Statistical Science
2024, Vol. 39, No. 2, 221–242
https://doi.org/10.1214/24-STS927

30 Years of Synthetic Data

Jörg Drechsler and Anna-Carolina Haensch

Abstract. The idea to generate synthetic data as a tool for broadening access to sensitive microdata has been proposed for the first time three decades ago. While first applications of the idea emerged around the turn of the century, the approach really gained momentum over the last ten years, stimulated at least in parts by some recent developments in computer science. We consider the 30th jubilee of Rubin's seminal paper on synthetic data (*J. Off. Stat.* **9** (1993) 462–468) as an opportunity to look back at the historical developments but also to offer a review of the diverse approaches and methodological underpinnings proposed over the years. We will also discuss the various strategies that have been suggested to measure the utility and remaining risk of disclosure of the generated data.

Key words and phrases: Access, confidentiality, data generation, disclosure, dissemination, privacy.

1. INTRODUCTION

We live in a data-driven world today. Data is collected, whenever we use our loyalty card in the supermarket, measure our physical activities through wearables, when we check the online weather forecast for our weekend trip, or when we stay in contact with our friends using social media. In the public sector, the ever-growing importance of data is reflected in concepts such as evidence-based policy and open data movements (see, e.g., [148] or [193]), and the fact that increasingly more countries explicitly define their own data strategies (see, e.g., [40] or [45] for the UK). In industry, the increased reliance on machine learning methods for decision-making results in ever-growing demands for more data to train these models.

However, the increased availability and storage of data also raises concerns regarding confidentiality and privacy. There is an increasing tension between the societal benefits of our digitized world and broad data access on one hand and the potential harms resulting from the misuse of data that have not been sufficiently protected on the other hand. For example, contact tracing apps played a crucial role in containing the spread of the Coronavirus [70]. Yet, it is well known that mobility information can be highly sensitive [189] and can pose a high risk of reidentification [41]. Thus, the providers of the tracing apps took great efforts to ensure that the privacy of the app users was maintained to increase trust in these tools (see, e.g., Apple and Google [214]).

Data providers have been concerned about these risks for decades, and various strategies have been developed over the years to avoid disclosing sensitive information when disseminating data to the public [63, 95]. Still, there is an inherent trade-off between data protection and data utility: increasing the level of protection will inevitably lead to lower utility, as some information will be lost. Besides, several prominent examples of confidentiality breaches both in the public and in the private sector [42, 86, 141, 171, 186] have demonstrated that risks of disclosure often still tend to be underestimated. Increasing computer power and the fact that more and more data are publicly available or are sold by private companies also imply that traditional protection strategies such as swapping, top-coding, or suppression are no longer sufficient to adequately protect the data.

A promising alternative to address the trade-off between broad data access and disclosure protection is the release of synthetic data. With this approach, a model is fitted to the original data[1] and draws from this model are

Jörg Drechsler is head of the Department for Statistical Methods at the Institute for Employment Research, Germany, Professor, Ludwig-Maximilians-Universität, Munich, Germany, and Associate Research Professor, Joint Program in Survey Methodology, University of Maryland, College Park, Maryland 20742, USA (e-mail: joerg.drechsler@iab.de). Anna-Carolina Haensch is Lecturer, Ludwig-Maximilians-Universität, Munich, Germany, and Assistant Research Professor at the Joint Program in Survey Methodology, University of Maryland, College Park, Maryland 20742, USA (e-mail: anna-carolina.haensch@stat.uni-muenchen.de).

[1] We note that there is some ambiguity regarding the meaning of *original data* and *confidential data* in the literature. We use both terms

used to replace the original values. Depending on the desired level of protection, only some records (partial synthesis) or the entire dataset (full synthesis) are replaced by synthetic values.

The idea of using synthetic data as a disclosure avoidance strategy is commonly attributed to [175] and [118] (although related ideas have been proposed earlier by [115]). Their approach to synthetic data was motivated by their own work on multiple imputation (MI) for nonresponse [119, 173]. Instead of imputing missing values, they suggested adopting the approach to replace sensitive values with imputed values. Similar to the nonresponse context, the release of multiple synthetic datasets would then allow obtaining valid variance estimates that also account for the uncertainty from the synthesis models (assuming the models are correctly specified). However, it took another ten years before the methodology was fully developed, and the first practical applications started to emerge.

Independent of the developments in the statistical community, the computer science community also proposed relying on synthetic data as a way of mitigating the risks of disclosure. The large body of work developed in this field has rarely aimed at ensuring valid statistical inference, including properly quantifying the uncertainty in the estimates obtained from the synthetic data. The research on synthetic data in computer science was (and still is) predominantly motivated by providing easier data access to train machine learning models. For example, a team of researchers at the University of Michigan successfully used synthetic pathology images to improve the accuracy of their machine learning tool for cancer prediction (https://tinyurl.com/3htnhe4z). Still, both approaches to synthetic data share the same core goal: ideally, any analysis run on the synthetic data should provide approximately the same answers that would have been obtained if the analysis were run on the original data.

While the body of research has grown steadily over the last thirty years and the first deployments of the idea date back to the turn of the century, the concept really gained momentum in the last five to ten years. Many statistical agencies, but also other government agencies such as tax authorities, ministries, or central banks, are exploring synthetic data approaches as a promising tool to broaden public access to their data. Especially within the health sector, the approach gained popularity with applications ranging from generating synthetic patient data [39] over synthetic electronic health records [206] to generating synthetic cell images [180]. More and more start-ups are offering synthetic data generation as a service, and in the industry, synthetic data are used in such diverse contexts as autonomous driving [142], classifying computed tomography images [74] or environmental monitoring [9].

Given this growing interest in the field, we consider the 30th jubilee of synthetic data as an opportunity to look back at the historical developments but also to offer a review of the diverse approaches and methodological underpinnings proposed over the years. We need to emphasize at this point that the diversity of the field and the exponential growth in literature in recent years makes it impossible to offer a detailed review of all methodological tweaks and use cases. We will therefore limit our review to synthetic data methods and applications that specifically aim at offering confidentiality protection. Other contexts in which ideas based on synthetic data have been exploited include, for example, microsimulation [140], which generates synthetic populations from various data sources, or applications in machine learning, where synthetic data are generated to increase the data pool for model training. Furthermore, we will only discuss and review strategies for the synthesis of regular datasets, that is, data structures in which the units are organized in rows and the columns contain the information collected about these units. We will not cover synthesis strategies for text data or images.

The remainder of the paper is organized as follows: In Section 2, we will provide a brief history of synthetic data. Although the bounds are sometimes blurry, we treat the developments in the statistical community separately from the developments in computer science. The inferential procedures for obtaining valid inferences for the multiple imputation inspired synthesis approaches are covered in Section 3. In Section 4, we provide a taxonomy for synthetic data and also discuss some extensions that have been proposed in the literature. Sections 5 and 6 discuss various approaches to measure the utility and remaining risks of disclosure. The paper concludes with a discussion of verification servers which might help enhance the usefulness of synthetic data in the future.

2. A BRIEF HISTORY OF SYNTHETIC DATA

2.1 The Statistical Approach

2.1.1 *Methodological developments.* The idea of releasing synthetic data instead of the real data was first proposed by Rubin [175]. In a discussion of another article, he suggested that his multiple imputation framework [173, 174] could be used as an innovative disclosure protection strategy. He proposed treating the units that were not sampled for the survey as missing data and to multiply impute this "missing" information. Simple random samples from these imputed populations should then be disseminated to the public. (We note that in practice the two-step procedure of imputing the full population first and then sampling from it is not necessary. It suffices to

interchangeably referring to the (potentially pre-processed) data that would be analyzed if there were no confidentiality or privacy concerns.

draw a new sample from the sampling frame and to generate synthetic values for the survey variables of the sampled units.) If the risk of releasing original records should be avoided completely, the records in the original sample can also be replaced by draws from the imputation model.

Similar to multiple imputation for nonresponse, valid inferences can be obtained from the synthetic datasets by analyzing each dataset separately and combining the estimates from each dataset using simple formulas to come up with the final estimates (see Section 3 for details).

An obvious advantage of the approach is that no original values are included in the released data (for this reason, this approach has been termed the *fully synthetic data* approach in the literature to distinguish it from the *partially synthetic data* approach described below). Furthermore, synthetic values are generated for units that never participated in the survey. Thus, the level of protection is very high. However, this high level of protection comes at a price. The synthetic data are drawn from a model fitted to the original data, and the quality of the synthetic data directly depends on the quality of that model. Finding a model that reflects all relationships in a complex dataset with hundreds of variables and complicated logical constraints between the variables can be challenging.

A closely related approach that overcomes the limitations of the fully synthetic approach was suggested by [118]. With this approach, only the sensitive records and/or records that could be used for re-identification are replaced with synthetic values. Since some true values remain in the dataset, the approach has been termed the *partially synthetic data* approach. The approach offers some flexibility over the fully synthetic data approach. The agency can decide which part of the data needs to be synthesized. The synthesis can range from synthesizing only some records for a single variable, for example, all income values for individuals with an income above a given threshold, to synthesizing all variables, basically mimicking the fully synthetic data approach (this connection between fully and partially synthetic data is further discussed in Section 3.3).

Ten years after the initial proposal by Rubin and Little, Raghunathan, Reiter and Rubin [156] and Reiter [160] developed the full methodology to enable valid inferences based on fully and partially synthetic data, respectively. Similar to multiple imputation for nonresponse, the multiple synthetic datasets are analyzed separately first, and the results from the different analyses are combined using simple combining rules to obtain estimates for the first two moments for the statistic of interest. However, these combining rules differ slightly from the rules in the nonresponse context, and they also differ between full and partial synthesis. In 2012, Reiter and Kinney [166] identified another difference between the two synthesis approaches: posterior draws of the model parameters which are necessary for full synthesis (and also in the context of multiple imputation for nonresponse) are not required for partial synthesis. Several years later, [153] developed combining rules for a variant of the fully synthetic approach that can also be used if only one synthetic copy of the original data is available (see Section 3 for further details).

While early illustrations [10, 161] and applications [4, 102] mostly relied on classical parametric modeling approaches for generating the synthetic data, the suite of modeling strategies has been extended over the years, incorporating ideas from the machine learning literature but also adopting strategies to properly account for the complex sampling designs found in most sample surveys. These will be reviewed in more detail in Section 4.

2.1.2 *Practical implementations.* The earliest application of the synthetic data idea dates back to 1997, when the U.S. Federal Reserve Board decided to replace monetary values at high risk of disclosure in the Survey of Consumer Finances with synthetic values [102]. [6, 7] demonstrated the usefulness of the approach for longitudinal, linked datasets using data from the French National Institute of Statistics and Economic Studies (INSEE). The most complex synthetic data product generated so far was first released by the U.S. Census Bureau in 2007: the SIPP synthetic beta [4]. It contains synthetic records of the Survey of Income Program Participation (SIPP) linked to administrative records from the Social Security Administration and the Internal Revenue Service. Almost all of the more than 600 variables in this longitudinal dataset are synthesized. Since its first release, the dataset has been updated regularly [19]. Another early application was OntheMap, a graphical interface that allows visualizing detailed commuting patterns for the entire United States [122]. This application was the first to offer formal privacy guarantees based on a concept called (ε, δ)-probabilistic differential privacy, a relaxation of the original definition of differential privacy proposed by [65]. Three years later, the U.S. Census Bureau released the Synthetic Longitudinal Business Database [108, 109], a partially synthetic copy of the Longitudinal Business Database, which is generated from administrative data at the U.S. Census Bureau and covers all businesses in the United States. The U.S. Census Bureau also uses synthetic data to protect sensitive information in the American Community Survey (ACS) [85]. Another large scale synthetic data project was conducted by the Maryland Longitudinal Data System Center (MLDSC), which houses longitudinal education data for the state of Maryland, combining data from various sources. The MLDSC launched the Synthetic Data Project in 2016, sponsored by the Institute of Education Sciences with the goal of facilitating access to this rich source of information [21, 79].

Outside the United States, the approach has first been used by Statistics New Zealand to release so-called synthetic unit record files (SURFs) for teaching purposes [73, 101]. A SURF has also been used more recently as input data for a micro-simulation model that estimates the uptake of and fluency in Te Reo Māori, the language of the Māori people, for various scenarios and policy interventions over the period from 2013 to 2040 [137]. The German Institute for Employment Research released a partially synthetic version of one wave of its Establishment Panel in 2011 [52]. The approach was also adopted to facilitate access to the Scottish Longitudinal Study [139]. This study links census data with other sensitive information from health records and death registers. Due to the high sensitivity, access to the data is highly restricted. To prepare their analyses, external researchers can request synthetic datasets that are tailor made to the research questions the users are trying to answer; that is, the synthetic datasets will always only contain those variables that are needed for the planned research. The R package *synthpop* [138], which is now a popular tool for generating synthetic datasets, was also developed as part of this project.

In 2015, a project under the leadership of Statistics Netherlands developed synthetic public use files for the EU Statistics on Income and Living Conditions (EU-SILC) [43]. These data, which are available for download at the Eurostat website [71], are not meant to provide valid statistical inferences. They can be used for training purposes or for developing analysis code while waiting for accreditation to get access to the restricted scientific use files. More recently, Statistics Canada generated synthetic data, which was used in a Hackathon hosted by Statistics Canada in 2020 [178].

Synthetic data are currently at the development stage at several agencies: Examples include the Urban Institute, which is developing synthetic tax data for the Internal Revenue Service [23, 22] and the Australian Bureau of Statistics that is currently evaluating synthetic data as a means of broadening access to its microdata [13].

Further practical applications have been discussed in the context of protecting data containing detailed geographical information [55, 143, 151, 152, 198], preserving and protecting longitudinal data structures [133, 157], small area estimation [176, 177], synthesizing business data [8, 38, 61, 62, 112, 188] dealing with nested data structures [92] or accounting for complex survey designs [47, 48, 94, 104, 134, 212]. Hu, Savitsky and Williams [93] proposed a strategy to reduce the risk of disclosure for partially synthetic data by down-weighting the contribution of high-risk records to the Likelihood function of the synthesizer, while [199] developed a synthesis strategy that preserves additive constraints.

2.2 The Computer Science Approach

The synthetic data approach did not get much attention in the literature on data privacy in computer science before the turn of the century, although [115] proposed a synthetic data approach for disclosure protection several years before Rubin's seminal paper. They outlined a three-step process: (1) independently estimate the univariate density for each variable that should be protected, (2) generate new data by randomly drawing new values from these densities, and (3) map each data point in the generated data to its corresponding point in the original data (i.e., sort the generated data and the original data in the same order and replace each element of the original data with the corresponding generated element) to preserve relationships between the variables.

We postulate that the lack of interest in data synthesis can be attributed (at least in part) to the fact that privacy standards play an important role in the computer science literature on data privacy and the privacy standards used before the advent of differential privacy ([65]) only focused on the properties of the data at hand. Popular standards such as k-anonymity [185], l-diversity [123], or t-closeness [114] all establish certain requirements regarding the properties of the data to consider the data safe from (certain types of) privacy attacks. k-anonymity is a privacy definition that requires that every unit in the dataset is indistinguishable from at least $k - 1$ other units with respect to certain identifying attributes. This means that when considering a combination of attributes (such as age, education, and marital status), each unique combination should occur at least k times in the dataset. The complementary principle l-diversity ensures that each equivalence class (a set of records that is indistinguishable based on certain attributes) contains at least l distinct values for the sensitive attributes. t-closeness requires that the distribution of a sensitive attribute in any equivalence class is close to the distribution of the variable in the overall dataset.

Synthetic data do not really fit into this notion of privacy. For example, even if a fully synthetic dataset does not fulfill k-anonymity for any $k > 1$, this does not automatically imply a risk. Unlike with the original data, a combination of attribute values that is unique in the synthetic data does not automatically pose a high risk of reidentification simply because the synthetic records cannot be sensibly linked to real units. Besides, a unique attribute combination in the synthetic data might not be unique in the original data or it might not even exist in the original data.

Differential privacy (DP) brought a fundamental change in the way computer scientists think about privacy, which paved the way for synthetic data applications in the computer science literature. DP requires that changing the database by one record has a strictly limited impact on the results of a mechanism run on the data (we will offer a more detailed review of DP in Section 4.5).[2] Note the

[2] We note that two definitions of DP exist in the literature: One focuses on the impact of changing the values of one existing record

change of focus from the data to the mechanism, which implies that it is no longer the data that needs to be adjusted to achieve privacy, but the mechanism. This concept of privacy aligns much better with the ideas of synthetic data. All that is required is to find a synthesis mechanism that satisfies the requirements of DP. Soon after the concept of DP was established in 2006, the first papers on differentially private synthetic data started to appear.

2.2.1 *Methodological developments.* One of the first approaches was developed by [16] who generated synthetic data using a Fourier transformation and linear programming for low-order contingency tables [64]. Other early applications include [20, 32, 67, 203]. Several papers also explicitly adapted the ideas from the statistical community to the DP context [5, 34, 122, 126]. The approach of [122] was later extended in [149] and [150]. [209, 210] proposed an approach that uses Bayesian networks to synthesize high-dimensional datasets, called PrivBayes. In parallel, [113] employed Copula functions to take into account the dependency structure of the data (DPCopula). DP guarantees were also integrated in Generative Adversarial Networks (GANs) [204, 207].

The advent of GANs proposed by [82] resulted in a boost in synthetic data research and applications in the computer science literature. This is probably not surprising, as synthetic data are generated as a by-product with any GAN model. We will review GANs in more detail in Section 4.4.1, but the basic setup of GANs consists of two neural networks, a generator and a discriminator. The generator produces fake data trying to fool the discriminator, which tries to distinguish the fake data from the real data. Both neural networks are improved in an iterative process. The final data produced by the generator can be seen as a variant of synthetic data. GANs turned out to be extremely successful for image and speech recognition and natural language understanding. Early applications of GANs for data synthesis also focused on generating synthetic images (see, e.g., [44]). However, the approach was quickly adapted for synthesizing microdata (microdata are often referred to as tabular data in the computer science literature. Thus, many approaches explicitly refer to tabular data in the title of the paper or the labeling of the algorithm to distinguish the approach from other applications that focus on images and other types of data). However, the adoption of GANs for microdata poses additional challenges. Microdata often have categorical variables that are sparse, and correlations among variables are often weaker than, for example, relationships between pixels that are located next to each other. The position of observations in a dataset is also only rarely informative for microdata, as the individual records are typically independent. Relationships between variable therefore have to be modeled without the help of any kind of spatial information.

Several of the early applications to microdata only focused on specific types of data, such as time series [69, 206] or count and binary data [39], medGAN. tableGAN [145] claims to be the first approach capable of handling continuous and categorical variables simultaneously. The approach is built on a GAN originally used for image data by converting records in the original table into a square matrix form. medGAN was extended to categorical variables and further refined in several works [15, 31]. corGAN uses Convolutional GANs and Convolutional Autoencoders to generate discrete and continuous health records [191]. Other applications relied on Wasserstein-GANs (WGANs) [111, 211]. In recent years, more focus has been put on modeling relationships between variables. Conditional tabular GAN (CTGAN) developed by [205] addresses challenges from imbalanced categorical and multi-modal continuous data. Causal Tabular GAN [200], Causal TGAN allows for modeling the causal relationships between variables in datasets.

Beyond approaches based on GANs, other synthesis strategies based on (Variational) Autoencoders [31, 121, 195], Bayesian Networks [210], copulas [99, 146], approaches based on latent normal variables and Gaussian processes [77], CLGP, or approaches that explicitly preserve certain marginal distributions [129, 130] have also been developed in recent years. For a short taxonomy of approaches, see Section 4.

2.2.2 *Practical implementations.* The earliest deployment of a differentially private synthesis strategy is OnTheMap [122] already mentioned in the previous section. The enforcement of DP for some of the publicly available tables from the US Decennial Census 2020 generated using the 2020 Disclosure Avoidance System [3] can also be seen as a synthetic data approach. At its core, the underlying algorithm ensures DP by cross-classifying all variables in the dataset and adding noise to each cell of the resulting table. The noisy table counts are turned into synthetic microdata from which the released tables are generated. Various strategies are implemented to improve the accuracy of the generated tables. For example, the noisy tables are generated in a hierarchical fashion. The algorithm starts by generating noisy counts at the national level. Next, noisy counts are produced on the state level under the constraint that the sum of noisy counts on this level matches the counts on the state level. This process is continued all the way down to the block level. We note that this strategy was only used for some of the released tables. The more granular tabulations used a different approach [98].

(bounded DP), while the other limits the impact of adding or removing one record (unbounded DP) (see, e.g., Kifer and Machanavajjhala [103] for further details.). We do not distinguish between these two definitions in the remainder of this paper.

The usefulness of various differentially private synthesis approaches in practical applications was also assessed in the three rounds of the Differential Privacy Synthetic Data Challenge organized by the National Institute of Standards and Technology (NIST) over the years 2018 to 2019 [33]. The winning teams relied on Bayesian networks or approaches to preserve pre-specified marginal distributions. See [26] for a review of the results from the competition.

Further applications of computer science approaches have been envisioned, proposed, and conducted by academic institutions and industry alike. GANs, for example, have been used to create financial time series data [201] and synthetic health records [69, 190, 191]. [35] also provides a validation study of a synthetic data generator for patient data with mixed results. Beyond the microdata context that is the focus of this review, GANs have also been used to create realistic images of, for example, skin lesions [78], pathology slides [124], and chest X-rays [197].

3. OBTAINING VALID INFERENCES FOR THE MI INSPIRED APPROACHES

As indicated in the Introduction, Rubin's initial proposal for data synthesis was motivated by his prior work on multiple imputation for nonresponse. Given the close relationships to those ideas, it seems natural to also adopt the simple combining procedures from the multiple imputation literature (Rubin's combining rules) to obtain valid point and variance estimates from the synthetic data. However, the synthetic data approaches differ in two important aspects from the original framework. With full synthesis as proposed by Rubin, synthetic data are only generated for a simple random sample of the population. This extra sampling step needs to be taken into account. With partial synthesis, the synthesis models are estimated using the full data and not only the fully observed subset of the data, as done in the nonresponse context. These deviations imply that the combining procedures also need to be adjusted. The correct rules for fully synthetic data were derived in [156], those for partially synthetic data are presented in [160]. Later, [162] also derived the multivariate analogs that can be used for multi-component testing based on Wald tests or Likelihood ratio tests. We will only review the combining rules for univariate estimates here, borrowing heavily from [50]. The interested reader is referred to [169], which offers a full review of all combining rules for synthetic data and also for the nonresponse context.

To understand the procedure of analyzing multiply imputed synthetic datasets, think of an analyst interested in an unknown scalar parameter Q, where Q could be, for example, the mean of a variable, the correlation coefficient between two variables, or a regression coefficient in a linear regression. For simplicity, assume that there are no data with items missing in the observed dataset. Inferences for Q derived from the original dataset usually are based on a point estimate q, an estimate for the variance of q, u, and a normal or Student's t reference distribution. For analysis of the synthetic datasets, let $q^{(i)}$ and $u^{(i)}$ for $i = 1, \dots, m$ be the point and variance estimates for each of the m synthetic datasets. The following quantities are needed for inferences for scalar Q:

$$\bar{q}_m = \sum_{i=1}^{m} q^{(i)}/m, \tag{1}$$

$$b_m = \sum_{i=1}^{m} (q^{(i)} - \bar{q}_m)^2/(m-1), \tag{2}$$

$$\bar{u}_m = \sum_{i=1}^{m} u^{(i)}/m. \tag{3}$$

3.1 Combining Rules for Fully Synthetic Data

The analyst can use $\bar{q}_m$ as an unbiased point estimate for Q under the assumption that the synthesis models are correctly specified (i.e., they match the true data generating process) and that q would be an unbiased estimate for Q based on the original data. Its variance can be estimated using

$$T_f = (1 + m^{-1})b_m - \bar{u}_m, \tag{4}$$

where b_m is an estimate for the variability of the point estimates between the synthetic datasets and $\bar{u}_m$ is an estimate for the sampling variance. When n is large, inferences for scalar Q can be based on t distributions with degrees of freedom $\nu_f = (m-1)(1 - \bar{u}_m/((1+m^{-1})b_m))^2$. Similar to the nonresponse context, these inferences are valid under the assumption that the point estimate q that would have been used on the original data approximately follows a normal distribution. Furthermore, valid inferences can only be obtained if the analysis model and the synthesis model are based on the same modeling assumptions (congeniality). We will come back to this point in Section 4.2. We note that similar assumptions are also required for the estimation procedures for partially synthetic data outlined below.

A disadvantage of the variance estimate for fully synthetic data is that it can become negative. For that reason, [158] suggested a slightly modified variance estimator that is always positive but will tend to overestimate the true variance, $T_f^* = \max(0, T_f) + \delta(\frac{n_{\text{syn}}}{n}\bar{u}_m)$, where $\delta = 1$ if $T_f < 0$ and $\delta = 0$ otherwise. Here, n_{syn} is the number of observations in the released datasets sampled from the synthetic population. Negative variance estimates can be avoided by increasing the number of synthetic datasets, as this helps to stabilize the estimate of the variance between the synthetic datasets, b_m. Given the large variability of this estimate in the fully synthetic data context, most researchers recommend generating more than the $m = 5$ datasets that are often found in the nonresponse literature. Suggestions for picking m range from $m = 10$ [155] to $20 < m < 40$ [158], while some illustrative applications use $m = 100$ [161].

3.2 Combining Rules for Partially Synthetic Data

Similar to fully synthetic data, the analyst can use $\bar{q}_m$ to estimate Q. The variance of $\bar{q}_m$ for partially synthetic data can be estimated using

$$T_p = b_m/m + \bar{u}_m. \tag{5}$$

When n is large, inferences for scalar Q can be based on t distributions with degrees of freedom $\nu_p = (m - 1)(1 + \bar{u}_m/(b_m/m))^2$. Note that the variance estimate T_p can never be negative, so no adjustments are necessary for partially synthetic datasets. Given that b_m/m is usually dominated by $\bar{u}_m$, choosing the number of imputations for partial synthesis has received limited attention so far, but $m = 5$ seems to be the default choice that is often used in practice.

3.3 An Alternative Variance Estimate for Fully Synthetic Data

When generating fully synthetic data, most researchers do not follow the protocol as initially envisioned by [175]. Rubin assumed that in addition to the survey variables Y some additional variables X would be available for the full population. In the survey context, these variables represent design variables available from the sampling frame. Under this assumption, fully synthetic data for Y would be generated by fitting a model for $f(Y|X)$ using the survey data and using this model to generate synthetic values for a new sample of design variables X^{new} by drawing from $f(Y|X^{\text{new}})$. Only the synthetic Y values would then be released to the public.

In practice, researchers typically only use the information in Y to generate synthetic data. In this setting, fully synthetic data can be seen as an extreme variant of partial synthesis for which the set of unsynthesized records is empty. This also implies that the combining rules for partial synthesis are still valid as first noted by [51]. Extending these ideas, [153] proposed an alternative variance estimator that can be used in this situation:

$$T_s = \left(\frac{n_{\text{syn}}}{n_{\text{org}}} + \frac{1}{m}\right)\bar{u}_m,$$

where n_{syn} is the number of synthetic records and n_{org} is the number of records in the original dataset. Note that this variance estimator does not rely on the between imputation variance b_m. This offers three important advantages compared to T_f, the variance estimator for fully synthetic data discussed above: (i) the estimator T_s can never be negative, (ii) it has less variability than T_f (b_m is only an estimate for the true variability between the datasets and the fact that it is based on a limited number of m synthetic datasets implies high uncertainty in this estimate, which is also the reason why T_f can sometimes be negative), and (iii) valid variance estimates can be obtained from a single synthetic dataset. The last point is especially important because previous research has shown that the risk of disclosure increases with the number of synthetic datasets [57, 165]. Of course, the price to pay is an increased level of uncertainty if only one synthetic dataset is released. Note that assuming $n_{\text{syn}} = n_{\text{org}}$, the variance can be reduced by 25% when releasing two datasets instead of one dataset. These accuracy gains are diminishing quickly with increasing m and the relative reduction in variance is bounded by 0.5 for $m \to \infty$. See [53] for further discussion of the advantages and disadvantages of the different synthesis strategies and which variance estimator is appropriate in which scenario.

An alternative approach for obtaining valid inferences from a single synthetic dataset was proposed by Klein and Sinha [110] under the assumption that the data follow a multivariate normal distribution. The authors also present analysis procedures for the linear regression context under the assumption that only the dependent variable is synthesized.

4. A TAXONOMY OF SYNTHETIC DATA APPROACHES

Given the broad range of synthetic data approaches and use cases, finding a one-dimensional taxonomy that fully covers all variants of synthetic data is difficult. Beyond the obvious distinction between approaches inspired by the ideas of multiple imputation (and their extensions) and approaches that originated in computer science, we suggest three alternative classification schemes: sequential vs. joint modeling approaches, parametric vs. machine learning inspired approaches, and approaches that offer formal privacy guarantees vs. those that do not. Obviously, other classifications, such as partial vs. full synthesis would be possible. However, we feel that classifying the approaches along these lines is obvious and does not require a separate discussion. Instead, we list a final class of synthesis approaches that are extensions of the MI-based approaches. These approaches are treated separately, as they typically require different procedures to obtain valid inferences compared to those discussed in Section 3.

4.1 Sequential vs. Joint Modeling

Most of the early applications of synthetic data relied on a sequential modeling approach, in which each variable is synthesized sequentially using models that condition on any variables that have been synthesized previously or variables that remain unchanged in the final data.

The underlying idea is that any joint distribution can be rewritten as a product of conditional distributions.

The sequential regression approach offers great flexibility, as different models can be used for each variable. These might include parametric models such as linear regression, logit models [161], or models based on quantile regressions [147], but also any machine learning tool that enables random draws from a conditional distribution, such as Classification and Regression Trees (CART, Reiter [163]) or random forests [30].

In contrast to the sequential modeling approach, joint modeling aims at directly specifying the joint distribution of the data. While early approaches such as the IPSO method [28] relied on a multivariate normality assumption that is seldom justified with real data, more flexible models have been proposed recently. For categorical data, [91] demonstrated that an approach based on a Dirichlet Process Mixture of Products of Multinomials (DPMPM) can offer high utility in real data applications. The approach was later extended to also allow for structural zeros, that is, impossible variable combinations such as married toddlers [125]. Synthesis approaches based on DPMPMs are implemented in the R package NPBayesImputeCat [89]. A related approach based on Quasi-Multinomial distributions was proposed by Hu and Hoshino [90], while Jackson et al. [96, 97] proposed saturated count models for easy synthesis of large databases with a-priori utility guarantees. [105] showed good performance of Dirichlet Process Normal Mixture Models for synthesizing continuous business data. This approach was later extended to also account for informative sampling designs that are common with business surveys [104]. Furthermore, many of the synthesis strategies used in computer science, such as Generative Adversarial Networks [82] or Bayesian Networks [210] can be subsumed under this category. We will review the literature based on these approaches in Section 4.4.

4.2 Parametric vs. Machine Learning Based

The methodology for obtaining valid inferences based on synthetic data reviewed in Section 3 above relies on the assumption that the synthesis models are correctly specified, that is, they match the true data generating process. An additional requirement is that the synthesis model is congenial to the analysis model to be run on the synthetic data. In broad terms, congeniality [131] means that the synthesis model is based on the same (modeling) assumptions as the analysis model.

To be fair, as it is impossible to anticipate all analyses that will be run on the synthetic data, achieving congeniality is typically a hopeless goal in practice. Still, it has been shown in the nonresponse context [131] that approximately valid inferences can be obtained if the synthesis model encompasses the analysis model, that is, it contains more variables than the analysis model. Intuitively, this makes sense: adding a predictor variable during synthesis that, in reality, is conditionally independent of the variable to be synthesized given the other predictors in the model will not do much harm. It might unnecessarily increase the variance from synthesis, but it will not introduce any bias. However, omitting important variables will introduce bias, as the relationship between the omitted variable and the synthetic variable will be attenuated in the synthetic data.

Based on this reasoning, it is generally recommended to use a rich set of predictors in the synthesis model, ideally conditioning on all other variables in the dataset and also including interaction and squared terms if possible (see [117] for a similar argument in the nonresponse context). However, this strategy is typically not feasible when using parametric models, as many datasets contain dozens of variables. Especially with categorical variables, multicollinearity issues and the problem of perfect prediction often imply that fitting parametric models containing many variables is no longer possible and uncongeniality becomes a major concern.

To overcome this problem, researchers started exploring alternative synthesis strategies, borrowing ideas from the machine learning literature. In 2005, [163] suggested using CART. [30] later extended these ideas to random forests, and [49] developed strategies to adapt Support Vector Machines for data synthesis. Synthesis strategies based on genetic algorithms were explored in [36] and [37]. All these approaches have the advantage that they let the data speak for itself, that is, they might automatically identify higher-order relationships that might easily be missed when specifying parametric models. Furthermore, they are not affected by multicollinearity or perfect prediction problems and can still be directly applied if the number of variables exceeds the number of observations. In an evaluation study, [59] compared the different approaches and found that CART models offered the best results in terms of preserving the information from the original data. As explained in more detail below, a possible downside of these models is the risk of exactly replicating some records from the original data even for continuous variables.

In the computer science approach to synthetic data, the problem of uncongeniality was never explicitly considered. Since from the beginning the expected use case was the training of machine learning models, the focus of the research was on machine learning models right from the start. Before we review the different approaches from the computer science literature, we briefly discuss some extensions of the MI based synthesis procedures.

4.3 Extensions of the MI Inspired Approaches

The approaches reviewed in this section offer various extensions to the classical synthesis problem. They differ from the other approaches in that they require different inferential procedures than those discussed in Section 3. We will not review all these procedures here for brevity. Instead, we refer to the various papers for further detail.

The first extension of the classical MI-based synthesis approach was proposed by [159]. The paper offers a strategy to deal with missing data and data confidentiality simultaneously. The author proposes a two-step procedure, in which missing values are imputed m times at the first stage, and r partially synthetic datasets are generated at the second stage within each first stage nest, that is, the final data comprises $m \cdot r$ datasets. The appropriate procedures for multi-component hypothesis testing under this scenario were derived in [107].

In a similar spirit, [165] proposed a two-stage synthesis, for which variables that have a higher risk of disclosure are synthesized at the first stage, and variables that require a larger number of synthetic datasets to reduce the model uncertainty are synthesized on the second stage. This approach was motivated by previous findings [57] that increasing the number of synthetic datasets can lead to increased risks of disclosure. The authors show that their approach offers better disclosure protection and similar utility compared to standard one-stage synthesis with the same total number of synthetic datasets.

A final type of extension proposes to use a (sub)sampling step before the synthesis. This approach is especially attractive for Census data, for which it is common practice that only random samples of the full data are released to the public. What makes this approach special in the synthesis context is that the synthesis models can be estimated using the full data even if only a (sub)sample is synthesized later. [58] present the methodology if the original data covers the full population. Using a real dataset, they illustrate that releasing synthetic samples can actually offer higher utility than releasing samples of the original data. This surprising result is due to the fact that the synthesis models are based on information from the full population. [60] extend the methodology to the context where the original data is itself already a sample.

4.4 MI Based vs. Computer Science Approaches

The methods covered in Sections 4.1 to 4.3 were mostly inspired by the multiple imputation framework, treating the synthesis process as a missing data problem. The goal is to "impute" synthetic values given the original data. As discussed previously, an important focus from the beginning was to be able to obtain valid point estimates based on the synthetic data and to accurately quantify the uncertainty of these estimates. A key distinction between the MI-based approaches and the concepts proposed in the computer science literature is that the latter never aimed at being able to quantify the uncertainty of the estimates obtained from the synthetic data. Thus, the idea of generating multiple synthetic datasets was never discussed and different modeling strategies were proposed. In computer science, machine learning and deep learning methods such as Generative Adversarial Networks [82], GANs and Variational Autoencoders [106], VAEs have been popular generative modeling frameworks in recent years. Thus, it is perhaps not surprising that a large body of work on synthetic data in computer science is based on one of these concepts. In this section, we offer a brief overview of the most popular variants of these two approaches. Due to the large body of work in the field, we discuss only the most influential contributions, excluding works that are targeted toward very narrow areas of application.

4.4.1 *Generative adversarial networks* (*GANs*). As indicated in the Introduction, we will only focus on GANs for microdata synthesis in this review. Compared with the abundance of literature on GANs and other deep learning approaches for text, audio, and visual data generation, literature on the use of deep generative learning approaches for the synthesis of microdata is relatively sparse but rapidly growing [31, 39, 111, 145, 205].

GANs consist of two neural networks that compete with each other: the so-called generator (network) is trained to generate synthetic data and outputs synthetic samples given a random noise input. The discriminator (network) is trained to discriminate between real and synthetic data. The discriminator tries to minimize the misclassification error while the generator loss is calculated from the discriminator's classification—it gets penalized if it does not fool the discriminator. The standard combined loss function was described by [82] and is also called minimax loss, since the generator tries to minimize it while the discriminator tries to maximize it. The training of the GAN is an iterative process in which each of the neural networks updates its parameters based on the feedback received from the other network, that is, GANs make use of adversarial feedback loops to learn how to generate synthetic data that is indistinguishable from real data.

In recent years, Wasserstein GANs (WGANs) [11] have become increasingly popular. WGANs use the Wasserstein distance for the cost function instead of the Kullback–Leibler (KL) and Jensen–Shannon (JS) Divergence to avoid the problem of vanishing gradients [11]. In the context of Generative Adversarial Networks (GANs), vanishing gradients can occur if the discriminator becomes too strong compared to the generator. This is because if the discriminator can easily distinguish between real and fake data, the gradients of the loss function with respect to the parameters of the neural network become very small as they propagate from the output layer to the earlier layers of the network. This means that the updates to the parameters during training become extremely small, leading to slow or stalled learning. The Wasserstein distance is also called Earth Mover's distance and is widely used to solve optimal transport problems, that is, problems where the goal is to move things from a given configuration to a desired configuration with the smallest cost possible. Early examples for the use of WGANs for data synthesis can be found in [31].

For WGANs, a Lipschitz constraint is usually enforced for the discriminator. A Lipschitz constraint limits the rate at which the output of a function can change with respect to its input. When applied to WGANs, this property helps stabilize the training process by controlling the rate of change in the discriminator's output function. This in turn ensures more controlled and stable updates to the model parameters, enhancing the overall performance of the network. To implement this constraint, [11] propose to clip the weights if necessary, but noted that this approach is not optimal. To overcome this problem, WGAN-gradient penalty (WGAN-GP) [83] uses a gradient penalty to fulfill the Lipschitz constraint. [15], [111], actGAN, [205] and [211], CTAB-GAN, all use different adaptations of WGAN or WGAN-GP for data synthesis. [205] use normal mixture distributions to improve the fit for continuous variables. They also use a conditional generator, aiming for proper conditional distributions for each variable.

There also exist alternatives to WGANs for data synthesis, for example, GANs based on the Cramér Distance [135].

Causal-TGAN is an approach that stands out from other GAN approaches, as it explicitly takes the potentially complex causal relationships between the variables into account. It is composed of two steps, first obtaining the causal graph that represents the causal relations of the original dataset and then using the causal graph when training the GAN [200].

4.4.2 *Variational autoencoders* (*VAEs*). Another approach based on deep neural networks that has been adapted for data synthesis lately are variational autoencoders [106], VAE. In comparison to GANs, a VAE has three instead of two networks, which learn complementary tasks: an encoder network, a decoder network, and a discriminator. The encoder network maps the data onto a latent representation, while the decoder network tries to reconstruct it. As with GANs, the discriminator network decides for each given sample whether it is real data or data generated by the decoder network. A VAE is trained to minimize the reconstruction error between the reconstructed data and the initial data. Data synthesis approaches that use VAE are discussed in Srivastava et al. [183], VEEGAN, [31, 195, 205], and [121].

4.5 Differentially Private vs. Nondifferentially Private Data Synthesis

In recent years, DP [65] has been widely adopted as a definition of privacy offering formal, that is, mathematically quantifiable privacy guarantees. DP requires that the impact that any single record can have on the probability of obtaining a specific result is strictly bounded. Specifically, ε-DP, often referred to as pure DP to distinguish it from relaxations such as (ε, δ)-DP, requires that the log-difference in the probability of obtaining a specific output computed on two neighboring datasets, that is, datasets that differ only in one record, is bounded between ε and $-\varepsilon$. In layman's terms, an algorithm is differentially private if someone seeing the output statistic cannot tell if the information on a specific individual was used in the computation or not. See [66] or [194] for an in-depth discussion of DP and some relaxations of the concept that have been proposed in the literature. The body of work on DP has grown exponentially in recent years and several tech companies, such as Apple [213], Google [68], and Microsoft [46] as well as the U.S. Census Bureau [2, 72] recently adopted the approach for some of their data.

The concept of DP has also stimulated research on generating differentially private synthetic data. In this case, the synthetic data is the output of the algorithm, and hence a synthesizer needs to ensure that the generated data would not change substantially if *any* possible original data (and not only the data at hand) is changed by one record. Without further adjustments, all synthesis approaches discussed in the previous sections do not satisfy this requirement. For example, if a standard linear regression approach is used to generate synthetic data, it is easy to come up with extreme examples, in which the estimated regression coefficients change substantially if the value of a single record is changed in the original data. Since this would imply that the distribution from which the synthetic data are sampled can change arbitrarily if one record is changed in the original data, such a synthesizer violates the requirements of DP.

The major advantage of differentially private synthetic data is that it also offers a strong formal privacy guarantee for any output computed on the synthetic data: DP is *immune to post-processing*, that is, any function of a differentially private output is guaranteed to also be differentially private with the same privacy guarantees as the original output (see, for example, Proposition 2.1 in Dwork and Roth [66]). This implies that researchers working with the differentially private synthetic data are more free to interact with the data and use any tools and workflows to process the data without the risk of accidentally or purposefully revealing any sensitive information.

Various approaches have been proposed in the literature for generating differentially private synthetic data (see [24] for a review of early approaches). Using marginal distributions for the synthesis has been one of the most popular approaches. Noise is added to either one-, two- or three-way marginal distributions [120, 129, 130]. Another popular approach for differentially private data synthesis are Bayesian networks [14], most prominently PrivBayes by [210]. It can be difficult to represent all important correlations in PrivBayes. Therefore, [29] propose a Markov random field (MRF) that models the correlations among the variables in the original datasets, and then uses the MRF for data synthesis (PrivMRF). Game-based approaches such as those by Hardt, Ligett and McSherry [84], MWEM, and Gaboardi et al. [76], Dual-Query, require a set of specified queries, optimizing the synthesis to ensure high validity for these queries. Yet another popular approach developed by Li, Xiong and Jiang [113], DPCopula, is based on Copula functions.

Finally, work on integrating DP into generative adversarial networks (GANs) has been growing fast in the last few years [18, 136, 192, 204, 207]. Since the generator commonly never accesses the real data directly, only the discriminator needs to be modified to ensure DP: [18] and [204] built on [1] for the private optimization, adding Gaussian noise to the gradient of the Wasserstein distance in the WGAN algorithm. The gradients are also clipped if necessary. [75] also proposes a private extension of WGAN. Conditional GANs [132], CGAN are adapted by [192]. [207] use the Private Aggregation of Teacher Ensembles (PATE) framework proposed by [144], which provides a differentially private method for classification tasks. The framework is used for the discriminator's task to differentiate real and fake data.

To provide greater robustness against low utility of generated DP data sets, [136] proposed a method combining weighted samples produced by a sequence of generators. Their approach can be applied to differentially private or nonprivate GANs for data synthesis.

5. UTILITY EVALUTION

There is a large body of literature on measuring the validity of data that has undergone some form of perturbation to protect confidentiality. Most of these methods can also be used to measure the validity of synthetic data. We will focus on the measures that are most relevant for synthetic data. Additional measures are discussed, for example, in [95] or [12].

Utility metrics can be broadly divided into three categories: The first category, commonly referred to as *global utility metrics* or *broad measures* of utility, tries to assess the utility by directly comparing the original data with the protected data. These measures offer the advantage that no assumptions regarding the types of analyses the synthetic data will be used for need to be made. On the downside, given that utility is measured on a very aggregated level, good results for these measures do not necessarily guarantee high utility for a specific type of analysis the user might be interested in. *Outcome-specific utility metrics* or *narrow measures* of utility sit on the other end of the spectrum. They measure the utility for a specific analysis, for example, the results of a linear regression model. A third class of measures that we label *fit-for-purpose measures* usually form the starting point of any utility assessment. In broad terms, they assess whether the synthetic data look reasonable. Examples of these measures would be graphical comparisons of the marginal and bivariate distributions of all variables or consistency checks to avoid implausible values such as negative age values in the synthetic data.

5.1 Global Utility Metrics

As discussed above, these measures try to evaluate the utility by directly comparing the synthetic data to the original data. One common approach in this context is to use some distance measure, such as Kulback–Leibler divergence [100] or Hellinger distance [80]. A downside of these general distance measures is that they can be difficult to compute for large datasets. An alternative strategy tries to assess how easy it is to discriminate between the original data and the synthetic data, borrowing ideas from the literature on propensity score matching [172]. Propensity scores are estimated by stacking the n_{org} original records and the n_{syn} synthetic records and adding an indicator, which is one if the record is from the synthetic data and zero otherwise. In the next step, a model is fitted using the information contained in the data to estimate the propensity scores, that is, to estimate the probability for each record to belong to the synthetic data. If the synthetic data would be an exact copy of the original data, the data would not offer any information to discriminate between the data sources and the distribution of the estimated propensity scores would be the same for both datasets. Thus, one way to measure the utility of the synthetic data is to evaluate the difference in the distribution of the propensity score between the original data and the synthetic data. Various metrics can be used for this purpose. [25] suggest estimating the Kolmogorov–Smirnov distance between the two distributions (they call this measure SPECKS for Synthetic data generation; Propensity score matching; Empirical Comparison via the Kolmogorov–Smirnov distance). Alternatively, the Mann–Whitney U test (Wilcoxon rank-sum test) can also be used.

A measure that gained popularity in recent years is the propensity score mean squared error (pMSE) as an evaluation metric [182, 202]. Let p_i, $i = 1, \dots, N$ with $N = n_{\text{org}} + n_{\text{syn}}$ denote the predicted value obtained from the model for record i in the stacked dataset. The pMSE is calculated as $1/N \sum_N (p_i - c)^2$, with $c = n_{\text{syn}}/N$. The smaller the pMSE the higher the analytical validity of the synthetic data (note that $p_i \to c$ if the model cannot discriminate between the original data and the synthetic data). A downside of the pMSE noted by [202] is that it increases with the number of predictors included in the propensity model. To overcome this problem, [182] derived the expected value and the standard deviation of the pMSE under the null hypothesis that the synthesis model is correctly specified and proposed two additional utility measures. The first measure is the pMSE ratio which is computed as the empirical pMSE divided by its expected value under the null. The second measure is the standardized pMSE, which is the empirical pMSE minus its expectation under the null divided by its standard deviation under the null. In a recent paper, [54] critically discussed the pMSE illustrating that the estimated scores are highly dependent on the specification of the propensity model and that even blatant differences in the utility between different synthesizers are sometimes not picked up by the pMSE.

5.2 Outcome-Specific Utility Measures

These measures explicitly focus on measuring the usefulness of the synthetic data for a specific analysis task. For example, a straightforward visualization of the analytical validity is to plot estimates of interest (means, regression coefficients, etc.) obtained from the original data against the same estimates obtained from the synthetic data. If the utility is high, the coefficients should cluster around the 45 degree line. Typically, the comparison plots between the original and synthetic datasets are only accessible to the entity or individual synthesizing the data, not the end user. This is because releasing these comparison plots could potentially reveal sensitive information about the original dataset.

A downside of this evaluation is that it does not account for the inherent uncertainty of the estimates. Larger deviations between the estimates might be acceptable, if the sampling error is large, for example, if the estimate of interest is based on a small subset of the data. The same deviation might be problematic for a statistic based on the entire sample. A popular measure that also takes the uncertainty of the estimates into account is the *confidence interval overlap* measure proposed by [100]. It measures the relative average overlap between the confidence interval obtained from the original data and the confidence interval obtained from the synthetic data. An overlap measure close to one indicates that approximately the same inferential conclusions will be drawn irrespective of whether the synthetic data or the original data were used for the analysis. The measure is defined in such a way that it punishes increased uncertainty in the synthetic data, that is, for two synthetic data intervals that fully contain the interval from the original data, the measure favors the shorter interval. A downside of the measure is that it becomes meaningless for very large datasets as very small biases in the point estimates will inevitably lead to no overlap between the confidence intervals.

Given the increased relevance of machine learning approaches, another utility metric gained popularity in recent years, especially in the computer science literature: *machine learning efficacy*. Utility measures of this type, which are also referred to as measures of *model comparability*, assess whether machine learning models trained on the synthetic data give similar results compared to when they were trained on the original data. For these evaluations, the models of interest are typically trained on both the synthetic data and the original data and then the performance of the models is compared based on the same set of test records, which is obtained from the original data. The utility of the synthetic data is considered high, if classical evaluation criteria such as accuracy, F1 score, etc., are similar irrespective of whether the models were trained using the original data or the synthetic data. Sometimes, utility is also evaluated by assessing whether using the synthetic data for model training would lead to the same ranking of various machine learning models. For example, if the original data would suggest that a classifier based on a multilayer perceptron performs better than a random forest and the random forest is better than logistic regression, the same ranking should be found if the synthetic data were used for model training.

5.3 Fit-for-Purpose Measures

These measures represent the first step when evaluating the usefulness of the generated data. We treat them separately from the other two measures, as they do not necessarily focus on measuring the validity of specific analyses that might be important for the users of the data. They also do not try to directly assess the similarity of the original and the synthetic data in one global metric. Their main aim is to get a first impression of the quality of the synthetic data, and, unlike the global measures, they can help to identify aspects of the synthesis process that might still need to be improved. These measures can be divided into three groups: graphical evaluations, plausibility checks, and computing various goodness-of-fit measures.

Graphical evaluations typically include strategies such as side-by-side plots of the marginal distributions of the synthetic and the original data or contour plots for comparing bi-variate distributions. They also include visual comparisons of conditional distributions such as the income distribution for males and females or for different age groups.

For the plausibility checks, it is important to involve subject-matter experts that regularly work with the data. This is crucial, as not all inconsistencies are immediately obvious. For example, while it might be straightforward to identify problems such as married two-year-olds, it is much more difficult to judge which year-to-year change in turnover would be considered plausible for an establishment in a given industry in a given year.

Finally, any goodness-of-fit measure can be used to assess the similarity for specific aspects of the original and synthetic data. For example, the Kolmogorov–Smirnov test statistic can be used for each continuous variable in the dataset. Cross-tabulations of several variables (discretizing continuous variables if necessary) can be evaluated using the χ^2 statistic or the likelihood ratio statistic. [196] discuss the advantages and disadvantages of various metrics. However, it must be noted that the statistics should not be used to test for statistically significant differences between the original and the synthetic data. Given that the synthetic data are generated based on information from the original data, the two samples cannot be treated as independent—an assumption underlying most goodness-of-fit tests. Thus, any p-values computed using the standard test procedure would be misleading. Nevertheless, the value of the test statistic can still be used to compare the performance of different synthesis strategies. Furthermore, the test statistic can also be used as a metric to identify potential problems with the quality of the synthetic data. For example, if the test statistic is high for many of the cross-tabulations involving age, this serves as an indicator that the synthesis of the age variable needs to be improved.

The pMSE measure discussed in Section 5.2 can also be used as a fit-for-purpose measure by only including the variables of interest when estimating the propensity score. An illustration of how this strategy can be used to visualize the utility for bi-variate distributions is presented in [154]. These graphical visualization tools are also implemented in the R package synthpop [153].

In [154], the authors empirically evaluate various goodness-of-fit measures and find a large correlation (> 0.9) between most of them. Noticeably, the adjusted χ^2 test proposed by [196], the Freeman–Tukey statistic, the Jensen–Shannon divergence (JSD), and the pMSE had an empirical correlation above 0.99, so did the Kolmogorov–Smirnoff test statistic, the Mann–Whitney test statistic, and two additional measures that we don't review here for brevity. In practice, this seems to imply that it is sufficient to only use one or two goodness-of-fit criteria when assessing the utility of the generated data.

6. RISK ASSESSMENT

From a risk perspective, there is a fundamental difference between disseminating partially or fully synthetic data. With partial synthesis, there still exists a one-to-one mapping between the original data and the synthetic data. With fully synthetic data, this is no longer the case. In fact, with this approach, the synthetic data does not have to be of the same size as the original data. This implies that measuring the risk of re-identification, as commonly done for other disclosure protection strategies [164, 179, 181], is not meaningful for fully synthetic data. However, this does not mean that fully synthetic data can be assumed to have no risk of spilling sensitive information. For example, [125] illustrate using real data that if a fully conditional specification approach (which is commonly applied when using multiple imputation in the nonresponse context) is used for CART-based synthesis, there is a risk that the synthesizer simply replicates most of the original records. The problem arises as the approach always conditions on all other variables in the dataset. With complex datasets containing many (categorical) variables, this can lead to situations in which the values of the variable to be synthesized are completely deterministic given the other variables. The CART synthesizer can get stuck in such a situation, simply replicating the records from the original data. While such a problem can easily be avoided by not using the fully conditional specification approach (the approach offers no advantages in the context of synthetic data), this example still highlights that it would be naïve to assume that fully synthetic data will never pose any threats of disclosing sensitive information. However, measuring these risks is challenging and research in this area is still limited.

We start this section by reviewing the approaches that have been proposed in the literature to assess risks for fully synthetic data. In principle, these measures can also be used to assess the risks for partially synthetic data, while the risk measures that we review in the second part of this section are only useful for partial synthesis as they try to assess the risk of re-identification for the generated data. We also refer the interested reader to Hu [88], which contains a detailed review of Bayesian risk measures for synthetic data.

6.1 Measuring the Risk of Disclosure for Fully Synthetic Data

Even though the link between the original and the synthetic data is broken with full synthesis, some agencies still evaluate how many synthetic records have a unique match in the original data. The reasoning behind this evaluation is that the agencies are concerned about perceived risks. Survey respondents might be concerned if they find a synthetic record that exactly matches their own record, especially if their combination of attributes makes them unique in the original data.

Some authors [145, 211] also compute the distance between the synthetic data records and their closest neighbors in the original data. The average of these distances across all synthetic records is then used as a risk measure. From a practical perspective, it is not obvious which risk this measure is supposed to quantify. Even if the average distance is small, the distance could be large for some records. A potential attacker would never know which records have small distance and even if the distance is small, this does not necessarily imply a risk if the closest record is in a high density area of the data distribution.

Another measure that evaluates risk by matching cases from the original and synthetic data was proposed by [187]. They suggest dividing the variables in the dataset into key variables, which are assumed to be known by the attacker, and target variables, which the attacker tries to infer. They assume that the attacker focuses on records with low l-diversity for the target variables within a given equivalence class given by the key variables. Let K denote the vector containing the key variables and T denote the vector of target variables. The authors define the Within Equivalence Class Attribution Probability (WEAP) as

$$\text{WEAP}_j = \Pr(T_j | K_j) = \frac{\sum_{i=1}^{n} I(T_i = T_j, K_i = K_j)}{\sum_{i=1}^{n} I(K_i = K_j)},$$

where $I(\cdot)$ is the indicator function that is one whenever the statement inside the parentheses is true and zero otherwise, and n is the size of the database. In their application, the authors focus on those synthetic records for which $\text{WEAP}_j = 1$. For those records, they compute the Targeted Correct Attribution Probability (TCAP):

$$\begin{aligned}\text{TCAP}_{sj} &= \Pr(T_{sj} | K_{sj})_o \\ &= \frac{\sum_{i=1}^{n} I(T_{o,i} = T_{s,j}, K_{o,i} = K_{s,j})}{\sum_{i=1}^{n} I(K_{o,i} = K_{s,j})},\end{aligned}$$

where the subscript s denotes synthetic data and o denotes the original data. The TCAP score is bounded between zero and one, with larger values indicating higher risks.

Another class of risk measures for fully synthetic data focuses on the fact that the synthesis models themselves can leak some information regarding the content of the original data. For example, when using a fully saturated log-linear model to synthesize a set of categorical variables combined with vague prior information, the existence of certain attribute combinations in the synthetic data reveals that the same combination must have been present in the original data. In the computer science literature, these types of risk evaluations are called membership attacks, as an attacker will learn that a certain record was present in the original data. Various strategies to estimate the risks from membership attacks have been proposed in the literature. Most of these approaches assume that the attacker already knows the true values for some target records and uses this information to learn whether these units are included in the original data [184]. These evaluations are based on the strong assumption that the attacker is not interested in learning something new about a unit contained in the data. Instead, the only goal is to learn whether the unit was part of the original data. There are situations in which learning this information is considered unacceptable: some laws explicitly state that such risks must be avoided. In addition, sometimes the fact that someone is contained in a database already reveals sensitive information, if the database only contains a specific subgroup of the population such as the Survey of Prison Inmates conducted by the Bureau of Justice Statistics in the United States.

However, there are also risk measures based on inferential attacks that do not make such strong assumptions. Borrowing ideas from the DP literature, [170] propose strategies to compute the posterior distribution $f(Y_i | D, X, M, d_{\text{org}}^{-i})$, where Y_i is the original value of some variable Y for unit i, D is the synthetic data, X might contain unchanged values from the original data (X will be empty for full synthesis), M contains information about the synthesis model and d_{org}^{-i} is the original data excluding record i. The approach evaluates how much an attacker can learn about an unknown value Y_i after seeing the synthetic data. If the posterior distribution for Y_i has low variability (especially if compared to the prior distribution before seeing the synthetic data) disclosure can occur. In principle, the strong assumption that the attacker knows all the information from the original data except for record i is not strictly necessary. However, in practice, it is typically unavoidable to make the problem computationally tractable. But even with these assumptions, this risk assessment is only feasible if the number of variables in the data is very limited (see [91] and [127] for illustrations).

In general, measuring disclosure risks for fully synthetic data remains challenging. While most researchers agree that fully synthetic data are not free from risk, more research is needed to quantify these risks under realistic settings. Another challenge in this context is the fact that the metrics used to assess risk must be interpretable for decision-makers such as disclosure review boards that will have to make the final decision whether the data are sufficiently protected before the release. If the metrics are too technical it can be difficult for the board to make this judgment call.

6.2 Measuring the Risk for Partially Synthetic Data

As indicated above, most of the risk measures from the previous section can also be used for partial synthesis. However, the fact that synthetic records are only generated for units that were already included in the original data implies that each record in the synthetic data has a unique match in the original data. Thus, one way to measure the risk with partially synthetic data is to evaluate whether an attacker would be able to reidentify some records in the synthetic data. Building on previous work in [164, 167] developed strategies to measure the risk of reidentification for partially synthetic data.

Borrowing from [58], the risk computations can be summarized as follows. Suppose the intruder has a vector of information, $\mathbf{t}$, on a particular target unit in the population $\mathbf{P}$. Let t_0 be the unique identifier of the target, and let P_{i0} be the (not released) unique identifier for record i in $\mathbf{d}_{\text{syn}}$, where $\mathbf{d}_{\text{syn}}$ denotes the synthetic data and

$i = 1, \dots, n$. Let $\mathcal{S}$ be any information released about the synthesis models.

The intruder's goal is to match unit i in $\mathbf{d}_{\text{syn}}$ to the target when $P_{i0} = t_0$. Let J be a random variable that equals i when $P_{i0} = t_0$ for $i \in \mathbf{d}_{\text{syn}}$. The intruder thus seeks to calculate $\Pr(J = i|\mathbf{t}, \mathbf{d}_{\text{syn}}, \mathcal{S})$ for $i = 1, \dots, n$. Because the intruder does not know the actual values of the synthesized variable Y^*, he or she should integrate over its possible values when computing the match probabilities. Hence, for each record he or she computes

$$\begin{aligned} &\Pr(J = i|\mathbf{t}, \mathbf{d}_{\text{syn}}, \mathcal{S}) \\ (6)\quad &= \int \Pr(J = i|\mathbf{t}, \mathbf{d}_{\text{syn}}, Y^*, \mathcal{S}) \Pr(Y^*|\mathbf{t}, \mathbf{d}_{\text{syn}}, \mathcal{S})\, dY^*. \end{aligned}$$

This construction suggests a Monte Carlo approach to estimating each $\Pr(J = i|\mathbf{t}, \mathbf{d}_{\text{syn}}, \mathcal{S})$. First, sample a value of Y^* from $\Pr(Y^*|\mathbf{t}, \mathbf{d}_{\text{syn}}, \mathcal{S})$. Let Y_{new} represent one set of simulated values. Second, compute $\Pr(J = i|\mathbf{t}, \mathbf{d}_{\text{syn}}, Y^* = Y_{\text{new}}, \mathcal{S})$ using a matching strategy such as nearest neighbor matching assuming Y_{new} are collected values. This two-step process is iterated h times, where ideally h is large, and (6) is estimated as the average of the resultant h values of $\Pr(J = i|\mathbf{t}, \mathbf{d}_{\text{syn}}, Y^* = Y_{\text{new}}, \mathcal{S})$. When $\mathcal{S}$ has no information, the intruder treats the simulated values as plausible draws of Y^*.

The disclosure risk can be measured using summaries of these identification probabilities. It is reasonable to assume that the intruder selects as a match for $\mathbf{t}$ the record i with the highest value of $\Pr(J = i|\mathbf{t}, \mathbf{d}_{\text{syn}}, \mathcal{S})$, if a unique maximum exists. [167] proposed three risk measures: the expected match risk, the true match rate, and the false match rate. Let c_i be the number of records with the highest match probability for the target $\mathbf{t_i}$; let $I_i = 1$ if the true match is among the c_i units and $I_i = 0$ otherwise. The expected match risk equals $\sum I_i/c_i$. When $I_i = 1$ and $c_i > 1$, the contribution of unit i to the expected match risk reflects the intruder randomly guessing at the correct match from the c_i candidates. Let $K_i = 1$ when $c_i I_i = 1$ and $K_i = 0$ otherwise and let N denote the total number of target records. The true match rate equals $\sum K_i/N$, which is the percentage of true unique matches among the target records. Finally, let $F_j = 1$ when $c_j(1 - I_j) = 1$ and $F_j = 0$ otherwise and let s equal the number of records with $c_i = 1$. The false match rate equals $\sum F_j/s$, which is the percentage of false matches among unique matches. Risk measures inspired by this methodology are available in the R package IdentificationRiskCalculation [87].

These risk assessments are based on the conservative assumption that the intruder knows that the target record is included in the released data. Extensions of the approach which also account for the extra uncertainty from sampling if the intruder does not know whether the individual he or she is looking for participated in the survey are given in [56].

7. CONCLUSION

The interest in synthetic data has been growing steadily over the last thirty years. While the focus was on methodological aspects and statistical properties during the first decade, first applications started to appear around the turn of the century. The great success of GANs, which always require generating synthetic data even if the final goal is not to disseminate these data, had a huge impact on the synthetic data movement, especially in the computer science community. The availability of easy-to-use software such as synthpop [153] or the synthetic data vault [146] also meant that more statistical agencies and other data disseminating organizations were able to explore the approach without the need to implement the synthesizers from scratch.

In this paper, we reviewed the historic developments of the synthetic data approach, offered a taxonomy of approaches, and discussed methods to measure risk and utility of the generated data. For organizational reasons, we treated the statistical approach separately from the computer science approach. While it is true that the developments in the two fields mostly happened independently with little exchange between the disciplines, the lines have always been blurry (e.g., [122] already integrate ideas from both fields), and the increasing number of collaborations between statisticians and computer scientists in recent years will hopefully make this distinction obsolete in the future.

Furthermore, most of the applications of the synthetic data approach do not use the synthetic data as the final product. The synthetic data are either used for training purposes [73] or to develop code in preparation for working with the real data [43, 154]. Even in those cases in which final access to the real data is not possible, the data providers typically guarantee that they will run the final results on the original data and report back the results if they can be released without violating confidentiality [19, 27]. This implies that procedures for obtaining valid variance estimates from the synthetic data as discussed in Section 3 are less relevant in practice, and the fact that many of the computer science approaches never achieved this goal is less of a concern.

For those cases in which access to the original data cannot be provided, verification servers can be a useful alternative. These servers hold both the synthetic and the original data. Researchers can submit their analysis of interest to the server, it runs the analysis on both datasets, and reports back some fidelity measure of how close the results from the synthetic data are to the results based on the original data. Compared to the guarantee of running the final models on the original data (sometimes called validation servers in the literature), verification servers have the advantage that the procedure can be automated.[3] Since the server only reports a fidelity measure and not the actual results, no manual output checking is required. This means that the server could also be used frequently during data preparation and not only for the final model. However, some care must be taken, as even fidelity measures might spill sensitive information. Developing measures that are informative but at the same time are guaranteed not to spill sensitive information is an area of active research [17, 128, 168, 208].

A systematic comparison between the approaches developed in the different fields is currently lacking, although Goncalves et al. [81] and [116] offer some first insights. The authors compared several synthesis strategies based on CART models, Bayesian Networks, various parametric and nonparametric models and three GAN implementations (medGAN, tableGAN and CTGAN). Although only some of the methods were considered in both papers, the general findings are comparable. Both papers found that the sequential-regression-based CART approach offered the highest utility, but also the highest risk. Goncalves et al. [81] also found high utility for DPMPM models and for CLGP, which only work for categorical data. Two of the GANs (tableGAN and medGAN) resulted in unacceptably low utility, while CTGAN and the approach based on Bayesian Networks performed almost similarly. However, these evaluations were based on only on a limited number of datasets and either relied on the default settings of the different synthesis implementations [116] or used limited hyperparameter tuning [81]. More extensive evaluations of the advantages and disadvantages of the various approaches that have been proposed in the literature would be an important area of future research.

Additionally, it will also be important to obtain a better understanding of the disclosure risk of (fully) synthetic data in the future. The measures that currently exist are either computationally too expensive to be useful, make unrealistically strong assumptions regarding the attacker, or only partially address the potential risks of the data release. Furthermore, many potential users especially from the scientific community have concerns and reservations against working with synthetic data. How can they be sure that the results that they obtained based on the synthetic data are reliable? The verification servers discussed above might be one strategy to address these concerns. Finally, DP synthetic data is still in its infancy. Many of the existing methods require so much noise to be infused that the utility of the resulting data would be too low, especially for the complex high-dimensional datasets that statistical agencies typically have to handle. It remains an open question whether the methodology can be sufficiently improved to be able to generate differentially private synthetic data with acceptable levels of utility for these complex data products in the future.

[3] We note that the terms “validation” and “verification” are not well-defined and are sometimes used exactly in the opposite meaning in the literature.

ACKNOWLEDGMENTS

The authors are grateful for helpful feedback on an earlier version of this paper from the FK2RG group at Mannheim University and LMU Munich. The authors also acknowledge very valuable feedback from three anonymous referees and an Associate Editor.

FUNDING

This work was partially supported by the German Federal Institute for Drugs and Medical Devices and a grant from the German Federal Ministry of Education and Research (grant number 16KISA096) with funding from the European Union—NextGenerationEU.

REFERENCES

[1] Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K. and Zhang, L. (2016). Deep learning with differential privacy. In *Proceedings of the* 2016 *ACM SIGSAC Conference on Computer and Communications Security* 308–318. ACM, Vienna, Austria.

[2] Abowd, J., Ashmead, R., Cumings-Menon, R., Garfinkel, S., Heineck, M., Heiss, C., Johns, R., Kifer, D., Leclerc, P. et al. (2022). The 2020 census disclosure avoidance system TopDown algorithm. *Harv. Data Sci. Rev.* **2**. Special Issue.

[3] Abowd, J., Ashmead, R., Simson, G., Kifer, D., Leclerc, P., Machanavajjhala, A. and Sexton, W. (2019). Census topdown: Differentially private data, incremental schemas, and consistency with public knowledge. U.S. Census Bureau, Washington, DC.

[4] Abowd, J. M., Stinson, M. and Benedetto, G. (2006). Final report to the social security administration on the SIPP/SSA/IRS public use file project Technical report, longitudinal employer–household dynamics program. U.S. Bureau of the Census, Washington, DC.

[5] Abowd, J. M. and Vilhuber, L. (2008). How protective are synthetic data? In *Privacy in Statistical Databases* (J. Domingo-Ferrer and Y. Saygın, eds.) **5262** 239–246. Springer, Berlin.

[6] Abowd, J. M. and Woodcock, S. D. (2001). Disclosure limitation in longitudinal linked data. In *Confidentiality, Disclosure, and Data Access: Theory and Practical Applications for Statistical Agencies* (P. Doyle, J. Lane, L. Zayatz and J. Theeuwes, eds.) 215–277. North-Holland, Amsterdam.

[7] Abowd, J. M. and Woodcock, S. D. (2004). Multiply-imputing confidential characteristics and file links in longitudinal linked data. In *Privacy in Statistical Databases* (J. Domingo-Ferrer and V. Torra, eds.) 290–297. Springer, New York.

[8] Alam, M. J., Dostie, B., Drechsler, J. and Vilhuber, L. (2020). Applying data synthesis for longitudinal business data across three countries. *Statist. Transition New Series* **21** 212–236.

[9] Allken, V., Handegard, N. O., Rosen, S., Schreyeck, T., Mahiout, T. and Malde, K. (2018). Fish species identification using a convolutional neural network trained on synthetic data. *ICES J. Mar. Sci.* **76** 342–349.

[10] AN, D. and LITTLE, R. J. A. (2007). Multiple imputation: An alternative to top coding for statistical disclosure control. *J. Roy. Statist. Soc. Ser. A* **170** 923–940. MR2408985 https://doi.org/10.1111/j.1467-985X.2007.00492.x

[11] ARJOVSKY, M., CHINTALA, S. and BOTTOU, L. (2017). Wasserstein GAN. Available at arXiv:1701.07875 [stat.ML].

[12] ARNOLD, C. and NEUNHOEFFER, M. (2020). Really useful synthetic data–a framework to evaluate the quality of differentially private synthetic data. Available at arXiv:2004.07740.

[13] AUSTRALIAN BUREAU OF STATISTICS (2021). Methodological news, Dec 2021. Available at https://www.abs.gov.au/statistics/research/methodological-news-dec-2021. Last accessed on 2022-05-17.

[14] BAO, E., XIAO, X., ZHAO, J., ZHANG, D. and DING, B. (2021). Synthetic data generation with differential privacy via Bayesian networks. *J. Priv. Confid.* **11**.

[15] BAOWALY, M. K., LIN, C.-C., LIU, C.-L. and CHEN, K.-T. (2019). Synthesizing electronic health records using improved generative adversarial networks. *J. Amer. Med. Inform. Assoc.* **26** 228–241.

[16] BARAK, B., CHAUDHURI, K., DWORK, C., KALE, S., MCSHERRY, F. and TALWAR, K. (2007). Privacy, accuracy, and consistency too: A holistic solution to contingency table release. In *Proceedings of the Twenty-Sixth ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems—PODS* '07 273–282. ACM, Beijing, China.

[17] BARRIENTOS, A. F., BOLTON, A., BALMAT, T., REITER, J. P., DE FIGUEIREDO, J. M., MACHANAVAJJHALA, A., CHEN, Y., KNEIFEL, C. and DELONG, M. (2018). Providing access to confidential research data through synthesis and verification: An application to data on employees of the U.S. federal government. *Ann. Appl. Stat.* **12** 1124–1156. MR3834297 https://doi.org/10.1214/18-AOAS1194

[18] BEAULIEU-JONES, B. K., WU, Z. S., WILLIAMS, C., LEE, R., BHAVNANI, S. P., BYRD, J. B. and GREENE, C. S. (2019). Privacy-preserving generative deep neural networks support clinical data sharing. *Circ. Cardiovasc. Qual. Outcomes* **12** e005122. https://doi.org/10.1161/CIRCOUTCOMES.118.005122

[19] BENEDETTO, G., STANLEY, J. C., TOTTY, E. et al. (2018). The creation and use of the SIPP synthetic beta version 7.0.

[20] BLUM, A., LIGETT, K. and ROTH, A. (2013). A learning theory approach to noninteractive database privacy. *J. ACM* **60** Art. 12, 25. MR3060810 https://doi.org/10.1145/2450142.2450148

[21] BONNÉRY, D., FENG, Y., HENNEBERGER, A. K., JOHNSON, T. L., LACHOWICZ, M., ROSE, B. A., SHAW, T., STAPLETON, L. M., WOOLLEY, M. E. et al. (2019). The promise and limitations of synthetic data as a strategy to expand access to state-level multi-agency longitudinal data. *J. Res. Educ. Eff.* **12** 616–647.

[22] BOWEN, C. M., BRYANT, V., BURMAN, L., CZAJKA, J., KHITATRAKUN, S., MACDONALD, G., MCCLELLAND, R., MUCCIOLO, L., PICKENS, M. et al. (2022). Synthetic individual income tax data: Methodology, utility, and privacy implications. In *International Conference on Privacy in Statistical Databases* 191–204. Springer, Berlin.

[23] BOWEN, C. M., BRYANT, V., BURMAN, L., KHITATRAKUN, S., MCCLELLAND, R., STALLWORTH, P., UEYAMA, K. and WILLIAMS, A. R. (2020). A synthetic supplemental public use file of low-income information return data: Methodology, utility, and privacy implications. In *International Conference on Privacy in Statistical Databases* 257–270. Springer, Berlin.

[24] BOWEN, C. M. and LIU, F. (2020). Comparative study of differentially private data synthesis methods. *Statist. Sci.* **35** 280–307. MR4106606 https://doi.org/10.1214/19-STS742

[25] BOWEN, C. M., LIU, F. and SU, B. (2021). Differentially private data release via statistical election to partition sequentially. *Metron* **79** 1–31. MR4239846 https://doi.org/10.1007/s40300-021-00201-0

[26] BOWEN, C. M. and SNOKE, J. (2021). Comparative study of differentially private synthetic data algorithms from the NIST PSCR differential privacy synthetic data challenge. *J. Priv. Confid.* **11**. https://doi.org/10.29012/jpc.748

[27] BURMAN, L. E., ENGLER, A., KHITATRAKUN, S., NUNNS, J. R., ARMSTRONG, S., ISELIN, J., MACDONALD, G. and STALLWORTH, P. (2019). Safely expanding research access to administrative tax data: creating a synthetic public use file and a validation server Technical report, Technical report US, Internal Revenue Service.

[28] BURRIDGE, J. (2003). Information preserving statistical obfuscation. *Stat. Comput.* **13** 321–327. MR2005433 https://doi.org/10.1023/A:1025658621216

[29] CAI, K., LEI, X., WEI, J. and XIAO, X. (2021). Data synthesis via differentially private Markov random fields. *Proc. VLDB Endow.* **14** 2190–2202.

[30] CAIOLA, G. and REITER, J. P. (2010). Random forests for generating partially synthetic, categorical data. *Trans. Data Priv.* **3** 27–42. MR2725418

[31] CAMINO, R., HAMMERSCHMIDT, C. and STATE, R. (2018). Generating multi-categorical samples with generative adversarial networks. Available at arXiv:1807.01202 [cs, stat].

[32] CANO, I., LADRA, S. and TORRA, V. (2010). Evaluation of information loss for privacy preserving data mining through comparison of fuzzy partitions. In *International Conference on Fuzzy Systems* 1–8 IEEE Press, Barcelona, Spain.

[33] CHALLENGE.GOV (2019). NIST differential privacy synthetic data challenge. Available at https://www.challenge.gov/?challenge=differential-privacy-synthetic-data-challenge. Last accessed on 2022-06-08.

[34] CHAREST, A.-S. (2011). How can we analyze differentially-private synthetic datasets? *J. Priv. Confid.* **2**.

[35] CHEN, J., CHUN, D., PATEL, M., CHIANG, E. and JAMES, J. (2019). The validity of synthetic clinical data: A validation study of a leading synthetic data generator (synthea) using clinical quality measures. *BMC Med. Inform. Decis. Mak.* **19** 1–9.

[36] CHEN, Y., ELLIOT, M. and SAKSHAUG, J. (2016). A genetic algorithm approach to synthetic data production. In *Proceedings of the* 1*st International Workshop on AI for Privacy and Security*. 1–4.

[37] CHEN, Y., ELLIOT, M. and SMITH, D. (2018). The application of genetic algorithms to data synthesis: A comparison of three crossover methods. In *International Conference on Privacy in Statistical Databases* 160–171. Springer, Berlin.

[38] CHIEN, C.-H., WELSH, A. H. and MOORE, J. D. (2020). Synthetic business microdata: An Australian example. *J. Priv. Confid.* **10**.

[39] CHOI, E., BISWAL, S., MALIN, B., DUKE, J., STEWART, W. F. and SUN, J. (2018). Generating multi-label discrete patient records using generative adversarial networks. Available at arXiv:1703.06490 [cs].

[40] COMMISSION, E. (2022). European data strategy. Available at https://ec.europa.eu/info/strategy/priorities-2019-2024/europe-fit-digital-age/european-data-strategy_en. Last accessed on 2022-05-03.

[41] DE MONTJOYE, Y.-A., HIDALGO, C. A., VERLEYSEN, M. and BLONDEL, V. D. (2013). Unique in the crowd: The privacy bounds of human mobility. *Sci. Rep.* **3** 1–5.

[42] DE MONTJOYE, Y.-A., RADAELLI, L., SINGH, V. K. and PENTLAND, A. S. (2015). Identity and privacy. Unique in the shopping mall: On the reidentifiability of credit card metadata. *Science* **347** 536–539. https://doi.org/10.1126/science.1256297

[43] DE WOLF, P.-P. (2015). Public use files of EU-SILC and EU-LFS data. Joint UNECE/Eurostat work session on statistical data confidentiality Helsinki, Finland, 1–10.

[44] DENTON, E. L., CHINTALA, S., FERGUS, R. et al. (2015). Deep generative image models using a Laplacian pyramid of adversarial networks. *Adv. Neural Inf. Process. Syst.* **28**.

[45] DEPARTMENT FOR DIGITAL, CULTURE, MEDIA & SPORT (2022). National data strategy. Available at https://www.gov.uk/government/publications/uk-national-data-strategy/national-data-strategy. Last accessed on 2022-05-03.

[46] DING, B., KULKARNI, J. and YEKHANIN, S. (2017). Collecting telemetry data privately. *Adv. Neural Inf. Process. Syst.* 3571–3580.

[47] DONG, Q., ELLIOTT, M. R. and RAGHUNATHAN, T. E. (2014). A nonparametric method to generate synthetic populations to adjust for complex sampling design features. *Surv. Methodol.* **40** 29–46.

[48] DONG, Q., ELLIOTT, M. R. and RAGHUNATHAN, T. E. (2014). Combining information from multiple complex surveys. *Surv. Methodol.* **40** 347–354.

[49] DRECHSLER, J. (2010). Using support vector machines for generating synthetic datasets. In *International Conference on Privacy in Statistical Databases* 148–161. Springer, Berlin.

[50] DRECHSLER, J. (2011). *Synthetic Datasets for Statistical Disclosure Control: Theory and Implementation. Lecture Notes in Statistics* **201**. Springer, New York. MR2809912 https://doi.org/10.1007/978-1-4614-0326-5

[51] DRECHSLER, J. (2011). Improved variance estimation for fully synthetic datasets. Proceedings of the joint UNECE/EUROSTAT work session on statistical data confidentiality.

[52] DRECHSLER, J. (2012). New data dissemination approaches in old Europe—synthetic datasets for a German establishment survey. *J. Appl. Stat.* **39** 243–265. MR2879819 https://doi.org/10.1080/02664763.2011.584523

[53] DRECHSLER, J. (2018). Some clarifications regarding fully synthetic data. In *International Conference on Privacy in Statistical Databases* 109–121. Springer, Berlin.

[54] DRECHSLER, J. (2022). Challenges in measuring utility for fully synthetic data. In *International Conference on Privacy in Statistical Databases* 220–233. Springer, Berlin.

[55] DRECHSLER, J. and HU, J. (2021). Synthesizing geocodes to facilitate access to detailed geographical information in large-scale administrative data. *J. Surv. Stat. Methodol.* **9** 523–548.

[56] DRECHSLER, J. and REITER, J. P. (2008). Accounting for intruder uncertainty due to sampling when estimating identification disclosure risks in partially synthetic data. In *Privacy in Statistical Databases* (J. Domingo-Ferrer and Y. Saygin, eds.) 227–238. Springer, New York.

[57] DRECHSLER, J. and REITER, J. P. (2009). Disclosure risk and data utility for partially synthetic data: An empirical study using the German IAB establishment survey. *J. Off. Stat.* **25** 589–603.

[58] DRECHSLER, J. and REITER, J. P. (2010). Sampling with synthesis: A new approach for releasing public use census microdata. *J. Amer. Statist. Assoc.* **105** 1347–1357. Supplementary materials available online. MR2796555 https://doi.org/10.1198/jasa.2010.ap09480

[59] DRECHSLER, J. and REITER, J. P. (2011). An empirical evaluation of easily implemented, nonparametric methods for generating synthetic datasets. *Comput. Statist. Data Anal.* **55** 3232–3243. MR2825406 https://doi.org/10.1016/j.csda.2011.06.006

[60] DRECHSLER, J. and REITER, J. P. (2012). Combining synthetic data with subsampling to create public use microdata files for large scale surveys. *Surv. Methodol.* **38** 73–79.

[61] DRECHSLER, J. and VILHUBER, L. (2014). Synthetic longitudinal business databases for international comparisons. In *International Conference on Privacy in Statistical Databases* 243–252. Springer, Berlin.

[62] DRECHSLER, J. and VILHUBER, L. (2014). A first step towards a German SynLBD: Constructing a German longitudinal business database. *Stat. J. IAOS* **30** 137–142.

[63] DUNCAN, G. T., ELLIOT, M. and SALAZAR-GONZÁLEZ, J.-J. (2011). *Statistical Confidentiality: Principles and Practice. Statistics for Social and Behavioral Sciences*. Springer, New York. MR3186259 https://doi.org/10.1007/978-1-4419-7802-8

[64] DWORK, (2008). Differential privacy: A survey of results. In *Theory and Applications of Models of Computation* (M. Agrawal, D. Du, Z. Duan and A. Li, eds.) 1–19. Springer, Berlin.

[65] DWORK, C., MCSHERRY, F., NISSIM, K. and SMITH, A. (2006). Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography. Lecture Notes in Computer Science* **3876** 265–284. Springer, Berlin. MR2241676 https://doi.org/10.1007/11681878_14

[66] DWORK, C. and ROTH, A. (2013). The algorithmic foundations of differential privacy. *Found. Trends Theor. Comput. Sci.* **9** 211–487. MR3254020 https://doi.org/10.1561/0400000042

[67] ENO, J. and THOMPSON, C. W. (2008). Generating synthetic data to match data mining patterns. *IEEE Internet Comput.* **12** 78–82.

[68] ERLINGSSON, Ú., PIHUR, V. and KOROLOVA, A. (2014). Rappor: Randomized aggregatable privacy-preserving ordinal response. In *Proceedings of the* 2014 *ACM SIGSAC Conference on Computer and Communications Security* 1054–1067.

[69] ESTEBAN, C., HYLAND, S. L. and RÄTSCH, G. (2017). Real-valued (medical) time series generation with recurrent conditional gans. Available at arXiv:1706.02633.

[70] EUROPEAN COMMISSION (2024). How contact tracing and warning apps helped during the COVID-19 pandemic. Available at https://commission.europa.eu/strategy-and-policy/coronavirus-response/travel-during-coronavirus-pandemic/contact-tracing-and-warning-apps-during-covid-19_en. Last accessed on 2024-01-12.

[71] EUROSTAT (2022). Statistics on income and living conditions. Available at https://ec.europa.eu/eurostat/web/microdata/statistics-on-income-and-living-conditions. Last accessed on 2022-05-16.

[72] FOOTE, A. D., MACHANAVAJJHALA, A. and MCKINNEY, K. (2019). Releasing earnings distributions using differential privacy: Disclosure avoidance system for post-secondary employment outcomes (PSEO). *J. Priv. Confid.* **9**.

[73] FORBES, S. and ZEALAND, S. N. (2008). Raising statistical capability: Statistics New Zealand's contribution. In *Government Statistical Offices and Statistical Literacy* 1–18.

[74] FRID-ADAR, M., KLANG, E., AMITAI, M., GOLDBERGER, J. and GREENSPAN, H. (2018). Synthetic data augmentation using GAN for improved liver lesion classification. In 2018 *IEEE* 15*th International Symposium on Biomedical Imaging* (*ISBI* 2018) 289–293.

[75] FRIGERIO, L., DE OLIVEIRA, A. S., GOMEZ, L. and DUVERGER, P. (2019). Differentially private generative adversarial networks for time series, continuous, and discrete open data.

[76] GABOARDI, M., ARIAS, E. J. G., HSU, J., ROTH, A. and WU, Z. S. (2014). Dual query: Practical private query release for high dimensional data. In *Proceedings of the* 31*st International Conference on Machine Learning* (E. P. Xing and T. Jebara, eds.). *Proceedings of Machine Learning Research* **32** 1170–1178. PMLR, Bejing, China.

[77] GAL, Y., CHEN, Y. and GHAHRAMANI, Z. (2015). Latent Gaussian processes for distribution estimation of multivariate categorical data. In *International Conference on Machine Learning* 645–654. PMLR.

[78] GHORBANI, A., NATARAJAN, V., COZ, D. and LIU, Y. (2020). DermGAN: Synthetic generation of clinical skin images with pathology. In *Proceedings of the Machine Learning for Health NeurIPS Workshop* (A. V. Dalca, M. B. A. McDermott, E. Alsentzer, S. G. Finlayson, M. Oberst, F. Falck and B. Beaulieu-Jones, eds.). *Proceedings of Machine Learning Research* **116** 155–170. PMLR.

[79] GOLDSTEIN, R., WOOLLEY, M. E., STAPLETON, L. M., BONNÉRY, D., LACHOWICZ, M., SHAW, T. V., HENNEBERGER, A. K., JOHNSON, T. L. and FENG, Y. (2020). Expanding MLDS data access and research capacity with synthetic data sets.

[80] GOMATAM, S. and KARR, A. F. (2003). Distortion measures for categorical data swapping Technical report, National Institute of Statistical Sciences, Research Triangle Park, NC.

[81] GONCALVES, A., RAY, P., SOPER, B., STEVENS, J., COYLE, L. and SALES, A. P. (2020). Generation and evaluation of synthetic patient data. *BMC Med. Res. Methodol.* **20** 1–40.

[82] GOODFELLOW, I. J., POUGET-ABADIE, J., MIRZA, M., XU, B., WARDE-FARLEY, D., OZAIR, S., COURVILLE, A. and BENGIO, Y. (2014). Generative adversarial networks. Available at arXiv:1406.2661 [cs, stat].

[83] GULRAJANI, I., AHMED, F., ARJOVSKY, M., DUMOULIN, V. and COURVILLE, A. (2017). Improved training of Wasserstein GANs.

[84] HARDT, M., LIGETT, K. and MCSHERRY, F. (2012). A simple and practical algorithm for differentially private data release. Available at arXiv:1012.4763 [cs].

[85] HAWALA, S. (2008). Producing partially synthetic data to avoid disclosure. In *Proceedings of the Joint Statistical Meetings* Amer. Statist. Assoc., Alexandria, VA.

[86] HOMER, N., SZELINGER, S., REDMAN, M., DUGGAN, D., TEMBE, W., MUEHLING, J., PEARSON, J. V., STEPHAN, D. A., NELSON, S. F. et al. (2008). Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays. *PLoS Genet.* **4** e1000167. https://doi.org/10.1371/journal.pgen.1000167

[87] HORNBY, R. and HU, J. (2021). Identification risks evaluation of partially synthetic data with the IdentificationRiskCalculation R package. *Trans. Data Priv.* **14** 37–52.

[88] HU, J. (2019). Bayesian estimation of attribute and identification disclosure risks in synthetic data. *Trans. Data Priv.* **12** 61–89.

[89] HU, J., AKANDE, O. and WANG, Q. (2021). Multiple imputation and synthetic data generation with NPBayesImputeCat. *R J.* **13**.

[90] HU, J. and HOSHINO, N. (2018). The quasi-multinomial synthesizer for categorical data. In *International Conference on Privacy in Statistical Databases* 75–91. Springer, Berlin.

[91] HU, J., REITER, J. P. and WANG, Q. (2014). Disclosure risk evaluation for fully synthetic categorical data. In *Privacy in Statistical Databases* (J. Domingo-Ferrer, ed.). *Lecture Notes in Computer Science* **8744** 185–199. Springer, Heidelberg.

[92] HU, J., REITER, J. P. and WANG, Q. (2018). Dirichlet process mixture models for modeling and generating synthetic versions of nested categorical data. *Bayesian Anal.* **13** 183–200. MR3737948 https://doi.org/10.1214/16-BA1047

[93] HU, J., SAVITSKY, T. D. and WILLIAMS, M. R. (2021). Risk-efficient Bayesian data synthesis for privacy protection. *J. Surv. Stat. Methodol.* (online-first).

[94] HU, J., SAVITSKY, T. D. and WILLIAMS, M. R. (2022). Private tabular survey data products through synthetic microdata generation. *J. Surv. Stat. Methodol.* **10** 720–752.

[95] HUNDEPOOL, A., DOMINGO-FERRER, J., FRANCONI, L., GIESSING, S., NORDHOLT, E. S., SPICER, K. and DE WOLF, P.-P. (2012). *Statistical Disclosure Control*. *Wiley Series in Survey Methodology*. Wiley, Chichester. MR3026260 https://doi.org/10.1002/9781118348239

[96] JACKSON, J., MITRA, R., FRANCIS, B. and DOVE, I. (2022). On integrating the number of synthetic data sets m into the a priori synthesis approach. In *Privacy in Statistical Databases* (J. Domingo-Ferrer and M. Laurent, eds.) 205–219. Springer, Cham.

[97] JACKSON, J., MITRA, R., FRANCIS, B. and DOVE, I. (2022). Using saturated count models for user-friendly synthesis of large confidential administrative database. *J. Roy. Statist. Soc. Ser. A* **185** 1613–1643. MR4537790 https://doi.org/10.1111/rssa.12876

[98] JANICKI, R., HOLAN, S. H., IRIMATA, K. M., LIVSEY, J. and RAIM, A. (2023). Spatial change of support models for differentially private decennial census counts of persons by detailed race and ethnicity. *J. Stat. Theory Pract.* **17** Paper No. 31, 20. MR4565882 https://doi.org/10.1007/s42519-023-00328-5

[99] KAMTHE, S., ASSEFA, S. and DEISENROTH, M. (2021). Copula flows for synthetic data generation. Available at arXiv:2101.00598 [cs, stat].

[100] KARR, A. F., KOHNEN, C. N., OGANIAN, A., REITER, J. P. and SANIL, A. P. (2006). A framework for evaluating the utility of data altered to protect confidentiality. *Amer. Statist.* **60** 224–232. MR2246755 https://doi.org/10.1198/000313006X124640

[101] KEEGAN, A. and TIDESWELL, A. (2013). Enabling learners to discover real stories in official statistics with a new synthetic unit record file of the New Zealand Income Survey 2011. Contributed paper to satellite: Statistics education for progress: Youth and official statistics.

[102] KENNICKELL, A. B. (1997). Multiple imputation and disclosure protection: The case of the 1995 survey of consumer finances. In *Record Linkage Techniques*, 1997 (W. Alvey and B. Jamerson, eds.) 248–267. National Academy Press, Washington, DC.

[103] KIFER, D. and MACHANAVAJJHALA, A. (2011). No free lunch in data privacy. In *Proceedings of the* 2011 *ACM SIGMOD International Conference on Management of Data* 193–204.

[104] KIM, H. J., DRECHSLER, J. and THOMPSON, K. J. (2021). Synthetic microdata for establishment surveys under informative sampling. *J. Roy. Statist. Soc. Ser. A* **184** 255–281. MR4204919 https://doi.org/10.1111/rssa.12622

[105] KIM, H. J., REITER, J. P. and KARR, A. F. (2018). Simultaneous edit-imputation and disclosure limitation for business establishment data. *J. Appl. Stat.* **45** 63–82. MR3736858 https://doi.org/10.1080/02664763.2016.1267123

[106] KINGMA, D. P. and WELLING, M. (2014). Auto-encoding variational bayes. Available at arXiv:1312.6114 [cs, stat].

[107] KINNEY, S. K. and REITER, J. P. (2010). Tests of multivariate hypotheses when using multiple imputation for missing data and disclosure limitation. *J. Off. Stat.* **26** 301–315.

[108] KINNEY, S. K., REITER, J. P. and MIRANDA, J. (2014). Synlbd 2.0: Improving the synthetic longitudinal business database. *Stat. J. IAOS* **30** 129–135.

[109] KINNEY, S. K., REITER, J. P., REZNEK, A. P., MIRANDA, J., JARMIN, R. S. and ABOWD, J. M. (2011). Towards unrestricted public use business microdata: The synthetic longitudinal business database. *Int. Stat. Rev.* **79** 362–384.

[110] KLEIN, M. and SINHA, B. (2015). Likelihood based finite sample inference for singly imputed synthetic data under the multivariate normal and multiple linear regression models. *J. Priv. Confid.* **7**.

[111] KOIVU, A., SAIRANEN, M., AIROLA, A. and PAHIKKALA, T. (2020). Synthetic minority oversampling of vital statistics data with generative adversarial networks. *J. Amer. Med. Inform. Assoc.* **27** 1667–1674. https://doi.org/10.1093/jamia/ocaa127

[112] LEE, J. H., KIM, I. Y. and O'KEEFE, C. M. (2013). On regression-tree-based synthetic data methods for business data. *J. Priv. Confid.* **5**.

[113] LI, H., XIONG, L. and JIANG, X. (2014). Differentially private synthesization of multi-dimensional data using Copula functions.

[114] LI, N., LI, T. and VENKATASUBRAMANIAN, S. (2007). t-closeness: Privacy beyond k-anonymity and l-diversity. In 2007 *IEEE* 23*rd International Conference on Data Engineering* 106–115.

[115] LIEW, C. K., CHOI, U. J. and LIEW, C. J. (1985). A data distortion by probability distribution. *ACM Trans. Database Syst.* **10** 395–411. MR0794552

[116] LITTLE, C., ELLIOT, M., ALLMENDINGER, R. and SAMANI, S. S. (2021). Generative adversarial networks for synthetic data generation: A comparative study. Available at arXiv:2112.01925.

[117] LITTLE, R. J. and RAGHUNATHAN, T. (1997). Should imputation of missing data condition on all observed variables. In *Proceedings of the Section on Survey Research Methods* 617–622. Amer. Statist. Assoc., Alexandria, VA.

[118] LITTLE, R. J. A. (1993). Statistical analysis of masked data. *J. Off. Stat.* **9** 407–426.

[119] LITTLE, R. J. A. and RUBIN, D. B. (1987). *Statistical Analysis with Missing Data. Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics*. Wiley, New York. MR0890519

[120] LIU, T., VIETRI, G., STEINKE, T., ULLMAN, J. and WU, S. (2021). Leveraging public data for practical private query release. In *International Conference on Machine Learning* 6968–6977. PMLR.

[121] MA, C., TSCHIATSCHEK, S., HERNÁNDEZ-LOBATO, J. M., TURNER, R. and ZHANG, C. (2020). VAEM: A deep generative model for heterogeneous mixed type data. Available at arXiv:2006.11941 [cs, stat].

[122] MACHANAVAJJHALA, A., KIFER, D., ABOWD, J. M., GEHRKE, J. and VILHUBER, L. (2008). Privacy: Theory meets practice on the map. In *IEEE* 24*th International Conference on Data Engineering* 277–286.

[123] MACHANAVAJJHALA, A., KIFER, D., GEHRKE, J. and VENKITASUBRAMANIAM, M. (2007). l-diversity: Privacy beyond k-anonymity. *ACM Trans. Knowl. Discov. Data* **1** 3–es.

[124] MAHMOOD, F., BORDERS, D., CHEN, R. J., MCKAY, G. N., SALIMIAN, K. J., BARAS, A. and DURR, N. J. (2019). Deep adversarial training for multi-organ nuclei segmentation in histopathology images. *IEEE Trans. Med. Imag.* **39** 3257–3267.

[125] MANRIQUE-VALLIER, D. and HU, J. (2018). Bayesian nonparametric generation of fully synthetic multivariate categorical data in the presence of structural zeros. *J. Roy. Statist. Soc. Ser. A* **181** 635–647. MR3807501 https://doi.org/10.1111/rssa.12352

[126] MCCLURE, D. and REITER, J. P. (2012). Differential privacy and statistical disclosure risk measures: An investigation with binary synthetic data. *Trans. Data Priv.* **5** 535–552. MR3018910

[127] MCCLURE, D. and REITER, J. P. (2016). Assessing disclosure risks for synthetic data with arbitrary intruder knowledge. *Stat. J. IAOS* **32** 109–126.

[128] MCCLURE, D. R. and REITER, J. P. (2012). Towards providing automated feedback on the quality of inferences from synthetic datasets. *J. Priv. Confid.* **4**.

[129] MCKENNA, R., MIKLAU, G. and SHELDON, D. (2021). Winning the NIST contest: A scalable and general approach to differentially private synthetic data. *J. Priv. Confid.* **11**.

[130] MCKENNA, R., SHELDON, D. and MIKLAU, G. (2019). Graphical-model based estimation and inference for differential privacy.

[131] MENG, X.-L. (1994). Multiple-imputation inferences with uncongenial sources of input (Disc: P558-573). *Statist. Sci.* **9** 538–558.

[132] MIRZA, M. and OSINDERO, S. (2014). Conditional generative adversarial nets. CoRR. Available at arXiv:1411.1784.

[133] MITRA, R., BLANCHARD, S., DOVE, I., TUDOR, C. and SPICER, K. (2020). Confidentiality challenges in releasing longitudinally linked data. *Trans. Data Priv.* **13** 151–170.

[134] MITRA, R. and REITER, J. P. (2006). Adjusting survey weights when altering identifying design variables via synthetic data. In *International Conference on Privacy in Statistical Databases* 177–188. Springer, Berlin.

[135] MOTTINI, A., LHERITIER, A. and ACUNA-AGOST, R. (2018). Airline passenger name record generation using generative adversarial networks. Available at arXiv:1807.06657 [cs, stat].

[136] NEUNHOEFFER, M., WU, Z. S. and DWORK, C. (2021). Private post-GAN boosting. Available at arXiv:2007.11934 [cs, stat].

[137] NICHOLSON CONSULTING & KŌTĀTĀ INSIGHT (2021). He Ara Poutama Mō te reo Māori Technical report.

[138] NOWOK, B., RAAB, G. M. and DIBBEN, C. (2016). Synthpop: Bespoke creation of synthetic data in R. *J. Stat. Softw.* **74** 1–26.

[139] NOWOK, B., RAAB, G. M. and DIBBEN, C. (2017). Providing bespoke synthetic data for the UK longitudinal studies and other sensitive data with the synthpop package for R. *Stat. J. IAOS* **33** 785–796.

[140] O'DONOGHUE, C. (2014). *Handbook of Microsimulation Modelling*. Emerald Group Publishing, Leeds, England.

[141] OHM, P. (2009). Broken promises of privacy: Responding to the surprising failure of anonymization. *UCLA Law Rev.* **57** 1701–1776.

[142] OSINSKI, B., JAKUBOWSKI, A., ZIECINA, P., MIŁOŚ, P., GALIAS, C., HOMOCEANU, S. and MICHALEWSKI, H. (2020). Simulation-based reinforcement learning for real-world autonomous driving. In 2020 *IEEE International Conference on Robotics and Automation* (*ICRA*) 6411–6418.

[143] PAIVA, T., CHAKRABORTY, A., REITER, J. and GELFAND, A. (2014). Imputation of confidential data sets with spatial locations using disease mapping models. *Stat. Med.* **33** 1928–1945. MR3256912 https://doi.org/10.1002/sim.6078

[144] PAPERNOT, N., SONG, S., MIRONOV, I., RAGHUNATHAN, A., TALWAR, K. and ERLINGSSON, Ú. (2018). Scalable private learning with PATE.

[145] PARK, N., MOHAMMADI, M., GORDE, K., JAJODIA, S., PARK, H. and KIM, Y. (2018). Data synthesis based on generative adversarial networks. *Proc. VLDB Endow.* **11** 1071–1083.

[146] PATKI, N., WEDGE, R. and VEERAMACHANENI, K. (2016). The synthetic data vault. In 2016 *IEEE International Conference on Data Science and Advanced Analytics* (*DSAA*) 399–410. IEEE Press, New York.

[147] PISTNER, M., SLAVKOVIĆ, A. and VILHUBER, L. (2018). Synthetic data via quantile regression for heavy-tailed and heteroskedastic data. In *International Conference on Privacy in Statistical Databases* 92–108. Springer, Berlin.

[148] PUBLICATIONS OFFICE OF THE EUROPEAN UNION (2022). data.europa.eu. Available at https://data.europa.eu/en. Last accessed on 2022-05-04.

[149] QUICK, H. (2021). Generating Poisson-distributed differentially private synthetic data. *J. Roy. Statist. Soc. Ser. A* **184** 1093–1108. MR4305573 https://doi.org/10.1111/rssa.12711

[150] QUICK, H. (2021). Improving the utility of Poisson-distributed, differentially private synthetic data via prior predictive truncation with an application to cdc wonder. *J. Surv. Stat. Methodol.* **10** 596–617. MR4305573 https://doi.org/10.1111/rssa.12711

[151] QUICK, H., HOLAN, S. H. and WIKLE, C. K. (2018). Generating partially synthetic geocoded public use data with decreased disclosure risk by using differential smoothing. *J. Roy. Statist. Soc. Ser. A* **181** 649–661. MR3807502 https://doi.org/10.1111/rssa.12360

[152] 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. *Spat. Stat.* **14** 439–451. MR3431050 https://doi.org/10.1016/j.spasta.2015.07.008

[153] RAAB, G. M., NOWOK, B. and DIBBEN, C. (2016). Practical data synthesis for large samples. *J. Priv. Confid.* **7** 67–97.

[154] RAAB, G. M., NOWOK, B. and DIBBEN, C. (2021). Assessing, visualizing and improving the utility of synthetic data. Available at arXiv:2109.12717.

[155] RAGHUNATHAN, T. E. (2021). Synthetic data. *Annu. Rev. Stat. Appl.* **8** 129–140. MR4243543 https://doi.org/10.1146/annurev-statistics-040720-031848

[156] RAGHUNATHAN, T. E., REITER, J. P. and RUBIN, D. B. (2003). Multiple imputation for statistical disclosure limitation. *J. Off. Stat.* **19** 1–16.

[157] RASHID, S., DRECHSLER, J. and MITRA, R. (2021). Accounting for longitudinal data structures when disseminating synthetic data to the public. In *UNECE Expert Meeting on Statistical Data Confidentiality* 2021.

[158] REITER, J. P. (2002). Satisfying disclosure restrictions with synthetic data sets. *J. Off. Stat.* **18** 531–544.

[159] REITER, J. P. (2004). Simultaneous use of multiple imputation for missing data and disclosure limitation. *Surv. Methodol.* **30** 235–242.

[160] REITER, J. P. (2005). Inference for partially synthetic, public use microdata sets. *Surv. Methodol.* **29** 181–189.

[161] REITER, J. P. (2005). Releasing multiply imputed, synthetic public use microdata: An illustration and empirical study. *J. Roy. Statist. Soc. Ser. A* **168** 185–205. MR2113234 https://doi.org/10.1111/j.1467-985X.2004.00343.x

[162] REITER, J. P. (2005). Significance tests for multi-component estimands from multiply imputed, synthetic microdata. *J. Statist. Plann. Inference* **131** 365–377. MR2139378 https://doi.org/10.1016/j.jspi.2004.02.003

[163] REITER, J. P. (2005). Using CART to generate partially synthetic, public use microdata. *J. Off. Stat.* **21** 441–462.

[164] REITER, J. P. (2005). Estimating risks of identification disclosure in microdata. *J. Amer. Statist. Assoc.* **100** 1103–1112. MR2236926 https://doi.org/10.1198/016214505000000619

[165] REITER, J. P. and DRECHSLER, J. (2010). Releasing multiply-imputed synthetic data generated in two stages to protect confidentiality. *Statist. Sinica* **20** 405–421. MR2640701

[166] REITER, J. P. and KINNEY, S. K. (2012). Inferentially valid, partially synthetic data: Generating from posterior predictive distributions not necessary. *J. Off. Stat.* **28** 583–590.

[167] REITER, J. P. and MITRA, R. (2009). Estimating risks of identification disclosure in partially synthetic data. *J. Priv. Confid.* **1** 99–110.

[168] REITER, J. P., OGANIAN, A. and KARR, A. F. (2009). Verification servers: Enabling analysts to assess the quality of inferences from public use data. *Comput. Statist. Data Anal.* **53** 1475–1482. MR2657106 https://doi.org/10.1016/j.csda.2008.10.006

[169] REITER, J. P. and RAGHUNATHAN, T. E. (2007). The multiple adaptations of multiple imputation. *J. Amer. Statist. Assoc.* **102** 1462–1471. MR2372542 https://doi.org/10.1198/016214507000000932

[170] REITER, J. P., WANG, Q. and ZHANG, B. (2014). Bayesian estimation of disclosure risks for multiply imputed, synthetic data. *J. Priv. Confid.* **6**.

[171] ROCHER, L., HENDRICKX, J. M. and DE MONTJOYE, Y.-A. (2019). Estimating the success of re-identifications in incomplete datasets using generative models. *Nat. Commun.* **10**. https://doi.org/10.1038/s41467-019-10933-3

[172] ROSENBAUM, P. R. and RUBIN, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika* **70** 41–55. MR0742974 https://doi.org/10.1093/biomet/70.1.41

[173] RUBIN, D. B. (1978). Multiple imputations in sample surveys-a phenomenological Bayesian approach to nonresponse. In *Proceedings of the Survey Research Methods Section of the American Statistical Association* **1** 20–34 Amer. Statist. Assoc., Alexandria, VA, USA.

[174] RUBIN, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. *Wiley Series in Probability and Mathematical Statistics*: *Applied Probability and Statistics*. Wiley, New York. MR0899519 https://doi.org/10.1002/9780470316696

[175] RUBIN, D. B. (1993). Discussion: Statistical disclosure limitation. *J. Off. Stat.* **9** 462–468.

[176] SAKSHAUG, J. W. and RAGHUNATHAN, T. E. (2010). Synthetic data for small area estimation. In *Privacy in Statistical Databases* (J. Domingo-Ferrer and E. Magkos, eds.) 162–173. Springer, Heidelberg.

[177] SAKSHAUG, J. W. and RAGHUNATHAN, T. E. (2014). Generating synthetic data to produce public-use microdata for small geographic areas based on complex sample survey data with application to the National Health Interview Survey. *J. Appl. Stat.* **41** 2103–2122. MR3292662 https://doi.org/10.1080/02664763.2014.909778

[178] SALLIER, K. (2020). Toward more user-centric data access solutions: Producing synthetic data of high analytical value by data synthesis. *Stat. J. IAOS* **36** 1059–1066.

[179] SHLOMO, N. (2014). Probabilistic record linkage for disclosure risk assessment. In *International Conference on Privacy in Statistical Databases* 269–282. Springer, Berlin.

[180] SIWICKI, B. (2021). Synthetic data boosts accuracy and speed of brain tumor surgery CDS. Available at https://www.healthcareitnews.com/news/synthetic-data-boosts-accuracy-and-speed-brain-tumor-surgery-cds. Last accessed on 2022-05-04.

[181] SKINNER, C. and SHLOMO, N. (2008). Assessing identification risk in survey microdata using log-linear models. *J. Amer. Statist. Assoc.* **103** 989–1001. MR2462887 https://doi.org/10.1198/016214507000001328

[182] SNOKE, J., RAAB, G. M., NOWOK, B., DIBBEN, C. and SLAVKOVIC, A. (2018). General and specific utility measures for synthetic data. *J. Roy. Statist. Soc. Ser. A* **181** 663–688. MR3807503 https://doi.org/10.1111/rssa.12358

[183] SRIVASTAVA, A., VALKOV, L., RUSSELL, C., GUTMANN, M. U. and SUTTON, C. (2017). VEEGAN: Reducing mode collapse in GANs using implicit variational learning.

[184] STADLER, T., OPRISANU, B. and TRONCOSO, C. (2021). Synthetic data—anonymisation groundhog day. Available at arXiv:2011.07018.

[185] SWEENEY, L. (2002). *k*-anonymity: A model for protecting privacy. *Internat. J. Uncertain. Fuzziness Knowledge-Based Systems* **10**. Aggregation and security assessment for inference control in statistical databases. MR1948199 https://doi.org/10.1142/S0218488502001648

[186] SWEENEY, L. (2013). Matching known patients to health records in Washington state data. Available at arXiv:1307.1370.

[187] TAUB, J. and ELLIOT, M. (2019). The synthetic data challenge. Joint UNECE/Eurostat work session on statistical data confidentiality, The Hague, The Netherlands.

[188] THOMPSON, K. and KIM, H. J. (2022). Incorporating economic conditions in synthetic microdata for business programs. *J. Surv. Stat. Methodol.* **10** 830–859.

[189] THOMPSON, S. A. and WARZEL, C. (2019). Twelve million phones, one dataset, zero privacy. Available at https://www.nytimes.com/interactive/2019/12/19/opinion/location-tracking-cell-phone.html. Last accessed on 2023-06-20.

[190] TORFI, A. (2020). Privacy-preserving synthetic medical data generation with deep learning. Virginia Tech.

[191] TORFI, A. and FOX, E. A. (2020). COR-GAN: Correlation-capturing convolutional neural networks for generating synthetic healthcare records. CoRR. Available at arXiv:2001.09346.

[192] TORKZADEHMAHANI, R., KAIROUZ, P. and PATEN, B. (2020). DP-CGAN: Differentially private synthetic data and label generation. Available at arXiv:2001.09700 [cs, stat].

[193] U. S. GENERAL SERVICES ADMINISTRATION (2022). Data.gov. Available at https://data.gov/. Last accessed on 2022-05-04.

[194] VADHAN, S. (2017). The complexity of differential privacy. In *Tutorials on the Foundations of Cryptography*. *Inf. Secur. Cryptography* 347–450. Springer, Cham. MR3837668

[195] VARDHAN, L. V. H. and KOK, S. (2020). Generating privacy-preserving synthetic tabular data using oblivious variational autoencoders. In *Proceedings of the Workshop on Economics of Privacy and Data Labor at the* 37*th International Conference on Machine Learning*.

[196] VOAS, D. and WILLIAMSON, P. (2001). Evaluating goodness-of-fit measures for synthetic microdata. *Geogr. Environ. Model.* **5** 177–200.

[197] WAHEED, A., GOYAL, M., GUPTA, D., KHANNA, A., AL-TURJMAN, F. and PINHEIRO, P. R. (2020). CovidGAN: Data augmentation using auxiliary classifier GAN for improved Covid-19 detection. *IEEE Access* **8** 91916–91923. https://doi.org/10.1109/ACCESS.2020.2994762

[198] WANG, H. and REITER, J. P. (2012). Multiple imputation for sharing precise geographies in public use data. *Ann. Appl. Stat.* **6** 229–252. MR2951536 https://doi.org/10.1214/11-AOAS506

[199] WEI, L. and REITER, J. P. (2016). Releasing synthetic magnitude microdata constrained to fixed marginal totals. *Stat. J. IAOS* **32** 93–108.

[200] WEN, B., COLON, L. O., SUBBALAKSHMI, K. P. and CHANDRAMOULI, R. (2021). Causal-TGAN: Generating tabular data using causal generative adversarial networks.

[201] WIESE, M., KNOBLOCH, R., KORN, R. and KRETSCHMER, P. (2020). Quant GANs: Deep generation of financial time series. *Quant. Finance* **20** 1419–1440. MR4149599 https://doi.org/10.1080/14697688.2020.1730426

[202] WOO, M. J., REITER, J. P., OGANIAN, A. and KARR, A. F. (2009). Global measures of data utility for microdata masked for disclosure limitation. *J. Priv. Confid.* **1** 111–124.

[203] XIAO, X., WANG, G. and GEHRKE, J. (2011). Differential privacy via wavelet transforms. *IEEE Trans. Knowl. Data Eng.* **23** 1200–1214.

[204] XIE, L., LIN, K., WANG, S., WANG, F. and ZHOU, J. (2018). Differentially private generative adversarial network. Available at arXiv:1802.06739 [cs, stat].

[205] XU, L., SKOULARIDOU, M., CUESTA-INFANTE, A. and VEERAMACHANENI, K. (2019). Modeling tabular data using conditional GAN. In *Advances in Neural Information Processing Systems* (H. Wallach, H. Larochelle, A. Beygelzimer, F. D. Alché-Buc, E. Fox and R. Garnett, eds.). **32**. Curran Associates, Red Hook.

[206] YAHI, A., VANGURI, R., ELHADAD, N. and TATONETTI, N. P. (2017). Generative adversarial networks for electronic health records: A framework for exploring and evaluating methods for predicting drug-induced laboratory test trajectories. Available at arXiv:1712.00164.

[207] YOON, J., JORDON, J. and SCHAAR, M. V. D. (2019). PATE-GAN: Generating synthetic data with differential privacy guarantees. In *International Conference on Learning Representations*.

[208] YU, H. and REITER, J. P. (2018). Differentially private verification of regression predictions from synthetic data. *Trans. Data Priv.* **11** 279–297.

[209] ZHANG, J., CORMODE, G., PROCOPIUC, C. M., SRIVASTAVA, D. and XIAO, X. (2014). PrivBayes: Private data release via Bayesian networks. In *Proceedings of the* 2014 *ACM SIGMOD International Conference on Management of Data*. 1423–1434.

[210] ZHANG, J., CORMODE, G., PROCOPIUC, C. M., SRIVASTAVA, D. and XIAO, X. (2017). PrivBayes: Private data release via Bayesian networks. *ACM Trans. Database Syst.* **42** Art. 25, 41. MR3730676 https://doi.org/10.1145/3134428

[211] ZHAO, Z., KUNAR, A., VAN DER SCHEER, H., BIRKE, R. and CHEN, L. Y. (2021). CTAB-GAN: Effective table data synthesizing. Available at arXiv:2102.08369 [cs].

[212] ZHOU, H., ELLIOTT, M. R. and RAGHUNATHAN, T. E. (2016). Synthetic multiple-imputation procedure for multistage complex samples. *J. Off. Stat.* **32** 231–256. https://doi.org/10.1515/JOS-2016-0011

[213] (2017). Learning with privacy at scale. *Apple Mach. Learn. J.* **1** 8.

[214] (2021). Exposure notification privacy-preserving analytics. White paper, available at https://covid19-static.cdn-apple.com/applications/covid19/current/static/contact-tracing/pdf/ENPA_White_Paper.pdf. Last accessed on 2023-06-21.