, in the tiered access system, researchers permitted access to the validation server are vetted and trusted to abide by the rules, setting a privacy budget for unreleased statistics serves as kind of a failsafe system to prevent abuse, and allows for the agency to choose to automate the review process in the future. The underlying philosophy embodies the Russian proverb, “trust, but verify.” The concern is that a bad actor who somehow gained access to the system could run repeated identical or similar analyses to derive a very precise measure of a statistic undermining the noise infusion process. The privacy budget prevents that.

The privacy budget will require analysts to make choices, much as a financial budget limits what and how much a shopper may put in their grocery cart. Researchers will have to be more thoughtful about how they plan out research projects. By design, data mining or p-hacking would quickly exhaust the privacy budget. This could encourage better research practice and would bolster the integrity of analyses deriving from a validation server, but it would also require a change in behavior for analysts accustomed to easily running hundreds of statistical analyses for exploratory purposes and selecting the one most consistent with their priors (Snoke et al. 2023; Leamer 1983).

## The current state of the validation server

The Urban Institute team completed the current prototype version of the validation server in February of 2024, which uses public Current Population Survey data from the Census Bureau as a stand-in for confidential data. The team is currently conducting outreach and soliciting feedback from early users, working on privacy-protection improvements, and hardening the security of the system to enable it to meet strict government requirements.

### Plans for development

We envision that, in the future, a researcher will be able to apply for access to confidential data via a validation server, be quickly vetted and approved by the relevant agencies, and be granted access to analyze a dataset directly through an agency or via a mechanism established by a future National Secure Data Service (NSDS).[[14]](#footnote-14) Once approved, users would gain access to a secure environment where they could conduct analyses on the validation server. Researchers would have developed and debugged code before accessing the server using synthetic data generated from the confidential data in the manner described earlier in this paper. Researchers would be able to spend a predesignated privacy budget to view and publicly release targeted statistics with as little or as much noise as they choose, within the boundary of that budget constraint.

A validation server could be designed to expand access to administrative tax data in new and important ways. For example, microsimulation models could be designed to run in the validation server interface, guaranteeing that model simulation results are statistically valid, even if noisy. In the context of the individual income tax, such a model could fact-check the results produced by a microsimulation model based on the synthetic PUF. And a corporate income tax model might run through a validation server. Corporate income tax microdata could probably never be publicly released in any useful form because of the high risk of disclosure about large firms in highly concentrated industries. However, a corporate tax model running in a validation server might produce reliable estimates for certain kinds of queries, such as estimating the revenue effect of changing depreciation rules or the research and experimentation tax credit. Moreover, while a synthetic corporate income tax microdata file would be dominated by noise, statistical programs developed using the synthetic data and submitted to the validation server could produce relatively accurate statistical estimates of the behavioral response to broad provisions such as changes in corporate income tax rates.

Despite the advances in the current validation server prototype, many challenges and limitations remain for future work. Those include: 1) How to design error messages that do not reveal sensitive information while still providing useful feedback that will allow researchers to effectively debug code on the confidential data; 2) Ensuring that the correct amount of noise is added for a given privacy budget for more complex statistical calculations; 3) Speeding up more complex, time intensive analyses on big data without compromising privacy; 4) Improving the privacy algorithm to allow for breakthroughs in privacy-preserving techniques while still maintaining a simple interface that makes it easy to interpret and interact with the privacy budget. (Taylor et al. forthcoming)

# Need for user education

ACDEB (2022) highlighted the need for educating users about data availability and the use of new tools such as synthetic data and validation servers. They proposed setting up a new concierge service that would help match users to data and point to the resources needed. There will also be a need for education to help users effectively use new tiers of access, which are expressly intended to provide just enough data to address a given need. Convincing users that they do not always need direct access to individual microdata and can instead find their answer in synthetic data or even tabular data will require a paradigm shift that must be supported across the research ecosystem. Students will also need to be trained to align their research practices with the privacy budget limitations described above. New statistical methods may also need to be developed to support the use of more restrictive tiers of access, such as tabular data, for causal analyses. All researchers will need to be more rigorous in accounting for non-sampling error from SDC measures when interpreting results.

The need for education is not new, but the impacts of legacy SDC methods may not have been as apparent or as quantifiable as newer, formally private methods. Some users of the traditional PUF may not be aware of its limitations, imposed as they were in a somewhat ad hoc manner as part of the SDC process. Traditional PUF data are distorted in ways that are hard to control for in statistical analysis, especially for those relying on parts of the distribution that are heavily altered such as very high-income taxpayers.

Users of the synthetic PUF and other synthetic data products will be warned that these data are not samples of tax returns but, rather, statistical replicas designed to look like the underlying data. As noted above, analysis based on the synthetic PUF could, in some cases, produce much different conclusions from the confidential microdata. For that reason, researchers should use the validation server before publishing or otherwise relying on statistical estimates derived from tax data.

In exchange for significantly democratizing access, the validation server will impose limits on researchers, who will need to hone new skills to effectively work within them. At least initially, it will require them to use a particular programming interface with a limited range of functions. Researchers will have to carefully test and debug their statistical programs before submitting them to the validation server to avoid exhausting their finite privacy budget. In addition, while the system will allow for an additional privacy budget to perform sensitivity analyses in response to feedback from reviewers and referees, the number of analyses will be limited. That said, some researchers may welcome a reason to tell referees that unlimited sensitivity tests are not permissible.

# Challenges in calibrating and applying formal privacy methods

The research discussed here is at the cutting edge of the relatively new field of formal privacy research. Differential privacy is only 18 years old, which is still young for a new mathematical concept. Perhaps unsurprisingly, there are unresolved issues.

One issue is how to define privacy and implement formal privacy mechanisms for real-world applications. Much has been written about differential privacy, a mathematically elegant approach to defining privacy. It is elegant in the sense that it produces a quantifiable privacy guarantee, but it builds on the extreme assumption that a potential intruder has every observation but one in a confidential dataset. Differential privacy and other formal privacy definition assumptions sometimes require a large amount of noise to be added to even simple statistics such as means, and may even produce bias (Williams and Bowen, 2023). Moreover, Barrientos et al. (2024) found that formally private methods perform well with a modest privacy budget for summary statistics, but formally private regression methods require much larger privacy-loss budgets—or an unacceptably large amount of noise infusion Moreover, the authors found that most formally private regression methods typically do not provide the associated standard errors (ones that are also formally private) for the coefficients, which makes hypothesis testing impossible.

For that reason, we use a local or observed sensitivity version of differential privacy in developing the validation server methodology—an extension of the methodology based on maximum observed sensitivity first described in Chetty and Friedman (2019). As noted in Section II.B, formally private methods often rely on the concept of global sensitivity or how robust the target statistic is to outliers, which helps determine the amount of noise that needs to be added to ensure privacy. Calculating the sensitivity based on the observed data instead of the universe of possible datasets is more flexible, but calculating the local or observed sensitivity is very computationally demanding. We are researching methods that could apply to complex estimators in large datasets.

An additional complication in applying differentially private, other formally private methods, or related methods, such as maximum observed sensitivity, is that many statistics published by government agencies are derived directly from confidential tax data with no infusion of noise to protect privacy, meaning that a portion of the user base for tax data implicitly has an infinite privacy budget. For example, JCT revenue estimators must often produce estimates under extremely tight deadline pressure. It would be infeasible to run their estimates through a validation server or some other mechanism designed to protect privacy. But there is also a question about whether blurring estimates would be necessary to protect privacy. JCT rounds their estimates and suppresses those based on small numbers of observations, which are both traditional SDC methods. In addition, the process of extrapolating tax data to future budget years alters the data substantially. Then the data are processed in very complex microsimulation models or in large spreadsheets that include many unpublished assumptions about key parameters. The process results in a kind of noise infusion that likely makes it nearly impossible to infer the underlying confidential administrative data.

However, this means that a validation server system would never have a completely accurate accounting of total draws on the privacy budget for all published statistics.

It should also be noted that the process of building statistical models for estimation substantially complicates the relationship between most published statistics and the underlying data. The complexity in statistical models adds an additional layer of privacy protection that is ignored in formal privacy models. The basic concern in formal privacy is whether an intruder could effectively glean the function that generates the data and calculate its inverse, conditional on every observation but one in a dataset, thus allowing inference about the one missing observation. Except for very simple statistics (e.g., sums, counts), calculating that inverse function would be extremely challenging if not impossible.

Formal privacy models do not account for the cost to an intruder of inferring information about particular data records. The concept of differential privacy assumes an intruder has almost infinite time available to hack the protection algorithm and infer a protected piece of information (Dwork and Roth 2014). In practice, the value of information declines quickly with time. Inferring an item from a celebrity’s tax return might be valuable now or next year, but not in a hundred, or a thousand years. An intruder is also assumed to have almost unlimited computing power, but super-computer time is expensive. A rational hacker would compare the costs of attacking a protected database against the benefits and conclude that even a small amount of noise added to statistics would make the cost far exceed the benefits. Even an irrational hacker has a finite lifespan, not “polynomial time” so would likely die before drawing a useful inference. Thus, even a modest amount of noise infusion could provide much more protection than implied in privacy models that ignore costs.

Finally, the privacy loss budget depends on the tradeoff between utility and privacy. The privacy literature offers little guidance about how data stewards should assess that tradeoff and set the budget. All these factors suggest a relaxed approach towards privacy budgets, both for individual projects, and for the total privacy loss permissible for a given dataset (before it is shut down). At least for the present, the main goal should be to prevent a bad actor from undermining privacy protections while allowing good faith researchers to derive the statistically valid estimates they need to educate the public and inform policymaking.

# Conclusions

The Commission on Evidence-Based Policymaking (2017) envisioned “a future in which rigorous evidence is created efficiently, as a routine part of government operations and used to construct effective public policy,” The Commission believed it was possible to increase the use of data without increasing privacy risk though the use of ”[m]odern technology and statistical methods.” In introducing a new data access system, SOI is attempting to realize the Commission’s vision.

Data used in tax administration offer particular promise for informing policies that benefit our nation and its citizens. SOI’s JSRP has produced important insights on the benefits and limitations of tax policies to lift individuals and families from poverty, impacts of taxes on business formation, and the challenges economic shocks impose of different groups. However, the JSRP access model is costly to support and therefore difficult to scale. In order to increase the use of tax data for evidence-building, new access methods are required. This paper has focused on two new tiers of access: synthetic data on individual income tax filers and a validation server that will enable users of the synthetic data to confirm findings based on the synthetic data. Developing these tools has required the application and improvement of new statistical methods that are proving the Commission’s optimism correct. When fully implemented, synthetic data paired with a validation server will greatly expand the safe, secure use of tax data by both traditional and new users and in doing so, will help to “revolutionize how government uses and protects the data it collects” (Commission on Evidence-Based Policymaking 2017). While we have made great progress in applying modern privacy methods to tax data, there is a lot yet to be done. A key goal is to further improve the quality of the synthetic data. In future research we will report on how different steps taken in the design of the synthetic data affect the utility and privacy of the data based on various metrics. This information will allow production of a safer and more useful dataset for policy analysis. A focus will be to improve the quality of the dataset for microsimulation modeling, which is the primary use to which the traditional PUF has been applied. We plan to develop a set of policy experiments that allow us to stress-test the synthetic PUF to provide confidence that simulations will produce estimates close to those generated by official scorekeepers using the administrative data. We are confident that the synthetic PUF will be at least as successful in that crucial role as the traditional PUF for reasons outlined in section III.A.

In addition, a significant challenge is how to expand the prototype validation server to be able to handle more complex statistical queries in large datasets. This is a high priority for future research.

Finally, we invite the tax research community to engage with SOI about the new data access system and to assist in setting priorities for future research and development. For example, what additional data items would be most useful for synthetic PUF users, which additional synthetic datasets should be the highest priority, and what are the most important capabilities to add to a validation server? We also welcome feedback on how best to educate the research community about the new tools, which would help to ensure that these tools can be best used to inform tax policy. Comments and suggestions may be emailed to sis@irs.gov.

ACKNOWLEDGEMENTS AND DISCLAIMERS

Claire McKay Bowen, Joshua Snoke, and Aaron R. Williams provided helpful comments. The work described in this paper would not have been possible without generous support from the Urban Institute’s funders, including Arnold Ventures, the Alfred P. Sloan Foundation, the Internal Revenue Service, and the National Science Foundation. The findings and conclusions are those of the authors and do not necessarily reflect positions or policies of the Internal Revenue Service, the Urban Institute, or its funders.

# References

Advisory Committee on Data for Evidence Building. 2021. “Advisory Committee on Data for Evidence Building: Year 1 Report.” <https://www.bea.gov/system/files/2021-10/acdeb-year-1-report.pdf>.

Advisory Committee on Data for Evidence Building. 2022. “Advisory Committee on Data for Evidence Building: Year 2 Report.” <https://www.bea.gov/system/files/2022-10/acdeb-year-2-report.pdf>.

Abowd, John M., and Ian M. Schmutte. 2015. “Economic Analysis and Statistical Disclosure Limitation.” *Brookings Papers on Economic Activity* 46 (1): 221–93.

Barrientos, Andrés F., Aaron R. Williams, Joshua Snoke, and Claire McKay Bowen. 2024. “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(545): 52-65..

Bowen, Claire, Leonard E. Burman, Robert McClelland, and Aaron R. Williams. 2024. “Safe Data Technologies: Safely Expanding Access to Administrative Tax Data.” in Jörg Drechsler, Daniel Kifer, Jerome Reiter, Aleksandra Slavkovic, eds. *Handbook of Sharing Confidential Data: Differential Privacy, Secure Multiparty Computation, and Synthetic Data* (Chapman & Hall/CRC Handbooks of Modern Statistical Methods)

Bowen, Claire, Victoria L. Bryant, Leonard E. Burman, Surachai Khitatrakun, Graham MacDonald, Robert McClelland, Philip Stallworth, Kyle Ueyama, Aaron R. Williams, and Noah Zwiefel. 2020. “A Synthetic Supplemental Public-Use File of Low-Income Information Return Data: Methodology, Utility, and Privacy Implications.” Tax Policy Center, Urban Institute. https://www.taxpolicycenter.org/publications/synthetic-supplemental-public-use-file-low-income-information-return-data-methodology

Bowen, Claire McKay, Victoria Bryant, Leonard Burman, Surachai Khitatrakun, Livia Mucciolo, Madeline Pickens, and Aaron R. Williams. 2022. “Synthetic Individual Income Tax Data: Promises and Challenges.” *National Tax Journal*, *75*(4): 767-790.

Bowen, Claire McKay and Simson Garfinkel. 2021. “Philosophy of Differential Privacy.” *Notices of the American Mathematical Society*, 68: 1727–1739.

Bryant, Victoria. 2022. “General Description Booklet for the 2015 Public Use Tax File.” Individual Statistics Branch, Statistics of Income Division, IRS. Updated February 2023.

Chetty, Raj, and John N. Friedman. 2019. “A Practical Method to Reduce Privacy Loss When Disclosing Statistics Based on Small Samples.” *AEA Papers and Proceedings*, 109: 414-420**.**

Commission on Evidence-Based Policymaking. 2017. “The Promise of Evidence-Based Policy Making.”

Drechsler, Jörg, and Jingchen Hu. "Synthesizing Geocodes to Facilitate Access to Detailed Geographical Information in Large-Scale Administrative Data." *Journal of Survey Statistics and Methodology*. 2020.

Dwork, Cynthia, and Aaron Roth. 2014. “The Algorithmic Foundations of Differential Privacy.” *Foundations and Trends® in Theoretical Computer Science.* *9*(3–4): 211-407.

Goldstein, Ross, Michael E. Woolley, Laura M. Stapleton, Daniel Bonnéry, Mark Lachowicz, Terry V. Shaw, Angela K. Henneberger, Tessa L. Johnson, and Yi Feng. 2020. "Expanding MLDS Data Access and Research Capacity with Synthetic Data Sets."

Hotz, V. Joseph, Christopher R. Bollinger, Tatiana Komarova, Charles F. Manski, Robert A. Moffitt, Denis Nekipelov, Aaron Sojourner, and Bruce D. Spencer. 2022. “Balancing Data Privacy and Usability in the Federal Statistical System.” *Proceedings of the National Academy of Sciences*, 119(31): e2104906119.

Leamer, Edward E. 1983. "Let's Take the Con out of Econometrics." *The American Economic Review* 73(1): 31-43.

Mitra, Robin, Stephanie Blanchard, Iain Dove, Caroline Tudor, and Keith Spicer. 2020. “Confidentiality Challenges in Releasing Longitudinally Linked Data.” *Transactions on Data Privacy* 13: 151–170.

National Academies of Sciences, Engineering, and Medicine. 2024. *A Roadmap for Disclosure Avoidance in the Survey of Income and Program Participation*. Washington, DC: The National Academies Press. https://doi.org/10.17226/27169.

National Academies of Sciences, Engineering, and Medicine. 2024. A Roadmap for Disclosure Avoidance in the Survey of Income and Program Participation. Washington, DC: The National Academies Press. https://doi.org/10.17226/27169.

O'Hara, Amy, Stephanie Straus, Ron Borzekowski, and Barry Johnson. 2024. “Exploring Improved Access to Individual Income Tax Data for Evidence-Building.” Paper at NBER, *Data Privacy Protection and the Conduct of Applied Research: Methods, Approaches and their Consequences, Spring 2024.*

Ruggles, Steven, and David Van Riper. 2022. “The Role of Chance in the Census Bureau Database Reconstruction Experiment.” *Population Research and Policy Review*, 41(3): 781–788.

Snoke, Joshua, Claire McKay Bowen, Aaron R. Williams, and Andrés F. Barrientos. 2023. “Incompatibilities Between Current Practices in Statistical Data Analysis and Differential Privacy.” *arXiv preprint arXiv* *2309. 16703.*

Taylor, Silke, Josh Miller, Erika Tyagi, Deena Tamaroff, Clayton Seraphin, Graham MacDonald, Claire McKay Bowen, and Aaron R. Williams. Forthcoming. “Building an Automated Validation Server Prototype,” *Submitted paper for the Privacy and Public Policy Conference*.

Williams, Aaron R., and Claire McKay Bowen. 2023. “The Promise and Limitations of Formal Privacy.” *WIREs Computational Statistics*, e1615. https://doi.org/10.1002/wics.1615

1. See https://www.irs.gov/statistics/soi-tax-stats-joint-statistical-research-program. [↑](#footnote-ref-1)
2. The Evidence-Based Policymaking Act of 2018 instructed the director of the Office of Management and Budget to create the advisory committee “to review, analyze, and make recommendations on how to promote the use of Federal data for evidence building.” (Public Law 115–435) Two coauthors, Leonard Burman and Barry Johnson, were members of the ACDEB. [↑](#footnote-ref-2)
3. https://ncses.nsf.gov/about/standard-application-process. [↑](#footnote-ref-3)
4. For an accessible discussion of data privacy and its limitations, see Williams and Bowen (2023). [↑](#footnote-ref-4)
5. The College Scorecard, which presents [↑](#footnote-ref-5)
6. The results follow from the sequential composition theorem and the parallel composition theorem, respectively (Bowen and Garfinkel 2021). [↑](#footnote-ref-6)
7. Hotz et al. (2022) have criticized this rule because it ignores the value of the additional statistic. They argue that a cost-benefit assessment should guide when additional formally private statistics may be released. [↑](#footnote-ref-7)
8. See <https://www.rstudio.com/conference/2022/talks/tidysynthesis-r-package/>. [↑](#footnote-ref-8)
9. The U.S. Census Bureau has produced a synthetic longitudinal business database, but nobody to our knowledge has attempted to produce a panel dataset as large and complex as a panel of individual income tax returns would be. [↑](#footnote-ref-9)
10. The U.S. Census Bureau developed a manual validation server that allowed researchers to submit statistical programs executed on a synthetic version of the Survey of Income and Program Participation to run on the confidential data. Census staff reviewed the program and output and returned the statistical results if they did not identify any privacy risks. The review process was slow and labor-intensive, taking a week or more. Census shut down the server on September 30, 2022 (National Academies of Sciences, Engineering, and Medicine 2024) and it is still dormant, pending review of “…various technical, funding, and process-related issues.” (See <https://www.census.gov/programs-surveys/sipp/guidance/sipp-synthetic-beta-data-product.html>.) [↑](#footnote-ref-10)
11. MacDonald, G., “New Tools Are Needed to Unlock Private Data for Better Policymaking,” *Data@Urban*. February 15, 2024. <https://medium.com/urban-institute/new-tools-are-needed-to-unlock-private-data-for-better-policymaking-64a088565b2b> [↑](#footnote-ref-11)
12. While Census is permitting access to some FSRDC data remotely (see <https://www.census.gov/about/adrm/fsrdc/about/secure-remote-access.html>), confidential income tax data are not currently part of that program. [↑](#footnote-ref-12)
13. As noted in section II.B, the maximum observed sensitivity algorithm we use is not formally private. Thus, ε is not the same as the ε in a differentially private system. But the basic concept is the same. [↑](#footnote-ref-13)
14. The 2016 Commission on Evidence-Based Policymaking recommended the creation of a NSDS. According to ACDEB (2021), the NSDS should “facilitate data access, enable data linkages, and develop privacy-enhancing techniques in support of increasing data for evidence building across the entire evidence-building ecosystem.” ACDEB (2022) fleshed out the concept with many specific recommendations. 2022 CHIPS and Science Act authorized a demonstration project under the auspices of the National Science Foundation National Center for Science and Engineering Statistics. See <https://ncses.nsf.gov/about/national-secure-data-service-demo>. [↑](#footnote-ref-14)