*Journal of the Royal Statistical Society Series A: Statistics in Society*,

2024, 1

–

23

doi: DOI HERE

Advance Access Publication Date: Day Month Year

Paper

Optimized Sequential Synthesis with an

Application to Legal Anonymization

# Cameron D. Bale[1](https://orcid.org/0000-0003-4361-2579),*∗* and Harrison Quick2

1Department of Marketing & Global Supply Chain, Brigham Young University, Utah, United

States and 2Division of Biostatistics & Health Data Science, University of Minnesota,

Minnesota, United States

∗Corresponding author. cameron.bale@byu.edu

FOR PUBLISHER ONLY Received on Date Month Year; revised on Date Month Year; accepted on Date Month

Year

## Abstract

Data synthesis is a promising approach to creating privacy-preserving datasets, but the complexity of tuning synthesis models may deter adoption by data stewards. To address this issue, we propose a novel tuning methodology that automates parameter selection using Bayesian optimization, reducing the need for manual trial-and-error adjustments to balance privacy and utility. We demonstrate its application by synthesizing South Korean COVID-19 patient data and public-use Current Population Survey data and assessing whether the resulting synthetic data meet GDPR legal anonymity standards. Our findings highlight three key insights: (1) our methodology effectively balances privacy and utility across multiple synthesis models, producing legally anonymous synthetic data with high analysis-specific utility; (2) legally anonymous synthetic data often yield more accurate regression estimates than differentially private regression models but increase privacy risks by exposing a full synthetic dataset; and (3) legally anonymous synthetic data do not fully protect against attribute disclosure attacks, emphasizing the need for further research on acceptable protection levels. These results have important implications for using synthetic data as a compliance tool and as a broader privacy solution.

Key words: Bayesian optimization, attribute disclosure, data privacy

# Introduction

Privacy-preserving data synthesis has emerged as a promising solution for data stewards tasked with disseminating confidential data (Hu and Bowen, 2024). Using generative models, which range from generative adversarial networks (Lee and Anand, 2020) to Bayesian models (Quick, 2021), data stewards can produce synthetic data that balance preserving insights from the confidential data with protecting identities and sensitive information of data subjects. Open-source programming packages such as *synthpop* (Nowok et al., 2016) facilitate the implementation of these methods. However, achieving the desired balance between privacy and utility often

© The Author 2024. Published by Oxford University Press. All rights reserved. For permissions, please

e-mail: journals.permissions@oup.com 1

requires manual tuning of model parameters, which can be time-consuming and complex for nonexperts.

Further complicating matters, data stewards must navigate an evolving landscape of privacy regulations. They can either implement robust security measures (failure to do so may result in severe penalties [Koch, 2018]) or employ anonymization methods, such as data synthesis, to produce *legally anonymous* data that fall outside the scope of privacy laws (European Data Protection Board, 2020). A recent Joint Research Centre report from the European Commission suggested that synthetic data can serve as a solution for producing legally anonymous data for informing policymaking (Hradec et al., 2022, p. 44), stating, “This artificially generated data is highly representative, yet completely anonymous. . . . [T]he risk of re-identification is effectively eliminated.” The company behind the synthesis, MOSTLY.AI, asserts that their synthetic data comply with privacy laws.[[1]](#footnote-1) This legal classification is advantageous for data stewards, as it enables broader data sharing and cross-border transfers, even under restrictive regulations such as the GDPR (Hintze and El Emam, 2020).

Despite these advantages, the complexity of implementing and fine-tuning data synthesis models may hinder adoption by data stewards. Moreover, even if synthesis models achieve a balance of privacy and utility, it remains unclear whether synthetic data generated from these models meets the legal criteria for anonymity under regulations like the GDPR. Addressing this gap is critical for the broader adoption of synthetic data as a privacy solution.

The goals of this paper are twofold. First, we aim to lower the barriers to implementing and fine-tuning data synthesis models, by using Bayesian optimization (Snoek et al., 2012) to automate the tuning of synthesis model parameters. Rather than manually testing different parameter combinations or performing a grid search, the method uses a Gaussian process to predict parameter combinations that will result in an optimized balance of privacy and utility. The method ensures that a synthetic dataset is produced in a reasonable amount of time without significantly over-fitting (resulting in high privacy risk) or under-fitting (resulting in low utility) and enables data stewards to spend more time improving the overall synthesis process (e.g., by comparing different model types) and less time optimizing the fit of individual models. The tuning method outlined in this paper uses metrics and an optimization function that we believe will work well for most tabular datasets. Furthermore, the method is highly flexible, and technically inclined data stewards can incorporate customized metrics and optimization functions that are tailored to produce synthetic datasets aligned with their specific requirements.

Second, we evaluate whether commonly used synthesis models can produce synthetic data that meets the legal definition of anonymity under the GDPR. To do so, we begin by outlining the privacy criteria used by MOSTLY.AI and show that these criteria are met in expectation when synthetic data are generated from the same distribution as the confidential data. Using the proposed tuning method, we train sequential synthesis models to approximate the data-generating distribution. We compare our optimized synthesis approach to the MOSTLY.AI method—treated as a baseline known to produce legally anonymous synthetic data—on two datasets: (1) location data from South Korean COVID-19 patients in 2020 and (2) the 1994 to 1996 Current Population Survey Annual Social and Economic Supplements (CPS ASEC). We also compare the utility of regression models applied to the synthetic CPS ASEC data to the utility of differentially private regression models applied through a validation server (Barrientos et al., 2023). Finally, we simulate an attribute disclosure attack to determine whether legally anonymous synthetic data also meet academic standards for privacy protection.

Our results show that the proposed method is effective at tuning synthesis models and that common synthesis models, such as classification and regression trees (CARTs) and multinomial logistic regression (MNL), are capable of producing legally anonymous synthetic data with high analysis-specific utility. Furthermore, the legally anonymous synthetic datasets can enable the estimation of more accurate regression coefficients than differentially private regression models on a validation server. However, this approach comes with increased privacy risk from releasing a full legally anonymous synthetic dataset, and we show that legally anonymous synthetic data do not guarantee protection against attribute disclosure.

The paper proceeds as follows: Section 2 provides a review of the relevant literature. Section 3 describes the optimized sequential synthesis method. Our empirical application in Section 4 contains several subsections: in 4.1, we outline the privacy criteria used by MOSTLY.AI to determine whether synthetic data are legally anonymous, and in 4.3 through 4.7, we apply tuned sequential synthesis methods and the MOSTLY.AI synthesis method to both the South Korean COVID-19 patient data and the CPS ASEC data and compare the privacy and utility results of both synthesis methods. We conclude and discuss opportunities for further research in Section 5.

# Literature Review

## Tuning the Privacy and Utility Trade-off

Bayesian models have been used to synthesize various types of sensitive data, such as time series (Schneider and Abowd, 2015) and location data (Quick et al., 2015). These models control the privacy–utility trade-off through prior distributions, where more informative priors increase privacy but reduce utility. Selecting appropriate values for these priors is crucial, as improper choices can yield synthetic data that either lack privacy protection or fail to retain sufficient utility.

CART models have also shown potential for producing synthetic data with desirable privacy and utility (Reiter, 2005; Drechsler and Reiter, 2011). These models have been used to synthesize health and medical survey data (El Emam et al., 2021; Azizi et al., 2021), individual-level income tax records (Bowen et al., 2022), and location data (Wang and Reiter, 2012; Drechsler and Hu, 2021). Often, CART parameters are set to default values that perform reasonably well, though some researchers have experimented with tuning the complexity parameter to balance privacy and utility. However, studies by Drechsler and Hu (2021) and Schneider et al. (2023) found that this approach produced inconsistent results, highlighting the need for more systematic tuning methods.

Typically, data stewards perform a post-hoc privacy assessment after generating synthetic data to ensure that privacy risk is within acceptable limits (e.g., Quick et al., 2015; Schneider et al., 2023). If the risk is too high, the synthesis model must be adjusted and the data regenerated. Conversely, if the risk is far below the acceptable limit, the model can be adjusted to improve utility. This iterative process is time-consuming, motivating the use of Bayesian optimization to automatically select parameters that balance privacy and utility.

Another approach to synthesis model parameter selection is to set conditions *a priori* that guarantee acceptable privacy or utility levels. For example, Jackson et al. (2022) derived expected values of risk and utility for synthetic data from saturated count models, allowing data stewards to define these levels before synthesis. In a similar vein, we show in Section 2 of the supplementary materials that if synthetic data are sampled from the same distribution as the confidential data, the privacy criteria for legal compliance will be met in expectation. Thus, our method allows data stewards to optimize models using the *pMSE* ratio to achieve the desired distributional similarity without manual tuning. More generally, data controllers can define any objective function based on privacy risk, utility, or both to tailor the synthesis process to their needs.

## Differential Privacy

Efforts to combine differential privacy with data synthesis aim to produce synthetic data with formal privacy guarantees. Bowen and Snoke (2021) analyzed submissions from the 2018–2019 NIST PSCR “Differential Privacy Synthetic Data Challenge” and found that the resulting differentially private synthetic datasets often had low global utility based on distributional similarity metrics such as the *pMSE* ratio (Snoke et al., 2018). However, differentially private synthetic data can preserve utility for specific cases, such as one-way and two-way positive conjunction queries (Asghar et al., 2020). More recent work has focused on synthesizing Poisson-distributed count data using Bayesian models with carefully selected priors to ensure differential privacy while improving utility using information from publicly available data (Quick, 2021, 2022).

An alternative approach involves allowing users to submit queries on sensitive data through a validation server, where outputs are perturbed using differential privacy. Barrientos et al. (2023) studied differentially private regression queries and found that these methods often yielded confidence intervals that were either significantly wider or inconsistent in sign and statistical significance when compared to their confidential data counterparts. As a result, the authors concluded that differentially private regression still faces substantial barriers before it can support accurate statistical inference.

## Legally Anonymous Data

Recital 26 of the GDPR (European Parliament and Council of European Union, 2016) states that data protection principles apply to any information linked to an identifiable individual but do not apply to data anonymized in such a way that the individual can no longer be identified. The concept of legally anonymous data is intuitive, but determining when data meet the legal standard is challenging.

Researchers have proposed formal definitions, such as the prevention of predicate singling out attacks (Cohen and Nissim, 2020). Cohen and Nissim (2020) showed that differential privacy protects against singling out, whereas syntactic approaches like *k*-anonymity do not provide the same guarantees. However, these definitions do not directly translate to synthetic data.

The European Data Protection Board has outlined a “reasonability test” for anonymization, emphasizing that anonymized data should not be able to be linked to an identifiable person using any “reasonable” effort, considering technical means, data context, and other factors (European Data Protection Board, 2020). In practice, determining whether synthetic data are legally anonymous involves assessing this reasonability. While differential privacy offers a strong theoretical guarantee, it may not always align with high-utility requirements, such as those for complex survey data (Drechsler, 2023).

Meanwhile, MOSTLY.AI has developed its own privacy metrics and standards for producing GDPR-compliant synthetic data using deep-learning-based methods. In this paper, we assess whether traditional synthesis models, such as CARTs, can generate legally anonymous synthetic data and whether such data provide higher utility for specific analyses compared to differentially private methods.

# Tuning Method: Bayesian Optimization of Synthesis Models

The goal of our proposed methodology is to eliminate the need for manual tuning of synthesis model parameters to balance utility and privacy in synthetic data. We achieve this goal by using Bayesian optimization based on a Gaussian process (Snoek et al., 2012) to automatically optimize the privacy parameters of our synthesis models. Bayesian optimization, commonly used for tuning black-box machine-learning models, optimizes an objective function by sequentially sampling inputs and updating based on observed outputs. This section draws on the comprehensive review by Brochu et al. (2010).

Bayesian optimization is used to maximize an objective *f*(*θ*), where *θ* denotes the parameters of the synthesis model in our case, and the objective *f*(*θ*) is modeled using a Gaussian process,

*f*(*θ*) ∼ *GP*(*m*(*θ*)*, k*(*θ,θ*′))*,* (1)

where *m*(*.*) and *k*(*.*) denote the mean and covariance functions of the Gaussian process, respectively. Suppose that we have a sequence of observed *θ* and *f*(*θ*) values, denoted *B*1:*t* = {*θ*1:*t,f*1:*t*}. The following improvement function,

*I*(*θ*) = max{0*,ft*+1(*θ*) − *f*(*θ*+)}*,* (2)

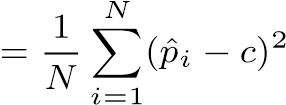
is positive when the predicted objective function value at *θ* is higher than the objective under *θ*+, the current best value. The next test point, *θt*+1, is chosen to maximize the expected improvement,

*θt*+1 = argmax*θ*E[*I*(*θ*)|*B*1:*t*]*,* (3)

which can be evaluated analytically (Brochu et al., 2010). Optimization is carried out by sequentially choosing values of *θ* that maximize the expected improvement in the objective function *f*(*θ*).

A key advantage of Bayesian optimization is its ability to handle nondifferentiable and nonconvex objective functions (Brochu et al., 2010), enabling users to choose diverse objective functions that capture various privacy risks and utility measures. Common utility metrics include the propensity score mean-squared error (*pMSE*) and *pMSE* ratio (Snoke et al., 2018), the Kolmogorov-Smirnov distance between propensity score distributions (Bowen et al., 2021), and differences in cross-tabulations between synthetic and confidential data (Drechsler and Hu, 2021; Schneider et al., 2023). We recommend using the *pMSE* ratio because it provides a single-valued measure of distributional similarity, and its theoretical expected value can help regularize the synthesis process (Snoke et al., 2018).

To compute the *pMSE* ratio, we first train a logistic regression model to discriminate between synthetic and confidential records in a dataset containing the combined synthetic and confidential data. The propensity score ˆ*pi* of each record in the combined data is the in-sample estimate of the probability that a given record is synthetic. The propensity score mean-squared error (*pMSE*) is computed as

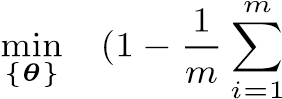
*pMSE* *,* (4)

where *N* is the number of combined synthetic and confidential records, and *c* is the fraction of the *N* records that are synthetic (typically *c* = 0*.*5).

A *pMSE* score close to zero indicates high utility, meaning the synthetic data are nearly indistinguishable from the confidential data. However, if the synthetic data replicate the confidential records, the synthetic data offer no privacy protection. To address this issue, Snoke et al. (2018) derived the expected *pMSE* from a logistic regression under two scenarios: (1) when the confidential data **Y** is fixed and the synthetic data **Z** is drawn from a correctly specified synthesis model, and (2) when both datasets are generated from the same model or data distribution D. The observed *pMSE* can be compared to the expected *pMSE* using the ratio

R = *pMSE/*E(*pMSE*)*,* (5)

with the goal of producing synthetic data with R ≈ 1. We define the objective function *f*(*θ*) using the *pMSE* ratio as follows:

*pMSEm/*E[*pMSE*])2*,* (6)

which minimizes the squared deviation from one of the average *pMSE* ratio across *m* synthetic datasets.

Our goal is not to minimize differences between synthetic and confidential data, as doing so would replicate confidential records and compromise privacy. Instead, we aim to match the level of similarity expected between two independent samples from the same distribution, ensuring any resemblance is due to chance, as it would be in another real dataset. This approach is conceptually similar to generative adversarial networks (GANs), which train a generator and discriminator to produce synthetic data that converge to the true distribution (Goodfellow et al., 2020). However, unlike GANs, which are complex and lack a single parameter to control the privacy-utility trade-off, our method uses Bayesian optimization to tune simpler models to achieve the desired balance.

We note that the proposed method is not designed to completely automate the synthesis process. Rather, it relieves data stewards of the need to manually tune the parameters of specific models and allows them to focus more on tuning the synthesis process as a whole. For example, a data steward could use the proposed method to tune various combinations of sequential synthesis models automatically, before spending time comparing the utility and privacy of the synthetic datasets from each method. This example is particularly relevant for data stewards who understand the characteristics and use cases of their data (and are positioned to evaluate its quality) but lack the subject matter expertise to effectively tune synthesis models. In cases where data stewards are well versed in data synthesis, the proposed method can still speed up the tuning process. For example, even if it is clear in what direction parameters should be adjusted (e.g., the complexity parameter of a CART can be increased to reduce over-fitting), trial and error is required to determine the correct magnitude of the adjustment, and the proposed method is likely to find an optimal solution more quickly than an individual manually testing different parameter values.

# Empirical Application

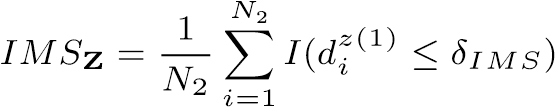
In this section, we evaluate whether commonly used synthesis models can produce synthetic data that meet the legal definition of anonymity under the GDPR. First, we outline the privacy criteria used by MOSTLY.AI and show that these criteria are met in expectation when synthetic data are generated from the same distribution as the confidential data. Using the proposed tuning method, we train sequential synthesis models to approximate the data-generating distribution. We compare the optimized synthesis approaches to the MOSTLY.AI method—treated as a baseline known to produce legally anonymous synthetic data—on two datasets: (1) location data from South Korean COVID-19 patients in 2020 and (2) the 1994 to 1996 CPS ASEC. We also compare the utility of regression models applied to the synthetic CPS ASEC data to the utility of differentially private regression models applied through a validation server (Barrientos et al., 2023). Finally, we simulate an attribute disclosure attack to determine whether legally anonymous synthetic data also meet academic standards for privacy protection.

## Privacy Criteria

Consider two confidential datasets, **Y** = (**y**1*,...,***y***N*1)T and **H** = (**h**1*,...,***h***N*1)T, both containing independently sampled records from the data-generating distribution D. A data steward releases a synthetic version of **Y**, denoted **Z** = (**z**1*,...,***z***N*2)T, where typically *N*1 = *N*2, though this is not required. We refer to **Y***,* **H***,* and **Z** as the confidential, holdout, and synthetic datasets, respectively. For **Z** to be legally anonymous, it must satisfy the three privacy criteria used by MOSTLY.AI (Hradec et al., 2022). Intuitively, synthetic data are considered private if, on average, they are no more similar to **Y** than **H** is. Thus, an ideal synthesis model will approximate D, making **Y** and **Z** interchangeable samples from the same distribution.

### Criterion One: Identical Match Share

Let *dzi,j* = dist(**z***i,***y***j*) denote a distance measure (e.g., Euclidean distance) between the *i*th and *j*th synthetic and confidential records, respectively. Furthermore, let denote the distance between the *i*th synthetic record and its *k*th nearest-neighbor confidential record. The *distance to closest record* (DCR) is the distance between the *i*th synthetic record and its nearest-neighbor confidential record, denoted. The *identical match share* (IMS) for the synthetic data is defined as

*,* (7)

where *I*(*.*) is the indicator function. The IMS is the proportion of synthetic records that are identical to their nearest-neighbor training record. For records with continuous attributes, there is zero probability of having an exactly identical record in a finite population. In this case, a small, nonzero *δIMS* may be specified as the threshold for records to be considered identical. Alternatively, continuous attributes can be discretized into bins, and a measure such as the Hamming distance may be used, in which case *δIMS* may be set to zero. Note that *IMS***H** can be defined similarly for the holdout dataset. The first privacy criterion requires that the IMS for the synthetic dataset be no more than the IMS for the holdout dataset:

*IMS***Z** ≤ *IMS***H***.* (8)

The ideal synthesis model would be identical to D such that sampling synthetic records would be no more likely to replicate a confidential record than sampling new confidential records from D. Privacy risks arise if the model produces a higher proportion of identical records than expected, indicating over-fitting and potentially revealing the presence of specific records in the confidential data. In such cases, the synthetic data fails to provide plausible deniability, as any similarity between synthetic and confidential records would exceed what is expected between two independent samples (Platzer and Reutterer, 2021).

### Criterion Two: Distance to Closest Record

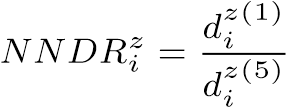
We let **F***DCR*(*z*) denote the empirical cumulative distribution function of the DCRs from the synthetic dataset. We denote the fifth percentile of **F**. Note that **F***DCR*(*h*) and can be defined similarly for the holdout dataset. The second privacy criterion requires that the fifth percentile of **F***DCR*(*z*) be no less than the fifth percentile of **F***DCR*(*h*):

*.* (9)

The ideal synthesis model would be identical to D to ensure that there is no systematic tendency for the synthesis model to produce synthetic records that are more similar to confidential records than holdout records to confidential records.

### Criterion Three: Nearest-Neighbor Distance Ratio

Let the *nearest neighbor distance ratio* (NNDR) for the *i*th synthetic record be denoted as

*,* (10)

which is bounded between [0*,*1] and denotes the ratio of the distances between the *i*th synthetic record and its nearest- and fifth nearest-neighbor confidential records. A ratio of one indicates that the synthetic record is equally similar to its nearest five confidential records. A ratio close to zero indicates that a synthetic record has a disproportionately higher similarity to its nearest-neighbor confidential record than to its fifth nearest-neighbor confidential record. This could indicate that a synthetic record is revealing the presence of an outlier in the confidential data. However, this is acceptable if the holdout data provide similar information on outliers, that is, if the distributions of the nearest-neighbor distance ratios are similar. Let **F***NNDR*(*z*) denote the empirical cumulative distribution function of the NNDRs of the synthetic dataset, and let  denote the fifth percentile of this distribution. Note that **F***NNDR*(*h*) and can be defined similarly for the holdout dataset.

The third privacy criterion requires that the fifth percentile of the nearest-neighbor distance ratios for synthetic records cannot be smaller than the fifth percentile for holdout records:

*pzNNDR* ≥ *phNNDR.* (11)

When compared to the holdout data, the ideal data synthesis model will produce synthetic records that are not overly similar to outliers compared to the other confidential records.

### Attribute Disclosure

The previous three privacy criteria do not account for adversarial attacks that could compromise privacy. Reidentification attacks are generally not a concern for fully synthetic data, as synthetic records do not correspond to real individuals (Hu, 2019). In contrast, attribute disclosure attacks, which use synthetic data to infer information about targeted individuals, remain a significant risk and are widely studied (Quick et al., 2015; Hittmeir et al., 2020; Guo and Hu, 2022). In our empirical application, we evaluate whether synthetic data deemed legally anonymous can protect against attribute disclosure.

We define an attribute disclosure attack as follows. Suppose an adversary seeks to deduce the value of a sensitive categorical or intervaled attribute *Sj* for a target record **y***j*, where the true value is *s*∗*j*. The adversary has access to the synthetic data **Z** but does not observe **y***j*. For the adversary, *Sj* is a random variable with possible values *sj* ∈ S and probability mass function *p*(*Sj* |**x***qj,***b**), where **x***qj* represents quasi-identifiers contained in **y***j* available to the adversary, and **b** includes other background information. After observing **Z**, the adversary updates their beliefs to *p*(*Sj* |**Z***,***x***qj,***b**) and predicts *Sj* as the value *s*′*j* with the highest probability,

*s*′*j* = argmax*sjp*(*Sj* = *sj* |**Z***,***x***qj,***b**)*.* (12)

To protect against attribute disclosure, the *increase* in probability of inferring the correct value *s*∗*j* for each sensitive record should be bounded, conditional on observing and analyzing **Z**. As suggested by Reiter et al. (2014), we measure the multiplicative increase and assume that the maximum increase over all *N*1 confidential records should be bounded as follows:

*p*(*Sj* = *s*∗*j*|**Z***,***x***qj,***b**)*/p*(*Sj* = *s*∗*j*|**x***qj,***b**) ≤ *c,* ∀ *j* = 1*,...,N*1*.* (13)

The value of *c* should be greater than or equal to one, and the strength of privacy protection decreases as *c* increases.

The prior probability *p*(*Sj* = *s*∗*j* |**x***qj,***b**) should be chosen to serve as a reasonable baseline for the ability of an adversary to infer values of the sensitive variable *Sj* without access to the synthetic dataset. Examples could be random guessing with equal probability over the possible values of *Sj*, or using publicly available population-level estimates of the probabilities of the possible values. To estimate *p*(*Sj* = *s*∗*j* |**Z***,***x***qj,***b**), we assume that the adversary constructs the following *δ*-neighborhood,

N(**x***qj*;*δAD,***Z**) = {*i* = 1*,...,N*2 : *d*(**x***qj*(*cont*)*,* **z***qi*(*cont*)) ≤ *δAD* ∧*I*(**x***qj*(*cat*) = **z***qi*(*cat*)) = 1}*.*

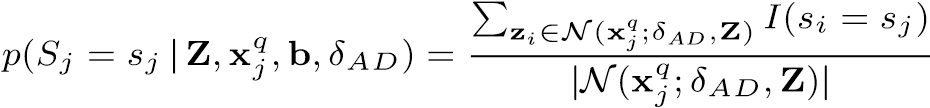
(14)

*q*(*cont*) *q*(*cont*)

The distance measure *d*(**x***j ,* **z***i* ) computes the similarity between the

*q*(*cont*) *q*(*cont*) continuous quasi-identifying attributes **x***j* and **z***i* . The expression *q*(*cat*) *q*(*cat*)

*I*(**x***j* = **z***i* ) is equal to one when the categorical and discrete quasi-identifying attributes of the external information and the *i*th synthetic record are identical. We assume that the adversary computes the probability mass function *p*(*Sj* = *sj* |**Z***,***x***qj,***b**) as follows:

*,* (15)

where *I*(*.*) is the indicator function. The probability *p*(*Sj* = *s*∗*j* |**Z***,***x***qj,***b***,δAD*) is the proportion of synthetic records within the *δ*-neighborhood of **x***qj* that match the true value *s*∗*j*. If no synthetic records fall within this neighborhood, the adversary must either abandon the attack or increase *δAD*. This estimate generalizes the “Type S” risk proposed by Quick et al. (2015). The choices of distance measure *d*(*.*) and threshold *δAD* are determined by the data steward and can be adjusted as needed based on data dimensionality.

## Tuning Method Applicability to Generating Legally Anonymous Synthetic Data

Our choice of optimization function in Section 3, including the use of logistic regression as a propensity score model and the pMSE ratio calculation, is designed to work well in general when synthesizing tabular data. Using the pMSE ratio-based objective function turns out to be highly appropriate for synthesizing legally anonymous synthetic data as well; our analysis in Section 2 of the supplementary materials shows that the legal privacy criteria from Section 4.1 will be met in expectation when synthetic data are sampled from the data-generating distribution D that produced the confidential data. Thus, in our objective function, we use the expected *pMSE* value from Snoke et al. (2018) that arises when the two datasets are drawn from the same data-generating distribution,

E(*pMSE*) = 2(*k* − 1)(1 − *c*)2*c/N,* (16)

where *k* is the number of variables that consist of synthesized values.

## The Datasets

We evaluate the performance of the optimized sequential synthesis method relative to the MOSTLY.AI method using two datasets. The first is a South Korean COVID-19 dataset, which includes location and demographic information for COVID-19 patients. This data, originally released by the “Data Science for COVID-19 (DS4C)” initiative on Kaggle, contains the following variables: Latitude/Longitude (location), Sex (male/female), Age (grouped into 10-year intervals from 0 to 99), and State (recovered or deceased). In total, there are N = 6,712 records, with 46.2% male and 55 deceased individuals.

Due to the sensitive nature of the previous dataset, we also analyze the publicly available 1994 to 1996 CPS ASEC data, which are accessible through IPUMS (Ruggles et al., 2021). This dataset allows other researchers to replicate our results and compare the utility results directly to those of the differentially private regression methods tested by Barrientos et al. (2023). To our knowledge, this is the first such comparison using the same underlying confidential data. The CPS ASEC dataset includes Income, Sex (male/female), Years of Education, Non-white (white/non-white racial status), and Potential Experience (age minus years of education minus six, minimum set to zero), with N = 197,756 records, 48.1% of which are female, and 28,072 of which are classified as “Non-white.”

## Synthesizer Implementation

We test two optimized sequential synthesis methods that each use a Gaussian mixture model to jointly synthesize continuous variables, and we then use either sequential CARTs or MNLs to synthesize the remaining categorical variables. We compare these methods to the proprietary method from MOSTLY.AI. We use Bayesian optimization to select the optimal parameters for the sequential methods, ensuring a balance between privacy and utility. For each synthesis strategy, we generate *m* = 20 synthetic datasets for both the South Korean COVID-19 data and the CPS ASEC data. Full implementation details, including model specifications and optimization procedures, are provided in the supplementary materials.

## Privacy Results

To calculate the privacy metrics, we use Euclidean distances between synthetic and confidential records, using between-record distances in the confidential dataset as a proxy for holdout data distances, which aligns with MOSTLY.AI’s approach for generating synthetic data privacy reports. Before calculating distances or simulating the attribute disclosure attack, we normalize the confidential data using the means and standard deviations of the synthetic variables.

### Legal Privacy Criteria

Figure 1 plots the average IMS across the *m* = 20 synthetic datasets from each synthesis method for a range of *δIMS* values. The lower the IMS value for a given value of *δIMS*, the better the privacy. For both datasets, the average IMS for all synthesis methods is less than or equal to the average IMS within the confidential data for all values of *δIMS*. In other words, on average, the proportion of records within the synthetic datasets that are identical to their nearest-neighbor confidential records is no higher than would be expected when sampling from the confidential data-generating distribution.

Figure 5 in the supplementary materials displays the distributions of the IMS (for *δIMS* = 0*.*001), DCR, and NNDR for the confidential and synthetic datasets from each synthesis method. The MOSTLY.AI datasets had higher IMS and lower DCR and NNDR values than the CART- and MNL-based datasets, indicating greater similarity to the confidential records. However, all synthetic datasets can be considered GDPR-compliant according to the MOSTLY.AI privacy criteria.

**Average IMS Across Synthetic Datasets**

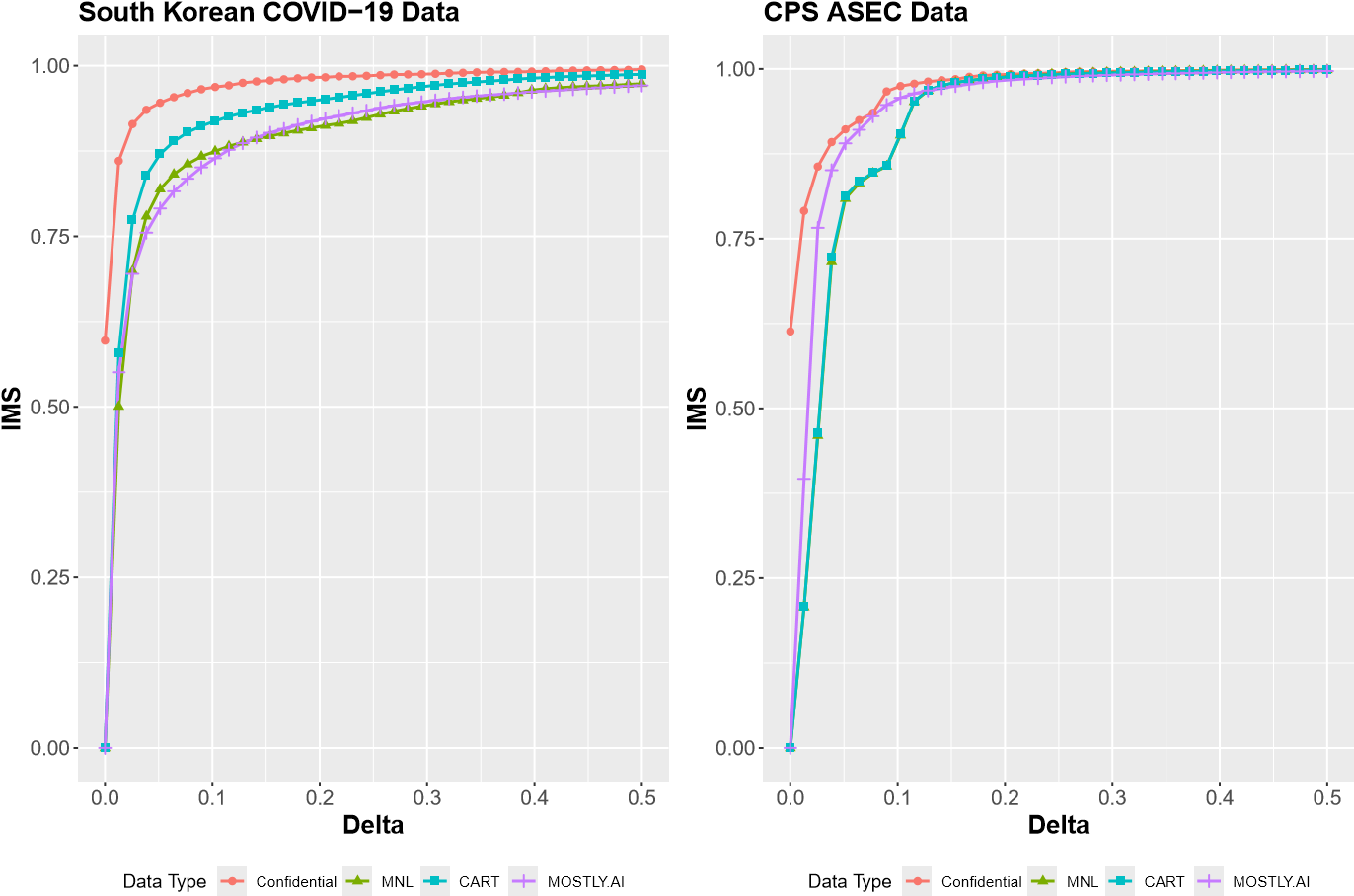


Fig. 1: Average IMS value across *m* = 20 synthetic datasets from each synthesis method. Lower IMS values (relative to the confidential dataset) indicate better privacy.

### Attribute Disclosure

To assess protection against attribute disclosure, we treat Status and Non-white as sensitive variables in the COVID-19 and CPS ASEC datasets, respectively, and compute the maximum of the probability ratio in Equation (13) in each synthetic dataset, assuming all other variables are known to the adversary, for a range of *δAD* values. The lower the probability ratio, the less likely an adversary can predict the value of a sensitive variable.

**Average Maximum Attribute Disclosure Probability Ratio**

0

20

40

60

0.0

0.1

0.2

0.3

0.4

0.5

**Delta**

**Probability Ratio**

Data Type

MNL

CART

MOSTLY.AI

**South Korean COVID−19 Data**

0

2

4

6

0.0

0.1

0.2

0.3

0.4

0.5

**Delta**

**Probability Ratio**

Data Type

MNL

CART

MOSTLY.AI

**CPS ASEC Data**

Fig. 2: Average maximum attribute disclosure probability ratio across synthetic datasets and *δAD* values for each synthesis method. Lower values equate to better protection against attribute disclosure.

For the COVID-19 data, we use publicly available data from the World Health Organization (WHO, 2024) detailing the total number of confirmed COVID-19 cases (9,583) and the total number of deaths (152) in South Korea as of March 23, 2020, to calculate the prior probabilities *p*(*Sj* = 1|**x***qj,***b**) = 152*/*9583 ≈ 0*.*016 and *p*(*Sj* = 0|**x***qj,***b**) = 1 − 0*.*016 = 0*.*984. Thus, the maximum value for the probability ratio in Equation (13) occurs when the *δ*−neighborhood N(**x***qj*;*δAD,***Z**) contains only synthetic records of individuals who passed away (*Sj* = 1), *i.e.*, *p*(*Sj* = 1|**Z***,***x***qj,***b**)*/p*(*Sj* = 1|**x***qj,***b**) = 1*.*00*/*0*.*016 = 62*.*5.

For the CPS ASEC data, we use *p*(*Sj* = 1|**x***qj,***b**) = 0*.*142 and *p*(*Sj* =

0|**x***qj,***b**) = 0*.*858, the proportions of confidential records that were labeled “Non-white”/“White,” respectively. The maximum probability ratio occurs when the *δ*−neighborhood N(**x***qj*;*δAD,***Z**) contains only “Non-white” synthetic records, that is, *p*(*Sj* = 1|**Z***,***x***qj,***b**)*/p*(*Sj* = 1|**x***qj,***b**) = 1*.*00*/*0*.*142 = 7*.*04.

Figure 2 shows the average maximum values of Equation (13) across all synthetic datasets for each method. None of the methods provide effective protection against attribute disclosure. For the COVID-19 data, values of *δAD* between 0.1 and 0.25 produce probability ratios close to the maximum of 62.5, indicating that an adversary could correctly predict *s*∗*j* with near certainty in a worst-case scenario. When *δAD* is near zero, the MNL-based method offers some protection, while the CART-based and MOSTLY.AI methods allow for larger increases in disclosure risk. This increased risk is especially concerning because adversaries may place higher confidence in predictions made with smaller *δAD*. A similar pattern is observed in the CPS ASEC data: almost all *δAD* values result in ratios near the maximum. For both datasets, the probability ratios decrease as *δAD* increases, with the numerator converging to the marginal distribution of *S* in the synthetic dataset conditional on the categorical quasi-identifiers once  for all synthetic records.

## Utility Results

We assess both the global utility (see Section 4.6.1) and the analysis-specific utility (see Section 4.6.2) of the synthetic datasets from all methods.

### Global Utility

We assess the global utility of the synthetic datasets using the *pMSE* ratio described in Section 3, which was used in the objective of the Bayesian optimization process. *pMSE* ratios near one are desired, as this result indicates that the synthesis method is accurately approximating the data-generating distribution. Figure 3 displays the *pMSE* ratio distributions for each synthesis method for each dataset.

**pMSE Ratio Distributions**

0.0

0.5

1.0

1.5

2.0

MNL

CART

MOSTLY.AI

**Data Type**

**pMSE Ratio**

**South Korean COVID−19 Data**

0

10

20

30

MNL

CART

MOSTLY.AI

**Data Type**

**pMSE Ratio**

**CPS ASEC Data**

Fig. 3: *pMSE* ratio distributions for each synthesis method for the South Korean COVID-19 Data (left) and the CPS ASEC data (right). Ratios near one are desired.

For the COVID-19 data, the averages of the *pMSE* ratio distributions under the MNL- and CART-based synthesis methods are approximately one, indicating that the synthesis methods accurately approximate the data-generating distribution. The MOSTLY.AI method had slightly worse utility on the COVID-19 data, with an average ratio of approximately 1.43. For the CPS ASEC data, the averages of the *pMSE* ratio distributions under the MNL- and CART-based synthesis methods—approximately five—and the average ratio from the MOSTLY.AI method, which is several orders of magnitude larger, suggest that the synthesis methods are biased and are producing synthetic records that are less similar to confidential records than would be expected when sampling from the data-generating distribution.

### Analysis Specific Utility

We evaluate analysis-specific utility by comparing the results of a data user’s model applied to both the confidential and synthetic datasets. The goal is for a data user to obtain results from analyzing the synthetic data that are similar to the results that would have been obtained from analyzing the confidential data. We measure utility using the sign, significance, and overlap (SSO) percentage (Barrientos et al., 2023), which represents the percentage of the time in which the sign and statistical significance (at the *α* = 0*.*05 level) of the synthetic and confidential coefficients match and in which the corresponding confidence intervals overlap. Synthetic data with a high SSO percentage enables inference similar to what would be obtained using the confidential data.

Sign, Significance, and Overlap for COVID−19 Data Coefficients

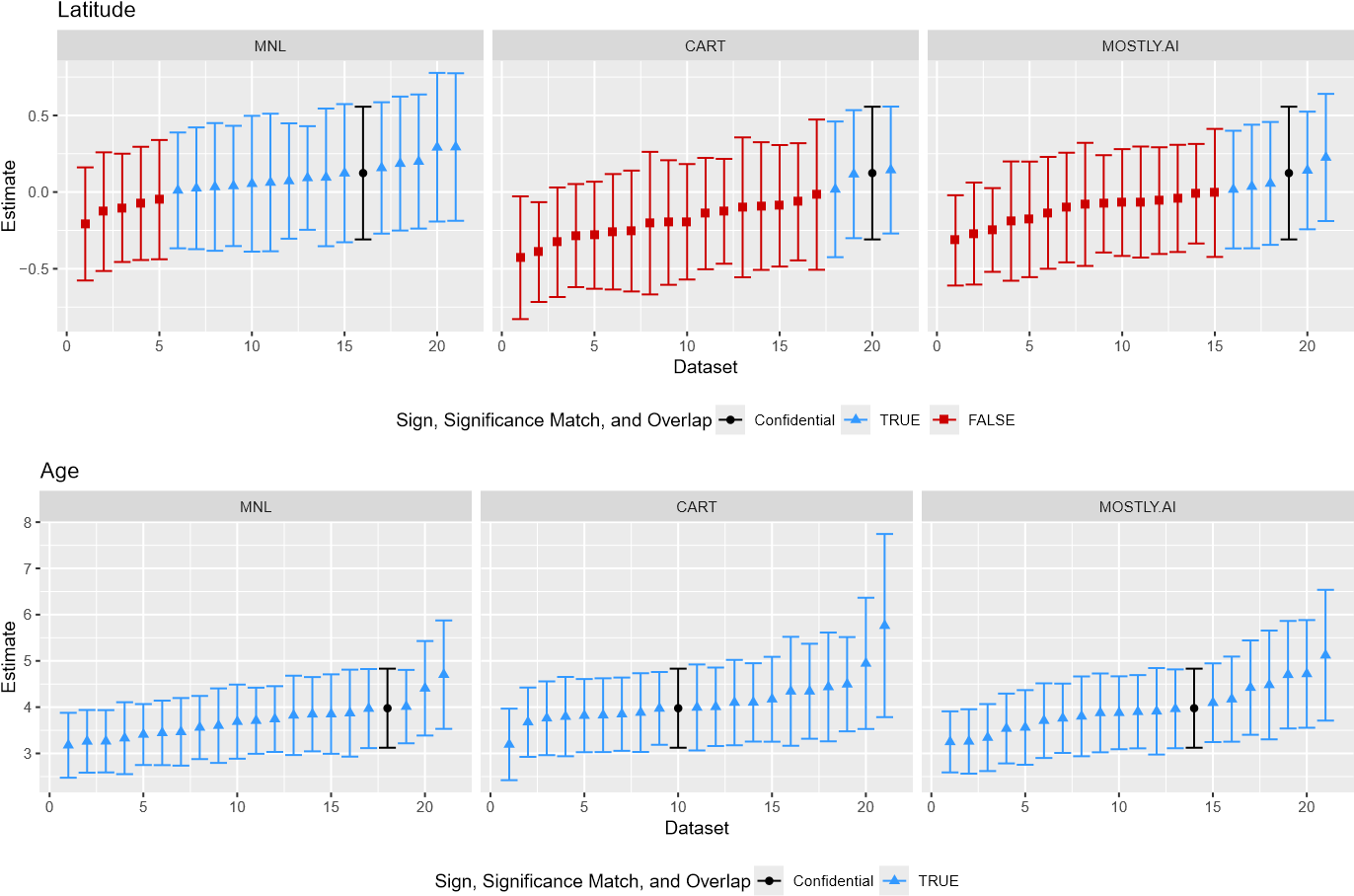


Fig. 4: Sign, Significance, and Overlap for the Latitude and Age point and confidence interval estimates from the *m* = 20 synthetic South Korean COVID-19 datasets for each synthesis method.

For the data user model for the COVID-19 data, we estimate a logistic regression that predicts Status as a function of Latitude, Longitude, Age, and Sex. To avoid quasi-complete separation, we binarize Age to indicate whether a record corresponds to an individual age 60 or older. We examine the SSO results for all *m* = 20 synthetic datasets for each synthesis method for the COVID-19 data by displaying the point estimates and confidence intervals for the Latitude and Age coefficients in Figure 4. The Latitude coefficient is not different from zero at a statistically significant level, and the vast majority of confidence intervals for all synthesis methods include zero. Even though many of the point estimates have a different sign, this would not affect the overall conclusion. Age (indicator for whether an individual is 60 or older), on the other hand, has a strong, positive, statistically significant effect on the probability that an individual passed away from COVID-19. This is reflected in the point estimates and confidence intervals for all synthesis methods. SSO results for the intercept and coefficients for Longitude and Sex are shown in Figure 6 in the supplementary materials. Overall, all synthesis methods produced synthetic datasets with good analysis-specific utility, illustrated by synthetic point estimates that maintained the inferential conclusions from the confidential data. The only exceptions were a few statistically significant estimates of the Longitude coefficient and non-statistically-significant estimates for the Sex coefficient from CART-based synthetic datasets.

For the CPS ASEC data, we estimate the regression models from Barrientos et al. (2023), which estimate the log of Income as a function of Years of Education, Non-white, and Potential Experience as well as its second- (Potential Experience Squared) and third-degree (Potential Experience Cubed) polynomials. This model is estimated for each value of Sex (male/female) separately. For comparability with the results from Barrientos et al. (2023), we present our results for the Years of Education and Non-white coefficients of the female regression in this section and include the results for the other coefficients from the female-based model and the male-based model in the supplementary materials.

Sign, Significance, and Overlap for CPS ASEC Data Coefficients (Female)

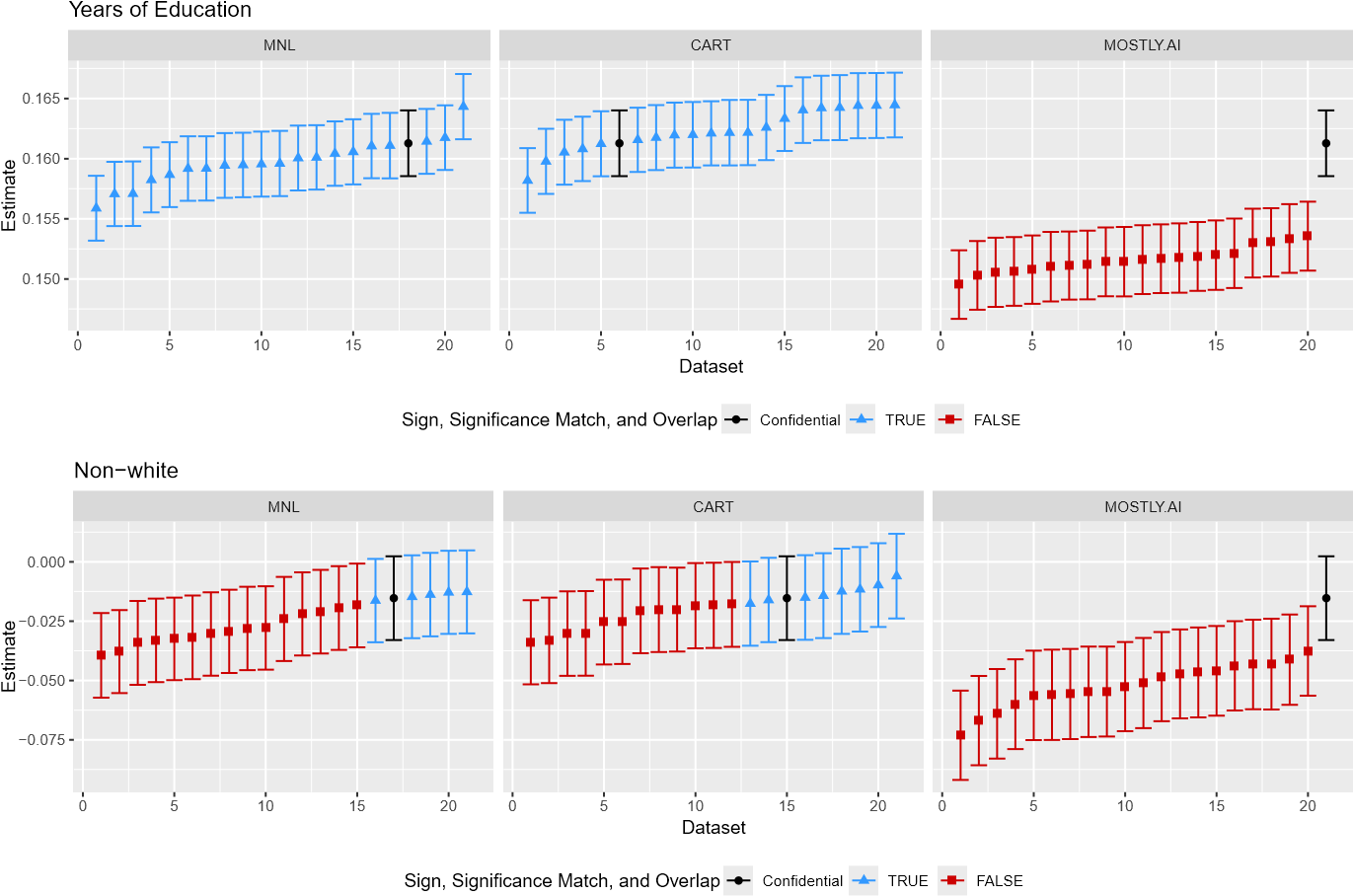


Fig. 5: Sign, Significance, and Overlap for the Years of Education and Non-White point and confidence interval estimates for the Female regression from the *m* = 20 synthetic CPS ASEC datasets for each synthesis method.

Figure 5 shows that the optimized sequential methods achieved 100% SSO coverage on the Years of Education coefficient, whereas the MOSTLY.AI method produced attenuated estimates that are smaller in magnitude than the confidential estimate but still illustrate a positive, statistically significant effect. The SSO results for the Non-white coefficient are relatively poor for all methods: only a small proportion of intervals for the MNL- and CART-based methods maintains the conclusion from the confidential data (no statistically significant effect), with most intervals indicating that “Non-white” status has a statistically significant negative effect on Income, particularly under the MOSTLY.AI method. SSO results for the model intercept and coefficients for Potential Experience and its polynomials are shown in Figure 7 in the supplementary materials. The MNL-based method had nearly 100% SSO coverage on all coefficients except Non-white. The CART-based method had high SSO coverage for the intercept but struggled to maintain correct estimates of the Potential Experience coefficients. On the other hand, the MOSTLY.AI method had high SSO coverage on the Potential Experience coefficients but low SSO coverage on the intercept.

## Comparison with Differentially Private Regression

Table 1 compares two differentially private regression methods assessed by Barrientos et al. (2023) to the synthesis methods from this paper on two metrics: the confidence interval ratio (CIR—ratio of the width of the confidence interval from protected data to the confidence interval from confidential data) and the SSO match percentage. Specifically, we select the analytic Gaussian mechanism (AGM) and Laplace mechanism (LM) with *ϵ* = 5 with confidence intervals estimated using the asymptotic approach, as these produced confidence interval ratios that were approximately one, which are comparable in width to the intervals estimated from the synthetic datasets. We note that setting *ϵ* = 5 resulted in the best utility (measured in terms of the CIR and SSO %) for the DP methods.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Coefficient** | **DP (AGM) DP (LM) MNL-Based CART-Based** | | | | | | **MOSTLY.AI** |
| *CIR SSO % CIR SSO % CIR SSO % CIR SSO %* | | | | | | *CIR SSO %* |
| Non-White | 1.00 53% | 1.01 | 43% | 1.00 | 24% | 1.01 38% | 1.07 0% |
| Years of Education | 1.00 62% | 1.01 | 42% | 0.99 | 95% | 0.99 95% | 1.05 0% |

Table 1. Confidence interval ratios (CIR) and SSO match percentage for the DP AGM and DP LM regression methods from Barrientos et al. (2023), with confidence intervals estimated using the asymptotic approach, compared to the results from all synthesis methods.

The DP methods have a higher SSO % than the synthetic datasets for the Non-White coefficient, whereas the optimized sequential synthesis methods produce higher SSO % for the Years of Education coefficient. Not shown here are the results for the intercept and Potential Experience coefficients, on which the DP methods performed very poorly (Barrientos et al. [2023] note that the results on the Non-White and Years of Education coefficients are a “good case”) but where the MNL-based synthesis method performed very well, achieving nearly 100% SSO match for the female-based regression. As noted previously, the CART-based and MOSTLY.AI methods struggled to maintain inferential conclusions for some of the other coefficients. Overall, none of the legally anonymous synthetic datasets or the DP regression produced coefficient estimates that consistently maintained the inferential conclusions from the confidential data.

Finally, we highlight that the SSO % results differ for the male-based regression, where all synthesis methods produced high SSO % for the Years of Education and Non-White coefficients. However, the optimized sequential methods had 0% SSO match for the intercept and Potential Experience coefficients, while the MOSTLY.AI method had better estimates for the intercept and main Potential Experience coefficient. While not tested here, the regression estimates would likely improve for the synthetic datasets if synthesis were performed on the male and female subsets separately to remove the noise induced from synthesizing the Sex variable prior to splitting the data and estimating the user regression models.

# Discussion

This paper proposed a tuning method that uses Bayesian optimization to select synthesis model parameters that balance privacy and utility, eliminating the need for manual tuning of these parameters. This method lowers the barriers to using and fine-tuning synthetic data methods, and was demonstrated to produce synthetic data with high analysis-specific utility for two datasets with very different numbers of observations. While it will always be necessary to evaluate synthetic data with various utility and privacy metrics, our method removes the need for data stewards to tune individual synthesis models and enables them to focus on improving the overall synthesis process, for example, by comparing the results of multiple optimized synthesis methods as we did in our empirical application. At a minimum, we have demonstrated the usefulness of the tuning method for synthesizing data when legal compliance is prioritized alongside data utility and for privacy nonexperts who want a more automated approach to training a synthesis model.

Given the complex landscape of privacy regulations, we evaluated whether the proposed method could tune sequential synthesis models to generate legally anonymous synthetic data based on the privacy criteria defined by MOSTLY.AI. We showed that when synthetic data are drawn from the same data-generating distribution as the confidential data, these legal privacy criteria are met in expectation. We generated legally anonymous versions of the South Korean COVID-19 location data and the CPS ASEC data and compared the results to synthetic data generated by MOSTLY.AI. Our findings confirmed that sequential synthesis models can produce legally anonymous synthetic data. However, we also demonstrated that meeting these legal criteria does not guarantee robustness against attribute disclosure attacks. We found that, given sufficient external information, sensitive values can still be inferred, raising the question of how much protection legally anonymous data should offer against such attacks. Addressing this question will require further research and will depend on the specific context and objectives of the anonymization process (Working Party, 2014).

Additionally, we compared regression results from legally anonymous synthetic CPS ASEC data to differentially private regression results from Barrientos et al. (2023). Our results showed that legally anonymous synthetic data can produce coefficient estimates that are more consistent with those from the original data. However, this improved utility comes with increased privacy risk because sharing a fully synthetic dataset poses a greater risk of attribute disclosure than a query-based validation server, and data stewards will need to weigh the pros and cons of each protection method. We note that both synthetic and differentially private data utility can be improved—by allowing greater similarity to the original data or increasing the privacy parameter *ϵ*—but these changes may compromise privacy to such an extent that legal anonymity cannot be claimed.

This study has some limitations that suggest avenues for future research. First, the *pMSE* ratio used to optimize the synthesis models is model-dependent, and relying on it alone may provide an inaccurate depiction of data utility (Drechsler, 2022). We showed that the tuning process works well with a *pMSE*-based objective function, but future work could explore alternative metrics and objectives. Such alternatives would be highly useful to data stewards who are interested in exploring customized objectives that suit their specific use cases for synthetic data. Future work should explore alternative distributional similarity measures, such as the SPEC metric (Bowen et al., 2021) or the multivariate Hellinger distance (El Emam et al., 2022), and examine how different objective functions influence the balance between privacy and utility. Additionally, while our method reduces the need for manual parameter selection, it still requires specifying feasible parameter ranges. The parameter ranges we supplied worked well for both datasets we examined in this paper and should serve as reasonable starting points for others using the tuning method. However, developing best practices for selecting these ranges or employing techniques like sequential domain reduction to refine the search space could further enhance the practicality of the method.

# References

Asghar, H. J., Ding, M., Rakotoarivelo, T., Mrabet, S. and Kaafar, D. (2020) Differentially private release of high-dimensional datasets using the Gaussian copula. *Journal of Privacy and Confidentiality*, **10**(2). doi: https://doi.org/10.29012/jpc.686.

Azizi, Z., Zheng, C., Mosquera, L., Pilote, L. and El Emam, K. (2021) Can synthetic data be a proxy for real clinical trial data? a validation study. *BMJ Open*, **11**(4), e043497.

Barrientos, A. F., Williams, A. R., Snoke, J. and Bowen, C. M. (2023) A feasibility study of differentially private summary statistics and regression analyses with evaluations on administrative and survey data. *J. Am. Statist. Ass.*, 1–14.

Bowen, C. M., Bryant, V., Burman, L., Czajka, J., Khitatrakun, S., MacDonald, G., McClelland, R., Mucciolo, L., Pickens, M., Ueyama, K., et al. (2022) Synthetic individual income tax data: Methodology, utility, and privacy implications. In *International Conference on Privacy in Statistical Databases*, pp. 191–204. Springer.

Bowen, C. M., Liu, F. and Su, B. (2021) Differentially private data release via statistical election to partition sequentially: Statistical election to partition sequentially. *Metron*, **79**(1), 1–31.

Bowen, C. M. and Snoke, J. (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). doi: 10.29012/jpc.748.

Brochu, E., Cora,V. M. and De Freitas, N. (2010) A tutorial on Bayesian optimization of expensive cost functions, with application to active user modeling and hierarchical reinforcement learning. *arXiv preprint arXiv:1012.2599*.

Cohen, A. and Nissim, K. (2020) Towards formalizing the gdpr’s notion of singling out. *Proceedings of the National Academy of Sciences*, **117**(15), 8344–8352.

Drechsler, . (2022) Challenges in measuring utility for fully synthetic data. In *International Conference on Privacy in Statistical Databases*, pp. 220–233. Springer.

Drechsler, . (2023) Differential privacy for government agencies—Are we there yet? *J. Am. Statist. Ass.*, **118**(541), 761–773.

Drechsler, J. and Hu, J. (2021) Synthesizing geocodes to facilitate access to detailed geographical information in large-scale administrative data. *Journal of Survey Statistics and Methodology*, **9**(3), 523–548.

Drechsler, J. and Reiter, J. P. (2011) An empirical evaluation of easily implemented, nonparametric methods for generating synthetic datasets. *Computational Statistics & Data Analysis*, **55**(12), 3232–3243.

El Emam, J., Mosquera, L., Jonker, E. and Sood, H.. (2021) Evaluating the utility of synthetic covid-19 case data. *JAMIA Open*, **4**(1), ooab012.

El Emam, K., Mosquera, L., Fang, X. and El-Hussuna, A. (2022) Utility metrics for evaluating synthetic health data generation methods: Validation study. *JMIR medical informatics*, **10**(4), e35734.

European Data Protection Board. (2020) Guidelines 04/2020 on the use of location data and contact tracing tools in the context of the COVID-19 outbreak. URL [https://edpb.europa.eu/our-work-tools/our-documents/ guidelines/guidelines-042020-use-location-data-and-contact-tracing\_en.](https://edpb.europa.eu/our-work-tools/our-documents/guidelines/guidelines-042020-use-location-data-and-contact-tracing_en)

European Parliament and Council of European Union. (2016) Regulation (EU) 2016/679 of the European Parliament and of the Council. URL [https://eur-lex.europa. eu/eli/reg/2016/679/oj.](https://eur-lex.europa.eu/eli/reg/2016/679/oj)

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. and Bengio, Y.. (2020) Generative adversarial networks. *Communications of the ACM*, **63**(11), 139–144.

Guo, S. and Hu, J. (2022) Data privacy protection and utility preservation through Bayesian data synthesis: A case study on Airbnb listings. *The American Statistician*, 1–9.

Hintze, M. and El Emam, K. (2020) Can synthetic data help organizations respond to ‘Schrems II’? iapp. URL [https://iapp.org/news/a/can-synthetic-data-help-organizations-respond-to-schrems-ii/.](https://iapp.org/news/a/can-synthetic-data-help-organizations-respond-to-schrems-ii/)

Hittmeir, M., Mayer, R. and Ekelhart, A. (2020) A baseline for attribute disclosure risk in synthetic data. In *Proceedings of the Tenth ACM Conference on Data and Application Security and Privacy*, pp. 133–143.

Hradec, J., Craglia, M., Di Leo, M., De Nigris, S., Ostlaender, N. and Nicholson, N. (2022) Multipurpose synthetic population for policy applications. *EUR 31116 EN, Publications Office of the European Union, Luxembourg*. doi: 10.2760/50072.

Hu, J. (2019) Bayesian estimation of attribute and identification disclosure risks in synthetic data. *Transactions on Data Privacy*, **12**, 61–89.

Hu, J. and Bowen, C. M. (2024) Advancing microdata privacy protection: A review of synthetic data methods. *Wiley Interdisciplinary Reviews: Computational Statistics*, e1636.

Jackson, J., Mitra, R., Francis, B. and Dove, I. (2022) Using saturated count models for user-friendly synthesis of large confidential administrative databases. *J. R. Statist. Soc.* A.

Koch, R. (2018) Data anonymization and GDPR compliance: The case of Taxa 4x35. URL [https://gdpr.eu/data-anonymization-taxa-4x35.](https://gdpr.eu/data-anonymization-taxa-4x35,)

Lee, C. and Anand, P. (2020) Using deep learning to overcome privacy and scalability issues in customer data transfer. *Available at SSRN 3769521*.

Nowok, B., Raab, G. M. and Dibben, C. (2016) synthpop: Bespoke creation of synthetic data in R. *Journal of Statistical Software*, **74**(11), 1–26. doi: 10.18637/jss.v074.i11.

Platzer, M. and Reutterer, T. (2021) Holdout-based empirical assessment of mixed-type synthetic data. In *Frontiers in big Data*, p. 43.

Quick, H. (2021) Generating Poisson-distributed differentially private synthetic data. *J. R. Statist. Soc.* A, **184**(3), 1093–1108.

Quick, H. (2022) Improving the utility of Poisson-distributed, differentially private synthetic data via prior predictive truncation with an application to cdc wonder. *Journal of Survey Statistics and Methodology*, **10**(3), 596–617.

Quick, H., Holan, S. H., Wikle, C. K. and Reiter, J. P. (2015) Bayesian marked point process modeling for generating fully synthetic public use data with point-referenced geography. *Spatial Statistics*, **14**, 439–451.

Reiter, J. P. (2005) Using cart to generate partially synthetic public use microdata. *J. Off. Statist.*, **21**(3), 441.

Reiter, J. P., Wang, Q. and Zhang, B. (2014) Bayesian estimation of disclosure risks for multiply imputed, synthetic data. *Journal of Privacy and Confidentiality*, **6**(1).

Ruggles, S., Flood, S., Foster, S., Goeken, R., Pacas, J., Schouweiler, M. and Sobek, M. (2021) Ipums usa: Version 11.0 [dataset]. URL [https://doi.org/10.18128/D030.V8.0.](https://doi.org/10.18128/D030.V8.0)

Schneider, M. J. and Abowd, J. M. (2015) A new method for protecting interrelated time series with Bayesian prior distributions and synthetic data. *J. R. Statist. Soc.* A, 963–975.

Schneider, M. J., Hu, J., Mankad, S. and Bale, C. D. (2023) Protecting the anonymity of online users through Bayesian data synthesis. *Expert Systems with Applications*, **216**, 119409.

Snoek, J., Larochelle, H. and Adams, R. P. (2012) Practical Bayesian optimization of machine learning algorithms. *Advances in neural information processing systems*, **25**.

Snoke, J., Raab, G. M., Nowok, B., Dibben, C. and Slavkovic, A.. (2018) General and specific utility measures for synthetic data. *J. R. Statist. Soc.* A, **181**, 663–688.

Wang, H. and Reiter, J. P. (2012) Multiple imputation for sharing precise geographies in public use data. *Ann. Appl. Statist.*, **6**(1), 229.

WHO. (2024) Number of covid-19 cases reported to WHO. WHO COVID-19 dashboard.. URL [https://data.who.int/dashboards/covid19/cases?n=c.](https://data.who.int/dashboards/covid19/cases?n=c)

Working Party. (2014) Opinion 05/2014 on anonymisation techniques.

1. See [Privacy & security in MOSTLY.AI’s synthetic data platform.](https://mostly.ai/privacy-and-security/) [↑](#footnote-ref-1)