UNITED NATIONS ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Expert meeting on Statistical Data Confidentiality

26–28 September 2023, Wiesbaden

ASSESSING THE UTILITY OF SYNTHETIC DATA: A DENSITY RATIO PERSPECTIVE

Thom Benjamin Volker (Utrecht University, the Netherlands; Statistics Netherlands, the Netherlands)
Peter-Paul de Wolf (Statistics Netherlands, the Netherlands)
Erik-Jan van Kesteren (Utrecht University, the Netherlands)
t.b.volker@uu.nl, pp.dewolf@cbs.nl, e.vankesteren1@uu.nl

Abstract

High quality synthetic data can be a solution to overcome disclosure risks that arise when disseminating research data to the public. However, for inferential purposes, the usefulness of the synthetic data largely depends on whether the synthetic data model approximates the distribution of the observed data close enough. Often, multiple candidate models are considered and improvements are made iteratively. Evaluating the utility of the synthetic data after the model building steps is a crucial but complicated endeavor, and although many methods exist, their results may tell an incomplete story. We propose to evaluate the utility of synthetic data using density ratio estimation techniques, which allows to embed both general and specific utility measures into a common framework. Using techniques from the density ratio estimation field, we show how an interpretable global utility measure can be obtained from the ratio of the observed and synthetic data densities that can even be used to express the utility of individual (synthetic) data points. Applying these techniques on (an approximation to the) posterior distributions of parameters gives rise to analysis-specific utility measures. Using empirical examples, we show that framing utility from a density ratio perspective improves on existing global utility measures. Lastly, we implemented the procedure along with existing utility measures in an R-package.

Introduction

In recent years, the academic interest in synthetic data has exploded. Synthetic data are increasingly being used as a solution to overcome privacy and confidentiality issues that are inherently linked to the dissemination of research data. National statistical institutes and other government agencies have started to disseminate synthetic data to the public while restricting access to the original data to protect sensitive information (e.g., Abowd, Stinson, and Benedetto 2006; Hawala 2008; Drechsler 2012). At the same time, researchers start to share a synthetic version of their research data to comply with open science standards (e.g., Wiel et al. 2023; Obermeyer et al. 2019; Zettler et al. 2021). Rather than sharing the original research data, a synthetic surrogate is shared to facilitate reviewing the data processing and analysis pipeline. Additionally, synthetic data is increasingly being used for training machine learning models (Nikolenko 2021). On a lower level, synthetic data can be used in model testing pipelines (before access to the real data is provided), for data exploration, and for educational purposes.

At its core, the idea of synthetic data is to replace values from the observed data with new values that are generated from a model. In this way, it is possible to generate an entirely new synthetic data set (commonly referred to as the *fully* synthetic data approach; Rubin 1993), but also to replace just those values that would yield a high risk of disclosure when released (called *partially* synthetic data; Little 1993). Both approaches essentially attempt to build a model that incorporates as much of the information in the real data as possible, given a pre-specified level of privacy risk that is still deemed acceptable. The models used to generate synthetic data were originally closely related to methods used for multiple imputation of missing data, such as fully conditional specification (Volker and Vink 2021) or sequential regression (Nowok, Raab, and Dibben 2016). Recently, significant improvements in generative modelling sparked the scientific interest in synthetic data in the computer science community, leading to novel synthesis methods (e.g., Patki, Wedge, and Veeramachaneni 2016; Xu et al. 2019). Combined with work on formal privacy guarantees, this resulted in new models that explicitly control the level of privacy risk in synthesis methods (Jordon, Yoon, and Schaar 2019; Torkzadehmahani, Kairouz, and Paten 2019). Through both methodological advances and practical implementations of data synthesis methods, the notion of synthetic data has developed into an increasingly popular solution to enhance data dissemination.

Regardless of these advances, the main challenge when generating synthetic data remains to adequately balance the privacy risk with the utility (i.e., quality) of the synthetic data. On the upper limit of this privacy-utility trade-off, the synthesis model captures the information in the observed data so precisely, that the real data is exactly reproduced, resulting in the same privacy loss as when disseminating the real data. In statistical terms, the synthesis model is overparameterized to such an extent that there are no degrees of freedom left, and there is thus no randomness involved in the generation of the synthetic values. On the lower limit of the trade-off, synthetic values are generated without borrowing any information from the real data. For example, we could place the value 0 or a random draw from a standard normal distribution for every record and every variable, such that the synthetic data contains only noise. Synthetic data sets sit somewhere between these two extremes: they contain some information from the real data, yielding some disclosure risk, but they also resemble the real data to some extent, yielding more than zero utility. At the same time, not all of the information is captured, and the utility of the synthetic data will thus always be lower than the utility of the real data. The question that naturally arises is where on the privacy-utility continuum the synthetic data is located: how much information is sacrificed, and what aspects of the real data is reproduced in the synthetic data. From the perspective of the data provider, it is important to know what privacy loss is incurred by disseminating the synthetic data, while the user wants to know whether their analysis can be reliably performed. Additionally, knowledge about the utility of the synthetic data can be used by the data provider to finetune the synthesis model and thus improving the synthetic data quality.

To evaluate the utility of synthetic data, three classes of utility measures have been distinguished in the synthetic data literature (for a thorough review of these measures, see Drechsler and Haensch 2023): fit-for-purpose measures, analysis-specific utility measures and global utility measures. Fit-for-purpose measures are typically the first step in assessing the quality of the synthetic data. They typically involve comparing the univariate distributions of the observed and synthetic data (for example using visualization techniques or goodness-of-fit measures). Although these measures provide an initial impression of the quality of the data synthesis models used, this picture is by definition limited, because only one or two variables are assessed at the same time. Hence, complex relationships between variables will always be out of scope. Global utility measures build on the fit-for-purpose measures, but attempt to capture the quality of the entire multivariate distribution of the synthetic data relative to the observed data in a single, global, indicator. This can be done using some distance measure (e.g., the Kullback-Leibler divergence; see Karr et al. 2006), but also by estimating how well a prediction model can distinguish between the observed and synthetic data, and using the predicted probabilities (propensity scores; Rosenbaum

and Rubin 1983) as a measure of discrepancy (e.g., the propensity score mean squared error, $pMSE$; Woo et al. 2009; Snoke et al. 2018). While global utility measures paint a rather complete picture, and provide information over the entire range of the data, they tend to be too general. That is, global utility measures can be so broad that important discrepancies between the real and synthetic data can be missed, and an a synthetic data set with high global utility might still yield analyses with results that are far from the results from real data analyses (see Drechsler 2022). Lastly, the analysis-specific utility measures quantify to what extent analyses performed on the synthetic data align with the same analyses on the observed data. These measures can evaluate to what degree the coefficients of a regression model are similar [e.g., using the confidence interval overlap; Karr et al. (2006)], but also to what extent prediction models trained on the synthetic and observed data perform similarly in terms of evaluation metrics. However, analysis-specific utility generally does not carry over, even not to closely related analyses: high specific utility for one analysis does not at all imply high utility for another analysis. Since data providers typically do not know which analyses will be performed with the synthetic data, it is impossible to provide analysis-specific utility measures for all potentially relevant analyses (for a more thorough discussion, see Drechsler 2022).

In this paper, we propose to use the framework of density ratio estimation (Sugiyama, Suzuki, and Kanamori 2012a) to place all above measures under a common umbrella. We show that these measures perform as least as good as varying existing utility measures, while providing a more fine-grained view of the degree of misfit of the synthetic data. At the same time, the density ratio estimation framework requires only little specification on the side of the user, as it is a non-parametric approach with automatic model selection. In short, density ratio estimation compares the multivariate distributions of two data sets (e.g., two different samples or groups) by directly estimation the ratio of their densities. Crucially, this method does not estimate the densities of the observed and synthetic data separately, subsequently taking their ratio, but rather estimates the ratio of the densities directly, which has been shown to yield better performance. The idea is that if two data sets are drawn from the same data-generating mechanism, the sampled data should be similar, and the ratio of their densities should thus be close to one at all possible points in the multivariate space. This approach easily extends from univariate to bivariate and multivariate densities. As such, we show how it bridges the gap between fit-for-purpose and global utility measures. Moreover, we discuss how density ratio estimation can also be used to compare the posterior distributions of parameters of observed and synthetic data, and thus also incorporates analysis-specific utility measures. The only requirement for density ratio estimation is that there are samples (e.g., observations, samples from a posterior) from the observed and synthetic data (or parameter) distributions. Hence, we show that it is a versatile approach that is useful in the entire domain of data utility.

Also from the privacy-side several promising advances have been made to quantify the amount of information leakage through the synthetic data. Important work has been done to build formal privacy guarantees into the synthesis models through differential privacy (Dwork 2006). In addition to these privacy-by-design mechanisms, some measures exist to quantify privacy loss of synthetic data after generation (e.g., McClure and Reiter 2016; Reiter and Mitra 2009; Hu 2019). However, the practical applicability of these measures depends on whether the data is fully or partially synthetic, and especially in case of the former, the practical applicability of these measures is often limited (for an extensive discussion of these issues, see Drechsler and Haensch 2023). More research on measures to evaluate disclosure risks in synthetic is thus certainly needed, but in this paper we focus exclusively on measuring utility of synthetic data.

In what follows, we describe the density ratio estimation framework by summarizing some of the work in this area, and show how it provides a useful framework for measuring utility of synthetic data. Subsequently, we illustrate how the method can be used in practice by providing multiple examples, and compare its performance to existing utility measures. Lastly, we will discuss how the method relates to those existing utility measures, discuss current shortcomings of the density ratio estimation framework and relate these shortcomings to avenues for future work.

Density ratio estimation

The framework of density ratio estimation was originally developed in the machine learning community for the comparison of two probability distributions (for an overview, see Sugiyama, Suzuki, and Kanamori 2012a). The framework has been shown to be applicable for prediction (Sugiyama et al. 2010; Sugiyama 2010), outlier detection (Hido et al. 2008), change-point detection in time-series (Liu et al. 2013), importance weighting under domain adaptation (or, in statistical terms, sample selection bias; Kanamori, Hido, and Sugiyama 2009), and, importantly, two-sample homogeneity tests (Sugiyama et al. 2011). The general idea of density ratio estimation is depicted in Figure 1, and boils down to comparing two distributions by estimating the density ratio $r(\boldsymbol{x})$ between the probability distributions of the numerator samples, which

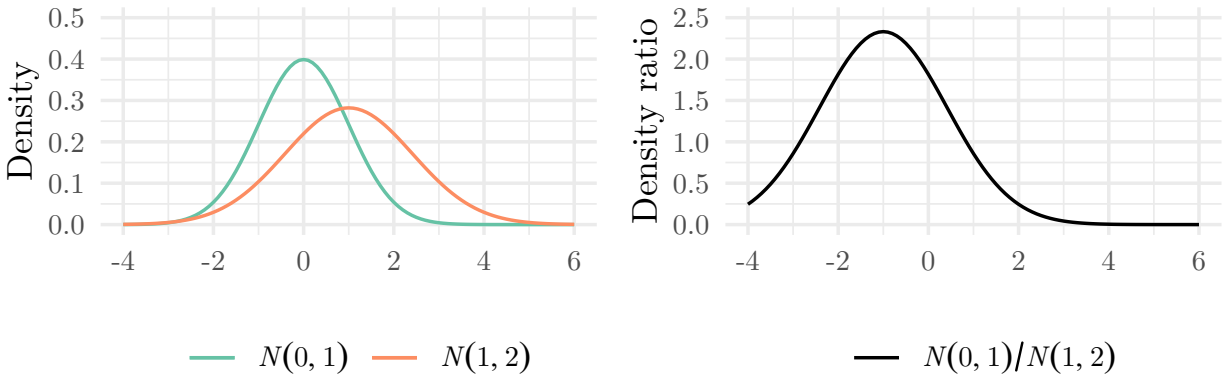

FIGURE 1. Example of the density ratio of two normal distributions with different means and variances (i.e., $N(0, 1)$ and $N(1, 2)$). Note that the density ratio is itself not a proper density.

we take to be the synthetic data samples, $p_{syn}(\boldsymbol{x})$, and the denominator samples, which we take to be the observed data samples, $p_{obs}(\boldsymbol{x})$, such that

$$r(\boldsymbol{x}) = \frac{p_{syn}(\boldsymbol{x})}{p_{obs}(\boldsymbol{x})}. \tag{1}$$

An advantage of this specification over its reciprocal is that, as will be shown later, we can typically weigh over the observed data, which is held constant, when calculating the density ratio relative to the synthetic data. An intuitive approach to estimating $r(\boldsymbol{x})$ from samples of $p_{obs}(\boldsymbol{x})$ and $p_{syn}(\boldsymbol{x})$ would be to estimate the observed and synthetic data density separately, for example using kernel density estimation (Scott 1992), and subsequently compute the ratio from these estimated densities. However, density estimation is one of the hardest tasks in statistical learning, unavoidably leading to estimation errors for both densities. When subsequently taking the ratio of the estimated densities, the estimation errors might be magnified, resulting in a poorer estimate of the density ratio than necessary as compared to direct estimation. An alternative is to specify and estimate a model directly for the ratio without first estimating the separate densities. Extensive simulations on a wide variety of tasks showed that this approach typically outperforms density ratio estimation through naive kernel density estimation, especially when the dimensionality of the data increases (e.g., Kanamori, Suzuki, and Sugiyama 2012; Hido et al. 2008; Kanamori, Hido, and Sugiyama 2009).

Over the past years, several methods for direct density ratio estimation have been developed. Typically, these methods aim to minimize some discrepancy $\mathscr{D}(r(\boldsymbol{x}), \hat{r}(\boldsymbol{x}))$ between the true density ratio and some estimated density ratio model. One commonly used discrepancy measure is the following squared error

$$\mathcal{S}_0(r(\boldsymbol{x}), \hat{r}(\boldsymbol{x})) = \frac{1}{2} \int (\hat{r}(\boldsymbol{x}) - r(\boldsymbol{x}))^2 p_{obs}(\boldsymbol{x}) d\boldsymbol{x}, \tag{2}$$

which can be considered as the expected discrepancy between the two functions over the observed data. One could also use other discrepancy measures, such as the binary or unnormalized Kullback-Leibler divergence or Basu's power divergence (which are all members of the family of Bregman divergences; for a detailed discussion, see Sugiyama, Suzuki, and Kanamori 2012b). It is convenient to model the density ratio with a linear model, such that

$$\hat{r}(\boldsymbol{x}) = \boldsymbol{\varphi}(\boldsymbol{x})\hat{\boldsymbol{\theta}}, \tag{3}$$

where $\boldsymbol{\varphi}(\boldsymbol{x})$ is a non-negative basis function vector that transforms the data from an $n \times p$ to an $n \times b$ matrix, and $\hat{\boldsymbol{\theta}}$ is the parameter vector that is estimated to give the estimated density ratio. Although the model is linear in the parameters, the density ratio itself is a non-linear function of the data if $\boldsymbol{\varphi}(\boldsymbol{x})$ is a non-linear transformation of the data, which it typically is.

To illustrate the the idea of density ratio estimation, we briefly review one method from the field: unconstrained least squares importance fitting (Kanamori, Hido, and Sugiyama 2009), which will also be used in our illustrations in the upcoming

section. The authors show that the squared error can be rewritten as

$$\begin{aligned}\mathcal{S}_0(r(\boldsymbol{x}), \hat{r}(\boldsymbol{x})) &= \frac{1}{2}\int \hat{r}(\boldsymbol{x})^2 p_{obs}(\boldsymbol{x})d\boldsymbol{x} - \int \hat{r}(\boldsymbol{x}) r(\boldsymbol{x}) p_{obs}(\boldsymbol{x})d\boldsymbol{x} + \frac{1}{2}\int r(\boldsymbol{x})^2 p_{obs}(\boldsymbol{x})d\boldsymbol{x} \\ &= \frac{1}{2}\int \hat{r}(\boldsymbol{x})^2 p_{obs}(\boldsymbol{x})d\boldsymbol{x} - \hat{r}(\boldsymbol{x}) p_{syn}(\boldsymbol{x})d\boldsymbol{x} + C,\end{aligned} \tag{4}$$

where $r(\boldsymbol{x})$ in the second term on the first line is written in terms of the ratio of $p_{syn}(\boldsymbol{x})$ over $p_{obs}(\boldsymbol{x})$. After dropping the irrelevant (with respect to the data) constant C, and substituting the density ratio model as defined in Equation 3, we have

$$\mathcal{S}(r(\boldsymbol{x}), \hat{r}(\boldsymbol{x})) = \frac{1}{2}\int \hat{\boldsymbol{\theta}}' \boldsymbol{\varphi}(\boldsymbol{x})' \boldsymbol{\varphi}(\boldsymbol{x}) \hat{\boldsymbol{\theta}} p_{obs}(\boldsymbol{x})d\boldsymbol{x} - \int p_{syn}(\boldsymbol{x})' \boldsymbol{\varphi}(\boldsymbol{x}) \hat{\boldsymbol{\theta}} \tag{5}$$

as the objective function to minimize. The integrals in Equation 5 are typically not available, but can be replaced by empirical averages, such that

$$\hat{\mathcal{S}}(r(\boldsymbol{x}), \hat{r}(\boldsymbol{x})) = \frac{1}{2}\boldsymbol{\theta}'\left(\frac{1}{n_{obs}}\boldsymbol{\varphi}(\boldsymbol{x}_{obs})'\boldsymbol{\varphi}(\boldsymbol{x}_{obs})\right)\hat{\boldsymbol{\theta}} - \left(\frac{1}{n_{syn}}\boldsymbol{\varphi}(\boldsymbol{x}_{syn})'\mathbf{1}'_{n_{syn}}\right)\hat{\boldsymbol{\theta}}. \tag{6}$$

It follows directly that the parameter vector $\hat{\boldsymbol{\theta}}$ can be estimated as

$$\hat{\boldsymbol{\theta}} = \left(\frac{1}{n_{obs}}\boldsymbol{\varphi}(\boldsymbol{x}_{obs})'\boldsymbol{\varphi}(\boldsymbol{x}_{obs})\right)^{-1}\left(\frac{1}{n_{syn}}\boldsymbol{\varphi}(\boldsymbol{x}_{syn})'\mathbf{1}_{n_{syn}}\right), \tag{7}$$

which shows the least-squares nature of the problem. Because one would expect the density ratio to be non-negative, a non-negativity constraint for $\hat{\boldsymbol{\theta}}$ can be added to the optimization problem, which would yield a convex quadratic optimization problem that can be solved with dedicated software. However, ignoring the non-negativity constraint has the advantage that Equation 6 has an analytical expression, which is numerically stable and computationally very efficient. The corresponding downside of having negative estimated density ratio values can be remedied by setting negative parameters in $\hat{\boldsymbol{\theta}}$ to 0.

From here, we are left with two remaining tasks. First, one typically wants to add a regularization parameter λ to the objective function to prevent overfitting and ensure positive-definiteness. In the unconstrained realm, a ridge penalty $\lambda/2\hat{\boldsymbol{\theta}}'\hat{\boldsymbol{\theta}}$ is typically added to the optimization problem in Equation 6. Adding this to the solution in Equation 7 yields

$$\hat{\boldsymbol{\theta}} = \left(\frac{1}{n_{obs}}\boldsymbol{\varphi}(\boldsymbol{x}_{obs})'\boldsymbol{\varphi}(\boldsymbol{x}_{obs}) + \lambda \boldsymbol{I}_b\right)^{-1}\left(\frac{1}{n_{syn}}\boldsymbol{\varphi}(\boldsymbol{x}_{syn})'\mathbf{1}_{n_{syn}}\right), \tag{8}$$

where $\boldsymbol{I}_b$ denotes a $b \times b$ identity matrix. The regularization parameter λ can be chosen via cross-validation. Conveniently, the *leave-one-out cross-validation* score can also be computed analytically when using unconstrained least-squares importance fitting. Second, we need to specify the basis functions used in the density ratio model. A common choice is to use a Gaussian kernel, which quantifies the similarity between observations as

$$\boldsymbol{\varphi}(\boldsymbol{x}) = \boldsymbol{K}(\boldsymbol{x}, \boldsymbol{c}) = \exp\left(\frac{\|\boldsymbol{x} - \boldsymbol{c}\|^2}{2\sigma^2}\right), \tag{9}$$

where $\boldsymbol{c}$ denote the Gaussian centers and σ controls the kernel width. The bandwidth parameter σ can also be selected using cross-validation. Typically a subset of the numerator samples are chosen as the Gaussian centers, because the density ratio tends to take large values at locations where the numerator density is dense. To estimate the density ratio accurately, we may take many kernels where the density ratio is expected to be large, whereas having few kernels might suffice in the locations where the density ratio is small. Accordingly, we take a sample of the synthetic data observations as Gaussian centers, with the sample size dependent on the computational resources available (but typically $\min(100, n_{syn}) \leq n_{centers} \leq \min(1000, n_{syn})$.

After estimating the density ratio, one can assess whether the numerator and denominator densities differ significantly via a permutation test. To this end, Sugiyama et al. (2011) propose a two-sample test that quantifies the discrepancy between the numerator (synthetic) and denominator (observed) samples through the density ratio, using the Pearson divergence $\mathcal{P}(p_{syn}(\boldsymbol{x}), p_{obs}(\boldsymbol{x}))$ as a test statistic:

$$\hat{\mathcal{P}}(p_{syn}(\boldsymbol{x}), p_{obs}(\boldsymbol{x})) = \frac{1}{2n_{syn}}\sum_{i=1}^{n_{syn}} \hat{r}(\boldsymbol{x}_{syn}) - \frac{1}{n_{obs}}\sum_{i=1}^{n_{obs}} \hat{r}(\boldsymbol{x}_{obs}) + \frac{1}{2}. \tag{10}$$

Intuitively, this discrepancy captures how different the synthetic data is from the observed data by measuring the distance from the density ratio at the observed data points to the density ratio at the synthetic data points. On itself, the statistic

$\hat{\mathcal{P}}(p_{syn}(\boldsymbol{x}), p_{obs}(\boldsymbol{x}))$ already quantifies the distance between the synthetic data and the true data. However, as will be shown later on, the value of this statistic is typically little informative in an absolute sense, but can be used to determine the relative fit of different synthetic data models. Additionally, the value of the test statistic can be compared to a reference distribution under the null hypothesis of identical distributions. In this way, it can be assessed whether the synthetic data model is misspecified, by comparing the observed value to what can be expected under a correctly specified synthesis model. This reference distribution is created by permuting the observed and synthetic data sample, fit the density ratio model (including the same optimization strategy of the hyperparameters), and calculating the test statistic given the permuted data sets. If the test statistic is large relative to the reference distribution, there are discrepancies between the observed and synthetic data distributions. Additionally, an empirical p-value can be calculated as the proportion of test statistics under the null model that are greater than the observed test statistic.

Density ratio estimation as a utility measure: Simulated and empirical examples

In the upcoming section, we illustrate density ratio estimation using unconstrained least-squares importance fitting. In a small simulation, we showcase that the method gives reasonable results when the goal is to estimate a density ratio in several parametric examples. Subsequently, we build on these examples to show how the results of density ratio estimation can be used as a measure of utility, and describe how a lack of fit of the synthesis model can be inferred from the density ratio. Starting with univariate examples, we compare the density ratio two-sample test with existing goodness-of-fit measures (the Kolmogorov-Smirnov test and the $pMSE$). As a final illustration, we build upon the work by Drechsler (2022), and showcase how density ratio estimation improves upon utility assessment through the $pMSE$ in a multivariate example.

First, we consider a simplified representation of a situation that is typical in the synthetic data field. Often when creating synthetic data, we have a complex, usually unknown, data distribution, that we want to approximate with a model. That is, we typically do not have enough information to correctly model real-world phenomena, and even if we would have the information theoretically, some important factors might be missing from the data, or the model might be so complex that it is unfeasible to actually simulate data from it. For the sake of illustrational clarity, we generate univariate data according to four *true* data-generating mechanisms, and approximate the true data generating mechanism with a normal model. Specifically, we generate 200 samples of size $n_{obs} = 250$ from the following *true* data-generating models (with corresponding parameters in parentheses): (1) a Laplace distribution (Laplace($\mu = 1, b = 1$)), a (2) lognormal distribution (Lognormal($\mu_{\log} = \log\{\mu^2/\sqrt{\mu^2 + \sigma^2}\}, \sigma^2_{\log} = \log\{1 + \sigma^2/\mu^2\}$), with $\mu = 1$ and $\sigma^2 = 2$, (3) a location-scale t-distribution ($lst(\mu = 1, \tau^2 = 1, \nu = 4)$, which boils down to a t_ν-distribution with $\nu = 4$ degrees of freedom centered around 1), and a normal distribution (Normal($\mu = 1, \sigma^2 = 2$)). Note that these four distributions all have the same population mean $\mu = 1$ and the same population variance $\sigma^2 = 2$. For all scenarios, we generate the synthetic data samples (also 200 samples of size $n_{syn} = 250$) from a normal distribution (Normal($\mu = 1, \sigma^2 = 2$)), such that we accurately model the mean and variance of each *true* data-generating distribution (see also Figure 2 for a graphical depiction of the *true* and synthetic data densities). We chose this approach because it is closely related to modelling some unknown distribution with a normal linear model. Note that in the fourth scenario, we thus model the *true* data-generating distribution correctly, which is included to get some intuition on how density ratio estimation performs when we specify the synthesis model correctly. All density ratios were estimated with the exact same model specifications: we used 100 observations from the synthetic data as Gaussian centers and performed cross-validation over 10 values of the Gaussian kernel width σ and 10 values of the regularization parameter λ.

Figure 3 shows the true density ratios for each of the scenarios (the black line in each plot) and the corresponding estimated density ratios in each of the 200 simulations. In each of the four figures, the estimated density ratios follow the general trend of the true density ratios. In the top-left plot, showing the ratio of the normal distribution over the Laplace distribution, the density ratio decreases at the sides, then increases when moving towards the center, but decreases again close to the center. The same can be observed in the bottom-left plot, which shows the normal distribution over the lst-distribution. In the top right panel, the estimated density ratios are typically large for negative values, very close to zero (or even negative) around the peak of the Laplace distribution, and subsequently increasing and later on decreasing again. In the bottom right panel, where both distributions are identical, the majority of the estimated density ratios are very flat, tending towards zero to some extent at the edges of the figure where only few data points are located. Moreover, all figures show some highly variable estimated density ratios due to modest overfitting regardless of the cross-validation scheme, whereas the normal versus Laplace figure shows many highly variable estimates outside of the center of the figure, due to the fact that either the

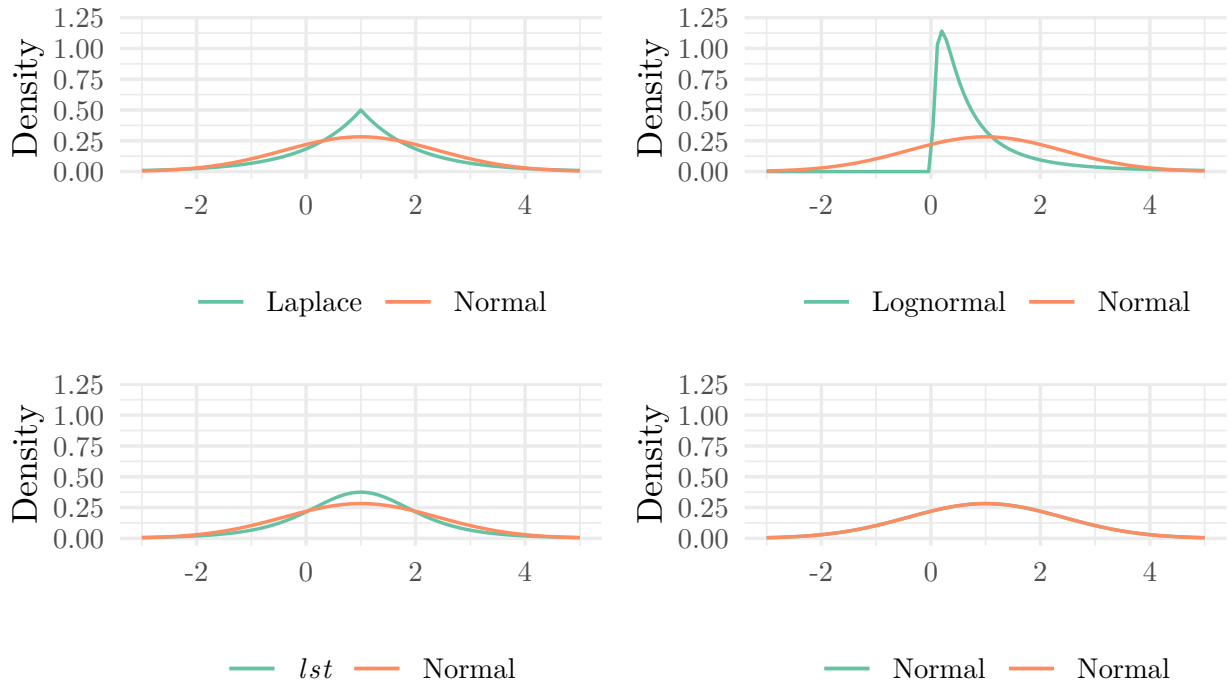

FIGURE 2. True and synthetic data densities for the examples considered (Laplace, Lognormal, t and Normal), all distributions have mean $\mu = 1$ and variance $\sigma^2 = 2$. Note that the true and synthetic data density in the bottom right plot are completely overlapping.

synthetic or the observed data has only few cases in these regions. Typically, the stability of the estimates increases with the sample size. Figure 3 clearly also shows one of the main advantages of density ratio estimation as a utility measure, in the sense that it provides a quantification of the fit for every data point. At those locations where the estimated density ratio takes large values, there too many synthetic observations compared to what should be expected based on the observed data, whereas at the points where the density ratio is close to zero, there are only few synthetic observations relative to the observed data.

As previously explained, the estimated Pearson divergence can be used as a measure of the lack of fit of the synthetic data. However, as the measure is not straightforward to interpret on an absolute scale, and we have no comparison of synthesis models here, we focus here on using the estimated Pearson divergence to perform a significance test to evaluate the fit of the synthetic data. To get an idea of the properties of the density ratio-based test, we compare it to the p-values obtained with a Kolmogorov-Smirnov test and with a $pMSE$-ratio-based test, obtained by performing a permutation test and assessing the proportion of times the permuted $pMSE$-ratios are larger than the observed $pMSE$-ratio (Snoke et al. 2018; see Table 1). The $pMSE$-ratios are calculated by using the `utility.tab()` function in the `R`-package `synthpop` (Nowok, Raab, and Dibben 2016).

TABLE 1. Proportion of significant tests for the fit of the synthetic data.

Data	Density ratio	Kolmogorov-Smirnov	S-$pMSE$
Laplace	0.620	0.375	0.615
Lognormal	1.000	1.000	1.000
lst	0.495	0.235	0.480
Normal	0.050	0.045	0.040

Discussion

Discuss relation with existing utility measures as pmse (i.e., similar through probabilistic classification) and eventually kl-divergence.

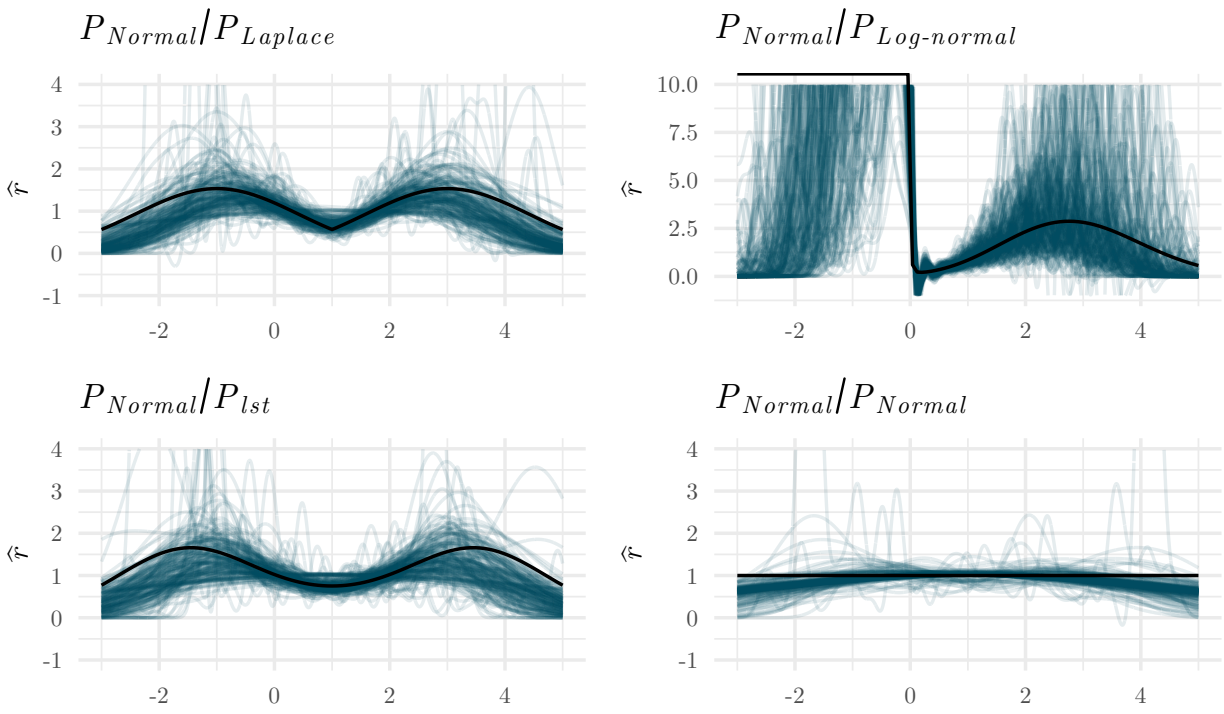

FIGURE 3. Estimated density ratios by unconstrained least-squares importance fitting in four univariate examples: A Laplace distribution, a log-normal distribution, a t-distribution and a normal distribution, all approximated by a normal distribution with the same mean and variance as the original distributions.

———. 2022. "Challenges in Measuring Utility for Fully Synthetic Data." In *Privacy in Statistical Databases*, edited by Josep Domingo-Ferrer and Maryline Laurent, 220–33. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-031-13945-1_16.

Drechsler, Jörg, and Anna-Carolina Haensch. 2023. "30 Years of Synthetic Data." https://doi.org/10.48550/ARXIV.2304.02107.

Dwork, Cynthia. 2006. "Differential Privacy." In *Automata, Languages and Programming*, edited by Michele Bugliesi, Bart Preneel, Vladimiro Sassone, and Ingo Wegener, 1–12. Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/11787006_1.

Hawala, Sam. 2008. *Producing Partially Synthetic Data to Avoid Disclosure*. http://www.asasrms.org/Proceedings/y2008/Files/301018.pdf.

Hido, Shohei, Yuta Tsuboi, Hisashi Kashima, Masashi Sugiyama, and Takafumi Kanamori. 2008. "Inlier-Based Outlier Detection via Direct Density Ratio Estimation." In *2008 Eighth IEEE International Conference on Data Mining*, edited by Fosca Giannotti, Dimitrios Gunopulos, Franco Turini, Carlo Zaniolo, Naren Ramakrishnan, and Xindong Wu, 223–32. https://doi.org/10.1109/ICDM.2008.49.

Hu, Jingchen. 2019. "Bayesian Estimation of Attribute and Identification Disclosure Risks in Synthetic Data." *Transactions on Data Privacy* 12: 61–89. http://www.tdp.cat/issues16/tdp.a313a18.pdf.

Jordon, James, Jinsung Yoon, and Mihaela van der Schaar. 2019. "PATE-GAN: Generating Synthetic Data with Differential Privacy Guarantees." In *International Conference on Learning Representations*. https://openreview.net/forum?id=S1zk9iRqF7.

Kanamori, Takafumi, Shohei Hido, and Masashi Sugiyama. 2009. "A Least-Squares Approach to Direct Importance Estimation." *Journal of Machine Learning Research* 10 (48): 1391–1445. http://jmlr.org/papers/v10/kanamori09a.html.

Kanamori, Takafumi, Taiji Suzuki, and Masashi Sugiyama. 2012. "Statistical Analysis of Kernel-Based Least-Squares Density-Ratio Estimation." *Machine Learning* 86 (3): 335–67. https://doi.org/10.1007/s10994-011-5266-3.

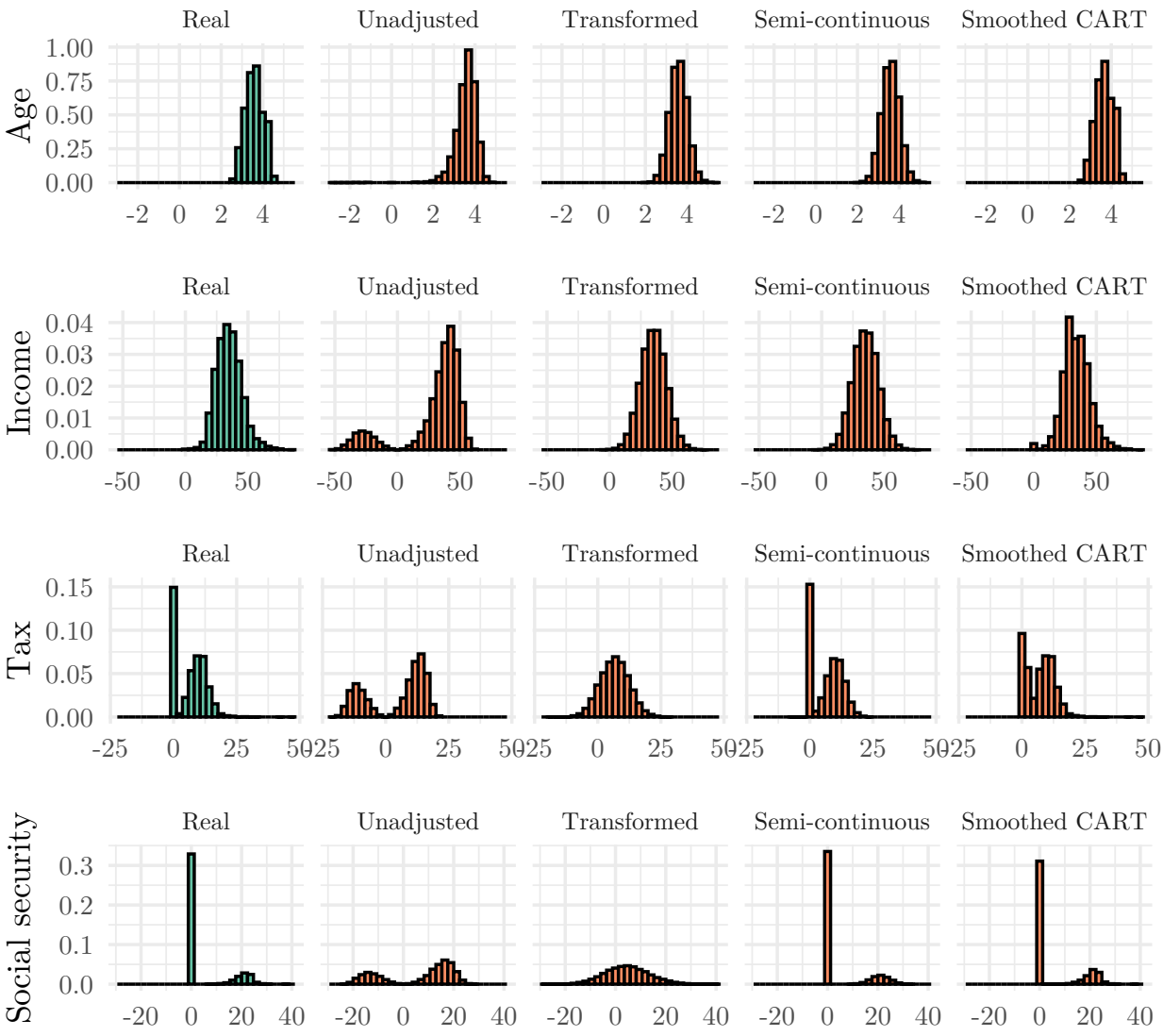

FIGURE 4. Real and synthetic data distributions for the variables age, household income (income), household property taxes (tax) and social security payments (social security).

Karr, Alan F., Christine N. Kohnen, Anna Oganian, Jerome P. Reiter, and Ashish P. Sanil. 2006. "A Framework for Evaluating the Utility of Data Altered to Protect Confidentiality." *The American Statistician* 60 (3): 224–32. https://doi.org/10.1198/000313006X124640.

Little, Roderick J. A. 1993. "Statistical Analysis of Masked Data." *Journal of Official Statistics* 9 (2): 407–7. https://www.scb.se/contentassets/ca21efb41fee47d293bbee5bf7be7fb3/statistical-analysis-of-masked-data.pdf.

Liu, Song, Makoto Yamada, Nigel Collier, and Masashi Sugiyama. 2013. "Change-Point Detection in Time-Series Data by Relative Density-Ratio Estimation." *Neural Networks* 43: 72–83. https://doi.org/10.1016/j.neunet.2013.01.012.

McClure, David, and Jerome P Reiter. 2016. "Assessing Disclosure Risks for Synthetic Data with Arbitrary Intruder Knowledge." *Statistical Journal of the IAOS* 32 (1): 109–26. https://doi.org/10.3233/SJI-160957.

Nikolenko, Sergey I. 2021. *Synthetic Data for Deep Learning*. Springer International Publishing. https://doi.org/10.1007/978-3-030-75178-4.

Nowok, Beata, Gillian M. Raab, and Chris Dibben. 2016. "**Synthpop**: Bespoke Creation of Synthetic Data in *R*." *Journal of Statistical Software* 74 (11). https://doi.org/10.18637/jss.v074.i11.

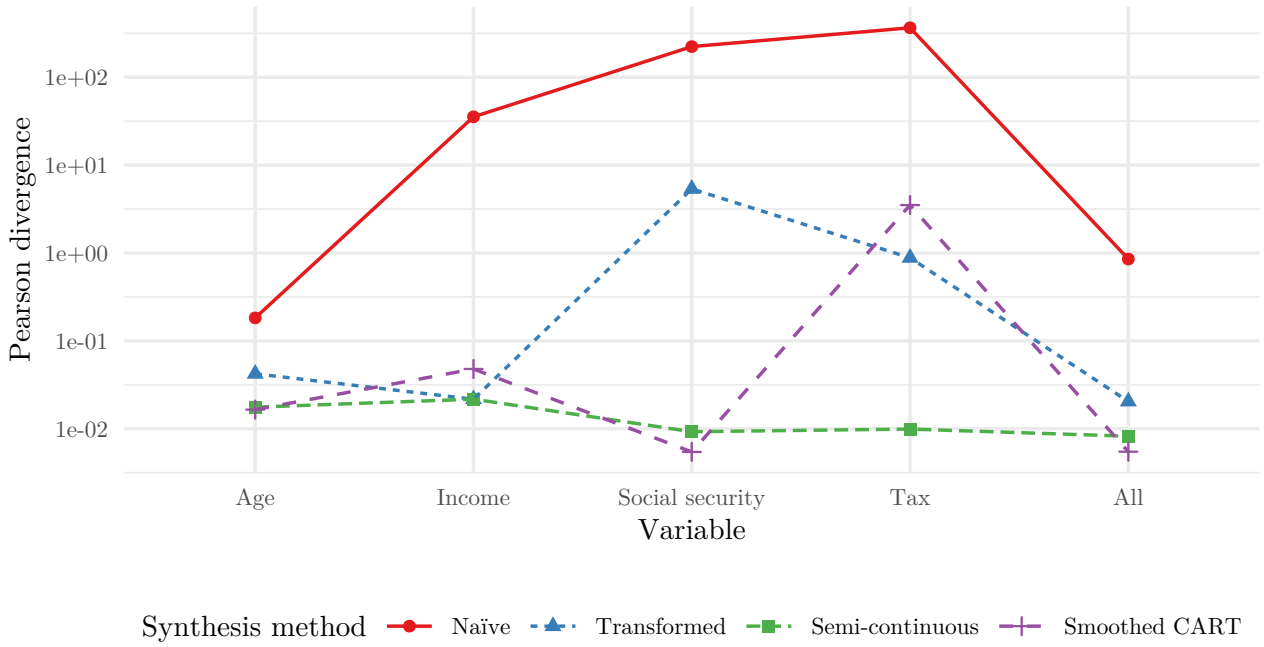

FIGURE 5. Pearson divergence estimates after different synthesis strategies for the separate variables and the synthetic data sets as a whole.

Obermeyer, Ziad, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. 2019. "Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations." *Science* 366 (6464): 447–53. https://doi.org/10.1126/science.aax2342.

Patki, Neha, Roy Wedge, and Kalyan Veeramachaneni. 2016. "The Synthetic Data Vault." *2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, October. https://doi.org/10.1109/dsaa.2016.49.

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.

Rosenbaum, Paul R., and Donald B. Rubin. 1983. "The Central Role of the Propensity Score in Observational Studies for Causal Effects." *Biometrika* 70 (1): 41–55. https://doi.org/10.1093/biomet/70.1.41.

Rubin, Donald B. 1993. "Statistical Disclosure Limitation." *Journal of Official Statistics* 9 (2): 461–68. https://www.scb.se/contentassets/ca21efb41fee47d293bbee5bf7be7fb3/discussion-statistical-disclosure-limitation2.pdf.

Scott, David W. 1992. *Multivariate Density Estimation: Theory, Practice, and Visualization*. Wiley. https://doi.org/10.1002/9780470316849.

Snoke, Joshua, Gillian M. Raab, Beata Nowok, Chris Dibben, and Aleksandra Slavkovic. 2018. "General and Specific Utility Measures for Synthetic Data." *Journal of the Royal Statistical Society. Series A (Statistics in Society)* 181 (3): pp. 663–688. https://doi.org/10.1111/rssa.12358.

Sugiyama, Masashi. 2010. "Superfast-Trainable Multi-Class Probabilistic Classifier by Least-Squares Posterior Fitting." *IEICE Transactions on Information and Systems* E93-D (10). https://doi.org/10.1587/transinf.E93.D.2690.

Sugiyama, Masashi, Taiji Suzuki, Yuta Itoh, Takafumi Kanamori, and Manabu Kimura. 2011. "Least-Squares Two-Sample Test." *Neural Networks* 24 (7): 735–51. https://doi.org/10.1016/j.neunet.2011.04.003.

Sugiyama, Masashi, Taiji Suzuki, and Takafumi Kanamori. 2012a. *Density Ratio Estimation in Machine Learning*. Cambridge University Press. https://doi.org/10.1017/CBO9781139035613.

———. 2012b. "Density-Ratio Matching Under the Bregman Divergence: A Unified Framework of Density-Ratio Estimation." *Annals of the Institute of Statistical Mathematics* 64 (5): 1009–44. https://doi.org/10.1007/s10463-011-0343-8.

Sugiyama, Masashi, Ichiro Takeuchi, Taiji Suzuki, Takafumi Kanamori, Hirotaka Hachiya, and Daisuke Okanohara. 2010. "Conditional Density Estimation via Least-Squares Density Ratio Estimation." In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, edited by Yee Whye Teh and Mike

Titterington, 9:781–88. Proceedings of Machine Learning Research. Chia Laguna Resort, Sardinia, Italy: PMLR. https://proceedings.mlr.press/v9/sugiyama10a.html.

Torkzadehmahani, Reihaneh, Peter Kairouz, and Benedict Paten. 2019. "DP-CGAN: Differentially Private Synthetic Data and Label Generation." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. https://doi.org/10.1109/cvprw.2019.00018.

Volker, Thom Benjamin, and Gerko Vink. 2021. "Anonymiced Shareable Data: Using Mice to Create and Analyze Multiply Imputed Synthetic Datasets." *Psych* 3 (4): 703–16. https://doi.org/10.3390/psych3040045.

Wiel, Mark A. van de, Gwenaël G. R. Leday, Jeroen Hoogland, Martijn W. Heymans, Erik W. van Zwet, and Ailko H. Zwinderman. 2023. "Think Before You Shrink: Alternatives to Default Shrinkage Methods Can Improve Prediction Accuracy, Calibration and Coverage." https://doi.org/10.48550/ARXIV.2301.09890.

Woo, Mi-Ja, Jerome P. Reiter, Anna Oganian, and Alan F. Karr. 2009. "Global Measures of Data Utility for Microdata Masked for Disclosure Limitation." *Journal of Privacy and Confidentiality* 1 (1). https://doi.org/10.29012/jpc.v1i1.568.

Xu, Lei, Maria Skoularidou, Alfredo Cuesta-Infante, and Kalyan Veeramachaneni. 2019. "Modeling Tabular Data Using Conditional GAN." In *Advances in Neural Information Processing Systems*, edited by H. Wallach, H. Larochelle, A. Beygelzimer, F. dAlché-Buc, E. Fox, and R. Garnett. Vol. 32. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2019/file/254ed7d2de3b23ab10936522dd547b78-Paper.pdf.

Zettler, Ingo, Christoph Schild, Lau Lilleholt, Lara Kroencke, Till Utesch, Morten Moshagen, Robert Böhm, Mitja D. Back, and Katharina Geukes. 2021. "The Role of Personality in COVID-19-Related Perceptions, Evaluations, and Behaviors: Findings Across Five Samples, Nine Traits, and 17 Criteria." *Social Psychological and Personality Science* 13 (1): 299–310. https://doi.org/10.1177/19485506211001680.