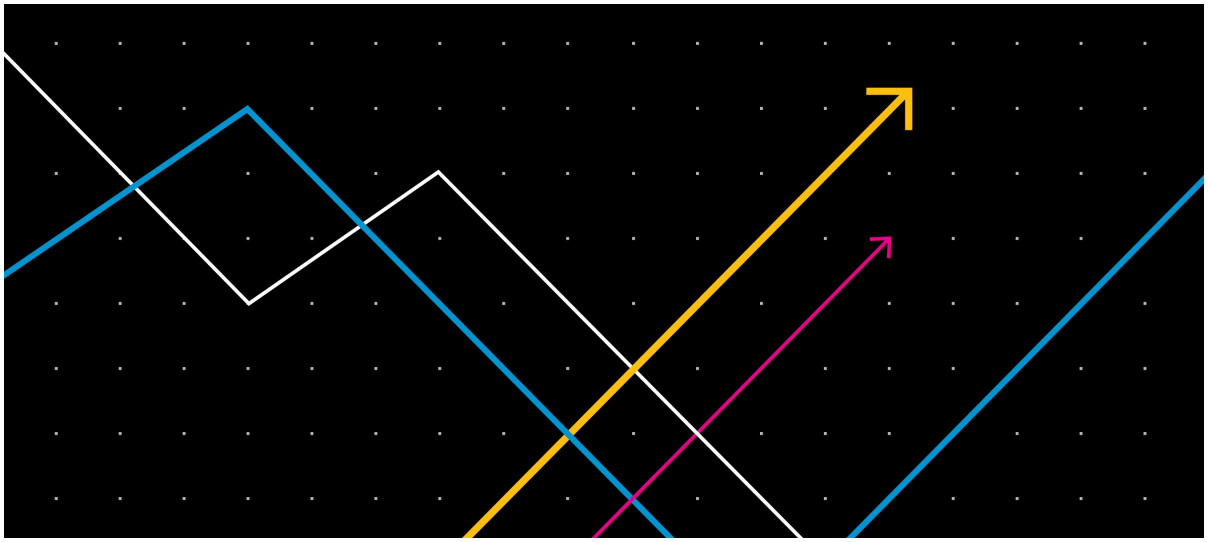


RESEARCH REPORT

Opt-In Statistical Disclosure Protections

Empowering Survey Respondents to Improve Data Quality

Aaron R. Williams

Jennifer Andre

September 2023

URBAN INSTITUTE · ELEVATE · THE · DEBATE

Table of contents

Executive Summary

Opt-In Statistical Disclosure Protections

Introduction

People generate and share data about themselves every day when they browse the web, interact with government services, and respond to questionnaires, and they should be empowered to make decisions about how these data are accessed and used. Currently, disclosure protection policies at the US Census Bureau do not allow for such empowerment – respondents to questionnaires like the decennial census and the American Community Survey are all subjected to disclosure protections. Data for all respondents are, by default, masked with some form of statistical disclosure control, and those who may wish to see themselves accurately reflected in the data are unable to do so. Statistical disclosure control may inflict greater damage to the accuracy of information for certain groups, such as smaller race and ethnicity groups, relative to others.

In this brief, we explore a new framework for disclosure protections that would require respondents to actively opt in to disclosure protections. Responses for those who opt in would be treated with disclosure protections, while responses for those who forego protections would remain unchanged in statistical product outputs. We present two demonstration studies, the first using an opt-in local differential privacy approach for the decennial census, and the second using an opt-in synthetic data approach for the American Community Survey. In both cases, we seek to explore the impact of varying the rate of opting in to disclosure protections on data quality and the associated privacy consequences. We especially examine the impact for small racial/ethnic groups, including the impact on quality and privacy if some groups opt in at higher rates than others.

We aim to test the feasibility of this potential solution path, contributing to ongoing public discussions and debate about disclosure protections and public data quality involving researchers, public data users, and other stakeholders. This solution would have wide-ranging implications, including operational changes, new outreach strategies, and many complex legal questions about Title 13 and other regulations. The findings we present here provide early evidence on the impact of turning privacy disclosure choices over to participants.

Background

The primary goal of this report is to present the opt in privacy framework and early evidence on the impact of such a framework on two demonstrations. The data privacy literature is extensive, and we assume a baseline knowledge of certain key concepts. In this section, we provide a brief overview of a few key

concepts, as well as a review of disclosure avoidance at the US Census Bureau.

Data Privacy Key Terms

Statistical Disclosure Control Methods

Statistical Disclosure Control (SDC) methods are used to release sensitive or confidential data products while preserving the confidentiality of the data. Traditional SDC methods include suppression, rounding, top- and bottom-coding, and synthetic data generation.

Differential Privacy

Differential Privacy (DP) is a formal definition of privacy, meaning that DP methods meet certain mathematical properties and guarantees. With DP, it is possible to quantify the worst-case disclosure risk of a data release using a "privacy-loss budget." The privacy-loss budget is quantified with parameters like and .

Utility-Privacy Trade-off

When applying disclosure avoidance methods to confidential data, there exists a central tension between the usefulness or quality of the resulting "noisy" data and the amount of privacy risk. Before any noise infusion, confidential data have the highest possible utility, but also have high privacy risks. Efforts to improve disclosure protections via SDC methods can reduce these privacy risks, but also may worsen overall data quality and usefulness for intended analyses or other applications. With DP methods, it is possible to "tune" the balance between utility and privacy by changing the value of , with lower values of corresponding to greater noise infusion, implying lower data quality and higher disclosure protection. With methods that are not formally private, the utility-privacy trade-off is ad hoc.

The US Census Bureau and Disclosure Avoidance

The US Census Bureau is tasked with providing high quality data about the US and its people. These data are of enormous consequence for the public, serving as the basis for political representation, community funding and planning, and key research. Given these use cases, the accuracy and quality of Census Bureau products is crucial.

Decennial census statistical products are used for congressional apportionment, redistricting, federal funding allocations, planning and decision-making for government and business organizations, and informing many other surveys (Mather and Scommegna 2019). The American Community Survey (ACS) is used to inform federal policymaking and program delivery, state and local service provision (e.g., roads and schools), and research and analysis by nongovernmental organizations (United States Census Bureau 2017). In fiscal year 2015, 132 federal programs used Census Bureau data to allocate more than $675 billion in funds to state

and local communities (Hotchkiss and Phelan 2017). Decennial census and ACS data are also foundational to racial equity analytics, enabling researchers to answer important research and policy questions (Axelrod, Ramos, and Bullied 2022).

In addition to conducting surveys and releasing high quality public data, the Census Bureau is also obligated to protect the confidentiality of individual respondents reflected in these data products. The Census Bureau's approach to safeguarding the identities of respondents in publicly released data is informed by its interpretation of Section 9 of Title 13 of the US Code, enacted in 1954. This approach has evolved over the years, especially in response to advances in computing technologies and attack methods (Hotz and Salvo 2022). In 2018, the Census Bureau announced its intention to "modernize how we protect respondent confidentiality," including the adoption of Differential Privacy (DP) (J. M. Abowd 2018). This move was motivated by certain benefits of DP over traditional disclosure limitation, including more robust protections and greater transparency (J. Abowd et al. 2022).

For the 2020 Decennial Census, the Census Bureau updated their Disclosure Avoidance System (DAS) from traditional swapping algorithms to the TopDown Algorithm (TDA), which refers to a system of DP mechanisms for privacy loss accounting, along with optimization algorithms and post-processing. The TDA satisfies zero-concentrated DP (-zCDP), a relaxation of pure DP. As a formally private method, the privacy protections can be quantified, and the Census Bureau does translate the -zCDP privacy parameter to the corresponding values of and (Bowen, Williams, and Pickens 2022; J. Abowd et al. 2022).

In contrast, the Census Bureau has conceded that the "science does not yet exist to comprehensively implement a formally private solution for the ACS" (Daily 2022). Instead, they are currently exploring the feasibility of a fully synthetic public-use microdata file and accompanying validation server. In both cases, all respondents are subjected to disclosure protections, even those who might otherwise prefer to see their data accurately reflected.

Opt-In Privacy Framework and Implications

In our demonstrations, we imagine a framework for disclosure protection in which respondents would be asked to actively opt into statistical disclosure control methods. Those who do not opt-in would simply contribute their true data to statistical products. The choices we make to build this framework may have significant ethical and legal implications, discussed in this section.

Opt In Framework Ethical Implications

The adoption of a framework requiring respondents to opt in to disclosure protections has various ethical considerations. First, a major challenge upon implementation of such a framework would be a significant knowledge gap for respondents. The Census Bureau or any other organization would need to carry out extensive outreach efforts and design plain-language explanations to ensure that respondents understand what they are or are not opting into. It is crucial for respondents to understand the implications of their opt-in choice not only for the privacy of their personal information, but also for the resulting data quality and the downstream impact on their community.

Second, there are ethical considerations in the framing of the opt-in decision and the administration of the question in a survey. For this project, we intentionally chose language of "opting in" to disclosure avoidance protections, treating the decision to forego protections as the default and requiring an additional step for those desiring additional protections. This is in contrast to language of "opting out", which would treat disclosure avoidance protections as the default position. Behavioral economists and psychologists have explored these systems of "explicit consent" (opt-in) and "presumed consent" (opt-out) in policy interventions (Thaler and Sunstein 2009), and some have found that individuals are less likely to take the additional step to opt in to a system (Johnson and Goldstein 2003; Thaler 2009). We intentionally define an opt-in framework, thereby assuming lower rates of opt-in to privacy protections.

Additionally, we made the simplifying assumption that the primary survey respondent, or householder, makes the opt in decision for all members of the household. This choice may be appropriate if the householder is the parent or legal guardian of a minor in the household, but may not be appropriate for other adults in the household who may disagree with the respondent's opt-in decision. We also only allow for unit opt-in disclosure protection (a single opt-in decision for all questions for a given respondent) and do not consider item opt-in disclosure protection (making a separate opt-in decision for each respondent for each question).

Finally, any application of an opt in framework to disclosure avoidance methods must consider the impact of one respondent choosing to forego protections of the privacy risks of respondents who do opt in. The mathematical guarantees of differential privacy allow for individuals to forego protections without decreasing other respondents' protection, but this is not guaranteed in non-formally private methods like synthetic data generation.

Opt In Framework Legal Implications

In addition to ethical considerations, the implementation of an opt in disclosure avoidance framework would also have significant legal implications. Most notably, the implementation of this type of framework would most likely require a change to Title 13 of the US Code, which could trigger complex legal arguments and

potential politicization.

Demonstration 1: Local Differential Privacy for the Decennial Census

The Census Bureau deployment of DP for the 2020 Decennial Census uses a global approach in which tabulated cells of confidential responses in a series of data tables are infused with noise by a central curator (the Census Bureau). All census respondents are automatically subjected to the DAS, even those who might otherwise wish to see their data reflected accurately, without any noise. Further, the noise that is injected into tabulated data cells is independent of the size of the population in the cell. In effect, there is more relative error added for small groups than for larger ones. This could lead to worse data quality for small groups such as some racial/ethnic groups and, as a result of inaccurate representation in the data, these groups could receive inadequate funding and incorrect research findings.

Global and Local Differential Privacy

To allow for individual-level opt-in, we move from a central DP approach, in which the Census Bureau as a data curator would add noise to all respondents, to a local DP approach, allowing for some respondents to opt in and others to forego disclosure protections. Typically, a local model assumes that a central curator cannot be trusted, and so a respondent adds noise to their data before sending it to the curator. For this use case, we can imagine a slight variation on this approach in which the trusted Census Bureau still receives all data in its confidential form, and then infuses noise only for respondents who opt in.

Many local DP mechanisms are based on the concept of Randomized Response, first proposed by Warner (1965). The central idea is that a survey respondent flips a coin, and the result of the coin flips determines if they answer a yes/no question with the true answer or not. The randomization of the coin flip infuses the noise that grants disclosure protections (Near and Abuah 2022).

We use Generalized Random Response (GRR) for our use case, allowing us to move from a binary coin flip to a setting with higher cardinality. With GRR, we turn the entire domain of potential responses into a histogram and randomly switch observations based on a rate determined by the privacy loss budget, . We then tabulate a resulting histogram of counts and apply an adjustment to account for the randomly perturbed responses (Wang et al. 2020). With GRR, individuals who opt in to disclosure protections report a true response with probability $p = \frac{e}{e+d1}$, where d is the overall cardinality of possible response values. Otherwise, the record is randomly replaced with another combination of fields.

We also deploy a global DP approach in order to compare our local model results. We use a Laplace

sanitizer, a global method in which cells are infused with noise from a Laplace distribution. The Laplace distribution is centered at zero and the variability is the ratio of the privacy loss budget, , over the l_1-global sensitivity of the statistic (Dwork et al. 2006; Williams and Bowen 2023). Though the TDA deployed by the Census Bureau is a more complicated series of algorithms and processing steps than the simple Laplace mechanism presented here, this simplification allows us to compare more easily with a simple local approach and focus specifically on the impact of the opt in framework on data quality.

Data

For this demonstration, we use person-level records from the 2010 Decennial Census Stateside Public Use Microdata Sample. This sample contains records representing 10 percent of housing units, and the people residing in them, along with 10 percent of people living in group quarters. We restrict our sample to Washington, DC and Iowa because the former has two large racial/ethnic groups, while the latter is more homogeneous. These data contain demographic and household characteristics about respondents, including age, race, ethnicity, and sex.

Simulations

For our simulation approach, we run iterations of a disclosure mechanism to generate noisy histograms of counts for a set of defined attributes. The specifications for our simulations are as follows. For each combination of specifications, we run 100 iterations of each disclosure mechanism to compare simulation-to-simulation variation.

- Scenarios: the set of grouping attributes for the resulting histogram frequencies
  - Scenario 1 (cardinality $= 2$)
    - Hispanicity: Hispanic or Latino, Not Hispanic or Latino
  - Scenario 2 (cardinality $= 24$)
    - Age bucket: Child (0-17), Adult (18-64), Senior (65+)
    - Race/Ethnicity: White alone, Black or African American alone, Other alone, or Hispanic or Latino (any race)
    - Sex: Male, Female
- Privacy loss budget,
  - 1
  - 5
  - 10
  - 20

- Opt-in rate: the probability that respondents opt in to disclosure protections
  - 0.01
  - 0.1
  - 0.5
  - 0.9
  - 1

Evaluation

We evaluate the results of these simulations using bias and accuracy metrics, comparing the noisy histograms generated under each privacy approach to each other and to the true values. In this report, we use mean percent error to evaluate bias, or the tendency for noisy estimates to systematically move in one direction relative to the true values, and absolute mean percent error to evaluate accuracy, or the closeness of noisy estimates to the true values. Complete results with all bias and accuracy metrics listed below are available in the Urban Institute Data Catalog.

- Bias
  - Mean error
  - Mean percent error
  - Median error
  - Median percent error
- Accuracy
  - Mean absolute error
  - Mean absolute percent error
  - Median absolute error
  - Median absolute percent error
  - Root Mean Square Error
  - Coefficient of Variation
  - Percent Difference Thresholds (percent of tabulations with percent error exceeding threshold of 10%)

Results and Discussion

Local DP Methods Result in Overall Lower Accuracy than Global Methods

For Scenario 1, we focus primarily on results allowing us to compare the performance of the local GRR method to the central Laplace method. With a cardinality of just 2 (Hispanic or Latino, Not Hispanic or

Latino), this scenario allows for the most similar comparison with the global method (in which is allocated to just one statistic).

Figure 1 shows the distribution of bias metrics from our simulations at the defined levels of and opt in rate. Both the local and central methods are unbiased, with the distribution of mean percent error values centered around zero.

FIGURE1

Local and Central DP Approaches are Similarly Unbiased

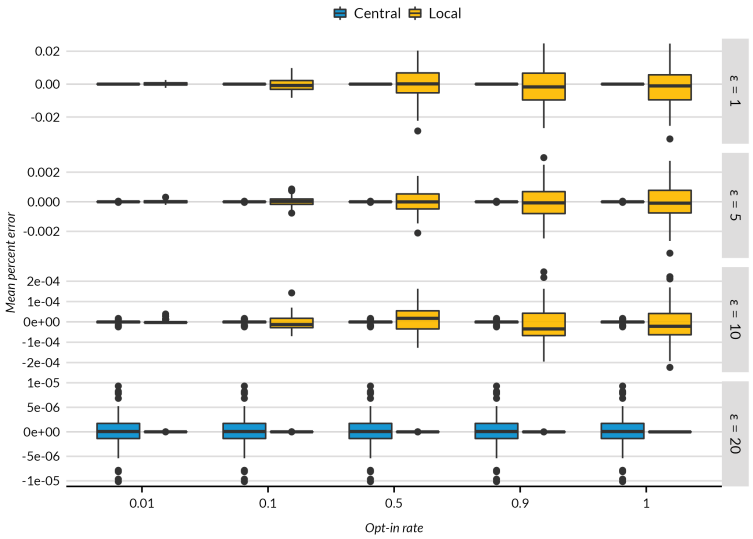


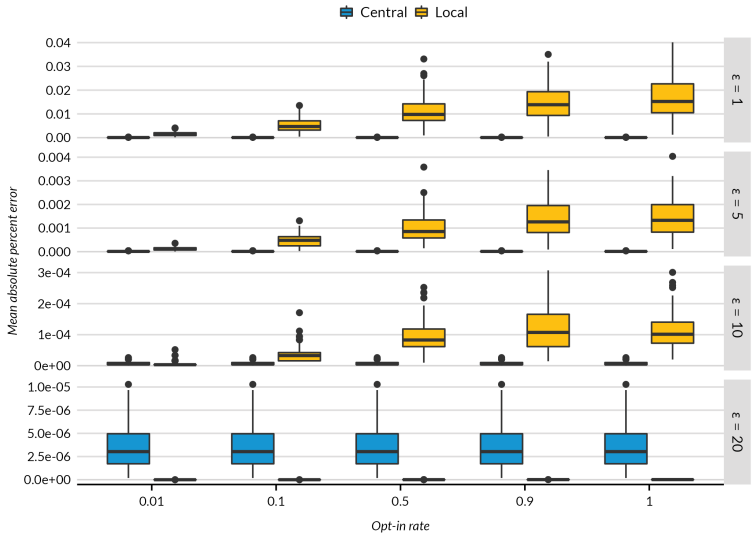
Figure 2 shows the distribution of accuracy metrics from our simulations at the defined levels of and opt in rate, demonstrating two key takeaways about the accuracy of these methods. First, the opt-in framework approach does improve the overall accuracy of the local GRR method. As we decrease the level of opt in, the width of the mean percent error distribution shrinks and moves closer to zero. However, the second key takeaway is that the central method significantly outperforms the local method in terms of accuracy at nearly every tested level of and opt in rate, even with very small opt in rates. The local method only outperforms the central method with a very high privacy loss budget of = 20, and the errors for both methods are very small for that level of privacy loss anyway.

While the opt-in local approach does improve the accuracy of estimates with lower levels of opt in, this improvement alone is unfortunately not enough to justify a switch from a central model to a local model. Existing local DP methods cannot offer the same level of accuracy as central methods, especially for datasets with even higher cardinality. However, the potential to improve data quality results with a local method and opt-in framework motivates greater focus on developing potentially more powerful local DP methods in the future.

Opt-in Privacy Offers the Potential to Improve Data Quality for Small Groups

FIGURE2

Local Method Outperforms Central Method Only with Very High Privacy Loss Budget



Although existing local DP methods may be disappointing for overall accuracy, an opt-in local DP framework still offers the potential to improve data quality for small groups. Data quality may be especially improved for groups that opt-in at relatively lower rates than others. For Scenario 2, we focus on results allowing us to compare differences in data quality by racial/ethnic group.

Figures 3 and 4 show the distribution of accuracy results for the specified opt in rates for each racial/ethnic group (using $= 1$), separately for Washington, DC and Iowa.

FIGURE3

Accuracy By Racial/Ethnic Group - Washington, DC

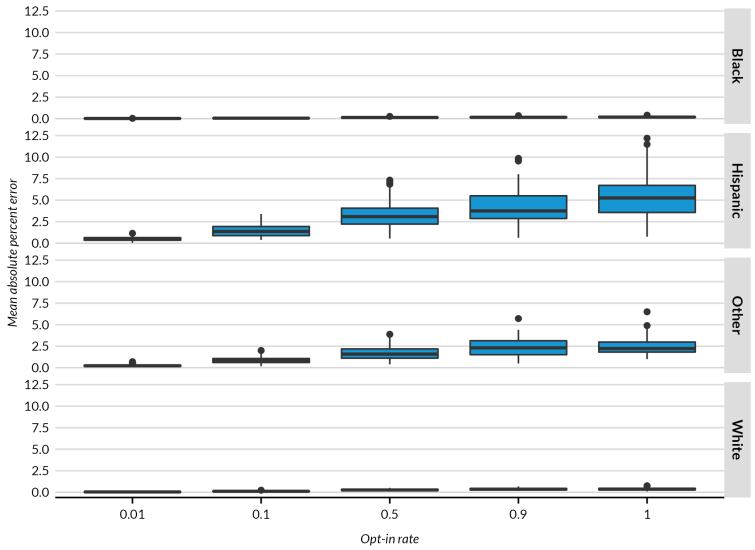
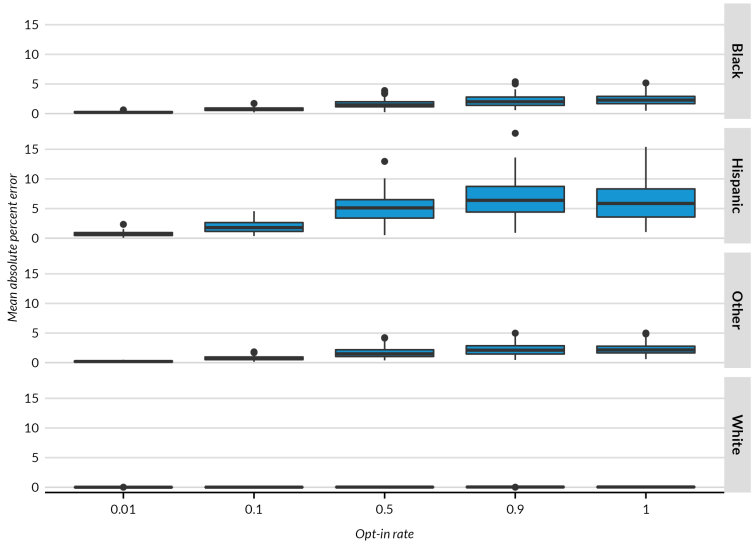


FIGURE4
Accuracy By Racial/Ethnic Group - Iowa



According to the 2010 Census Redistricting Data (Public Law 94-171) Summary File, the population of Washington DC was 38% white, non-Hispanic and 51% Black, non-Hispanic, and the population of Iowa was 91% white, non-Hispanic. For both states, mean percent error is smallest for these relatively large groups, reflecting the larger sample sizes. Error tends to be relatively larger, and with larger spreads, for the smaller groups in both places.

The opt-in framework offers a solution path to improve data accuracy for these smaller racial/ethnic groups. For example, the median absolute percent error for the Hispanic group is about 5.3% in Washington, DC and 5.9% in Iowa when there is 100% opt in, or when all respondents are subjected to disclosure protections. These error values shrink to about 1.4% and 1.8%, respectively, with a Hispanic group opt-in rate of 10%. Given the properties of formal privacy, the privacy protections afforded to those who opt-in are unaffected by those who choose to forego protections.

With an opt-in disclosure framework, the US Census Bureau and community groups could engage in outreach efforts, especially to smaller groups, to help respondents understand the implications of foregoing disclosure protections, both for their privacy but also for the wide-ranging impacts of improving their data quality. This type of outreach could result in better data quality for these groups, with positive downstream impacts on representation and funding allocations to communities.

All in all, existing local DP methods generate protected data of overall lower accuracy than data generated by central models. However, this demonstration shows the potential of local DP to improve data accuracy for small groups while still protecting privacy with an opt-in DP framework. This use case motivates further development of local DP methods and opt in experimentation to improve accuracy results relative to central

models.

Demonstration 2: Synthetic Data for the American Community Survey

The Census Bureau is investigating the creation of a fully synthetic American Community Survey paired with a validation server[1]. The Census Bureau intended to create formally private data by 2025 but later conceded that the science does not exist yet to comprehensively implement a formally private solution for the ACS[2]. Next, the Census Bureau intended to release a non-formally private ACS in 2024 but their timeline has been delayed because of legitimate concerns about the impact of synthetic data.[3]

Synthetic data could potentially harm the usefulness of the American Community Survey. Groups with fewer observations, including smaller racial and ethnic groups, could see the biggest losses in data quality. An opt-in disclosure protection methodology paired with synthetic data could mitigate the harms of data synthesis and give outreach organizations a tool to improve data quality.

Synthetic Data

Demonstration 1 focused on summary statistics calculated on microdata. We now pivot to a demonstration where the goal is to produce high-quality microdata that can be used for a range of valid analyses. We will pursue this goal with synthetic data generation.

Synthetic data generation is a statistical disclosure control method that replaces confidential microdata with pseudo microdata that can maintain the statistical properties of the confidential data while limiting disclosure risks. United Nations (2022) and Hu and Bowen (2023) offer thorough introductions to synthetic data. We will briefly introduce topics central to this demonstration.

There are two main flavors of synthetic data. With partially synthetic data, some but not all variables are synthesized (Little 1993). With fully synthetic data, all variables are synthesized (Rubin 1993). Fully synthetic data generally provides stronger disclosure protection.

Partially synthetic data maintains a one-to-one mapping between observations in the synthetic data and observations in the GSDS. This creates identity disclosure risks and increases attribute disclosure risks. It also means it is possible to calculate disclosure metrics based on re-identification (J. P. Reiter and Mitra 2009). Fully synthetic data don't have a one-to-one mapping because every record is fully generated. This minimizes identity disclosure risks and attribute disclosure risks, but dramatically reduces approaches for evaluating disclosure risks. We create a fully synthetic version of the ACS for this demonstration.

It is possible to create formally private synthetic data in idealized situations (Bowen and Snoke 2021). Model-based approaches generally require discretizing categorical variables. Promising approaches generate 1-, 2-, and 3-way marginals and then use graphical models to generate synthetic data (McKenna, Sheldon, and Miklau 2019). Other approaches use GANs but generally don't work well outside of idealized situations with modest privacy budgets (Tao et al. 2021). Like the Census Bureau, we abandon formal privacy for this demonstration because it is currently infeasible for the ACS. This means we no longer have a provable bound on the worst-case privacy loss.

We use a fully conditional specification (FCS) to generate synthetic data. FCS uses a sequential approach to model the joint distribution of the ACS as a sequence of univariate conditional distributions. We use non-parametric decision trees and regression trees because they are easy to fit and model relatively complex distributions with ease (J. Reiter 2005). We modify the decision trees and regression trees so that predictions are samples from the final nodes instead of using the means and modes of the nodes. This increases the sample variances in the synthetic data.

We use the tidysynthesis[4] and syntheval R packages to synthesize data and evaluate synthetic data.

Data

For this demonstration, we use person-level records from the 2021 1-Year American Community Survey (ACS) to implement an opt-in disclosure protection demonstration using synthetic data. The ACS is a roughly 1-in-100 sample of households in the US and is the premier source of information for small area estimation. The 2021 ACS Public Use Microdata Sample (PUMS) contains about 3.25 million individuals and 1.44 million households, and we access these ACS data through IPUMS (Ruggles et al., n.d.).

We restrict the data in five important ways to minimize computation and simplify comparisons. First, we restrict our data to observations from Florida, Michigan, and Pennsylvania. We chose these states because they are large states with different types of racial and ethnic diversity. Second, we restrict our data to heads of households ages 18 or older who are not in group quarters in order to limit structurally missing values. Furthermore, synthesizing relationships within households is a major technical challenge (Benedetto and Totty 2020).

Third, we synthesize only the following subset of variables.

- State (categorical)
- Sex (binary)
- Age (numeric)
- Marital status (categorical)
- Race (categorical)

- Hispanicity (categorical)
- Any health insurance coverage (binary)
- Educational attainment (categorical)
- Employment status (categorical)
- Labor force participation (binary)
- Total family income (numeric)
- Total personal income (numeric)
- Wage and salary income (numeric)
- Welfare income (numeric)
- Income residual (numeric)

The income residual is personal income minus wages and salary and welfare income. The income residual captures all other sources of income including business income; Social Security income; SSI income; interest, dividend, and rental income; retirement income; and other income.

Fourth, we ignore weights during synthesis and when calculating metrics. Synthesis with weights is an open area of research. Fifth and finally, we randomly partition the data into two halves. The half we apply disclosure protection to is the gold standard data set (GSDS). The other half is a holdout data set (HDS) for calculating disclosure metrics.

PUMS include statistical disclosure limitation to protect the confidentiality of responses. For example, high incomes are top-coded. The Census Bureau has more observations and unaltered variables that could improve the quality of the synthesis but also worsen utility metrics because our metrics do not reflect the current SDC techniques like sampling and top-coding.

Simulations

We implement a simulation study to evaluate the benefits and costs of opt-in disclosure protection in synthetic data generation. We vary two key parameters for each specification - the overall opt-in rate and the potentially differential nature of the opt-in.

First, we vary the opt-in rate with the values 0.1, 0.5, 0.9, and 1. For example, 90% of observations are unaltered and 10% of observations are synthesized when the opt-in rate is set to 0.1. All observations are synthesized when the opt-in rate is 1. The opt-in decision for each observation is stochastic. We use a function to set the probability of opt-in and then randomly select the opt-in decision for an individual record based on that probability. We call the result of this SDC approach "candidate data".

Second, we vary the nature of the opt-in. Under the default settings, all observations have the same opt-in probability within a specification. It's plausible that different racial and ethnic groups would opt into

disclosure protections at different rates, so we vary individual opt-in propensities based on race/ethnicity. In addition to equal opt-in probabilities, we implement a white multiplier such that white individuals opt in at half the rate of other individuals (0.5) and twice the rate of other individuals (2). The white multiplier is approximate because the opt-in decisions are stochastic and because the multiplier is imprecise at high levels of opt in.

Given the sequential approach of our synthetic data generation, we need to pick an order to synthesize the variables. Typically, more information is preserved for a synthetic variable by moving it from later in the synthesis order to earlier in the synthesis order. As such, we prioritize race and ethnicity. Next, we synthesize the remaining categorical variables, then age and educational attainment (both treated as numeric variables), then all of the income variables.

We implement a simple synthesis without much customization for the sake of this demonstration. We use the rpart implementation of decision trees and regression trees (Therneau and Atkinson 2022) using their default values (minsplit $= 20$, minbucket $= 8$, and cp $= 0.01$). These parameters keep the trees from going too deep and overfitting the data. We run each specification five times to evaluate the simulation-to-simulation variation in metrics.

Evaluation

We evaluate the results of the simulations with general utility metrics, specific utility metrics, and disclosure risk metrics. Descriptions of each metric are available in the appendix and complete results are available in the Urban Institute Data Catalog.

General utility measures the univariate and multivariate distributional similarity between the gold standard data and the candidate data (e.g., comparing the medians for all numeric variables). Specific utility measures the similarity of results for a specific analysis of the gold standard data and candidate data (e.g., comparing the coefficients in regression models). Disclosure risk metrics *estimate* the risk of attribute and membership inference attacks.

We calculate the following metrics for all simulations:

General utility

- Absolute error for proportions for all categorical variables
- Absolute error for proportions for all categorical variables by simplified race/ethnicity and detailed race
- Absolute proportion error for means for all numeric variables
- Absolute proportion error for means for all numeric variables by simplified race/ethnicity and

detailed race

- Proportion error in percentiles for numeric variables
- k-marginal score for all 1-way, 2-way, and 3-way marginals for categorical variables
- Mean absolute error in pairwise correlation coefficients for numeric variables
- Discriminator ROC AUC

Specific utility

- Regression confidence interval overlap for a Mincer model

Disclosure risk

- -diversity for regression trees summarized for each numeric variable
- Attribute inference test on welfare income
- Membership inference test

Results and Discussion

We have two main concerns with opt-in fully synthetic data. First, the approach may lead to untenable disclosure risks. Second, the utility improvements from including unaltered records in released data may be dampened by synthetic data with worse utility because the synthesizer is trained on smaller and non-random subsets of the GSDS. Essentially, as opt-in decreases, the number of observations in the training data decreases, which could lead to terrible models for the synthetic data.

Our demonstration suggests that opt-in fully synthetic data do not significantly increase disclosure risks. We also observe that all utility metrics improve with the introduction of an opt-in and improve as the opt-in rate decreases. Most of the utility metrics are surprisingly resilient to differential rates of opt-in for white vs. non-white respondents. Our results support the idea that opt-in approaches to synthetic data can improve data quality with minimal changes to disclosure risks.

Multivariate Utility

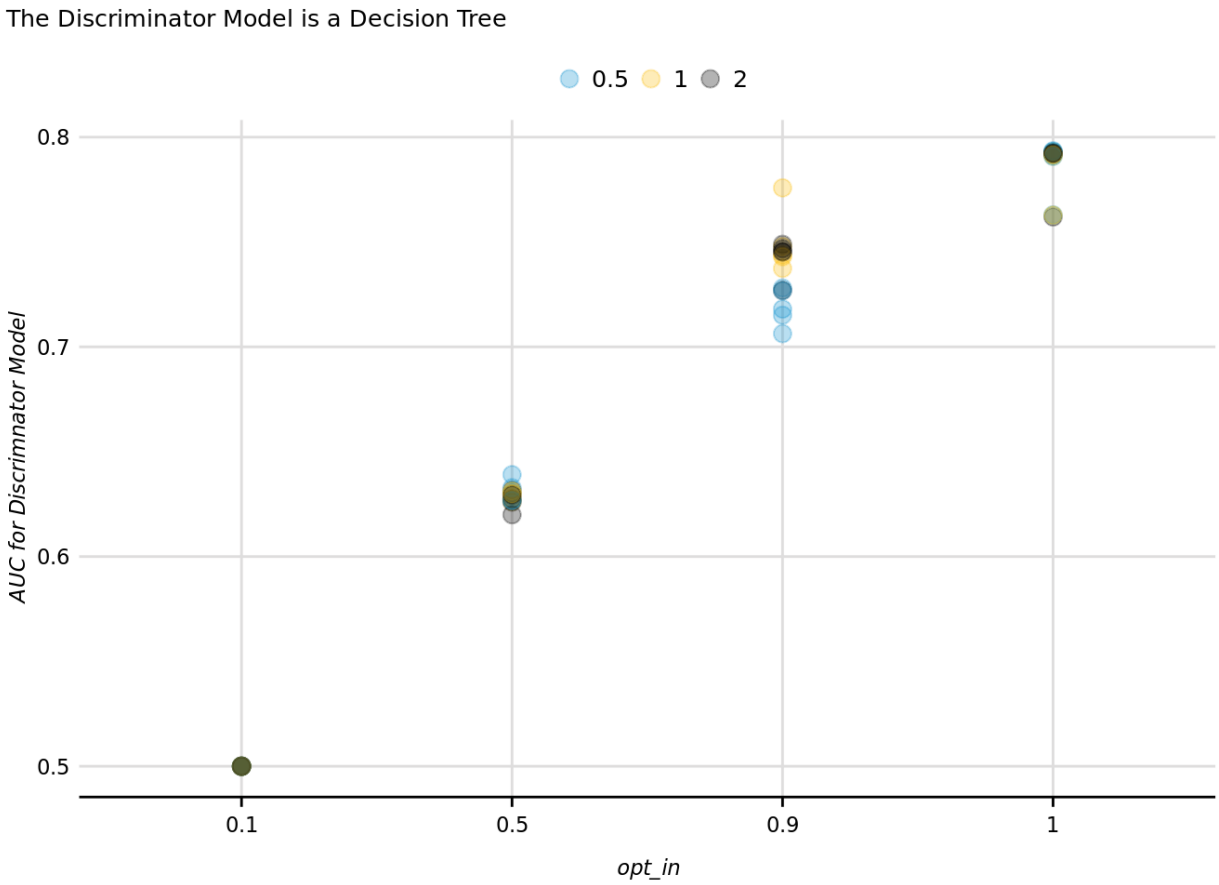
We start with multivariate utility. The discriminator AUC, shown in figure 5, is high when all observations opt-in to disclosure protections and quickly declines with lower opt-in rates. This suggests that a simple predictive model has difficulty distinguishing between the synthetic data and the GSDS after the introduction of opt-in synthesis.

The correlation difference, shown in figure 6 is modest for all opt-in rates and improves with low levels of opt-in.

Finally, regression estimates dramatically improve with low levels of opt-in. Figure 7 shows that the gap

FIGURE5

Low Opt-In Rates Dramatically Improve Discriminant Metrics



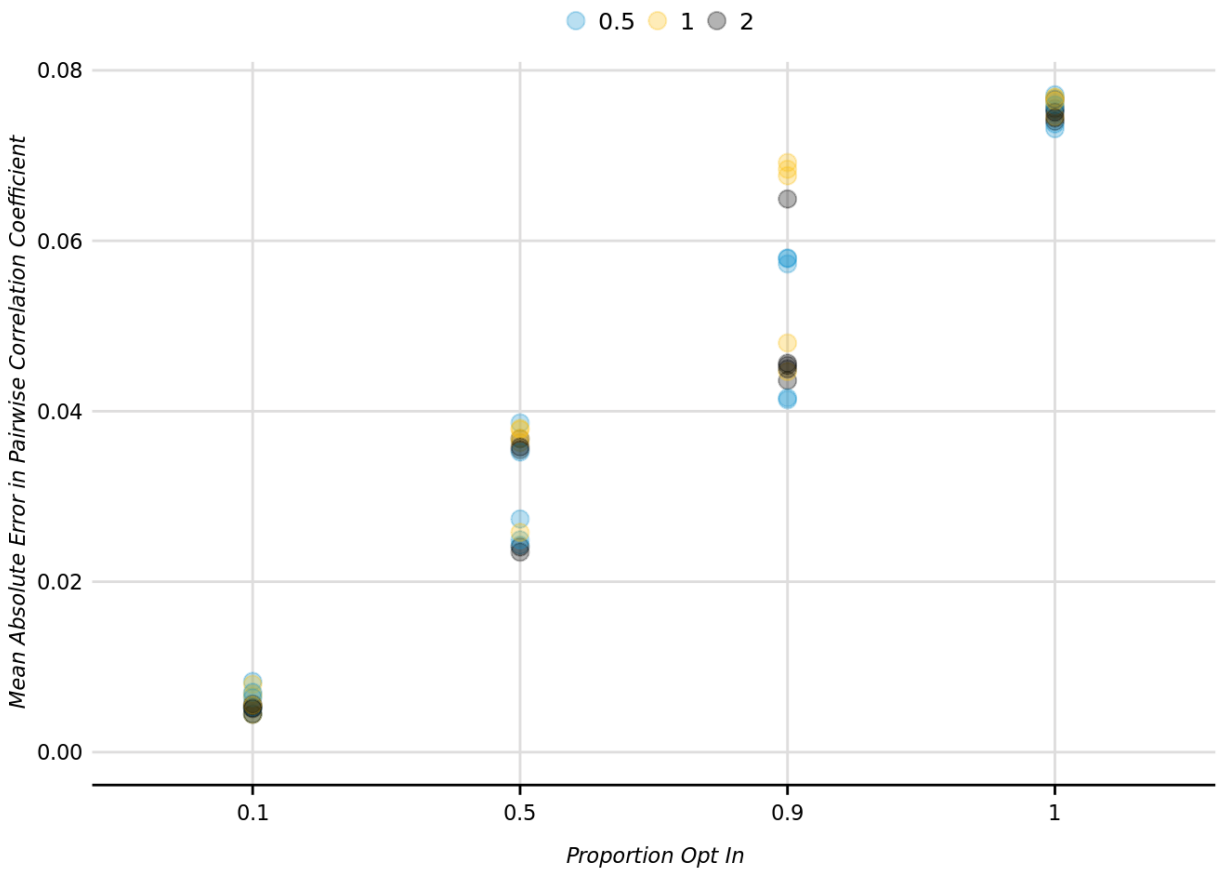
between confidence intervals, which is about six times the length of the confidence interval with full opt-in disclosure protection, fully disappears with low levels of opt-in.

Univaritate Utility

We next zoom in on specific variables and subsets of variables. The categorical utility of the synthetic data is very high for all scenarios but improves marginally with low opt-in rates.

Figure 8 shows the error in proportions for all categories of all categorical variables. Figure 9 shows the proportions for all classes of all categorical variables disaggregated into four race/ethnicity groups. Some race/ethnicity groups are poorly represented in the synthetic data. For example, the Black proportions for "Married, spouse present" and "Never married, spouse not present" are very wrong in the synthetic data because the estimates are attenuated to the proportions in the majority group. Some of these differences can be refined by using different hyperparameters during the synthesis process.

FIGURE6
Low Opt-In Improves Synthetic Correlations



We see a similar pattern when reviewing the proportion of individuals with Salaries or Wages. Figure 10 shows that the synthetic data generally overestimates the number of white people with salaries and wages and underestimates for Hispanic and Other Races and Ethnicities.

The 1-marginal score for the synthetic data is always better than the 1-marginal score for the holdout data. The results flip for the 2-marginal score and 3-marginal score. In all three cases, low levels of opt-in dramatically improve the results. Figure 11 shows the k-marginal score for 1-marginals and 3-marginals.

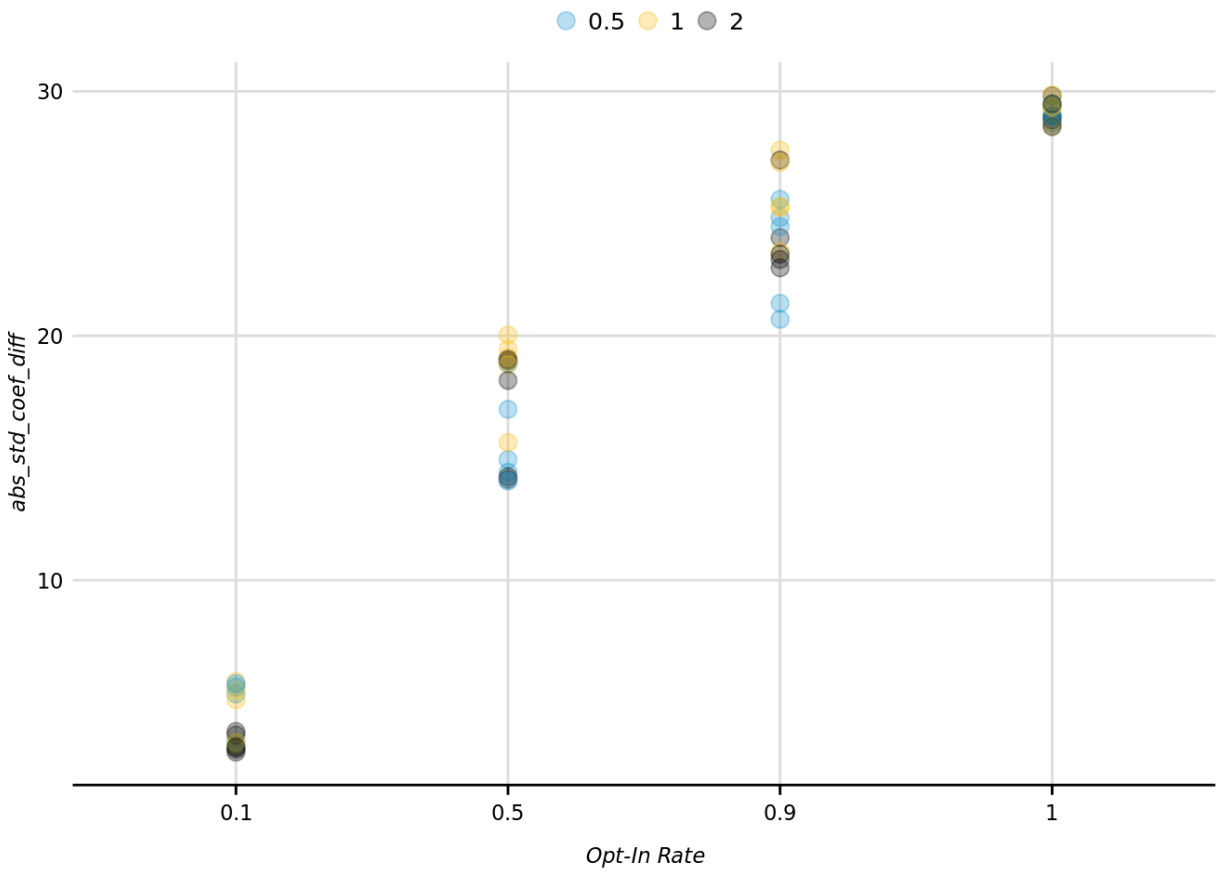
The syntheses have mixed results for numeric variables. Figure 12 shows that all syntheses closely match the distribution of total family income. The only major errors are for the minimum and maximum values.

The syntheses miss the distribution of salary and wage income, shown in figure 13, but the results improve dramatically with low levels of opt-in.

Disclosure Risks

FIGURE7

Coefficient Estimates are Closer with Low Opt-In Levels



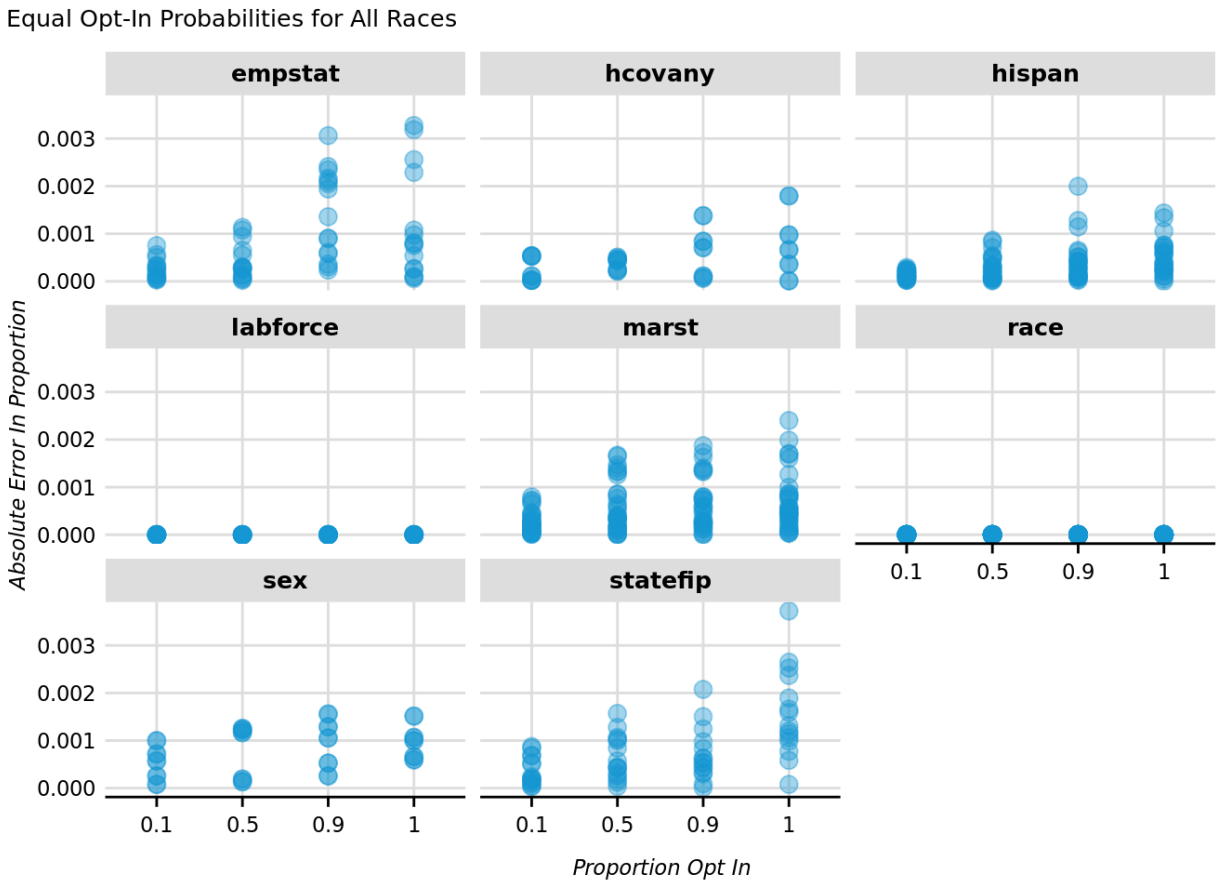
The synthesizer demonstrates promising -diversity for all syntheses. This is unsurprising since we use conservative hyperparameters for the decision trees and regression trees. In the worst-case situation, less than 0.5% of values for one variable come from nodes in regression trees with fewer than three unique values.

In Figure 14, the AUC for the membership inference attack is close to 0.5. Given the synthetic data, an attacker would struggle to tell if a confidential record is from the GSDS or holdout data using distance-based matching.

Finally, the RMSE for the attribute inference test from the synthetic data is always lower than the RMSE for a model trained on the holdout data. Using a simple predictive model, this suggests that the synthetic data are not copying chance features of the GSDS. These disclosure metrics enable us to infer the disclosive properties of the synthetic but they don't not provide a bound or guarantee on the disclosure risks of the synthetic data. This is a major disadvantage of non-formally private synthetic data when compared with

FIGURE8

Low Opt-In Modestly Improves Univariate Proportion Estimates



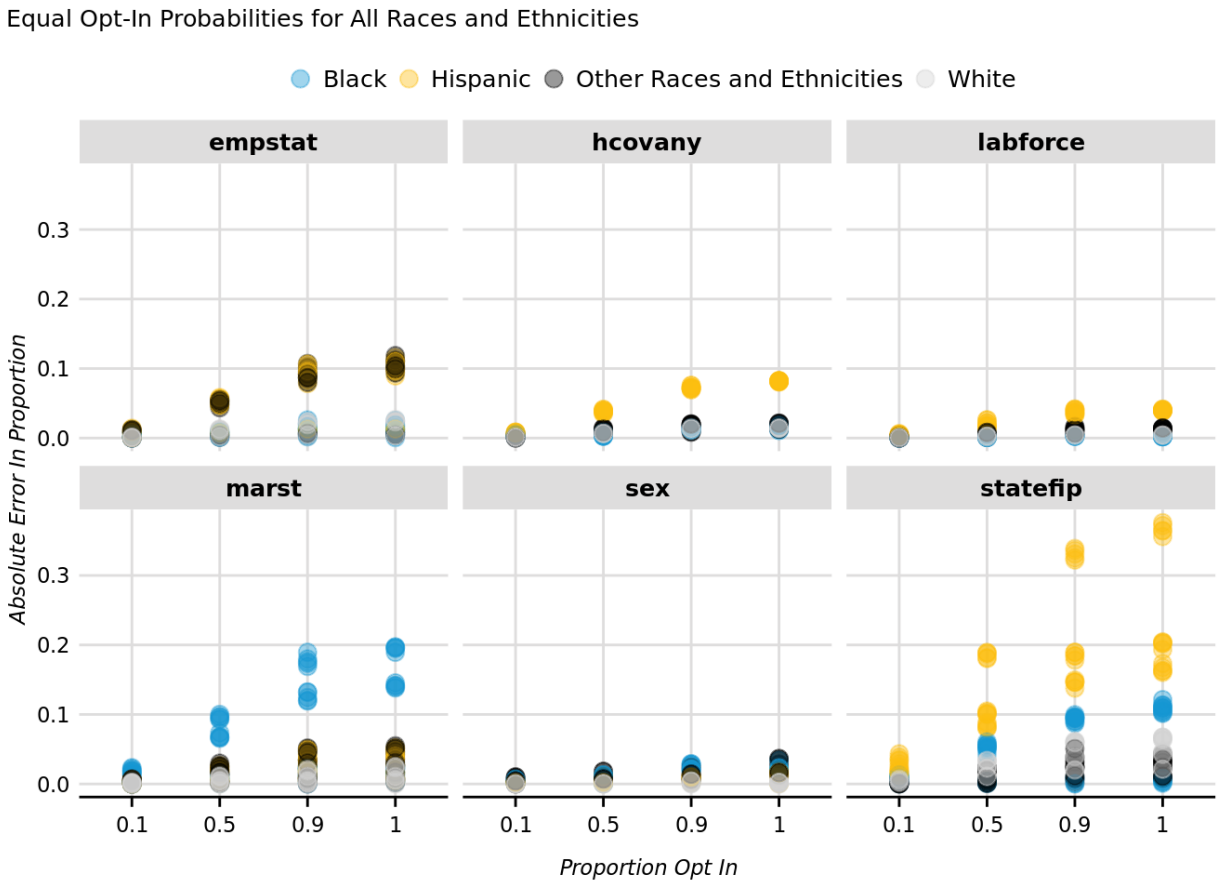
formally private tabulations like in the first demonstration.

Nonetheless, we believe the results are promising. Allowing individuals to forego disclosure protections increases data utility without increasing disclosure risks for the individuals who still want disclosure protection.

The idea of opt-in disclosure protection with synthetic microdata rests on a major implementation assumption: we do not label observations that forego disclosure protections. Doing so could increase disclosure risks by making differencing attacks easier. Data stewards need to be careful to not implicitly label the opt-in decision. For example, the GSDS and observations without opt-in could have precision to the closest integer when the synthetic records have precision to three decimal places. In this case, it would be trivial to identify the records with and without opt-in disclosure protection.

FIGURE9

Low Opt-In Reduces Major Racial/Ethnic Errors in Estimates



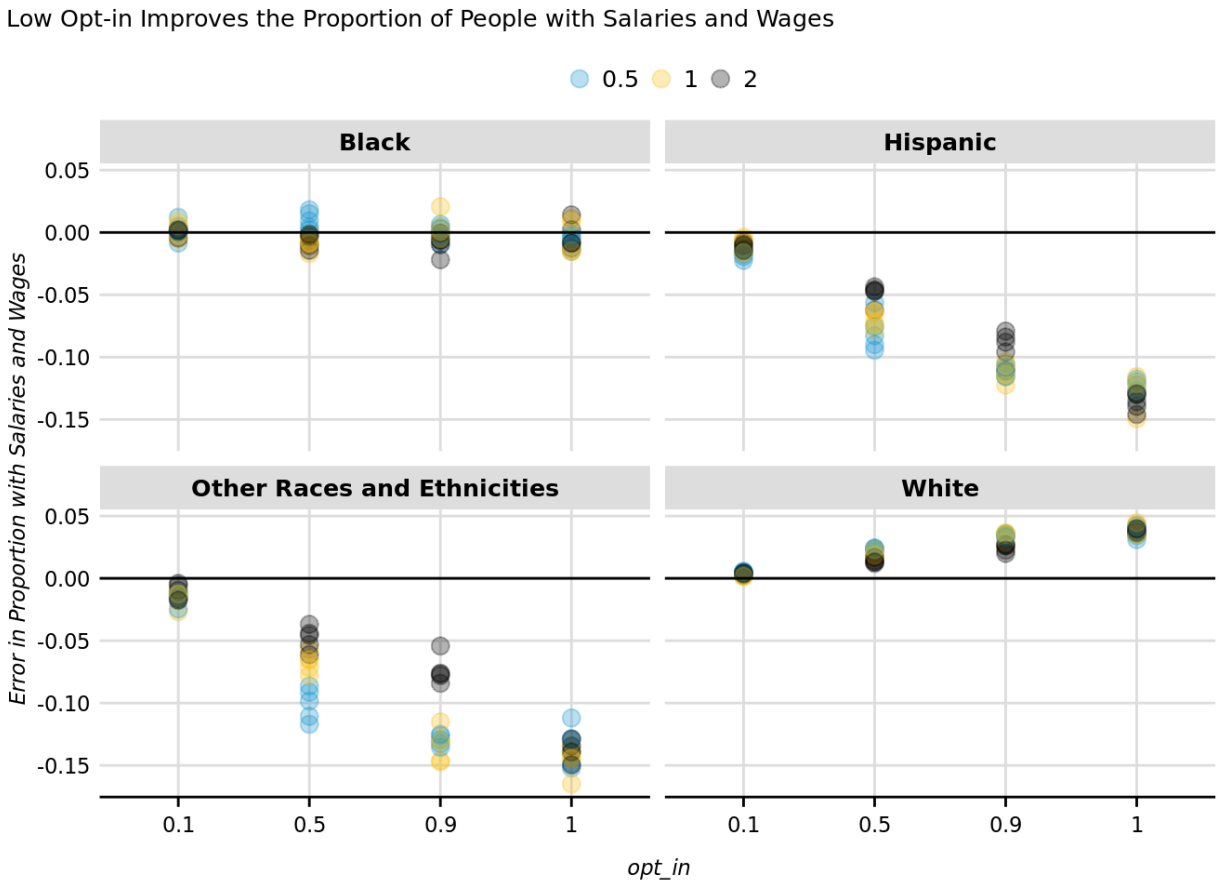
Conclusion and Future Work

Our results demonstrate the impacts of applying a novel opt-in privacy framework to formally and non-formally private SDC methods on resulting data quality and disclosure protections. Our demonstrations show that the opt-in framework can generally improve data quality without worsening privacy protections, however significant questions about implementation and comparisons to alternative methods remain.

The first demonstration shows the impact of opt-in disclosure protections for a differentially private Decennial Census. Although the opt-in framework shows smaller error with lower rates of opt-in relative to higher rates of opt-in, the overall level of error is still much higher than the centralized DP alternative. The errors introduced when switching from a centralized DP approach to a local DP approach dramatically outweigh the benefits of embracing opt-in disclosure protections. Major methodological breakthroughs for local DP would be needed for opt-in disclosure protections to improve the utility-disclosure risk trade off for

FIGURE10

Low Opt-In Improves Estimates of the Proportions of People with Salaries and Wages



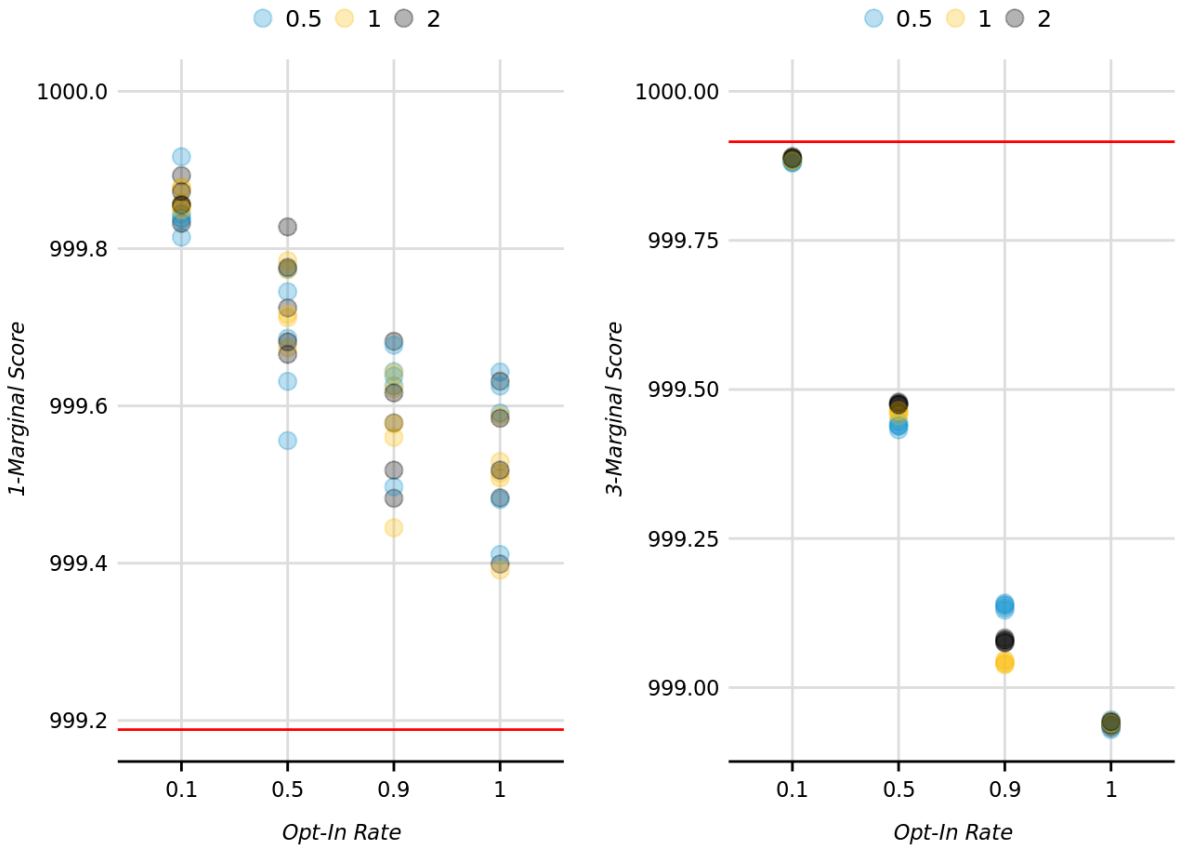
the Decennial Census.

The second demonstration shows the impact of opt-in disclosure protections for a fully-synthetic American Community Survey. In all cases, low levels of opt-in improve the utility of the synthetic data without worsening disclosure risk metrics. Our synthesis methodology was simple and suffered from some errors including major errors for certain race and ethnicity groups. Low levels of opt-in mitigated many of these issues.

Ample followup work is needed to explore the implementation of opt-in disclosure protection. We do not address questionnaire design and giving respondents clear information so they can provide informed consent. These implementation decisions inform the ethics of opt-in disclosure protection and would unambiguously affect the opt-in rate. We also only allow for unit opt-in disclosure protection and do not consider item opt-in disclosure protection. Item opt-in disclosure protection would be more difficult to implement but would give respondents even more control over their data.

FIGURE11

Low Opt-In Improves Higher-Dimensional Distributions

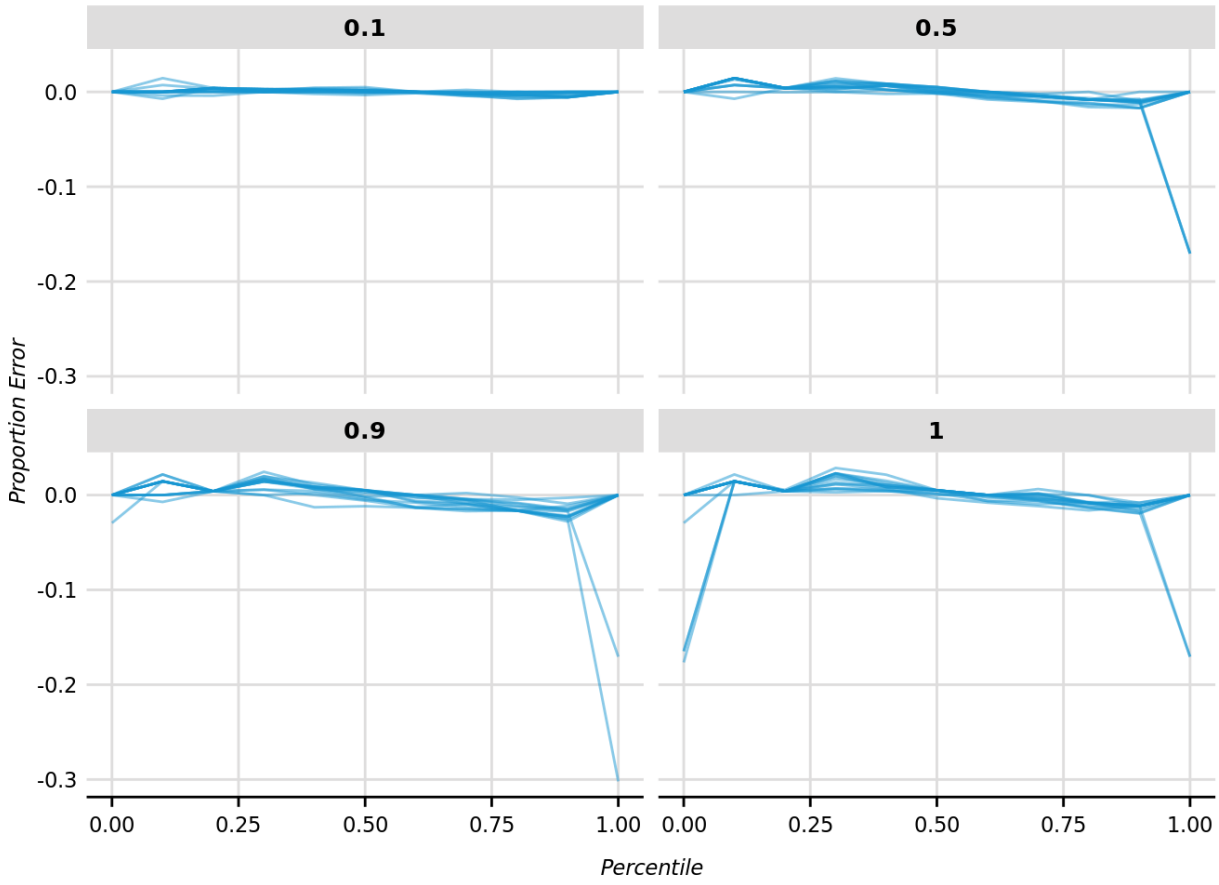


Further, implementation may be impeded by significant technical challenges and legal questions. The Census Bureau is currently wrestling with significant outstanding technical challenges for producing a fully synthetic ACS. Furthermore, census data products are protected by Title 13 and the Bureau would likely need legal changes to allow for an opt-in approach to disclosure protection.

The Survey of Earned Doctorates (SED) could be an ideal candidate for further evaluating this approach. The survey is a census of persons who recently earned doctorate degrees, and this population would be easier to inform because of their research backgrounds. The SED has remarkably high response rates and we speculate that the population would opt-in to disclosure protections at very low rates. Furthermore, the file avoids many of the technical challenges currently facing a synthetic ACS. Applying an opt-in framework to the SED could inform implementation questions about opt-in rates, differences by groups, and outreach strategies, along with further empirical evidence on data utility and privacy outcomes.

FIGURE12

All Syntheses Recreate the Total Family Income Distribution



Metrics Appendix

FIGURE13

Low Opt-In Improves the Distribution of Salary and Wage Income

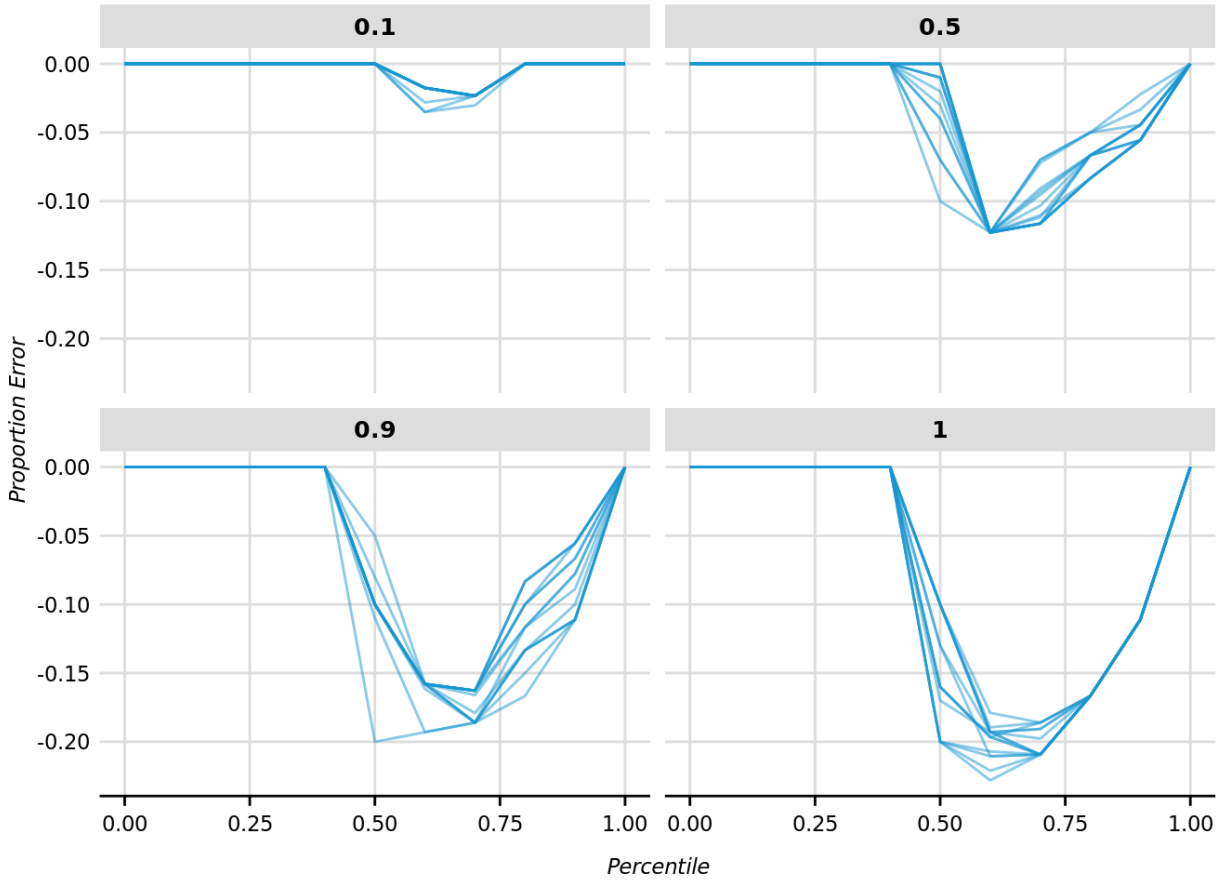
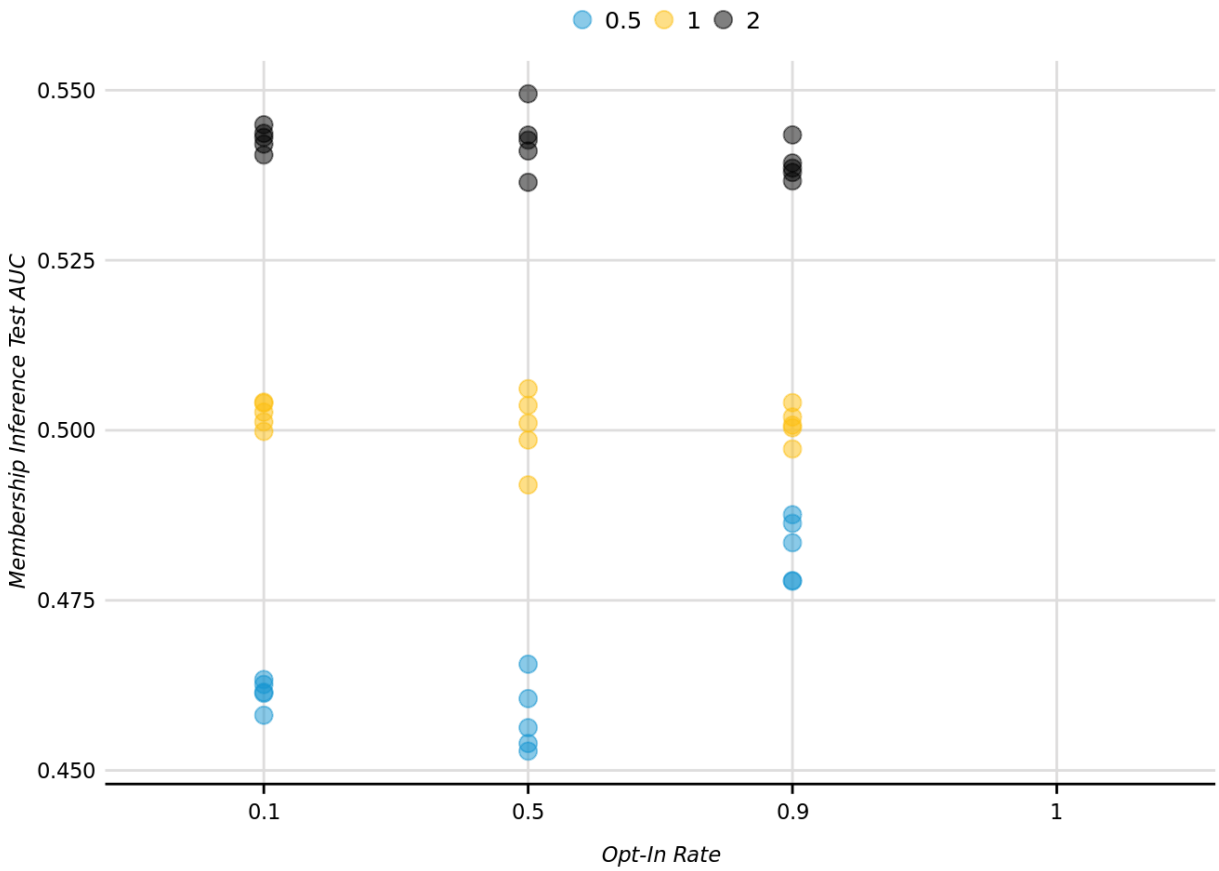


FIGURE14

No SYnthesis Demonstrates High Membership Inference Risks



A. Metrics Descriptions

General utility

- **Absolute error for proportions for all categorical variables:** For each categorical variable, calculate the proportion of observations with each class for the GSDS and candidate data. Calculate the absolute value of the difference between the GSDS and candidate data.

- **Absolute error for proportions for all categorical variables by simplified race/ethnicity and detailed race:** For each categorical variable, calculate the proportion of observations with each class for each demographic group for the GSDS and candidate data. Calculate the absolute value of the difference between the GSDS and candidate data.

- **Absolute proportion error for means for all numeric variables:** For each numeric variable, calculate the mean for the GSDS and candidate data. Calculate the absolute value of the difference between the GSDS and candidate data.

- **Absolute proportion error for means for all numeric variables by simplified race/ethnicity and detailed race:** For each numeric variable, calculate the mean for the GSDS and candidate data for each demographic group. Calculate the absolute value of the difference between the GSDS and candidate data.

- **Proportion error in percentiles for numeric variables:** For each numeric variable, calculate the minimum, 10th percentile, 20th percentile, 30th percentile, 40th percentile, 50th percentile, 60th percentile, 70th percentile, 80th percentile, 90th percentile, and maximum. Calculate the proportion difference between the GSDS and candidate data.

- **k-marginal score for all 1-way, 2-way, and 3-way marginals for categorical variables:** [@raab2021; @sen2023] Select k. Let v be the number of categorical variables. Then there are $\binom{v}{k}$ k-marginals. Exhaustively calculate the proportion of observations in each cell for all k-way marginals for categorical variables. For each k-marginal, calculate the mean absolute error (this is called mean absolute difference between distributions). Finally, make the metric ascending and rescale to max at 1,000 with $(1\ mean(MADD))\ 1000$.

- **Mean absolute error in pairwise correlation coefficients for numeric variables:** For all numeric variables, calculate a Pearson's linear correlation matrix on the GSDS and candidate data. Difference the lower triangle of the matrix for the GSDS and the matrix for the candidate file. Calculate the mean absolute error.

- **Discriminator ROC AUC:** Discriminator metrics measure how well a predictive model can distinguish between observations from the GSDS and candidate data. Ideally, the models perform poorly. The p-MSE ratio [@woo2009; @snoke2018] and SPECKS [@bowen2021a] summarize the propensity scores from

discriminant models. We take a different approach and look at the ROC AUC for the discriminant models. The three measures are highly correlated, the values for ROC AUC are more familiar to predictive modelers, and the measure doesn't require bootstrapping.

Specific utility

- **Regression confidence interval overlap** measures the overlap of confidence estimated on the gold standard data and synthetic data [@karr2006; @snoke2018]. 1 represents perfect overlap, 0 represents adjacency with no overlap, and negative values represent gaps between the confidence intervals. We regress log wages and salary income on potential experience, potential experience squared, an indicator for white, and sex and calculate the regression confidence interval overlap. This is a simplified version of a "Mincer model" [@card1999].

Disclosure risk

- **-diversity for regression trees summarized for each numeric variable:** [@bowen2020] We are concerned that our synthesizer could memorize the confidential data or too closely fit the confidential data. We calculate -diversity [@machanavajjhala2007] on the nodes from decision trees and regression trees to measure the heterogeneity of values in the nodes. We calculate the proportion of synthetic values for each variable in the candidate data that come from nodes with low heterogeneity < 10.

- **Attribute inference test on welfare income:** [@elemam2020] We are concerned that an attacker could train predictive models on the synthetic data and make precise inferences about observations in the confidential data. We train a regression tree to predict welfare income on the synthetic data using all other variables as predictors. We assume that an attacker knows all variables but welfare income for the confidential data. We make predictions and calculate the RMSE. We repeat this process using the holdout data instead of the synthetic data to benchmark against the RMSE of models trained when an attacker has access to another random sample from the population instead of the synthetic data.

- **Membership inference test:** [@zhang2022] We are concerned that an attacker could determine if a confidential observation is in the training data for the candidate file. We construct a data set that is about 10,000 observations from the GSDS and 10,000 observations from the holdout data. For each observation, we measure Gower's distance [@gower1971] to all synthetic observations. Traditionally, analysts pick different thresholds to predict membership and then calculate precision. We normalize the distances so they are in

$$0, 1$$

and are like probabilities. We then calculate the ROC AUC to measure the tradeoff between true positives and false positives for different thresholds. ROC AUC closer to 0.5 indicates difficulty assessing membership.

Notes

[1] https://www.ipums.org/changes-to-census-bureau-data-products

[2] https://www.census.gov/newsroom/blogs/random-samplings/2022/12/disclosure-avoidance-protections-acs.html

[3] https://acsdatacommunity.prb.org/discussion-forum/m/2021-acs-conference-files/147/download

[4] "The tidysynthesis R package." Presentation given at rstudio::conf(2022), Washington, DC, July 25 – 28.

About the Authors

Aaron R. Williams but the rest of the text lightface. Use Author Bios–First style for the introductory paragraph of each bio. You can paste your bio from your author page on the Urban website (and condense it if needed) here.

If your bio is more than one paragraph long, use Author Bios–Additional for any subsequent paragraphs. This style suppresses spacing between paragraphs.

Author bios no longer include photos.

Jennifer Andre but the rest of the text lightface. Use Author Bios–First style for the introductory paragraph of each bio. You can paste your bio from your author page on the Urban website (and condense it if needed) here.

If your bio is more than one paragraph long, use Author Bios–Additional for any subsequent paragraphs. This style suppresses spacing between paragraphs.

Author bios no longer include photos.

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.

Abowd, John M. 2018. "Protecting the Confidentiality of Americas Statistics: Adopting Modern Disclosure Avoidance Methods at the Census Bureau." https://www.census.gov/newsroom/blogs/research-matters/2018/08/protecting_the_confi.html.

Abowd, John, Robert Ashmead, Ryan Cumings-Menon, Simson Garfinkel, Micah Heineck, Christine Heiss, Robert Johns, et al. 2022. "The 2020 Census Disclosure Avoidance System TopDown Algorithm." *Harvard Data Science Review*, no. Special Issue 2 (June). https://doi.org/10.1162/99608f92.529e3cb9.

Axelrod, Judah, Karolina Ramos, and Rebecca Bullied. 2022. "Opportunities and Challenges in Using Private-Sector Data for Racial Equity Analysis."

Benedetto, Gary, and Evan Totty. 2020. "Synthesizing Familial Linkages for Privacy in Microdata." CED-DA Working Paper.

Bowen, Claire McKay, and Joshua Snoke. 2021. "Comparative Study of Differentially Private Synthetic Data Algorithms from the NIST PSCR Differential Privacy Synthetic Data Challenge." *Journal of Privacy and Confidentiality* 11 (1). https://doi.org/10.29012/jpc.748.

Bowen, Claire McKay, Aaron R Williams, and Madeline Pickens. 2022. "Decennial Disclosure: An Explainer on Formal Privacy and the TopDown Algorithm."

Daily, Donna. 2022. "Disclosure Avoidance Protections for the American Community Survey." https://www.census.gov/newsroom/blogs/random-samplings/2022/12/disclosure-avoidance-protections-acs.html.

Dwork, Cynthia, Frank McSherry, Kobbi Nissim, and Adam Smith. 2006. "Calibrating Noise to Sensitivity in Private Data Analysis." In, edited by Shai Halevi and Tal Rabin, 3876:265–84. Berlin, Heidelberg: Springer Berlin Heidelberg. http://link.springer.com/10.1007/11681878_14.

Hotchkiss, Marisa, and Jessica Phelan. 2017. "Uses of Census Bureau Data in Federal Funds Distribution: A New Design for the 21st Century," September. https://www2.census.gov/programs-surveys/decennial/2020/program-management/working-papers/Uses-of-Census-Bureau-Data-in-Federal-Funds-Distribution.pdf.

Hotz, V. Joseph, and Joseph Salvo. 2022. "A Chronicle of the Application of Differential Privacy to the 2020 Census." *Harvard Data Science Review*, June. https://doi.org/10.1162/99608f92.ff891fe5.

Hu, Jingchen, and Claire McKay Bowen. 2023. "Advancing Microdata Privacy Protection: A Review of Synthetic Data." https://doi.org/10.48550/ARXIV.2308.00872.

Johnson, Eric J., and Daniel Goldstein. 2003. "Do Defaults Save Lives?" *Science* 302 (5649): 1338–39. https://doi.org/10.1126/science.1091721.

Little, Roderick JA. 1993. "Statistical Analysis of Masked Data." *Journal of Official Statistics* 9 (2): 407. https://www.proquest.com/docview/ca21efb41fee47d293bbee5bf7be7fb3/statistical-analysis-of-masked-data.pdf.

Mather, Mark, and Paola Scommegna. 2019. "Why Is the u.s. Census so Important?" March.

https://www.prb.org/resources/importance-of-u-s-census/.

McKenna, Ryan, Daniel Sheldon, and Gerome Miklau. 2019. "Graphical-Model Based Estimation and Inference for Differential Privacy." *Proceedings of the 36th International Conference on Machine Learning* 97: 4435–44. http://proceedings.mlr.press/v97/mckenna19a/mckenna19a.pdf.

Near, Joseph P, and Chiké Abuah. 2022. *Programming Differential Privacy*.

Reiter, Jerome P., and Robin Mitra. 2009. "Estimating Risks of Identification Disclosure in Partially Synthetic Data." *Journal of Privacy and Confidentiality* 1 (1). https://doi.org/10.29012/jpc.v1i1.567.

Reiter, JP. 2005. "Using CART to Generate Partially Synthetic Public Use Microdata." *Journal of Official Statistics* 21 (3): 441. https://www.proquest.com/docview/1266792149/abstract/B4CAD6D1A888424FPQ/1?accountid=11091.

Rubin, Donald B. 1993. "Statistical Disclosure Limitation." *Journal of Official Statistics* 9 (2): 461–68.

Ruggles, Steven, Sarah Flood, Matthew Sobek, Danika Brockman, Grace Cooper, Stephanie Richards, and Megan Schouwiler. n.d. "IPUMS USA: Version 13.0." https://doi.org/10.18128/D010.V13.0.

Tao, Yuchao, Ryan McKenna, Michael Hay, Ashwin Machanavajjhala, and Gerome Miklau. 2021. "Benchmarking Differentially Private Synthetic Data Generation Algorithms." https://doi.org/10.48550/ARXIV.2112.09238.

Thaler, Richard H. 2009. "Opting in Vs. Opting Out." *The New York Times*, September. https://www.nytimes.com/2009/09/27/business/economy/27view.html.

Thaler, Richard H., and Cass R. Sunstein. 2009. *Nudge: Improving Decisions about Health, Wealth, and Happiness*. Rev. and expanded ed. New York: Penguin Books.

Therneau, Terry M., and Elizabeth J. Atkinson. 2022. "An Introduction to Recursive Partitioning Using the RPART Routines." https://cran.r-project.org/web/packages/rpart/vignettes/longintro.pdf.

United Nations, ed. 2022. *Synthetic Data for Official Statistics: A Starter Guide*. Geneva: United Nations.

United States Census Bureau. 2017. "American Community Survey Information Guide," October.

Wang, Teng, Xuefeng Zhang, Jingyu Feng, and Xinyu Yang. 2020. "A Comprehensive Survey on Local Differential Privacy Toward Data Statistics and Analysis." *Sensors* 20 (24): 7030. https://doi.org/10.3390/s20247030.

Warner, Stanley L. 1965. "Randomized Response: A Survey Technique for Eliminating Evasive Answer Bias." *Journal of the American Statistical Association* 60 (309): 63–69. https://doi.org/10.1080/01621459.1965.10480775.

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

500 L'Enfant Plaza SW
Washington, DC 20024

www.urban.org