# 30 years of synthetic data

Jörg Drechsler[1,2] and Anna-Carolina Haensch[2,3]

[1]Institute for Employment Research, Germany
[2]The Joint Program in Survey Methodology, University of Maryland, College Park, USA
[3]Ludwig-Maximilians-Universität, Munich, Germany

**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 upcoming 30th jubilee of Rubin's seminal paper on synthetic data (Rubin, 1993) 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.

## 1 Introduction

We live in a data-driven world today. Data are collected whenever we interact with our environment, 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, for example, Publications Office of the European Union (2022) or U.S. General Services Administration (2022)), and the fact that increasingly more countries explicitly define their own data strategies (see, for example, European Commission (2022) or Department for Digital, Culture, Media & Sport (2022) 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 misuse of data that have not been sufficiently protected on the other hand. 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 (Domingo-Ferrer et al., 2012; Duncan et al., 2011). Still, several prominent examples of confidentiality breaches both in the public and in the private sector (Homer et al., 2008; Ohm, 2009; Sweeney, 2013; De Montjoye et al., 2015) 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 and draws from this model are 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 Rubin (1993) and Little (1993) (although related ideas have been proposed earlier by Liew et al. (1985)). Their approach to synthetic data was motivated by their own work on multiple imputation (MI) for nonresponse (Rubin, 1978; Little and Rubin, 1987). 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 to obtain 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 only 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. 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 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 (Choi et al., 2018) over synthetic electronic health records (Yahi et al., 2017) to generating synthetic cell images (Siwicki, 2021). 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 (Osinski et al., 2020), classifying computed tomography images (Frid-Adar et al., 2018) or environmental monitoring (Allken et al., 2018).

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, micro simulation (O'Donoghue, 2014), 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 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. We offer different approaches how to classify various synthesis strategies 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

The idea of releasing synthetic data instead of the real data was first proposed by Rubin (1993). In a discussion of another article, he suggested that his multiple imputation framework (Rubin, 1978, 1987) 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 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 Little (1993). 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.

A related idea for generating synthetic data was proposed by Fienberg (1994). He suggested using a smoothed estimate of the empirical cumulative function of the data and releasing bootstrap samples from this distribution. This approach was further developed in Fienberg et al. (1998).

Ten years after the initial proposal by Rubin and Little, Raghunathan et al. (2003) and Reiter (2003) 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 identified another difference between the two synthesis approaches: posterior draws of the model parameters which are necessary for full synthesis (and also in the con-

text of multiple imputation for nonresponse) are not required for partial synthesis. Several years later, Raab et al. (2016) 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 (Reiter, 2005b; An and Little, 2007) and applications (Kennickell, 1997; Abowd et al., 2006) 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.

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 (Kennickell, 1997). Abowd and Woodcock (2001, 2004) 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 (Abowd et al., 2006), 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 (Benedetto et al., 2018). Another early application was OntheMap, a graphical interface that allows visualizing detailed commuting patterns for the entire United States (Machanavajjhala et al., 2008). This application was the first to offer formal privacy guarantees based on a concept called $(\varepsilon, \delta)$-probabilistic differential privacy, a relaxation of the original definition of differential privacy proposed by Dwork et al. (2006). Three years later, the U.S. Census Bureau released the Synthetic Longitudinal Business Database (Kinney et al., 2011, 2014), 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) (Hawala, 2008). 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 (Bonnéry et al., 2019; Goldstein et al., 2020).

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 (Forbes and Zealand, 2008; Keegan and Tideswell, 2013). 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 vari-

ous scenarios and policy interventions over the period from 2013 to 2040 (Nicholson Consulting & Kōtātā Insight, 2021). The German Institute for Employment Research released a partially synthetic version of one wave of its Establishment Panel in 2011 (Drechsler, 2011b). The approach was also adopted to facilitate access to the Scottish Longitudinal Study (Nowok et al., 2017). 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* (Nowok et al., 2016), 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) (de Wolf, 2015). These data, which are available for download at the Eurostat website (Eurostat, 2022), 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 (Sallier, 2020).

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 (Bowen et al., 2020, 2022) and the Australian Bureau of Statistics that is currently evaluating synthetic data as a means of broadening access to its microdata (Australian Bureau of Statistics, 2021).

Further practical applications have been discussed in the context of protecting data containing detailed geographical information (Wang and Reiter, 2012; Paiva et al., 2014; Quick et al., 2015, 2018; Drechsler and Hu, 2021), preserving and protecting longitudinal data structures (Rashid et al., 2021; Mitra et al., 2020), small area estimation (Sakshaug and Raghunathan, 2010, 2014), synthesizing business data (Lee et al., 2013; Drechsler and Vilhuber, 2014b,a; Alam et al., 2020; Chien et al., 2020; Thompson and Kim, 2022) dealing with nested data structures (Hu et al., 2018) or accounting for complex survey designs (Mitra and Reiter, 2006; Dong et al., 2014b,a; Zhou et al., 2016; Kim et al., 2021; Hu et al., 2022). Hu et al. (2021) 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 Wei and Reiter (2016) developed a synthesis strategy, which 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 Liew et al. (1985) proposed a synthetic data approach for disclosure protection (even though

they never called it that way) several years before Rubin's seminal paper. We postulate that this can be attributed (at least in part) to the fact that privacy standards play an important role in the literature on data privacy and the privacy standards used before the advent of differential privacy (DP, Dwork et al. (2006)) only focused on the properties of the data at hand. Popular standards such as $k$-anonymity (Sweeney, 2002), $l$-diversity (Machanavajjhala et al., 2007), or $t$-closeness (Li et al., 2007) all establish certain requirements regarding the properties of the data to consider the data safe from (certain types of) privacy attacks. Synthetic data do not really fit into this notion of privacy. For example, a fully synthetic dataset might not fulfill $k$-anonymity for any $k > 1$, but intuitively still offers a high level of privacy protection.

Differential privacy 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. In a nutshell, differential privacy requires that changing one record in the database has a strictly limited impact on the results of a mechanism run on the data (we will offer a more detailed review of differential privacy in Section 4.5). Note the 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 differential privacy. Soon after the concept of DP was established in 2006, the first papers on differentially private synthetic data started to appear.

One of the first approaches was developed by Barak et al. (2007) who generated synthetic data using a Fourier transformation and linear programming for low-order contingency tables (Dwork, 2008). Other early applications include Eno and Thompson (2008); Cano et al. (2010); Blum et al. (2011); Xiao et al. (2011). Several papers also explicitly adapted the ideas from the statistical community to the DP context (Abowd and Vilhuber, 2008; Machanavajjhala et al., 2008; Charest, 2011; McClure and Reiter, 2012a). The approach of Machanavajjhala et al. (2008) was later extended in Quick (2021) and Quick (2022). Zhang et al. (2014, 2017) proposed an approach that uses Bayesian networks to synthesize high-dimensional datasets, called PrivBayes. In parallel, Li et al. (2014) employed Copula functions to take into account the dependency structure of the data (DPCopula). DP guarantees were also integrated in Generative Adversarial Networks (GANs) later (Xie et al., 2018; Yoon et al., 2019).

The advent of GANs proposed by Goodfellow et al. (2014) 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, for example, Denton et al. (2015)). 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 this 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 variable are often weaker than, for example, relationships between pixels that are located next to each other. The position of variables 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 (Esteban et al., 2017; Yahi et al., 2017) or count and binary data (Choi et al., 2018, medGAN). tableGAN (Park et al., 2018) 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 (Camino et al., 2018; Baowaly et al., 2019). Later applications relied on Wasserstein-GANs (WGANs) (Koivu et al., 2020; Zhao et al., 2021). In recent years, more focus has been put on modeling relationships between variables. Conditional tabular GAN (CTGAN) developed by Xu et al. (2019) addresses challenges from imbalanced categorical and multi-modal continuous data. Causal Tabular GAN (Wen et al., 2021, Causal TGAN) allows for modeling the causal relationships between variables in datasets.

Beyond approaches based on GANs other synthesis strategies based on (Variational) Autoencoders (Camino et al., 2018; Vardhan and Kok, 2020; Ma et al., 2020), Bayesian Networks (Zhang et al., 2017), copulas (Patki et al., 2016; Kamthe et al., 2021), or approaches that explicitly preserve certain marginal distributions (McKenna et al., 2019, 2021) have also been developed in recent years. For a short taxonomy of approaches, see Section 4.

The earliest deployment of a differentially private synthesis strategy is OntheMap (Machanavajjhala et al., 2008) already mentioned in the previous section. The enforcement of differential privacy for the Decennial Census 2020 can also be seen as a synthetic data approach, as the noisy table counts produced by the TopDown algorithm (Abowd et al., 2019) are turned into synthetic microdata from which the actually released tables are generated.

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 (Challenge.gov, 2019). The winning teams relied on Bayesian networks or approaches to preserve pre-specified marginal distributions. See Bowen and Snoke (2019) 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 synthetic health records by Esteban et al. (2017); Torfi (2020) and Torfi and Fox (2020). Chen et al. (2019) 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 (Ghorbani et al., 2020), pathology slides (Mahmood et al., 2019), and chest X-rays (Waheed et al., 2020).

# 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 Raghunathan et al. (2003), those for partially synthetic data are presented in Reiter (2003). Later, Reiter (2005c) 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 Drechsler (2011c). The interested reader is referred to Reiter and Raghunathan (2007), 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 $i = 1, ..., m$ be the point and variance estimates for each of the $m$ synthetic datasets. The following quantities are needed

for inferences for scalar $Q$:

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

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

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

## 3.1 Combining rules for fully synthetic data

The analyst can use $\bar{q}_m$ as an unbiased point estimate for $Q$. Its variance can be estimated using

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

When $n$ is large, inferences for scalar $Q$ can be based on $t$ distributions with degrees of freedom $\nu_f = (m-1)(1 - \bar{u}_m/((1+m^{-1})b_m))^2$. A disadvantage of this variance estimate is that it can become negative. For that reason, Reiter (2002) suggested a slightly modified variance estimator that is always positive but will tend to overestimate the true variance, $T_f^* = \max(0, T_f) + \delta(\frac{n_{syn}}{n}\bar{u}_m)$, 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.

## 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

$$T_p = b_m/m + \bar{u}_m. \quad (5)$$

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.

## 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 Rubin (1993). 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 $f(Y|X^{new})$. 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 Drechsler (2011a). Extending these ideas, Raab et al. (2016) proposed an alternative variance estimator that can be used in this situation:

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

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), (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 (Drechsler and Reiter, 2009; Reiter and Drechsler, 2010). Of course, the price to pay is an increased level of uncertainty, if only one synthetic dataset is released. Note that assuming $n_{syn} = n_{org}$, 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 relative reduction in variance is bounded by 0.5 for $m \to \infty$. See Drechsler (2018) for further discussion of the advantages and disadvantages of the different synthesis strategies and which variance estimator is appropriate in which scenario.

# 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 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 re-

quire 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 (Reiter, 2005b), or models based on quantile regressions (Pistner et al., 2018), but also any machine learning tool that enables random draws from a conditional distribution, such as CART or random forests.

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 (Burridge, 2003) relied on a multivariate normality assumption that is seldom justified with real data, more flexible models have been proposed recently. For categorical data, Hu et al. (2014) 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 (Manrique-Vallier and Hu, 2018). Synthesis approaches based on DPMPMs are implemented in the R package NPBayesImputeCat (Hu et al., 2021). A related approach based on Quasi-Mutinomial distributions was proposed by Hu and Hoshino (2018), while Jackson et al. (2022a) and Jackson et al. (2022b) proposed saturated count models for easy synthesis of large databases with a-prori utility guarantees. Kim et al. (2018) 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 (Kim et al., 2021). Furthermore, many of the synthesis strategies used in computer science such as Generative Adversarial Networks(Goodfellow et al., 2014) or Bayesian Networks (Zhang et al., 2017) can be subsumed under this category. We will review the literature based on these approaches in Section 4.4.

## 4.2 Parametric vs. ML 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 (Meng, 1994) 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 (Meng, 1994) 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 Little and Raghunathan (1997) 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, Reiter (2005d) suggested using Classification and Regression Trees (CART). Caiola and Reiter (2010) later extended these ideas to random forests, and Drechsler (2010) developed strategies to adapt Support Vector Machines for data synthesis. Synthesis strategies based on genetic algorithms were explored in Chen et al. (2016) and Chen et al. (2018). All these approaches have the advantage that they "let the data speak for themselves", 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, Drechsler and Reiter (2011) compared the different approaches and found that CART models offered the best results in terms of preserving the information from the original data.

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 Reiter (2004). 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 mutli-component hypothesis testing under this scenario were derived in Kinney and Reiter (2010).

In a similar spirit, Reiter and Drechsler (2010) 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 (Drechsler and Reiter, 2009) 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. Drechsler and Reiter (2010) 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 the information from the full population. Drechsler and Reiter (2012) extend the methodology to the context where the original data is itself already a sample.

## 4.4 Computer Science approaches

In computer science, machine learning and deep learning methods such as Generative Adversarial Networks (Goodfellow et al., 2014, GANs) and Variational Autoencoders (Kingma and Welling, 2014, 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 towards 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 synthesis of microdata is relatively sparse but rapidly growing (Park et al., 2018; Choi et al., 2018; Camino et al., 2018; Koivu et al., 2020; Xu et al., 2019).

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 Goodfellow et al. (2014) 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) (Arjovsky et al., 2017) 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 (Arjovsky et al., 2017). 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 Camino et al. (2018).

However, for WGANs, the discriminator usually has to obey a Lipschitz constraint. To enforce this constraint, the weights of the discriminator must be within a certain range controlled by a hyperparameter. Arjovsky et al. (2017) propose to clip the weights if necessary, but noted that this approach is not optimal. To overcome this problem, WGAN-gradient penalty (WGAN-GP) Gulrajani et al. (2017) uses a gradient penalty to fulfill the Lipschitz constraint. Baowaly et al. (2019),(Koivu et al., 2020, actGAN), Xu et al. (2019) and (Zhao et al., 2021, CTAB-GAN) all use different adaptations of WGAN or WGAN-GP for data synthesis. Xu et al. (2019) 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 (Mottini et al., 2018).

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 (Wen et al., 2021).

### 4.4.2 Variational Autoencoders (VAEs)

Another approach based on deep neural networks that has been adapted for data synthesis lately are variational autoencoders (Kingma and Welling, 2014, VAE). In comparison to GANs, a VAE has three instead of two networks, which learn complimentary 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. 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. (2017, VEEGAN) Camino et al. (2018); Xu et al. (2019); Vardhan and Kok (2020), and Ma et al. (2020).

## 4.5 Differential private data synthesis

In recent years, differential privacy (Dwork et al., 2006, DP) 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, pure $\varepsilon$-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 $\epsilon$ and $-\epsilon$. In layman's term, 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 Dwork and Roth (2014) or Vadhan (2017) for an in-depth discussion of differential privacy 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 (Apple's Differential Privacy Team, 2017), Google (Erlingsson et al., 2014), and Microsoft (Ding et al., 2017) as well as the U.S. Census Bureau (Foote et al., 2019; Abowd et al., 2022) recently adopted the approach for some of their data.

The concept of DP has also stimulated research on generating differentially private synthetic data. Differentially private synthetic data have an advantage over other approaches for private data analysis: 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. 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 Bowen and Liu (2020) 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 (McKenna et al., 2019, 2021; Liu et al., 2021). Another popular approach for differentially private data synthesis are Bayesian networks (Bao et al., 2021), most prominently PrivBayes by Zhang et al. (2017). It can be difficult to represent all important correlations in PrivBayes. Therefore, Cai et al. (2021) 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 et al. (2012, MWEM) and Gaboardi et al. (2014, 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 et al. (2014, DPCopula) is based on Copula functions.

Finally, work on integrating differential privacy into generative adversarial networks (GANs) has been growing fast in the last few years (Beaulieu-Jones et al., 2019; Xie et al., 2018; Yoon et al., 2019; Torkzadehmahani et al., 2020; Neunhoeffer et al., 2021). Since the generator commonly never accesses the real data directly, only the discriminator needs to be modified to ensure DP: Beaulieu-Jones et al. (2019) and Xie et al. (2018) built on Abadi et al. (2016) 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. Frigerio et al. (2019) also proposes a private extension of WGAN. Conditional GANs (Mirza and Osindero, 2014, CGAN) are adapted by Torkzadehmahani et al. (2020). Yoon et al. (2019) use the Private Aggregation of Teacher Ensembles (PATE) framework proposed by Papernot et al. (2018), 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, Neunhoeffer et al. (2021) proposed a method combining weighted samples produced by a sequence of generators. Their approach can be applied to different private or non-private 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 Domingo-Ferrer et al. (2012) or Arnold and Neunhoeffer (2020).

Utility metrics can be broadly divided into three categories: The first category, commonly referred to as *global utility metrics*, 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* 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. 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 (Karr et al., 2006) or Hellinger distance (Gomatam and Karr, 2003). 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 (Rosenbaum and Rubin, 1983). 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. Bowen et al. (2021) 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 (Woo et al., 2009; Snoke et al., 2018). Let $p_i$, $i = 1, \ldots, N$ with $N = n_{org} + n_{syn}$ 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 $c = n_{syn}/N$. 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 Woo et al. (2009) is that it increases with the number of predictors included in the propensity model. To overcome this problem, Snoke et al. (2018) 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, Drechsler (2022) critically discussed the $pMSE$ illustrating that the estimated scores are highly dependent on the specification of the propensity model and that even baltant 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.

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 Karr et al. (2006). 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.

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 statistic. 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 $\chi^2$ statistic or the likelihood ratio statistic. Voas and Williamson (2001) 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 bivariate distributions is presented in Raab et al. (2021). These graphical visualization tools are also implemented in the R package synthpop (Raab et al., 2016).

In Raab et al. (2021), the authors empirically evaluate various goodness-of-fit

measures and find a large correlation ($> 0.9$) between most of them. Noticeably, the adjusted $\chi^2$ test proposed by Voas and Williamson (2001), 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 (Reiter, 2005a; Skinner and Shlomo, 2008; Shlomo, 2014), 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, Manrique-Vallier and Hu (2018) 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 (2019), 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 (Park et al., 2018; Zhao et al., 2021) 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 Taub and Elliot (2019). 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

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

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 $WEAP_j = 1$. For those records, they compute the Targeted Correct Attribution Probability (TCAP):

$$TCAP_{sj} = Pr(T_{sj}|K_{sj})_o = \frac{\sum_{i=1}^{n} I(T_{o,i} = T_{s,j}, K_{o,i} = K_{s,j})}{\sum_{i=1}^{n} I(K_{o,i} = K_{s,j})},$$

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 (Stadler et al., 2021). 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 differential privacy literature, Reiter et al. (2014) propose strategies to compute the posterior distribution $f(Y_i|D, X, M, d_{org}^{-i})$, 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 Hu et al. (2014) and McClure and Reiter (2016) 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.

## 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 re-identify some records in the synthetic data. Building on previous work in Reiter (2005a), Reiter and Mitra (2009) developed strategies to measure the risk of re-identification for partially synthetic data.

Borrowing from Drechsler and Reiter (2010), 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 $\mathbf{d}_{syn}$, where $\mathbf{d}_{syn}$ denotes the synthetic data and $i = 1, \ldots, n$. Let $\mathcal{S}$ be any information released about the synthesis models.

The intruder's goal is to match unit $i$ in $\mathbf{d}_{syn}$ to the target when $P_{i0} = t_0$. Let $J$ be a random variable that equals $i$ when $P_{i0} = t_0$ for $i \in \mathbf{d}_{syn}$. The intruder thus seeks to calculate $Pr(J = i|\mathbf{t}, \mathbf{d}_{syn}, \mathcal{S})$ for $i = 1, \ldots, n$. 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

$$Pr(J = i|\mathbf{t}, \mathbf{d}_{syn}, \mathcal{S}) = \mathbf{d}_{syn}, Y^*, \mathcal{S})Pr(Y^*|\mathbf{t}, \mathbf{d}_{syn}, \mathcal{S})dY^*.$$

This construction suggests a Monte Carlo approach to estimating each $Pr(J = i|\mathbf{t}, \mathbf{d}_{syn}, \mathcal{S})$. First, sample a value of $Y^*$ from $Pr(Y^*|\mathbf{t}, \mathbf{d}_{syn}, \mathcal{S})$. Let $Y_{new}$ represent one set of simulated values. Second, compute $Pr(J = i|\mathbf{t}, \mathbf{d}_{syn}, Y^* = Y_{new}, \mathcal{S})$ 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 (6.2) is estimated as the average of the resultant $h$ values of $Pr(J = i|\mathbf{t}, \mathbf{d}_{syn}, Y^* = Y_{new}, \mathcal{S})$. 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 $Pr(J = i|\mathbf{t}, \mathbf{d}_{syn}, \mathcal{S})$, if a unique maximum exists. Reiter and Mitra (2009) 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 (Hornby and Hu, 2021).

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 Drechsler and Reiter (2008).

# 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 (Raab et al., 2016) or the synthetic data vault (Patki et al., 2016) 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 (for example, Machanavajjhala et al. (2008) 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 (Forbes and Zealand, 2008) or to develop software code in preparation for working with the real data (de Wolf, 2015; Raab et al., 2021). 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 (Benedetto et al., 2018; Burman et al., 2019). 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 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, verification servers have the advantage that the procedure can be automated. 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 sen-

sitive information is an area of active research (Reiter et al., 2009; McClure and Reiter, 2012b; Barrientos et al., 2018; Yu and Reiter, 2018).

A systematic comparison between the approaches developed in the different fields is currently lacking, although Little et al. (2021) offers some first insights. The authors compared several synthesis strategies based on CART models, Bayesian Networks, and two GAN implementations (tableGAN and CTGAN). They found that the sequential-regression-based CART approach offered the highest utility, but also the highest risk. One of the GANs (tableGAN) resulted in unacceptably low utility, while CTGAN and the approach based on Bayesian Networks performed almost similarly. However, this evaluation was based on only one dataset and also relied on the default settings of the different software packages that were used for the different synthesis approaches. 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.

# Acknowledgements

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). The authors are grateful for helpful feedback on an earlier version of this paper from the fk2 reading group at Mannheim Unversity and LMU Munich.

# References

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

Abowd, J., R. Ashmead, G. Simson, D. Kifer, P. Leclerc, A. Machanavajjhala, and W. Sexton (2019). Census topdown: Differentially private data, incremental schemas, and consistency with public knowledge. *US Census Bureau*.

Abowd, J. M., R. Ashmead, R. Cumings-Menon, S. Garfinkel, M. Heineck, C. Heiss, R. Johns, D. Kifer, P. Leclerc, A. Machanavajjhala, et al. (2022). The 2020 census disclosure avoidance system topdown algorithm. *arXiv:2204.08986*.

Abowd, J. M., M. Stinson, and G. Benedetto (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.

Abowd, J. M. and L. Vilhuber (2008). How protective are synthetic data? In J. Domingo-Ferrer and Y. Saygın (Eds.), *Privacy in statistical databases*, Volume 5262, pp. 239–246. Berlin, Heidelberg: Springer Berlin Heidelberg.

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

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

Alam, M. J., B. Dostie, J. Drechsler, and L. Vilhuber (2020). Applying data synthesis for longitudinal business data across three countries. *Statistics in Transition New Series 21*(4).

Allken, V., N. O. Handegard, S. Rosen, T. Schreyeck, T. Mahiout, and K. Malde (2018). Fish species identification using a convolutional neural network trained on synthetic data. *ICES Journal of Marine Science 76*(1), 342–349.

An, D. and R. J. A. Little (2007). Multiple imputation: An alternative to top coding for statistical disclosure control. *Journal of the Royal Statistical Society, Series A 170*, 923–940.

Apple's Differential Privacy Team (2017). Learning with privacy at scale. Apple Machine Learning Journal **1** 8.

Arjovsky, M., S. Chintala, and L. Bottou (2017). Wasserstein GAN. *arXiv:1701.07875 [stat.ML]*.

Arnold, C. and M. Neunhoeffer (2020). Really useful synthetic data–a framework to evaluate the quality of differentially private synthetic data. *arXiv:2004.07740*.

Australian Bureau of Statistics (2021). Methodological news, dec 2021. `https://www.abs.gov.au/statistics/research/methodological-news-dec-2021`. Last accessed on 2022-05-17.

Bao, E., X. Xiao, J. Zhao, D. Zhang, and B. Ding (2021). Synthetic data generation with differential privacy via bayesian networks. *Journal of Privacy and Confidentiality 11*(3).

Baowaly, M. K., C.-C. Lin, C.-L. Liu, and K.-T. Chen (2019). Synthesizing electronic health records using improved generative adversarial networks. *Journal of the American Medical Informatics Association 26*(3), 228–241.

Barak, B., K. Chaudhuri, C. Dwork, S. Kale, F. McSherry, and K. Talwar (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*, Beijing, China, pp. 273–282. ACM Press.

Barrientos, A. F., A. Bolton, T. Balmat, J. P. Reiter, J. M. de Figueiredo, A. Machanavajjhala, Y. Chen, C. Kneifel, and M. DeLong (2018). Providing access to confidential research data through synthesis and verification: An application to data on employees of the us federal government. *The Annals of Applied Statistics 12*(2), 1124–1156.

Beaulieu-Jones, B. K., Z. S. Wu, C. Williams, R. Lee, S. P. Bhavnani, J. B. Byrd, and C. S. Greene (2019). Privacy-preserving generative deep neural networks support clinical data sharing. *Circulation: Cardiovascular Quality and Outcomes 12*(7), e005122.

Benedetto, G., J. C. Stanley, E. Totty, et al. (2018). The creation and use of the sipp synthetic beta version 7.0.

Blum, A., K. Ligett, and A. Roth (2011). A Learning Theory Approach to Non-Interactive Database Privacy.

Bonnéry, D., Y. Feng, A. K. Henneberger, T. L. Johnson, M. Lachowicz, B. A. Rose, T. Shaw, L. M. Stapleton, M. E. Woolley, and Y. Zheng (2019). The promise and limitations of synthetic data as a strategy to expand access to state-level multi-agency longitudinal data. *Journal of Research on Educational Effectiveness 12*(4), 616–647.

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

Bowen, C. M., V. Bryant, L. Burman, S. Khitatrakun, R. McClelland, P. Stallworth, K. Ueyama, and A. R. Williams (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*, pp. 257–270. Springer.

Bowen, C. M. and F. Liu (2020). Comparative study of differentially private data synthesis methods. *Statistical Science 35*(2).

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

Bowen, C. M. and J. Snoke (2019). Comparative study of differentially private synthetic data algorithms from the nist pscr differential privacy synthetic data challenge.

Burman, L. E., A. Engler, S. Khitatrakun, J. R. Nunns, S. Armstrong, J. Iselin, G. MacDonald, and P. Stallworth (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.

Burridge, J. (2003). Information preserving statistical obfuscation. *Statistics and Computing 13*(4), 321–327.

Cai, K., X. Lei, J. Wei, and X. Xiao (2021). Data synthesis via differentially private markov random fields. *Proceedings of the VLDB Endowment 14*(11), 2190–2202.

Caiola, G. and J. P. Reiter (2010). Random forests for generating partially synthetic, categorical data. *Transactions on Data Privacy 3*, 27–42.

Camino, R., C. Hammerschmidt, and R. State (2018). Generating multi-categorical samples with generative adversarial networks. *arXiv:1807.01202 [cs, stat]*.

Cano, I., S. Ladra, and V. Torra (2010). Evaluation of information loss for privacy preserving data mining through comparison of fuzzy partitions. In *International Conference on Fuzzy Systems*, Barcelona, Spain, pp. 1–8. IEEE.

Challenge.gov (2019). NIST differential privacy synthetic data challenge. `https://www.challenge.gov/?challenge=differential-privacy-synthetic-data-challenge`. Last accessed on 2022-06-08.

Charest, A.-S. (2011). How can we analyze differentially -private synthetic datasets? *Journal of Privacy and Confidentiality 2*(2).

Chen, J., D. Chun, M. Patel, E. Chiang, and J. James (2019). The validity of synthetic clinical data: a validation study of a leading synthetic data generator (synthea) using clinical quality measures. *BMC medical informatics and decision making 19*(1), 1–9.

Chen, Y., M. Elliot, and J. Sakshaug (2016). A genetic algorithm approach to synthetic data production. In *Proceedings of the 1st International Workshop on AI for Privacy and Security*, pp. 1–4.

Chen, Y., M. Elliot, and D. Smith (2018). The application of genetic algorithms to data synthesis: a comparison of three crossover methods. In *International Conference on Privacy in Statistical Databases*, pp. 160–171. Springer.

Chien, C.-H., A. H. Welsh, and J. D. Moore (2020). Synthetic business microdata: an australian example. *Journal of Privacy and Confidentiality 10*(2).

Choi, E., S. Biswal, B. Malin, J. Duke, W. F. Stewart, and J. Sun (2018). Generating Multi-label Discrete Patient Records using Generative Adversarial Networks. *arXiv:1703.06490 [cs]*.

De Montjoye, Y.-A., L. Radaelli, V. K. Singh, and A. S. Pentland (2015). Unique in the shopping mall: On the reidentifiability of credit card metadata. *Science 347*(6221), 536–539.

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.

Denton, E. L., S. Chintala, R. Fergus, et al. (2015). Deep generative image models using a laplacian pyramid of adversarial networks. *Advances in neural information processing systems 28*.

Department for Digital, Culture, Media & Sport (2022). National Data Strategy. `https://www.gov.uk/government/publications/uk-national-data-strategy/national-data-strategy`. Last accessed on 2022-05-03.

Ding, B., J. Kulkarni, and S. Yekhanin (2017). Collecting telemetry data privately. In *Advances in Neural Information Processing Systems*, pp. 3571–3580.

Domingo-Ferrer, J., L. Franconi, S. Giessing, E. Nordholt, K. Spicer, P. de Wolf, and A. Hundepool (2012). *Statistical Disclosure Control*. Wiley Series in Survey Methodology. Wiley.

Dong, Q., M. R. Elliott, and T. E. Raghunathan (2014a). Combining information from multiple complex surveys. *Survey methodology 40*(2), 347–354.

Dong, Q., M. R. Elliott, and T. E. Raghunathan (2014b). A nonparametric method to generate synthetic populations to adjust for complex sampling design features. *Survey methodology 40*(1), 29–46.

Drechsler, J. (2010). Using support vector machines for generating synthetic datasets. In *International Conference on Privacy in Statistical Databases*, pp. 148–161. Springer.

Drechsler, J. (2011a). Improved variance estimation for fully synthetic datasets. *Proceedings of the Joint UNECE/EUROSTAT Work Session on Statistical Data Confidentiality*.

Drechsler, J. (2011b). New data dissemination approaches in old Europe – synthetic datasets for a German establishment survey. *Journal of Applied Statistics*, 243–265.

Drechsler, J. (2011c). *Synthetic Datasets for Statistical Disclosure Control: Theory and Implementation. Lecture Notes in Statistics 201*. New York: Springer.

Drechsler, J. (2018). Some clarifications regarding fully synthetic data. In *International Conference on Privacy in Statistical Databases*, pp. 109–121. Springer.

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

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

Drechsler, J. and J. P. Reiter (2008). Accounting for intruder uncertainty due to sampling when estimating identification disclosure risks in partially synthetic data. In J. Domingo-Ferrer and Y. Saygin (Eds.), *Privacy in Statistical Databases*, pp. 227–238. New York: Springer.

Drechsler, J. and J. P. Reiter (2009). Disclosure risk and data utility for partially synthetic data: An empirical study using the German IAB Establishment Survey. *Journal of Official Statistics 25*, 589–603.

Drechsler, J. and J. P. Reiter (2010). Sampling with synthesis: A new approach for releasing public use census microdata. *Journal of the American Statistical Association 105*, 1347—1357.

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

Drechsler, J. and J. P. Reiter (2012). Combining synthetic data with subsampling to create public use microdata files for large scale surveys. *Survey Methodology 38*, 73–79.

Drechsler, J. and L. Vilhuber (2014a). A first step towards a german synlbd: Constructing a german longitudinal business database. *Statistical Journal of the IAOS 30*(2), 137–142.

Drechsler, J. and L. Vilhuber (2014b). Synthetic longitudinal business databases for international comparisons. In *International Conference on Privacy in Statistical Databases*, pp. 243–252. Springer.

Duncan, G. T., M. Elliot, and J.-J. Salazar-González (2011). *Statistical confidentiality*. Springer.

Dwork, C. (2008). Differential privacy: A survey of results. In M. Agrawal, D. Du, Z. Duan, and A. Li (Eds.), *Theory and Applications of Models of Computation*, Berlin, Heidelberg, pp. 1–19. Springer Berlin Heidelberg.

Dwork, C., F. McSherry, K. Nissim, and A. Smith (2006). Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography conference*, pp. 265–284. Springer.

Dwork, C. and A. Roth (2014). The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science 9*(3-4), 211–407.

Eno, J. and C. W. Thompson (2008). Generating synthetic data to match data mining patterns. *IEEE Internet Computing 12*(3), 78–82.

Erlingsson, Ú., V. Pihur, and A. Korolova (2014). Rappor: Randomized aggregatable privacy-preserving ordinal response. In *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security*, pp. 1054–1067.

Esteban, C., S. L. Hyland, and G. Rätsch (2017). Real-valued (medical) time series generation with recurrent conditional gans. *arXiv:1706.02633*.

European Commission (2022). European data strategy. `https://ec.europa.eu/info/strategy/priorities-2019-2024/europe-fit-digital-age/european-data-strategy_en`. Last accessed on 2022-05-03.

Eurostat (2022). Statistics on income and living conditions. `https://ec.europa.eu/eurostat/web/microdata/statistics-on-income-and-living-conditions`. Last accessed on 2022-05-16.

Fienberg, S. E. (1994). A radical proposal for the provision of micro-data samples and the preservation of confidentiality. Technical report, Department of Statistics, Carnegie-Mellon University.

Fienberg, S. E., U. E. Makov, and R. J. Steele (1998). Disclosure limitation using perturbation and related methods for categorical data. *Journal of Official Statistics 14*, 485–502.

Foote, A. D., A. Machanavajjhala, and K. McKinney (2019). Releasing earnings distributions using differential privacy: Disclosure avoidance system for post-secondary employment outcomes (pseo). *Journal of Privacy and Confidentiality 9*(2).

Forbes, S. and S. N. Zealand (2008). Raising statistical capability: Statistics new zealand's contribution. *Government statistical offices and statistical literacy*, 1–18.

Frid-Adar, M., E. Klang, M. Amitai, J. Goldberger, and H. Greenspan (2018). Synthetic data augmentation using gan for improved liver lesion classification. In *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, pp. 289–293.

Frigerio, L., A. S. de Oliveira, L. Gomez, and P. Duverger (2019). Differentially Private Generative Adversarial Networks for Time Series, Continuous, and Discrete Open Data.

Gaboardi, M., E. J. G. Arias, J. Hsu, A. Roth, and Z. S. Wu (2014). Dual query: Practical private query release for high dimensional data. In E. P. Xing and T. Jebara (Eds.), *Proceedings of the 31st International Conference on Machine Learning*, Volume 32 of *Proceedings of Machine Learning Research*, Bejing, China, pp. 1170–1178. PMLR.

Ghorbani, A., V. Natarajan, D. Coz, and Y. Liu (2020). DermGAN: Synthetic Generation of Clinical Skin Images with Pathology. In A. V. Dalca, M. B. McDermott, E. Alsentzer, S. G. Finlayson, M. Oberst, F. Falck, and B. Beaulieu-Jones (Eds.), *Proceedings of the Machine Learning for Health NeurIPS Workshop*, Volume 116 of *Proceedings of Machine Learning Research*, pp. 155–170. PMLR.

Goldstein, R., M. E. Woolley, L. M. Stapleton, D. Bonnéry, M. Lachowicz, T. V. Shaw, A. K. Henneberger, T. L. Johnson, and Y. Feng (2020). Expanding mlds data access and research capacity with synthetic data sets.

Gomatam, S. and A. F. Karr (2003). Distortion measures for categorical data swapping. Technical report, National Institute of Statistical Sciences, Research Triangle Park, NC.

Goodfellow, I. J., J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio (2014). Generative Adversarial Networks. *arXiv:1406.2661 [cs, stat]*.

Gulrajani, I., F. Ahmed, M. Arjovsky, V. Dumoulin, and A. Courville (2017). Improved training of Wasserstein GANs.

Hardt, M., K. Ligett, and F. McSherry (2012). A simple and practical algorithm for differentially private data release. *arXiv:1012.4763 [cs]*.

Hawala, S. (2008). Producing partially synthetic data to avoid disclosure. In *Proceedings of the Joint Statistical Meetings*. Alexandria, VA: American Statistical Association.

Homer, N., S. Szelinger, M. Redman, D. Duggan, W. Tembe, J. Muehling, J. V. Pearson, D. A. Stephan, S. F. Nelson, and D. W. Craig (2008). Resolving individuals contributing trace amounts of dna to highly complex mixtures using high-density snp genotyping microarrays. *PLoS genetics 4*(8).

Hornby, R. and J. Hu (2021). Identification risks evaluation of partially synthetic data with the identificationriskcalculation r package. *Transactions on Data Privacy 14*, 37–52.

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

Hu, J., O. Akande, and Q. Wang (2021). Multiple imputation and synthetic data generation with npbayesimputecat. *R Journal 13*(2).

Hu, J. and N. Hoshino (2018). The quasi-multinomial synthesizer for categorical data. In *International Conference on Privacy in Statistical Databases*, pp. 75–91. Springer.

Hu, J., J. P. Reiter, and Q. Wang (2014). Disclosure risk evaluation for fully synthetic categorical data. In J. Domingo-Ferrer (Ed.), *Privacy in Statistical Databases*, Number 8744 in Lecture Notes in Computer Science, pp. 185–199. Heidelberg: Springer.

Hu, J., J. P. Reiter, and Q. Wang (2018). Dirichlet process mixture models for modeling and generating synthetic versions of nested categorical data. *Bayesian Analysis 13*(1), 183–200.

Hu, J., T. D. Savitsky, and M. R. Williams (2021). Risk-efficient bayesian data synthesis for privacy protection. *Journal of Survey Statistics and Methodology*, (online–first).

Hu, J., T. D. Savitsky, and M. R. Williams (2022). Private Tabular Survey Data Products through Synthetic Microdata Generation. *Journal of Survey Statistics and Methodology 10*(3), 720–752.

Jackson, J., R. Mitra, B. Francis, and I. Dove (2022a). On integrating the number of synthetic data sets m into the a priori synthesis approach. In J. Domingo-Ferrer and M. Laurent (Eds.), *Privacy in Statistical Databases*, Cham, pp. 205–219. Springer International Publishing.

Jackson, J., R. Mitra, B. Francis, and I. Dove (2022b). Using saturated count models for user-friendly synthesis of large confidential administrative databases. *Journal of the Royal Statistical Society: Series A*, 1613–1643.

Kamthe, S., S. Assefa, and M. Deisenroth (2021). Copula flows for synthetic data generation. *arXiv:2101.00598 [cs, stat]*.

Karr, A. F., C. N. Kohnen, A. Oganian, J. P. Reiter, and A. P. Sanil (2006). A framework for evaluating the utility of data altered to protect confidentiality. *The American Statistician 60*, 224–232.

Keegan, A. and A. Tideswell (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*.

Kennickell, A. B. (1997). Multiple imputation and disclosure protection: The case of the 1995 Survey of Consumer Finances. In W. Alvey and B. Jamerson (Eds.), *Record Linkage Techniques, 1997*, pp. 248–267. Washington, DC: National Academy Press.

Kim, H. J., J. Drechsler, and K. J. Thompson (2021). Synthetic microdata for establishment surveys under informative sampling. *Journal of the Royal Statistical Society: Series A (Statistics in Society) 184*(1), 255–281.

Kim, H. J., J. P. Reiter, and A. F. Karr (2018). Simultaneous edit-imputation and disclosure limitation for business establishment data. *Journal of Applied Statistics 45*(1), 63–82.

Kingma, D. P. and M. Welling (2014). Auto-Encoding Variational Bayes. *arXiv:1312.6114 [cs, stat]*.

Kinney, S. K. and J. P. Reiter (2010). Tests of multivariate hypotheses when using multiple imputation for missing data and disclosure limitation. *Journal of Official Statistics 26*(2), 301–315.

Kinney, S. K., J. P. Reiter, and J. Miranda (2014). Synlbd 2.0: improving the synthetic longitudinal business database. *Statistical Journal of the IAOS 30*(2), 129–135.

Kinney, S. K., J. P. Reiter, A. P. Reznek, J. Miranda, R. S. Jarmin, and J. M. Abowd (2011). Towards unrestricted public use business microdata: The synthetic longitudinal business database. *International Statistical Review 79*(3), 362–384.

Koivu, A., M. Sairanen, A. Airola, and T. Pahikkala (2020). Synthetic minority oversampling of vital statistics data with generative adversarial networks. *Journal of the American Medical Informatics Association 27*(11), 1667–1674.

Lee, J. H., I. Y. Kim, and C. M. O'Keefe (2013). On regression-tree-based synthetic data methods for business data. *Journal of Privacy and Confidentiality 5*(1).

Li, H., L. Xiong, and X. Jiang (2014). Differentially private synthesization of multi-dimensional data using copula functions.

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

Liew, C. K., U. J. Choi, and C. J. Liew (1985). A data distortion by probability distribution. *ACM Transactions on Database Systems (TODS) 10*(3), 395–411.

Little, C., M. Elliot, R. Allmendinger, and S. S. Samani (2021). Generative adversarial networks for synthetic data generation: A comparative study. *arXiv:2112.01925*.

Little, R. J. and T. Raghunathan (1997). Should imputation of missing data condition on all observed variables. In *Proceedings of the Section on Survey Research Methods*, pp. 617–622. American Statistical Association Alexandria.

Little, R. J. and D. B. Rubin (1987). Statistical analysis with missing data. *New York: Wiley*.

Little, R. J. A. (1993). Statistical analysis of masked data. *Journal of Official Statistics 9*, 407–426.

Liu, T., G. Vietri, T. Steinke, J. Ullman, and S. Wu (2021). Leveraging public data for practical private query release. In *International Conference on Machine Learning*, pp. 6968–6977. PMLR.

Ma, C., S. Tschiatschek, J. M. Hernández-Lobato, R. Turner, and C. Zhang (2020). VAEM: a Deep Generative Model for Heterogeneous Mixed Type Data. *arXiv:2006.11941 [cs, stat]*.

Machanavajjhala, A., D. Kifer, J. M. Abowd, J. Gehrke, and L. Vilhuber (2008). Privacy: Theory meets practice on the map. In *IEEE 24th International Conference on Data Engineering*, pp. 277–286.

Machanavajjhala, A., D. Kifer, J. Gehrke, and M. Venkitasubramaniam (2007). l-diversity: Privacy beyond k-anonymity. *ACM Transactions on Knowledge Discovery from Data (TKDD) 1*(1), 3–es.

Mahmood, F., D. Borders, R. J. Chen, G. N. McKay, K. J. Salimian, A. Baras, and N. J. Durr (2019). Deep adversarial training for multi-organ nuclei segmentation in histopathology images. *IEEE transactions on medical imaging 39*(11), 3257–3267.

Manrique-Vallier, D. and J. Hu (2018). Bayesian non-parametric generation of fully synthetic multivariate categorical data in the presence of structural zeros. *Journal of the Royal Statistical Society: Series A (Statistics in Society) 181*(3), 635–647.

McClure, D. and J. P. Reiter (2012a). Differential privacy and statistical disclosure risk measures: An investigation with binary synthetic data. *Trans. Data Privacy 5*(3), 535–552.

McClure, D. and J. P. Reiter (2016). Assessing disclosure risks for synthetic data with arbitrary intruder knowledge. *Statistical Journal of the IAOS 32*(1), 109–126.

McClure, D. R. and J. P. Reiter (2012b). Towards providing automated feedback on the quality of inferences from synthetic datasets. *Journal of Privacy and Confidentiality 4*(1).

McKenna, R., G. Miklau, and D. Sheldon (2021). Winning the NIST Contest: A scalable and general approach to differentially private synthetic data. *Journal of Privacy and Confidentiality 11*(3).

McKenna, R., D. Sheldon, and G. Miklau (2019). Graphical-model based estimation and inference for differential privacy.

Meng, X.-L. (1994). Multiple-imputation inferences with uncongenial sources of input (disc: P558-573). *Statistical Science 9*, 538–558.

Mirza, M. and S. Osindero (2014). Conditional generative adversarial nets. *CoRR abs/1411.1784*.

Mitra, R., S. Blanchard, I. Dove, C. Tudor, and K. Spicer (2020). Confidentiality challenges in releasing longitudinally linked data. *Transactions on Data Privacy 13*(2), 151–170.

Mitra, R. and J. P. Reiter (2006). Adjusting survey weights when altering identifying design variables via synthetic data. In *International Conference on Privacy in Statistical Databases*, pp. 177–188. Springer.

Mottini, A., A. Lheritier, and R. Acuna-Agost (2018). Airline passenger name record generation using generative adversarial networks. *arXiv:1807.06657 [cs, stat]*.

Neunhoeffer, M., Z. S. Wu, and C. Dwork (2021). Private post-GAN boosting. *arXiv:2007.11934 [cs, stat]*.

Nicholson Consulting & Kōtātā Insight (2021). He ara poutama mō te reo māori. Technical report.

Nowok, B., G. M. Raab, and C. Dibben (2016). synthpop: Bespoke creation of synthetic data in R. *Journal of statistical software 74*, 1–26.

Nowok, B., G. M. Raab, and C. Dibben (2017). Providing bespoke synthetic data for the uk longitudinal studies and other sensitive data with the synthpop package for R. *Statistical Journal of the IAOS 33*(3), 785–796.

O'Donoghue, C. (2014). *Handbook of microsimulation modelling*. Emerald Group Publishing.

Ohm, P. (2009). Broken promises of privacy: Responding to the surprising failure of anonymization. *UCLA l. Rev. 57*, 1701–1776.

Osinski, B., A. Jakubowski, P. Ziecina, P. Miloś, C. Galias, S. Homoceanu, and H. Michalewski (2020). Simulation-based reinforcement learning for real-world autonomous driving. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 6411–6418.

Paiva, T., A. Chakraborty, J. Reiter, and A. Gelfand (2014). Imputation of confidential data sets with spatial locations using disease mapping models. *Statistics in Medicine 33*(11), 1928–1945.

Papernot, N., S. Song, I. Mironov, A. Raghunathan, K. Talwar, and Ú. Erlingsson (2018). Scalable private learning with PATE.

Park, N., M. Mohammadi, K. Gorde, S. Jajodia, H. Park, and Y. Kim (2018). Data Synthesis based on Generative Adversarial Networks. *Proceedings of the VLDB Endowment 11*(10), 1071–1083.

Patki, N., R. Wedge, and K. Veeramachaneni (2016). The synthetic data vault. In *2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, pp. 399–410. IEEE.

Pistner, M., A. Slavković, and L. Vilhuber (2018). Synthetic data via quantile regression for heavy-tailed and heteroskedastic data. In *International Conference on Privacy in Statistical Databases*, pp. 92–108. Springer.

Publications Office of the European Union (2022). data.europa.eu. `https://data.europa.eu/en`. Last accessed on 2022-05-04.

Quick, H. (2021). Generating poisson-distributed differentially private synthetic data. *Journal of the Royal Statistical Society: Series A (Statistics in Society) 184*(3), 1093–1108.

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

Quick, H., S. H. Holan, and C. K. Wikle (2018). Generating partially synthetic geocoded public use data with decreased disclosure risk using differential smoothing. *Journal of the Royal Statistical Society: Series A 181*, 649–661.

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

Raab, G. M., B. Nowok, and C. Dibben (2016). Practical data synthesis for large samples. *Journal of Privacy and Confidentiality 7*(3), 67–97.

Raab, G. M., B. Nowok, and C. Dibben (2021). Assessing, visualizing and improving the utility of synthetic data. *arXiv:2109.12717*.

Raghunathan, T. E., J. P. Reiter, and D. B. Rubin (2003). Multiple imputation for statistical disclosure limitation. *Journal of Official Statistics 19*, 1–16.

Rashid, S., J. Drechsler, and R. Mitra (2021). Accounting for longitudinal data structures when disseminating synthetic data to the public. In *UNECE Expert Meeting on Statistical Data Confidentiality 2021*.

Reiter, J. P. (2002). Satisfying disclosure restrictions with synthetic data sets. *Journal of Official Statistics 18*, 531–544.

Reiter, J. P. (2003). Inference for partially synthetic, public use microdata sets. *Survey Methodology 29*, 181–189.

Reiter, J. P. (2004). Simultaneous use of multiple imputation for missing data and disclosure limitation. *Survey Methodology 30*, 235–242.

Reiter, J. P. (2005a). Estimating risks of identification disclosure in microdata. *Journal of the American Statistical Association 100*(472), 1103–1112.

Reiter, J. P. (2005b). Releasing multiply-imputed, synthetic public use microdata: An illustration and empirical study. *Journal of the Royal Statistical Society, Series A 168*, 185–205.

Reiter, J. P. (2005c). Significance tests for multi-component estimands from multiply-imputed, synthetic microdata. *Journal of Statistical Planning and Inference 131*, 365–377.

Reiter, J. P. (2005d). Using CART to generate partially synthetic, public use microdata. *Journal of Official Statistics 21*, 441–462.

Reiter, J. P. and J. Drechsler (2010). Releasing multiply-imputed synthetic data generated in two stages to protect confidentiality. *Statistica Sinica 20*, 405–421.

Reiter, J. P. and S. K. Kinney (2012). Inferentially valid, partially synthetic data: Generating from posterior predictive distributions not necessary. *Journal of Official Statistics 28*(4), 583–590.

Reiter, J. P. and R. Mitra (2009). Estimating risks of identification disclosure in partially synthetic data. *Journal of Privacy and Confidentiality 1*, 99–110.

Reiter, J. P., A. Oganian, and A. F. Karr (2009). Verification servers: Enabling analysts to assess the quality of inferences from public use data. *Computational Statistics & Data Analysis 53*(4), 1475–1482.

Reiter, J. P. and T. E. Raghunathan (2007). The multiple adaptations of multiple imputation. *Journal of the American Statistical Association 102*, 1462–1471.

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

Rosenbaum, P. R. and D. B. Rubin (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika 70*(1), 41–55.

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*, Volume 1, pp. 20–34. American Statistical Association Alexandria, VA, USA.

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. John Wiley & Sons.

Rubin, D. B. (1993). Discussion: Statistical disclosure limitation. *Journal of Official Statistics 9*, 462–468.

Sakshaug, J. W. and T. E. Raghunathan (2010). Synthetic data for small area estimation. In J. Domingo-Ferrer and E. Magkos (Eds.), *Privacy in Statistical Databases*, Heidelberg, pp. 162–173. Springer.

Sakshaug, J. W. and T. E. Raghunathan (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. *Journal of Applied Statistics 41*(10), 2103–2122.

Sallier, K. (2020). Toward more user-centric data access solutions: Producing synthetic data of high analytical value by data synthesis. *Statistical Journal of the IAOS 36*(4), 1059–1066.

Shlomo, N. (2014). Probabilistic record linkage for disclosure risk assessment. In *International Conference on Privacy in Statistical Databases*, pp. 269–282. Springer.

Siwicki, B. (2021). Synthetic data boosts accuracy and speed of brain tumor surgery cds. `https://www.healthcareitnews.com/news/synthetic-data-boosts-accuracy-and-speed-brain-tumor-surgery-cds`. Last accessed on 2022-05-04.

Skinner, C. and N. Shlomo (2008). Assessing identification risk in survey microdata using log-linear models. *Journal of the American Statistical Association 103*(483), 989–1001.

Snoke, J., G. M. Raab, B. Nowok, C. Dibben, and A. Slavkovic (2018). General and specific utility measures for synthetic data. *Journal of the Royal Statistical Society: Series A (Statistics in Society) 181*(3), 663–688.

Srivastava, A., L. Valkov, C. Russell, M. U. Gutmann, and C. Sutton (2017). VEEGAN: Reducing Mode Collapse in GANs using Implicit Variational Learning.

Stadler, T., B. Oprisanu, and C. Troncoso (2021). Synthetic data–anonymisation groundhog day. *arXiv:2011.07018*.

Sweeney, L. (2002). K-anonymity: A model for protecting privacy. *World Scientific Publishing Co., Inc. 10*(5), 557—-570.

Sweeney, L. (2013). Matching known patients to health records in washington state data. *arXiv:1307.1370*.

Taub, J. and M. Elliot (2019). The synthetic data challenge. *Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality, The Hague, The Netherlands*.

Thompson, K. and H. J. Kim (2022). Incorporating economic conditions in synthetic microdata for business programs. *Journal of Survey Statistics and Methodology 10*(3), 830–859.

Torfi, A. (2020). Privacy-preserving synthetic medical data generation with deep learning. Virginia Tech.

Torfi, A. and E. A. Fox (2020). COR-GAN: correlation-capturing convolutional neural networks for generating synthetic healthcare records. *CoRR abs/2001.09346*.

Torkzadehmahani, R., P. Kairouz, and B. Paten (2020). DP-CGAN: Differentially private synthetic data and label generation. *arXiv:2001.09700 [cs, stat]*.

U.S. General Services Administration (2022). Data.gov. `https://data.gov/`. Last accessed on 2022-05-04.

Vadhan, S. (2017). The complexity of differential privacy. In *Tutorials on the Foundations of Cryptography*, pp. 347–450. Springer.

Vardhan, L. V. H. and S. Kok (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*.

Voas, D. and P. Williamson (2001). Evaluating goodness-of-fit measures for synthetic microdata. *Geographical and Environmental Modelling 5*(2), 177–200.

Waheed, A., M. Goyal, D. Gupta, A. Khanna, F. Al-Turjman, and P. R. Pinheiro (2020). Covidgan: data augmentation using auxiliary classifier gan for improved covid-19 detection. *Ieee Access 8*, 91916–91923.

Wang, H. and J. Reiter (2012). Multiple imputation for sharing precise geographies in public use data. *Annals of Applied Statistics 6*, 229–252.

Wei, L. and J. P. Reiter (2016). Releasing synthetic magnitude microdata constrained to fixed marginal totals. *Statistical Journal of the IAOS 32*(1), 93–108.

Wen, B., L. O. Colon, K. P. Subbalakshmi, and R. Chandramouli (2021). Causal-TGAN: Generating Tabular Data Using Causal Generative Adversarial Networks.

Woo, M. J., J. P. Reiter, A. Oganian, and A. F. Karr (2009). Global measures of data utility for microdata masked for disclosure limitation. *Journal of Privacy and Confidentiality 1*, 111–124.

Xiao, X., G. Wang, and J. Gehrke (2011). Differential privacy via wavelet transforms. *IEEE Transactions on Knowledge and Data Engineering 23*(8), 1200–1214.

Xie, L., K. Lin, S. Wang, F. Wang, and J. Zhou (2018). Differentially private generative adversarial network. *arXiv:1802.06739 [cs, stat]*.

Xu, L., M. Skoularidou, A. Cuesta-Infante, and K. Veeramachaneni (2019). Modeling Tabular data using Conditional GAN. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d. Alché-Buc, E. Fox, and R. Garnett (Eds.), *Advances in Neural Information Processing Systems*, Volume 32. Curran Associates, Inc.

Yahi, A., R. Vanguri, N. Elhadad, and N. P. Tatonetti (2017). Generative adversarial networks for electronic health records: A framework for exploring and evaluating methods for predicting drug-induced laboratory test trajectories. *arXiv:1712.00164*.

Yoon, J., J. Jordon, and M. v. d. Schaar (2019). PATE-GAN: Generating synthetic data with differential privacy guarantees. In *International Conference on Learning Representations*.

Yu, H. and J. P. Reiter (2018). Differentially private verification of regression predictions from synthetic data. *Trans. Data Priv. 11*(3), 279–297.

Zhang, J., G. Cormode, C. M. Procopiuc, D. Srivastava, and X. Xiao (2014). *PrivBayes: private data release via bayesian networks*, pp. 1423–1434.

Zhang, J., G. Cormode, C. M. Procopiuc, D. Srivastava, and X. Xiao (2017). PrivBayes: Private data release via bayesian networks. *ACM Transactions on Database Systems 42*(4), 1–41.

Zhao, Z., A. Kunar, H. Van der Scheer, R. Birke, and L. Y. Chen (2021). CTAB-GAN: Effective table data synthesizing. *arXiv:2102.08369 [cs]*.

Zhou, H., M. R. Elliott, and T. E. Raghunathan (2016). Synthetic multiple-imputation procedure for multistage complex samples. *Journal of Official Statistics 32*(1), 231–256.