Systematic Review of Generative Modelling Tools and Utility Metrics for Fully Synthetic Tabular Data

ANTON DANHOLT LAUTRUP, Department of Mathematics and Computer Science, University of Southern Denmark, Odense, Denmark
TOBIAS HYRUP, Department of Mathematics and Computer Science, University of Southern Denmark, Odense, Denmark
ARTHUR ZIMEK, Department of Mathematics and Computer Science, University of Southern Denmark, Odense, Denmark
PETER SCHNEIDER-KAMP, Department of Mathematics and Computer Science, University of Southern Denmark, Odense, Denmark

Sharing data with third parties is essential for advancing science, but it is becoming more and more difficult with the rise of data protection regulations, ethical restrictions, and growing fear of misuse. Fully synthetic data, which transcends anonymisation, may be the key to unlocking valuable untapped insights stored away in secured data vaults. This review examines current synthetic data generation methods and their utility measurement. We found that more traditional generative models such as Classification and Regression Tree models alongside Bayesian Networks remain highly relevant and are still capable of surpassing deep learning alternatives like Generative Adversarial Networks. However, our findings also display the same lack of agreement on metrics for evaluation, uncovered in earlier reviews, posing a persistent obstacle to advancing the field. We propose a tool for evaluating the utility of synthetic data and illustrate how it can be applied to three synthetic data generation models. By streamlining evaluation and promoting agreement on metrics, researchers can explore novel methods and generate compelling results that will convince data curators and lawmakers to embrace synthetic data. Our review emphasises the potential of synthetic data and highlights the need for greater collaboration and standardisation to unlock its full potential.

CCS Concepts: • **General and reference** → **Surveys and overviews**; **Evaluation**; • **Mathematics of computing** → **Resampling methods**; • **Computing methodologies** → **Machine learning**.

Additional Key Words and Phrases: Synthetic Data, Generative Modelling, Tabular Data, Models and Metrics, Utility and Privacy, Model benchmark, Privacy Enhancing Technologies

1 Introduction

Data are an integral part of how we today conduct our science. Countless examples of data-driven research opposing dogma have changed our world, and today, new discoveries are made at an unprecedented rate, thanks to computers, big data, and machine learning, enabling us to gather and analyse data at scales unimaginable to our predecessors [48, 63].

Authors' Contact Information: Anton Danholt Lautrup, Department of Mathematics and Computer Science, University of Southern Denmark, Odense, Denmark; e-mail: lautrup@imada.sdu.dk; Tobias Hyrup, Department of Mathematics and Computer Science, University of Southern Denmark, Odense, Denmark; e-mail: hyrup@imada.sdu.dk; Arthur Zimek, Department of Mathematics and Computer Science, University of Southern Denmark, Odense, Denmark; e-mail: zimek@imada.sdu.dk; Peter Schneider-Kamp, Department of Mathematics and Computer Science, University of Southern Denmark, Odense, Denmark; e-mail: petersk@imada.sdu.dk.

ACM 1557-7341/2024/11-ART
https://doi.org/10.1145/3704437

However, despite an abundance of data, they are by no means freely available in every field, owing to a combination of data protection regulations, fear of misuse, and intellectual property rights. These considerations are valid, and necessary to protect the rights of individuals, but can also slow down research, even stop it in its tracks, in important fields including healthcare, finance, administration, and governance [1, 7, 10, 16, 35, 58]. These domains possess datasets that are unlikely to ever become publicly available, however, it is not impossible that third parties may yet gain valuable insights by studying (pseudo)-anonymised or fully synthetic versions of them [50, 94, 101, 103]. Although anonymisation techniques have long been the preferred method of disclosure control when sharing sensitive data, this approach has many flaws and is unsuitable for widespread dataset publication [5, 87, 97, 99, 100]. Instead, researchers have begun exploring the idea of modelling dataset distributions to create a substitute for real data, so-called synthetic data, that can be sampled independently of the original data, and therefore has a lower disclosure risk than pseudo anonymisation and anonymisation [88, 96, 123].

This is, however, only true for *fully* synthetic data, which has no one-to-one correspondence with the original samples. Partially synthetic data, on the other hand, are essentially full or partial samples of real data that are mapped into a "synthetic data space" using some transformation mechanism. Here, each synthetic sample results from a real sample, making this approach not much different from anonymisation in its resistance to adversarial attacks [5, 123]. In this work, we focus purely on fully synthetic data and the mechanisms used for making it.

Frameworks that model real data rely on statistics or machine learning and are generally referred to as generative models. Generative modelling has broad implications and encompasses image- and text synthesis, technologies that have made an impression on the general public recently, as well as tools that assist research into subjects like drug discovery, physics experiments, astronomy, climate modelling, and many others [11, 13, 18, 19, 42]. While creating synthetic images, text, and music has been largely solved using modern deep learning frameworks such as diffusion, transformers, and LSTMs, a final challenge on the front of synthetic data remains: structured tabular data. Tabular data have perhaps the greatest positive potential for developing new predictive models, knowledge discovery tasks, and decision-making but also present maybe the greatest challenge for privacy, in the case of highly detailed records [22, 50, 99, 109].

In this review, we investigate the most popular methods for generating fully synthetic tabular data in the current literature and how they are compared. Furthermore, since synthetic data must provide valid statistical analyses to be useful, we study how synthetic data utility is quantified, while also taking note of privacy metrics. This is done across the computer science literature as well as the multidisciplinary data science literature since the concept of synthetic data has captured the attention of many areas of research, and we aim to provide a broad perspective on the field.

1.1 Motivation

The systematic review of methods for creating and evaluating fully synthetic tabular data is motivated by the vast positive potential of sharing record-level data in health informatics, finance, and policy-making. The creation and sharing of synthetic data, which accurately mimics the properties of real sensitive data while preserving privacy, open up new opportunities for data science in these fields [8, 20, 82, 109]. In healthcare, electronic tabular health records offer the opportunity to make better decision-making tools, improve treatments, and explore previously unknown causal relationships [10, 101, 109, 120]. In finance, open data allows for refinements in fraud detection, and identification of sales trends, risks, and opportunities [2, 7, 88]. In governance and administration, more access to population statistics can help policymakers make more informed decisions and improve public services, identify areas for improvement, and develop evidence-based policies [14, 31, 108]. Finally, emerging uses of synthetic data for data amplification and augmentation are also seeing lots of interest, such as mitigating biases to make more robust and fair models for downstream tasks, decreased cost, and the possibility of simulating scenarios not included in the original sample [34, 79, 82, 110].

Table 1. **Different features of related works.** The table presents an overview of some existing surveys that review synthetic data, along with information on each paper's focus. A "-" means no, "✓" is yes and "(✓)" means that the paper *does* seemingly fulfil the requirement, but does not explicitly state so. To get the checkmark in the "details models (metrics)", the paper is required to dedicate detailed sections to individual models (metrics) or families of models (metrics), and include more than a single model (metric).

		Focused?		Restrictions on data?			Details:		Includes
Year	Author	on domain	on model	fully synthetic	tabular	snapshot	models	metrics	experiment(s)
2021	Bond-Taylor et al. [13]	-	-	-	-	-	✓	-	✓
2021	Coutinho-Almeida et al. [25]	Health	GAN	(✓)	✓	✓	-	-	-
2022	Abedi et al. [1]	Health	GAN	✓	✓	✓	✓	-	✓
2022	Dankar et al. [28]	-	-	✓	✓	✓	-	✓	✓
2022	Figueira et al. [41]	-	GAN	(✓)	-	-	✓	✓	-
2022	Hernandez et al. [50]	Health	GAN	✓	✓	-	✓	-	-
2023	Hernandez et al. [51]	Health	-	✓	✓	✓	-	✓	✓
2023	McDuff et al. [82]	Health	-	(✓)	-	-	-	-	-
2023	Murtaza et al. [86]	Health	-	-	-	-	✓	✓	-
-	Lautrup et al. [This study]	-	-	✓	✓	✓	✓	✓	✓

In summary, sharing synthetic data for use in deep learning and data mining can bring significant benefits to society [68, 97]. However, ensuring the quality of synthetic data that meets the expectations of data curators and lawmakers remains an open question and a challenge due to the potential consequences if privacy guarantees should prove insufficient [5, 22].

1.2 Contributions

This work is a comprehensive systematic review of the developments in generating and evaluating synthetic tabular data. Many reviews and surveys of synthetic data generation exist[1], with several different approaches (some are outlined in Table 1). However, the field is in rapid development, and moving in many different directions; a review simultaneously addressing utility metrics and generation methods has to the best of our knowledge not been done rigorously before and may yield some much-needed insights for newcomers and veterans alike. In addition, many previous works have domain- or model-focused research questions, which is a limitation in watching the whole field develop.

Specifically, this work makes the following contributions;

- Overview of established tools for fully synthetic tabular data generation.
- Discussion and overview of metrics for utility and privacy evaluation of synthetic tabular data, focusing on metrics that generalise well and therefore can be used for comparing and benchmarking more widely.
- Benchmark of three synthetic data generation models using an evaluation tool based on the discovered metrics[2]

In particular, three studies have been of influence in the creation of this work. A study by Bond-Taylor et al. [13] already investigates synthetic data generation models but leaves out most quality and utility assessments in favour of a detailed theoretical presentation of the models themselves and their efficiency. Hernandez et al. [50] review generation methods and show that Generative Adversarial Network (GAN)-based approaches perform well in generating synthetic data with good utility and privacy. They name metrics used in the papers they compare but do not discuss the metrics in a wider context. Finally, Dankar et al. [28] investigate the relationships

[1] At the time of writing we are aware of at least 30 papers since 2018.

[2] The tool "SynthEval" [73] is available for Python from https://github.com/schneiderkamplab/syntheval

between metrics used in the evaluation of fully synthetic tabular data. They have some quite interesting ideas, but only thoroughly assess four metrics.

Another recent study by Murtaza et al. [86] details many generative modelling tools that did not make it into our study for various reasons. Namely, they have much broader restrictions on the type of data that they allow their generative models to model. Accommodating longitudinal, time-series, and text data, requires different methodological adaptations in the model architecture than just focusing on generating snapshot (or cross-sectional) records. Finally, some high-level overviews focused on the healthcare domain are worth mentioning, El Emam [34], Marwala et al. [79], and McDuff et al. [82], provide a good overview of the current narrative in synthetic data for healthcare applications while highlighting opportunities and pitfalls of the technology.

1.3 Research questions

This paper will analyse the different approaches to generating fully synthetic tabular data and evaluating data quality, as guided by the following research questions:

RQ1: What are the most reliable solutions for generating high-fidelity fully synthetic tabular data?
RQ2: What methods are used in the evaluation of synthetic data utility?
RQ3: How can generative models be compared in an objective and universal manner?

1.4 Structure of the paper

The remainder of this paper is structured in the following way; in Section 2 below, we go over the methods of the systematised review and explain the search strategy. Next in Section 3 the results are displayed in tabular form, and in the following Sections 4-6, we present the findings relevant to each of the three research questions, including an empirical experiment. In Section 7 we discuss the various results, implications, and limitations and provide directions for future research. Finally, we summarise and conclude the review in Section 8.

2 Systematic Review Process

For executing a systematic review, we first define a systematic review research protocol. This is done in the following steps; define search strategy and search limits, define inclusion and exclusion criteria, and build a data synthesis plan (quality assessments of the included studies) [67, 69]. The goal of the research protocol is to solidify this research in an unambiguous and reproducible manner.

2.1 Search strategy

For searching the vast literature in a manageable way, we elected to limit ourselves to searching four databases[3]. We included Scopus and Web of Science since they were the databases with the most unique papers in previous studies [41, 50]. In addition, we picked the IEEE Xplore, and the ACM Digital Library databases, since they show up in many of the related literature surveys and cover a wide range of applied fields.

From our pre-investigation, we identified relevant keywords for searching literature on synthetic data generation and limited our search to papers matching the following search string;

```
TITLE((synthe* OR generat*) AND (data OR model*)) AND
TITLE-ABS(("synthetic data" OR "generative model")
    AND (utility OR usability OR efficiency OR resemblance)
    AND (evaluat* OR metric* OR statistic*))
```

In Figure 1 we show the number of papers from each year that show up in the selected databases using these key phrases. To further limit the results, we looked for papers published from 1st of January 2020 to the 1st of April

[3] https://www.scopus.com/, https://www.webofscience.com/, https://ieeexplore.ieee.org/, https://dl.acm.org/

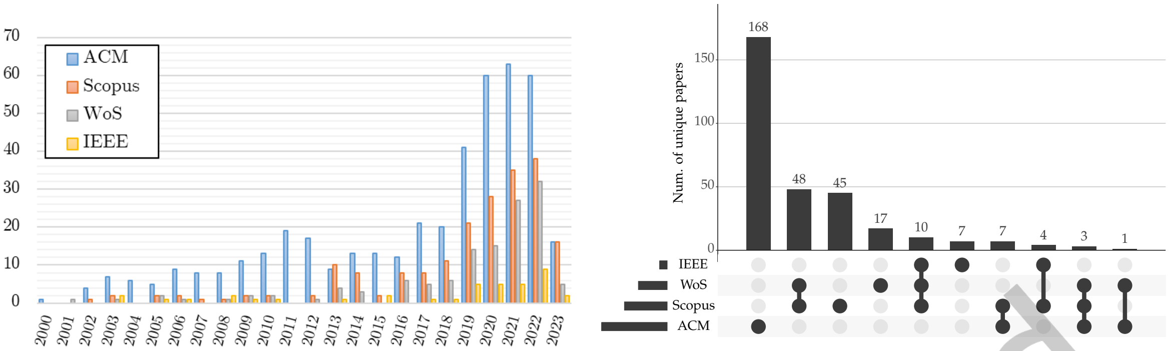


Fig. 1. **Literature search results.** Left: The histogram shows the growth of the field of synthetic data generation, based on the number of papers showing up using our search string. Our search was last conducted on the 1st of April 2023, why the results are cut short for 2023. Right: UpSet plot [23] showing database intersections: many of the papers included in the study show up in multiple databases, with some found only in one. This UpSet plot shows the distribution of unique papers identified in each database and those that appeared more than once.

2023 (the date of conducting the latest search). We chose this interval to focus on actively used (state-of-the-art) and matured ideas, but also to gain some semblance of development of different field directions. Also, we required our results to be written in English and be from a peer-reviewed journal or conference.

2.2 Inclusion and Exclusion criteria

The criteria used for selecting the papers were as listed below. First, papers are selected for inclusion if they are deemed helpful for answering at least one of the proposed research questions. That is, a retrieved paper should either present a method to create synthetic tabular data (RQ1), cover evaluation tools (RQ2), or discuss comparison methodologies (RQ3). To ensure this requirement, papers are only included if they satisfy one or more of the following inclusion criteria:

IC1: The paper presents tools for generating synthetic tabular data.
IC2: The paper discusses how to evaluate the utility of synthetic tabular data.
IC3: The paper compares existing generative frameworks for tabular data.

This judgement was most often done based on title and abstract, but in cases of doubt, the full paper was retrieved. Next, to ensure no irrelevant papers are retrieved, we formulate a series of exclusion criteria which we apply to the included papers. Papers were removed if they violated at least one of the following exclusion criteria:

EC1: The study does not discuss the evaluation or generation of synthetic data.
EC2: The study deals with data that are not mainly tabular.
EC3: The study does not include empirical results or measures of utility.
EC4: The mechanism for generating synthetic data requires real data as input.
EC5: The paper is removed for another reason.

Each record was assessed by a primary reviewer, who conferred with at least one secondary reviewer in all non-trivial cases. The final set of selected papers was validated by all co-authors. The exclusion criteria were motivated as follows: Papers should detail workable models and metrics. Thus, we excluded works that did not evaluate primary model results and papers that proposed metrics which were not shown to be applied. Moreover, we required the main body of work in the paper to be about tabular data because detailing models that work on different or multi-modal data would obfuscate the primary objective of this review. Additionally, after training or fitting the generation mechanism in a paper, the resulting model had to be able to generate data exclusively from

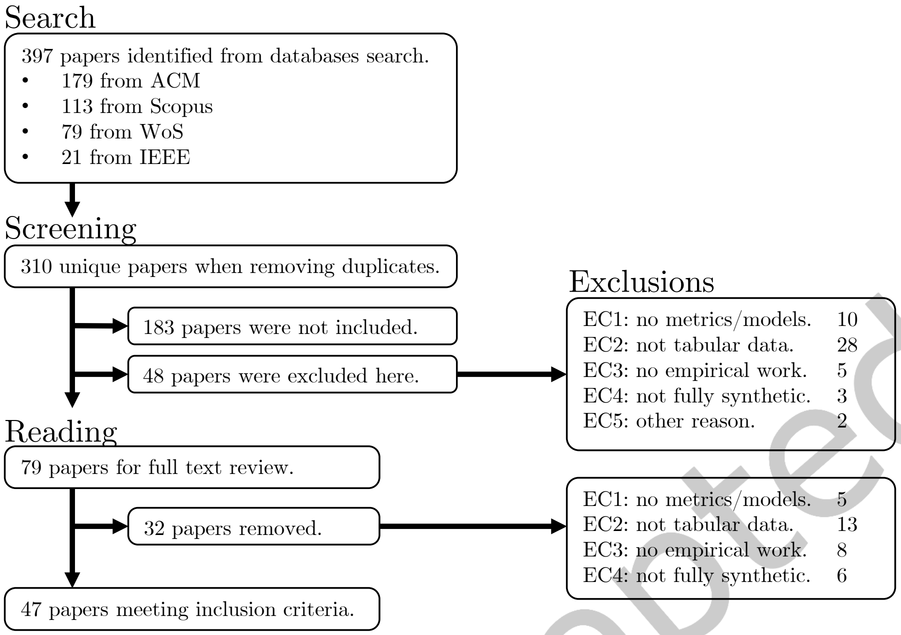

Search
397 papers identified from databases search.
• 179 from ACM
• 113 from Scopus
• 79 from WoS
• 21 from IEEE
Screening
310 unique papers when removing duplicates.
183 papers were not included.
48 papers were excluded here.
Exclusions
EC1: no metrics/models. 10
EC2: not tabular data. 28
EC3: no empirical work. 5
EC4: not fully synthetic. 3
EC5: other reason. 2
Reading
79 papers for full text review.
32 papers removed.
EC1: no metrics/models. 5
EC2: not tabular data. 13
EC3: no empirical work. 8
EC4: not fully synthetic. 6
47 papers meeting inclusion criteria.

Fig. 2. **Overview of paper selection procedure.** The flowchart illustrates how the 397 papers collected from the database searches are sifted down to the 47 papers selected for review. The primary cause for not being included and/or excluded was that many of the papers consider synthetic data generation in the context of image, text, sound, and graph data. In terms of exclusion, only the primary (first) reason for exclusion is shown (papers may be subject to more than one reason for exclusion).

the model parameters without using the training data as input — that is, retrieved papers should concern *fully* synthetic data only and papers on methods that merely perturb or otherwise transform the existing data should be excluded. Exclusion criterion EC5 was invoked two times to remove a journal news article and a conference poster.

2.3 Data collection process

The data were collected by reading through the included papers. Information about the domain, stated purpose, and datatypes used was gathered in a spreadsheet; we also noted information on the median number of entries and attributes of the datasets used in a paper. The details on utility and privacy metrics were collected, and for generative models, we also noted the types, names, and ranking (if the paper performed a ranking), and whether the model was the one proposed by the paper. The results are presented below.

3 Overview of selected publications

Here, we present the results of the systematic review process, a breakdown of the paper selection process can be seen in Figure 2. In total, from conducting the search on the 1st of April 2023, we identified 397 papers in the four databases (179 on ACM, 113 in Scopus, 79 in WoS, and 21 on IEEE, an overview of the search results can be seen in Figure 1 on the right). When duplicates were removed, we were left with 310 unique papers, from which we selected 47 papers for review. We include a brief characterisation of each paper in Tables 2 and 3. We made a division of the papers since about a third of them did not try out different approaches for synthetic data generation, which makes it more difficult to make objective conclusions based on their recommendations. Thus, Table 2 shows all the papers that only discussed a single method for generating synthetic data, whereas Table 3 shows all the papers where at least two methods are compared. The latter allows for a basic comparison to be carried out.

In the tables, we present authors, the domain of the paper along with their stated purpose. We indicate if the papers work on discrete, numerical, or mixed datatypes, and of what scales. The final two columns contain the paper's method(s) for generating synthetic data and the evaluation metrics chosen. In Table 3 we assign the letter (W) to the model that was reported to be the most successful within the paper, while we assign (L) to the worst-performing model. We also let the letter (p) signify which model the authors reported was the most private. The letters in front of the metrics Q, R, U, and P signify which family we assign it to; quality, resemblance, usability and privacy (see Section 5).

* The Wang 2021 [113] paper is included in Table 2 because the two SDG approaches they use, are not compared or used in the same context in the study.

** In the Yale 2020b [118] study, the winning model was actually one of their baseline models; the KDE Parzen window approach, which had better overall utility, but the authors handed the win to HealthGAN since KDE had an unacceptable footprint.

4 RQ1: Reliable methods for generating high fidelity synthetic data

In this section, we look at the many synthetic data generation approaches explored by the selected papers. Our primary focus lies in identifying proven methods that are currently being applied and have received appraisal from many authors for their ability to generate useful and private synthetic data. We intentionally emphasise scientifically established approaches to provide a composed narrative for new researchers, excluding highly experimental methods and legacy techniques. As we shall see in later sections, comparing and evaluating generative methods is far from trivial and since the models and evaluations across the selected papers differed significantly (e.g., different datasets used, no common baseline, structural variations of models, and differences in hyperparameters), we were unable to perform a proper meta-analysis. Instead, we will rely on a simple count of "best/worst performance" of models from across papers that make comparisons/rankings. The intention is not to select an indisputable best model, but to uncover kinds of models that are often successful at arbitrary tasks, and therefore notable to new researchers in the field.

To begin, we tally up the number of papers looking into each method, not counting multiple occurrences within a paper (this count is based on the "SDG approaches" columns in Tables 2 and 3 — references for GAN, BN, and CART methods are provided further down):

21	Generative adversarial network (GAN)
18	Bayesian network (BN)
18	Classification and regression trees (CART)
8	Variationel Autoencoder (VAE) [39, 40, 44, 75, 106, 107, 112, 124]
7	Copula technique [27, 28, 39, 78, 81, 112, 113]
4	Multiple imputation (MI) [75, 102, 118, 121]
3	Multivariate probability densities (PDF) [76, 92, 118]
2	DP Group Fields [15, 43]
1	Poisson saturated counts [62]
1	Deep Boltzmann machines (DBM) [75]
1	Graphical model [15]
1	Probabilistic database framework (KAMINO) [44]
1	Differentially Private Mean Embeddings with Random Features (DP-MERF) [49]

Based on the list, some models receive a lot of attention, while others are yet to establish themselves within the community. The methods on the right-hand side each deserves an honourable mention, but will not be treated further here, since we do not have enough unbiased evidence in their support to recommend them over more well-established frameworks.

To proceed from here, we restrict our count to the appearances within Table 3, this time allowing for duplicates internally. Looking at only this group of papers and models, we can get a general picture of each model's success

Table 2. **Selected papers with only a single synthetic data generation approach.** Since the papers in this table only present a single tool for generating synthetic data, it is difficult to assert how well the models perform outside of that specific context. The papers are still useful in other regards. For the number of entries: small datasets with < 1000 entries are denoted with a single #, intermediate datasets with ##, and datasets with ≥ 10000 entries with ###. For attributes, the categories are: small (#) < 20, medium (##), $60 \leq$ large (###). In the cases where more than one dataset is tried, we take the median number of entries/attributes.

Author Year	Domain	Stated purpose	Datatype	#ents	#atts	SDG approach	Evaluation metrics
Bowen 2020 [14]	Economics	make synthetic data to protect sensitive data	*mixed*	###	#	CART (synthpop)	(R) visual, (R) pMSE (synthpop), (Q) correlation (matrix), (P) Hit. rate (rare items), (P) cloning severity
Chen 2020 [21]	Genomics	propose new model (data augmentation)	*discrete*	##	###	GAN (PG-cGAN) (proposed)	(R) PCA, (R) MSE, (U) domain-specific metrics
Deeva 2020 [30]	Comp. Sci.	make synthetic data for education purposes	*mixed*	###	#	BN (generic)	(R) visual, (R) SRMSE
Ho 2020 [53]	Comp. Sci.	propose new model (privacy protection)	*num.*	#	#	GAN (DP-GAN) (proposed)	(R) RMSE
Holmes 2020 [54]	Comp. Sci.	evaluate synthetic data for unknown future use	*mixed*	###	##	BN (PrivBayes)	(Q) mutual information, (Q) Jensen–Shannon divergence, (U) ML-accuracy (Acc, F1-diff (%))
Taub 2020 [108]	Comp. Sci.	evaluate feasibility of synthetic data	*mixed*	###	#	CART (synthpop)	(R) CIO (number), (U) ML-accuracy (RoE), (U) reproducibility test
Yale 2020a [116]	Health	HealthGAN tutorial	*mixed*	##	###	GAN (HealthGAN)	(R) NNAA, (U) reproducibility test, (P) ADR, (P) MDR
Azizi 2021 [8]	Health	evaluate feasibility of synthetic data	*mixed*	##	###	CART (rpart)	(R) KL-divergence, (R) CIO (intervals), (U) domain-specific metrics, (U) reproducibility test
El Emam 2021a [37]	Health	evaluate feasibility of synthetic data	*mixed*	###	#	CART (generic)	(R) CIO (number, intervals), (R) pMSE, (U) variable importance, (U) reproducibility test, (P) ADR, (P) MDR
El Emam 2021b [38]	Health	evaluate sensitivity of model to permuted data	*mixed*	###	#	CART (generic)	(R) Hellinger distance, (R) pMSE, (U) ML-accuracy (AUROC)
Harder 2021 [49]	Comp. Sci.	propose new model (privacy protection)	*mixed*	###	##	DP-MERF (proposed)	(Q) Negative LL, (U) ML-accuracy (F1-diff, AUROC)
Hornby 2021 [55]	Comp. Sci.	present new tools for evaluating privacy	*mixed*	##	#	CART (synthpop)	(R) pMSE, (P) Hit. rate
Montevechi 2021 [84]	Comp. Sci.	make synthetic data to model input data	*num.*	##	#	GAN (generic)	(R) visual, (R) pMSE (C2S variant), (R) CIO (intervals)
Park 2021 [88]	Economics	make synthetic data for education purposes	*mixed*	###	##	GAN (generic)	(R) visual, (Q) mutual information, (U) ML-accuracy (Pr, Re, F1)
Wang 2021 [113]*	Health	propose evaluation framework	*mixed*	#	##	Copula (generic) BN (unknown)	(Q) KS-test (dist), (Q) correlation (matrix), (R) Consult Experts, (U) ML-accuracy (AUROC), (U) reproducibility test, (P) Hit. rate (rare items)
Chandra 2022 [20]	Health	evaluate feasibility of synthetic data	*mixed*	##	##	CART (synthpop)	(R) visual, (Q) correlation (matrix), (Q) mutual information, (Q) attribute entropy, (Q) KS-test (p-values), (U) ML-accuracy (F1, Pr, Re)
Jackson 2022 [62]	Admin.	propose new model (access private data)	*discrete*	###	#	Saturated counts (proposed)	(R) CIO (number)
Lenatti 2023 [74]	Health	assert if XAI can be used to evaluate synthetic data	*mixed*	#	#	GAN (DR-GAN)	(Q) correlation (number), (R) Hellinger distance, (R) MMD, (R) pMSE (C2S variant), (U) XAI rule similarity
Montevechi 2023 [85]	Comp. Sci.	assert if GANs can be used to evaluate synthetic data	*mixed*	##	#	GAN (generic)	(R) pMSE (C2S variant), (Q) Equivalence test (Power)
Quick 2022 [92]	Health	propose new model (improve utility)	*mixed*	###	#	multivariate PDF (proposed)	(R) visual, (U) domain-specific metrics
Tai 2022 [106]	Comp. Sci.	evaluate feasibility of privacy framework	*mixed*	###	#	VAE (generic)	(R) visual, (U) ML-accuracy

Table 3. **Selected papers with multiple synthetic data generation approaches.** This table characterises the papers that allow for some degree of comparison or consistent ranking between the models presented. For entries: small (#) < 1000, medium (##), 10000 ≤ large (###); for attributes: small (#) < 20, medium (##), 60 ≤ large (###). Again, in the case of multiple datasets compared: we take the median number. The signifying letters W, L and p, indicate the winning-, losing- and most private model within each paper. The letters Q, R, U, and P in front of the evaluation metrics indicate which category of utility, that we assign it to (see section 5)

Author Year	**Domain**	**Stated purpose**	**Datatype**	**#ents**	**#atts**	**SDG approachs**	**Evaluation metrics**
Alharbi 2020 [3]	Comp. Sci.	propose new model (data imputation)	*num.*	#	#	(W) GAN (randomGAN) (proposed) (L) GAN (mesh-GAN) (proposed)	(U) MAE, (R) RMSE
Fan 2020 [40]	Comp. Sci.	conduct experimental study of GAN architectures	*mixed*	###	##	(W) GAN (generic) (p) GAN (DPGAN) (p) BN (PrivBayes) (L) VAE (generic)	(U) ML-accuracy (F1-diff.), (P) Hit. rate, (P) DCR
Galloni 2020 [43]	Comp. Sci.	propose new evaluation methodology	*mixed*	###	#	(Wp) BN (PrivBayes) (Lp) DPGroupFields	(U) ML-accuracy, (Q) Pearson test
Mayer 2020 [81]	Economics	evaluate synthetic data for anomaly detection	*mixed*	###	##	(W) CART (synthpop) (p) BN-DP (DataSynthesizer) BN (DataSynthesizer) (L) Copula (Synthetic data vault)	(U) ML-accuracy (Pr, Re, F2)
Rankin 2020 [96]	Health	evaluate feasibility of synthetic data	*mixed*	#	#	(W) CART non-param. (synthpop) CART param. (synthpop) (Lp) BN (DataSynthesizer)	(Q) mutual information, (U) ML-accuracy, (U) ML-accuracy (F1-diff.)
Yale 2020b [118]	Health	propose new model (protect privacy)	*mixed*	###	###	(Wp) GAN (HealthGAN) (proposed) KDE (Parzen Windows)** Multiple imputation (L) Gaussian multivariate	(R) NNAA, (R) PCA, (R) CIO (intervals), (U) ML-accuracy (AUROC), (U) reproducibility test, (P) Privacy loss
Bowen 2021 [15]	Comp. Sci.	evaluation of synthetic data and generators	*mixed*	###	##	(Wp) Graphical model (p) BN (PrivBayes) (p) GAN (Team UCLANESL) (Lp) DPGroupFields	(R) pMSE, (Q) KS-test (dist), (R) CIO (number)
Dankar 2021 [27]	Comp. Sci.	evaluate effect of user settings on data quality	*mixed*	#	###	(W) CART non-param. (synthpop) Copula (Synthetic data vault) BN (DataSynthesizer) (L) CART param. (synthpop)	(R) pMSE, (U) ML-accuracy
Ge 2021 [44]	Comp. Sci.	propose new model (protect privacy)	*mixed*	###	#	(Wp) KAMINO (proposed) (p) BN (PrivBayes) (p) GAN (PATE-GAN) (Lp) VAE (DP-VAE)	(U) ML-accuracy, (U) Var. dist.
Kaur 2021 [66]	Health	evaluate feasibility of methods	*discrete*	###	###	(Wp) BN (bnlearn) (L) GAN (medBGAN)	(Q) KS-test (p-values), (U) association rules, (U) ML-accuracy (Pr, Re), (Q) correlation (matrix), (R) pMSE, (Q) rare results, (P) ADR
Lenz 2021 [75]	Genomics	implement model in federated learning setting	*binary*	#	##	(Wp) DBM (DataSHIELD) (proposed) GAN (generic) VAE (generic) (L) Multiple imputation	(R) RMSE, (U) reproducibility test, (P) MDR, (P) privacy loss
Takagi 2021 [107]	Comp. Sci.	propose new model (protect privacy)	*mixed*	###	###	(Wp) VAE (P3GM) (proposed) VAE (generic) (p) BN (PrivBayes) (Lp) VAE (DP-GM)	(U) ML-accuracy (AUROC), (R) L1-dist
Branddon 2022 [16]	Health	evaluate feasibility of synthetic data	*mixed*	###	##	(W) CART non-param. (synthpop) (Lp) CART param. (synthpop)	(R) visual, (R) CIO (intervals), (U) ML-params., (P) Hit. rate (normal, rare items)
Dankar 2022 [28]	Comp. Sci.	comparison of synthetic data generators	*mixed*	#	###	(W) CART non-param. (synthpop) Copula (Synthetic data vault) BN (DataSynthesizer) (L) CART param. (synthpop)	(R) Hellinger distance, (R) pMSE, (U) ML-accuracy (F1), (Q) correlation (number)
El Emam 2022 [36]	Health	evaluate effectiveness of utility metrics	*mixed*	#	##	(W) CART (generic) GAN (generic) (Lp) BN-DP (DataSynthesizer)	(R) MMD, (R) Hellinger distance, (R) Wasserstein distance, (U) clustering metric, (U) ML-accuracy (AUROC), (R) pMSE (C2S variant)
Endres 2022 [39]	Comp. Sci.	identify most effective generation method	*mixed*	###	##	(W) CART non-param. (synthpop) BN (DataSynthesizer) GAN (generic) GAN (synthetic data vault) Copula (synthetic data vault) (L) VAE (generic)	(Q) correlation (matrix), (U) SD metrics (SDV metric)

Table 3. **Selected papers with multiple synthetic data generation approaches (cont.)**

Author Year	**Domain**	**Stated purpose**	**Datatype**	**#ents**	**#atts**	**SDG approachs**	**Evaluation metrics**
Llugiqi 2022 [78]	Comp. Sci.	evaluate synthetic data for anomaly detection	*mixed*	###	##	(W) CART (synthpop) Copula (synthetic data vault) (L) BN (DataSynthesizer)	(U) ML-accuracy (F1, F2)
Pezoulas 2022 [89]	Health	propose extension of synthetic data generator	*mixed*	#	##	(W) BN (BGMM-OCE) (proposed) BN (generic) (L) CART (generic)	(Q) correlation (number), (Q) KS-test (GoF), (Q) coefficient variations, (R) KL-divergence
Venugopal 2022 [111]	Health	propose new model (protect privacy)	*mixed*	##	#	(W) GAN (HealthGAN) (p) GAN (pGAN) (proposed) (L) GAN (CTGAN)	(R) PCA, (R) cos-sim., (R) NNAA, (U) ML-accuracy (F1), (P) privacy loss
Visani 2022 [112]	Comp. Sci.	propose new evaluation framework	*mixed*	###	#	(W) GAN (CTGAN) GAN (CopGAN) VAE (TVAE) (Lp) Copula (Synthetic data vault)	(Q) correlation (number), (Q) χ^2-test, (R) MMD, (R) pMSE, (U) ML-accuracy (AUROC), (U) information value, (P) Hit. rate, (P) ADR, (P) MDR
Yan 2022 [119]	Health	propose new benchmarking framework	*mixed*	###	###	(W) GAN (EMR-WGAN) GAN (medWGAN) GAN (medBGAN) GAN (medGAN) (Lp) GAN (DPGAN)	(R) NNAA, (R) Wasserstein distance, (Q) correlation (number), (U) clustering metric, (U) ML-accuracy (AUROC), (U) important features, (U) domain-specific metrics, (P) Hit. rate, (P) ADR, (P) MDR, (P) privact loss
Yu 2022 [121]	Math.	propose new model (protect privacy)	*mixed*	###	###	(W) Multiple imputation (proposed) CART (synthpop) (L) CART (IVEware)	(R) pMSE, (Q) CIO (intervals, number), (Q) coefficient variations
Zhu 2022 [124]	Comp. Sci.	evaluate sensitivity of model to permuted data	*mixed*	###	###	(W) GAN (CTAB-GAN) GAN (AE-GAN) (proposed) GAN (CTGAN) VAE (TVAE) (L) GAN (table-GAN)	(R) Wasserstein distance, (Q) correlation (number), (U) ML-accuracy (acc-diff)
Duan 2023 [32]	Comp. Sci.	propose federated learning framework	*mixed*	###	##	(Wp) GAN (HT-Fed-GAN) (proposed) (L) GAN (DP-FedAvg-GAN)	(R) visual, (U) ML-accuracy (F1, MAE) (P) MDR
Li 2023 [76]	Comp. Sci.	propose new model (protect privacy)	*mixed*	###	#	(Wp) GMM (MC-GEN) (proposed) BN (PrivBayes) GMM (NoIFS) (L) GMM (RonGauss)	(U) ML-accuracy
Smith 2023 [102]	Admin.	evaluate feasibility of synthetic data	*mixed*	###	#	(W) CART (synthpop) (L) Multiple imputation	(R) Wasserstein distance, (Q) mutual information, (P) Hit. rate

rate by counting the number of times authors rank a model as the best, worst, and most private. In Table 4 the results are shown. It is seen that across the 26 papers that include ranking aspects, GAN and CART models were attributed the most success (best in 8/27 and 8/16 cases respectively). It is worth noting that GAN models were often only compared to other deep-learning approaches, including other GANs. BN models were rarely a top-ranking framework in terms of utility, however, for privacy, BN were often the recommended model. Additionally, BN was only ranked as the worst model in 2/17 papers, which is the lowest "failure rate" across the table. While GAN, BN, and CART show an indication of being consistently average or above average models, it is more difficult to make fair recommendations on behalf of the other models based on the fewer occurrences. In the following parts, we provide a brief summary of GANs, BNs, and CART techniques.

Table 4. **Models' performances in comparisons.** The table provides an overview of the seven most investigated generative model families in the papers retrieved (Table 3). We show the number of times a model type was evaluated as the best, worst, and most private in the included papers.

	GAN	BN	CART	VAE	Copula	PDF	MI
Total count	27	17	16	9	6	5	4
Best model	8	3	8	1	0	1	1
Worst model	6	2	5	4	2	2	2
Neither	13	12	3	4	4	2	1
Most private	7	9	1	2	1	1	0

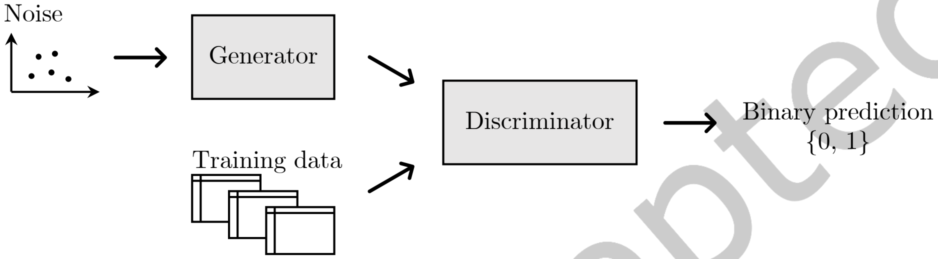


Fig. 3. **Sketch of generative adversarial network structure.** In the figure is seen how the generator learns to produce acceptable samples from noise by observing feedback from the discriminator, which simultaneously trains to classify real and fake samples. In principle, the boxes need not be neural networks but can be represented by any constellation of learners.

4.1 Generative adversarial networks (GAN)

Generative adversarial network models, as introduced by Goodfellow et al. [46], describe a zero-sum game between two neural networks; a generator and a discriminator, in which one model's gain is the other's loss. The discriminator, a classifier, trains to distinguish between real data samples and fake samples proposed by the generator model. The generator, on the other hand, takes noise as input and uses it to produce sample proposals. By observing the feedback from the discriminator, the generator gradually improves its ability to imitate the real samples without ever seeing them (see Figure 3).

Numerous adjustments and add-ons exist for this framework, many of which aim to control the training process, which can be challenging due to issues such as mode collapse, convergence failure, vanishing gradients and catastrophic forgetting [13, 42, 45]. One of the most widespread modifications is the Wasserstein GAN (WGAN) model, which replaces the training objective (commonly the binary cross-entropy) with the Wasserstein loss. Consequently, the objective of the discriminator is changed from labelling samples as either real or fake into assigning a "critic score" of the perceived authenticity. In effect, training is stabilised by more reliable gradients, and a loss metric that follows convergence and correlates with sample quality [6]. Many of the popular GAN tools seen in the literature use this approach (e.g. [115, 118]). In cases of classed or highly imbalanced data, one can use the conditional version of GAN, which adds labels to the generator and discriminator [83]. This allows specifying which samples we want to be generated, making it especially useful for oversampling of minority classes [1, 40], and encouraging more stable representations to be learnt [13].

Although GAN-based approaches have been largely replaced for image generation and manipulation in favour of the more malleable methods such as vision transformers and diffusion, they remain very relevant for high-dimensional tabular data [41, 50, 111, 124], with apt modifications for accommodating mixed and discrete data [13, 115]. In terms of privacy, differentially private versions of GAN (e.g. DPGAN [114]) exist, however, their privacy guarantee often significantly compromises on utility [44, 114]. Regular GAN models, if trained properly, should have similarly low disclosure risks compared to other methods discussed below [40].

Retrieved papers on GANs: [3, 15, 21, 32, 36, 39, 40, 44, 53, 66, 74, 75, 84, 85, 88, 111, 112, 116, 118, 119, 124].
Notable adaptations: HealthGAN [118], ADS-GAN [120], CTGAN [115], DPGAN [114].

4.2 Bayesian Networks (BN)

Bayesian networks are a class of probabilistic graphical models consisting of interconnected nodes, where each node represents a feature of the data, and edges define conditional relationships. Specifically, they are based on directed acyclic graphs (DAGs) where the direction of conditional relationships is only in one direction (similar to a decision tree). For sampling, the nodes in BNs can take on specific values or represent a stochastic decision based on inputs and learned conditional probability distributions (see Figure 4) [31, 66].

BNs are particularly well-suited for datasets with a limited number of features and high dependencies among them. However, for larger high-dimensional datasets, optimising the DAG structure using meta-heuristics can be exceedingly taxing on computational resources. Possible solutions are breaking down the dataset into smaller chunks and discretising numerical variables [30, 105].

Despite being challenged by the combinatorial explosion with increasing dimensionality of the data, BNs remain a valid and recommended option in datasets with fewer features, an option worth considering before turning towards larger machine-learning frameworks [113]. This is because BNs are highly interpretable and can be fine-tuned using domain knowledge, which makes them particularly useful in fields such as the medical domain, where they have been employed for a considerable time [72, 113]. Furthermore, BNs have also been proposed as a solution for preserving privacy, as they are only trained to model conditional probabilities stochastically [30, 105, 109]. Additionally, BNs can also be made to satisfy provable differential privacy constraints, while retaining an acceptable degree of utility [54, 122].

In summary, BNs are a flexible and efficient method for generating synthetic datasets with fewer features, capturing dependencies in an interpretable fashion, and with a decent option for privacy.

Retrieved papers on BNs: [15, 27, 28, 30, 36, 39, 40, 43, 44, 54, 66, 76, 78, 81, 89, 96, 107, 113].
Notable adaptations: DataSynthesizer [90], PrivBayes [122].

4.3 Classification and regression trees (CART)

In this study, CART models demonstrated the most success among the models included, being ranked the best model in half of the cases where it was compared to others. Originally proposed for synthetic data generation in Reiter et al. [98], the method involves creating a chain of simple classification and regression trees where each tree feeds into the next, and makes a stochastic prediction of a feature conditioned on the previously found

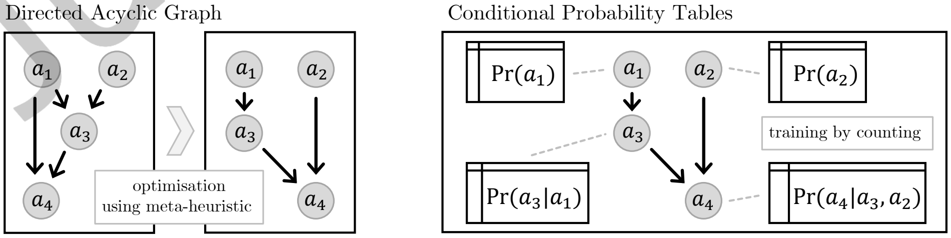


Fig. 4. **Sketch of Bayesian network structure.** The figure illustrates the two parts of Bayesian network training, on the left is seen a directed acyclic graph optimisation step by some meta-heuristic (e.g. hill climb). On the right, the final Bayesian network is seen with conditional probability tables defined for each node in the network. Sampling is done stochastically from the BN's joint probability distribution.

variables (see Figure 5). Since the design can be flexible in choosing between classification and regression trees as required by the datatypes, this design has shown impressive results for high-dimensional datasets with mixed datatypes and missing values [8, 16]. It is worth mentioning, that there appears to be some variability in result quality dependent on the synthesis order of features [38]. However, since individual trees can be relatively simple, modelling conditional distributions can be done efficiently and at a lower cost than with artificial neural networks [20, 108], making order optimisation of the trees a possibility.

Privacy-preserving mechanisms have not been extensively explored for CART models, and outside of the stochastic nature of the predictions, most privacy is introduced by tree hyperparameters such as pruning of leaves with few records. However, it has been suggested by El Emam et al. [35, 37] that CART models can have an acceptably low risk to privacy, which can be further minimised by the addition of data sanitation prior to training [39].

CART models have several different implementations, including parametric and non-parametric versions, where the latter are generally considered superior [16, 28, 39]. In the literature, the CART functionality of the R package synthpop [87] is particularly noteworthy and held in high esteem by the generative modelling community.

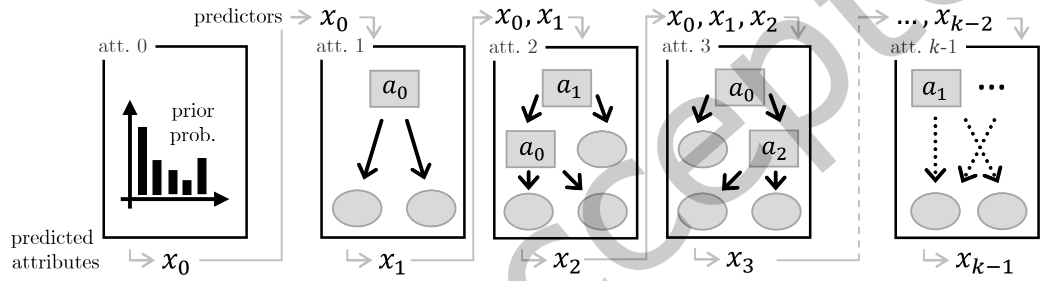


Fig. 5. **Sketch of classification and regression model structure.** The figure shows how the first attribute is sampled from its prior probability. Subsequent attributes are found by decision/regression tree models, using the already created variables as possible predictors. Randomness is introduced in the decision- and terminal nodes to alleviate overfitting/disclosure risk.

In summary, CART models have proven to be an effective and efficient choice for synthetic data generation, especially for small, high-dimensional datasets. With its flexibility and low privacy risk, it is a valuable model for future research.

Retrieved papers on CART: [8, 14, 16, 20, 27, 28, 36–39, 55, 78, 81, 89, 96, 102, 108, 121].
Notable adaptations: Synthpop [87], IVEware [95], Conditional inference trees [57].

5 RQ2: Methods for evaluating synthetic data

As evidenced by Tables 2 and 3, a multitude of metrics is used to evaluate synthetic data. A key takeaway is that no evaluation framework seems universally accepted, making it difficult to compare and rank existing synthetic data generation models across different contexts [22, 28, 50, 61]. This is problematic for research, since it spawns duplicate science, and heightens the barrier of entry for newly proposed generative methods.

Utility and privacy are the two key concepts in the evaluation of synthetic data[4]. Striking the correct balance is what allows synthetic data to serve as a beneficial alternative to sensitive real-world data [94]. Without privacy, sensitive data may be compromised, and without utility, synthetic data may not accurately model the real world.

[4] Some would argue that efficiency is equally important, but since it relates to the models themselves rather than the data, and is only treated in a few of the reviewed papers, we will not treat it explicitly. For considerations on GANs, VAE, and more deep learning approaches we refer the reader to Bond-Taylor et al. [13].

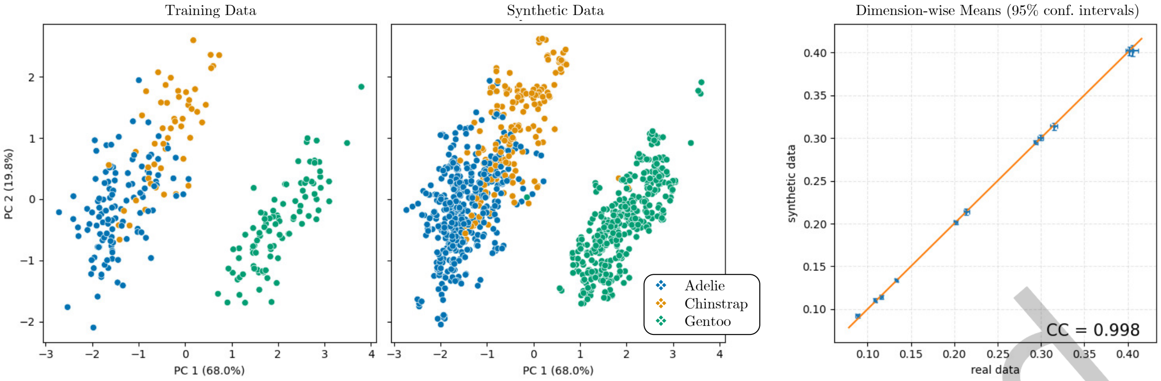


Fig. 6. **Early utility analysis figures.** Left: Scatters of real and synthetic records projected onto the first two principal components of the real Palmer penguins dataset [56]. Right: shows dimension-wise means of synthetic data compared to real data; in this case, the white wine dataset [24].

In the following, we will first review the methods used for quantifying utility, and next dedicate a part to the discussion of the collected privacy metrics.

5.1 Utility evaluation

The categorisation of utility metrics is a common practice in the field, with Hernandez et al. [50, 51] dividing them into utility, resemblance and performance dimensions, and Coutinho et al. [25] grouping them into 'dimension-wise probability', 'cross-testing', and 'distance metrics'. Others, like Hu et al. [59], Guo et al. [47], and Chandra et al. [20] have simplified their categorisation into two groups; analysis-specific and global utility while Dankar et al. [28] introduce further subdivisions of these, i.e., attribute-, bivariate-, population- and application fidelity.

In this study, we find it natural based on the findings to group utility into three larger categories: *Quality*, which assesses how well the statistical relationships are preserved in the synthetic data, *Resemblance*, which measures how hard it is to distinguish the synthetic data from real data, and *Usability*, which evaluates how well the synthetic data can serve as a substitute for real data in various tasks. In Table 5 all the utility metrics we identified are presented and roughly categorised into these three categories. Below, instead of merely promoting the most widely used metrics, we will highlight some of the more practical metrics from each category, with a focus on flexibility and popularity.

5.1.1 Early utility analysis. In many branches of generative modelling, especially those using models that are taxing to train, having good preliminary tools to gauge utility can save time during fine-tuning and training. Although these metrics may not be prominently featured in academic papers, they serve an important role in sanity-checking the results. One obvious choice is to compare the dimension-wise means and variances of the synthetic data with those of the training data, this can even be made into informative figures [41, 66, 103]. However, it is always recommended to supplement summary statistics with visual inspections of the synthetic data, if feasible, by utilising techniques such as histograms, 2D embeddings, and PCA components; summary statistics can sometimes show a good result for the wrong reasons [4, 80]. Some examples of early utility visualisation techniques are provided in Figure 6. Ultimately, if synthetic data are not even visually convincing, they are unlikely to be of much further use.

Table 5. **Table of identified utility metrics.** Presentation of the utility metrics found in the review. Not all of them can be applied to every situation, and some of them may be slightly overlapping. Examples of domain-specific metrics are minor allele frequency used in genomics [21] and clinical knowledge violation [119].

Quality: preserving statistical relationships
General statistics
Visualisations [14, 16, 20, 30, 32, 84, 88, 92, 106], Coefficient variation [89, 121], Attribute entropy [20].
Pairwise statistics
Correlation matrix difference (matrix: [14, 20, 39, 66, 113], number: [28, 74, 89, 112, 119, 124]), Mutual information difference [20, 54, 88, 96, 102], Jensen–Shannon divergence [54].
Statistical likelihoods
Kolmogorov–Smirnov test (test: [20, 66, 89], dist: [15, 113]), Pearson's independence test [43], χ^2-test [112], Negative log-likelihood [49], Equivalence test power [85], Variation distance [44, 107].
Resemblance: distinguishability
Distance metrics
RMSE [3, 53, 75], SRMSE [30], MSE [21], Cosine similarity [111].
Distribution similarity
Confidence interval overlap (number: [15, 37, 62, 108, 121], intervals: [8, 16, 37, 84, 118, 121]), Hellinger-distance [28, 36, 38, 74], Wasserstein distance [36, 102, 119, 124], Maximum mean discrepancy [36, 74, 112], KL-divergence [8, 89].
Distinguishability
Propensity MSE score (regular: [14, 15, 27, 28, 37, 38, 55, 66, 112, 121], C2S variant: [36, 74, 84, 85]), Compare PCA [21, 111, 118], Nearest-neighbour adversarial accuracy [111, 116, 118, 119], experts try to discriminate [113].
Usability: using synthetic data
ML-accuracy
Prediction accuracy / F1-score [20, 27, 28, 32, 43, 44, 54, 76, 78, 88, 96, 106, 111], AUROC [36, 38, 49, 107, 112, 113, 118, 119], F1-difference [40, 49, 54, 96, 124], Precision & Recall [20, 66, 81, 88], Clustering metric [36, 119], F2-score [78, 81], MAE [3, 32], Ratio of estimates [108].
analysis specific
Reproduce results of previous works [8, 37, 75, 108, 113, 116, 118], Domain-specific utility metrics [8, 21, 92, 119], Feature importance [37, 112, 119], Association rule mining [66], ML fit parameters comparison [16], Rare results analysis [66], XAI rule similarity [74].

5.1.2 Quality metrics. Conducting a full-quality analysis requires the assessment of pairwise statistics to determine if the relationships between variables are preserved in the synthetic data. What makes this especially challenging, is the requirement to accommodate the mixed datatypes of most real-world tabular data.

Correlation matrix. With the issue of heterogeneity in mind, correlation matrices may not seem like an optimal solution, as Spearman's ρ is only defined for numerical variables. Nevertheless, correlation matrices have been adopted to accommodate mixed datatypes, as is demonstrated in papers [66, 112, 113, 124]. The methods involved are using different "correlation-like" measures (Kendall's τ, Goodman and Kruskal's γ, or significance testing) and combining them into a single matrix [60]. On the other hand, if the dataset has only a few nominal attributes, constructing a correlation matrix from scraps may not be altogether productive. It may be more sensible in this

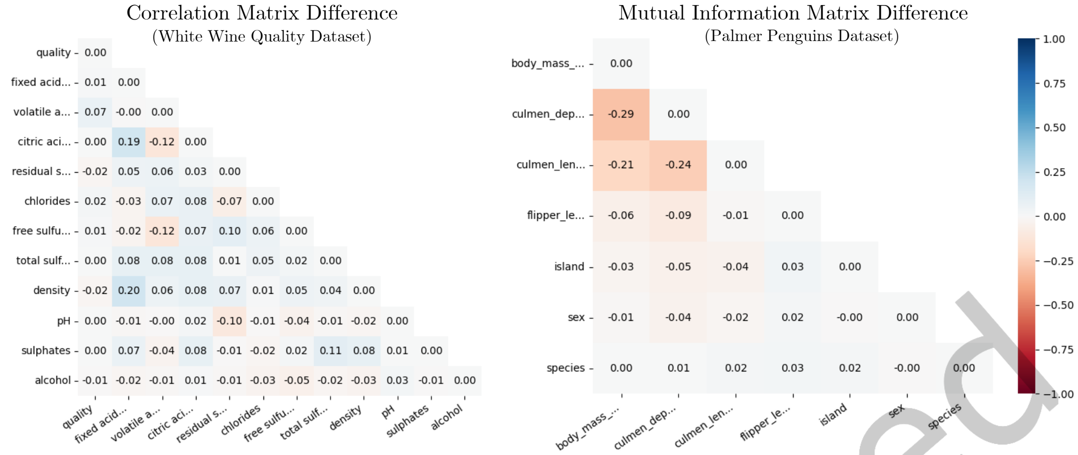


Fig. 7. **Examples of matrix difference heatmaps.** LHS: correlation matrix difference of the white wine data [24] with synthetic data. RHS: mutual information matrix difference of the palmer penguin data [56] minus synthetic version.

case to omit the nominal attributes from this calculation [20]. In any case, calculating the matrix difference like

$$\text{Corr. diff.} = \text{Corr}(\textit{real}) - \text{Corr}(\textit{synth.}), \tag{1}$$

provides an indication of which pairwise relationships are modelled accurately / poorly (an example is provided in Figure 7). Alternatively, the Frobenius norm can be utilised to condense the matrix differences into a single number [28].

Mutual information matrix difference. A potential alternative metric that addresses the challenge of varying data types is the pairwise mutual information matrix. Mutual information measures the shared information between two distributions and, thus, the extent to which the occurrence of one variable helps in predicting the other variable. In information theory mutual information between two discrete[5] random variables X and Y can be computed as follows:

$$\begin{aligned} I(X,Y) &= D_{KL}(p_{X,Y}(x,y)||p_X(x) \otimes p_Y(y)), & (2)\\ &= \sum_{x\in X, y\in Y} p_{X,Y}(x,y) \log \frac{p_{X,Y}(x,y)}{p_X(x)p_Y(y)}, & (3)\end{aligned}$$

where D_{KL} denotes the KL-divergence and $p_{X,Y}(x,y)$, $p_X(x)$, and $p_Y(y)$ represent joint and marginal probability distributions respectively [26]. Therefore, the mutual information is essentially the relative entropy of the two marginal PMF/PDFs with the joint distribution, (or alternatively, the entropy of the intersection $H(X \cap Y)$ [20]).

The mutual information score implemented in `scikit-learn`, which is employed in papers [54, 96], is an empirical approximation of the above formula. Given two sets of observations U and V, their mutual information can be calculated[6]:

$$MI(U,V) = \sum_{i=1}^{|U|} \sum_{j=1}^{|V|} \frac{|U_i \cap V_j|}{N} \log \frac{N|U_i \cap V_j|}{|U_i||V_j|}. \tag{4}$$

[5] For numerical variables upgrade sums to integrals.

[6] The full documentation is available at; https://scikit-learn.org/stable/modules/generated/sklearn.metrics.mutual_info_score.html

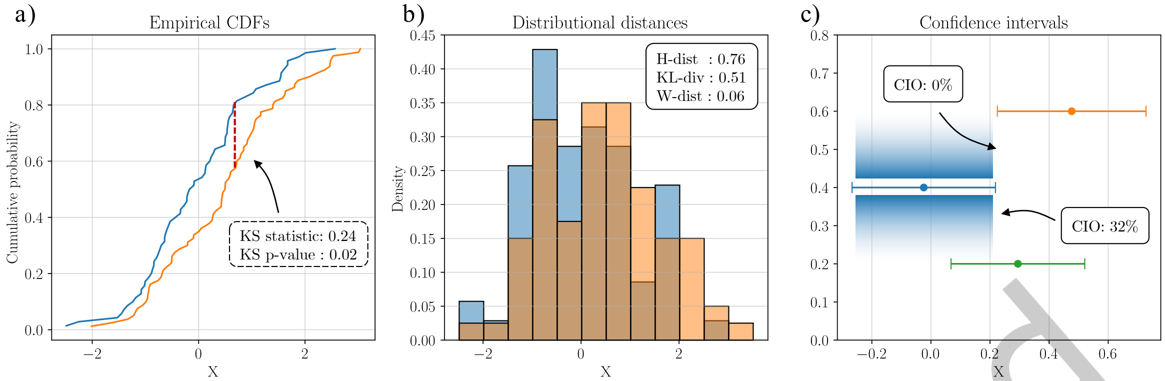


Fig. 8. **Illustrations of empirical distribution metrics.** a) Shows empirical cumulative distribution functions of samples from two random variables. The dotted line is at the point of maximum separation, which is the KS statistic. In this case, this difference is enough to reject the null hypothesis that the samples are from the same distribution. b) Shows histograms of the same samples on top of each other. Different distributional distance measures are calculated. c) 95% confidence intervals of the two previous random variables (they do not overlap), and a new one that overlaps with both.

where $|U|$ and $|V|$ represent the sizes of U and V, respectively, and N is the total number of observations. Similarly, to the correlation matrix difference, the pairwise mutual information matrix difference can be shown as a heatmap (see Figure 7) or reduced to a scalar value using the Frobenius norm. Alternatively, in the papers by Chandra et al. [20] and Smith et al. [102], they simply declare that no notable differences in the individual values were observed.

Kolmogorov–Smirnov test. Hypothesis testing is a commonly used technique to assess the statistical alignment of datasets. The pairwise Kolmogorov–Smirnov test (KS-test) is the most popular test used in synthetic data evaluation and tests whether two one-dimensional distributions are significantly different. Although multivariate versions of the KS-test are available [9, 12], the pairwise test is more commonly used to compare the marginal distributions of attributes (see Figure 8a). P-values can be used as the metric value, or the fraction/number of significant tests can double as a higher level metric [66, 112]. Others prefer to avoid the often inconclusive p-value, and instead use the test statistic as a distance measure or a goodness-of-fit indicator [15, 89, 113]. The marginal two-sample KS-test statistic is found using:

$$D_{n,m} = \sup_x |F_{1,n}(x) - F_{2,m}(x)|, \tag{5}$$

where $F_{1,n}, F_{2,m}$ are the empirical cumulative distribution functions. Intuitively the KS statistic is the height of the largest separation between the eCDFs (see Figure 8). To reject the null hypothesis the statistic must satisfy:

$$D_{n,m} > \sqrt{-\ln\left(\frac{\alpha}{2}\right) \cdot \left(1 + \frac{m}{n}\right) \cdot \frac{1}{2m}}, \tag{6}$$

where m and n are the sizes of the samples being tested, and α is the significance level. To better capture categoricals, the KS test can be combined with analysis of total variation distance and permutation testing [29, 44, 73, 107].

5.1.3 Resemblance metrics. Resemblance metrics quantify how easily we can distinguish two distributions. Obvious measures are distances, distributional overlap, and population overlap metrics. One major challenge is to reliably measure distances for heterogeneous data without favouring either numerical or categorical values. In the papers we selected, there was little innovation on this issue, and most authors tended to ignore the issue or altered the data to fit the metric. Other ways for measuring resemblance were calculating *distributional* distance of the marginal distributions, or quantifying overlap between the multivariate distributions in a different way, e.g., by using a discriminative model to see if distinguishing the datasets is possible.

Confidence interval overlap. Measuring the overlap of confidence intervals practically involves comparing the estimates of means and the 95% confidence intervals, as outlined in Karr et al. [64] (an example is provided in Figure 8c). This only makes sense for numerical variables, however, making this another metric that is not as flexible as we would like but still too useful to disregard entirely. The confidence interval for an attribute may be calculated using:

$$\mathrm{CIO} = \frac{1}{2}\left(\frac{\min(U_r, U_s) - \max(L_r, L_s)}{U_r - L_r} + \frac{\min(U_r, U_s) - \max(L_r, L_s)}{U_s - L_s}\right), \tag{7}$$

where U and L denote the upper and lower bound of the confidence interval for the real r and synthetic s data, respectively. A single number that may be used as a summary metric can be obtained by taking the average across all variables, while recording how many and which non-overlaps are found may also give valuable insights [108].

Hellinger-distance. In recent studies of utility measures, the univariate similarity or attribute fidelity is measured using the Hellinger distance [28, 36, 74]. The authors of these studies prefer the Hellinger due to its interpretability being constrained to the unit interval. Hellinger distance is calculated in the following way:

$$H(x, x') = \frac{1}{\sqrt{2}}\sqrt{\sum_i \left(\sqrt{q_i} - \sqrt{p_i}\right)^2}, \tag{8}$$

where q_i and p_i are the probabilities of every distinct result in x and x' variable spaces, respectively. A Hellinger distance of zero indicates fully overlapping distributions, while a Hellinger distance of one indicates disjoint distributions. This is unlike other options such as the Kullback–Leibler divergence, which measures relative entropy, or Wasserstein distance, which quantifies the amount of distribution weight that must be moved and how far. The Hellinger distance reads out much clearer and averages nicely across multiple attributes. An example of the three measures is shown in Figure 8b.

Propensity mean square error. The fundamental concept of this metric involves assessing if a classifier can distinguish real from synthetic samples. This has some similarities to how the discriminator works in a GAN but is usually done using a simple classifier post-training. The score is calculated using mean square error, as indicated below:

$$\mathrm{pMSE} = \frac{1}{N}\sum_i (p_i - 0.5)^2, \tag{9}$$

where p_i is the prediction indicator. This metric, which ranges from 0 (best) to 0.25 (worst), is regarded by some as the most practical and expressive utility measure [28, 93, 103]. However, the use of pMSE as a metric is subject to some limitations: Namely there is no convention for which classifier to use, though the logistic regression classifier seems to be the most common [55, 93, 103]. In extreme cases, deficient classifiers or targeted examples can show artificially good or bad results (e.g., a lattice of alternating samples cannot be discriminated effectively by a simple KNN classifier). However, in most typical use scenarios with proper validation, artificially good results should be improbable.

Some authors use a similar approach to a discriminatory classifier but do not calculate a propensity score; instead, they use the base classifier accuracy which is perhaps easier to interpret. This variant is sometimes referred to as the C2S (classifier two-samples) test [74, 85].

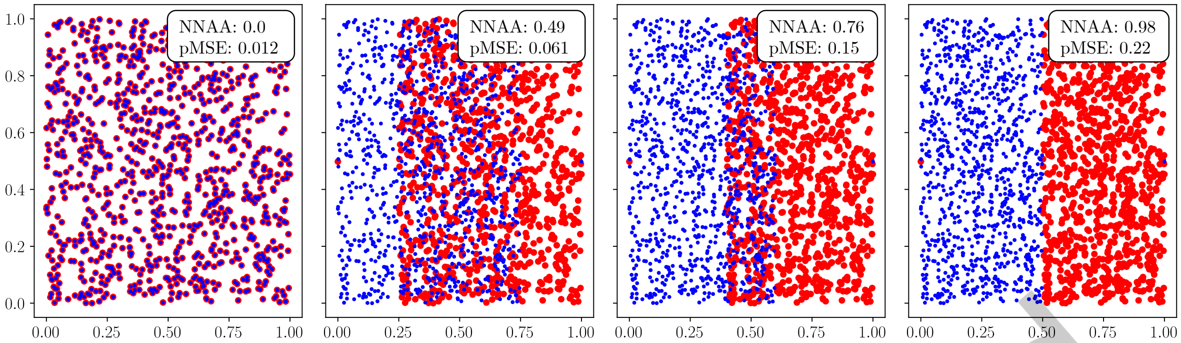


Fig. 9. **Dataset distinguishability metrics examples.** The row of images shows two copies of the same dataset being translated (they are min-max normalised as directed by the paper where NNAA is proposed [118], but we added in a single outlier point in each set to force the translation) from fully overlapping to mostly disentangled. In this case with uniform data, the NNAA score closely mirrors the overlapping percentage. The pMSE also behaves mostly as expected, being smallest when the sets fully overlap and approaching 0.25 for disjoint sets.

Nearest neighbour adversarial accuracy. Another option for assessing the distinguishability of data is using metrics such as the nearest neighbour adversarial accuracy (NNAA), first proposed by Yale et al. [117]. This metric aims to capture how well a highly competent classifier could perceivably differentiate the data, by summarising the possible true positive and true negative rates:

$$\text{NNAA} = \frac{1}{2}\left[\frac{1}{N}\sum_{i}^{N} u(d_{RS}(i) > d_{RR}(i)) + \frac{1}{N}\sum_{i}^{N} u(d_{SR}(i) > d_{SS}(i))\right], \tag{10}$$

where $u(\cdot)$ is the unit step function, $d_{RS}(i)$ is the distance from real data point i to its nearest synthetic neighbour, and $d_{RR}(i), d_{SS}(i)$ are distances (using some measure of distance) to the closest neighbour internally in the real or synthetic data [111, 118]. Intuitively, NNAA quantifies how much of the parameter space has "areas" in which either synthetic or real samples are overrepresented (hence summarising the possible true positive and true negative rates). Examples are shown above in Figure 9. Since the data in the top row is uniform, NNAA resembles the fraction of points outside of the mixed region. In the original paper, the data is min-max normalised prior to calculating NNAA.

It is worth noting that there seems to be some confusion over the ideal value of this metric in the original paper [117] and many of the follow-ups [50, 111, 118, 119]. The original paper considers a value of NNAA = 0.5 as indicative of two indistinguishable datasets, while sometimes writing that NNAA = 0 is good resemblance and NNAA = 0.5 is poor resemblance. In our experiments (see Figure 9), we find that a lower value indicates higher resemblance, but too low shows that the model is overfitting. A value of NNAA = 0.5 is however by no means indicative of indistinguishable datasets as evidenced by the figure.

5.1.4 Usability metrics. In determining if synthetic data suits the purpose for which it was created, most studies employ empirical validation techniques [41, 96]. These techniques often involve generic classification tasks or clustering, with accuracy measures or area under the receiver operating characteristic curve (AUROC) being used as evaluation metrics, and represents perhaps the most important aspect of the utility evaluation. Researchers commonly employ a selection of `scikit-learn` classifiers and generate several synthetic datasets to average out empirical fluctuations. Some studies use hold-out sets for validation (e.g. [40, 81, 119]), while others prefer k-fold cross-validation on the training set [27, 38, 49], and some use both techniques (e.g. [20, 28, 36]). Additionally, researchers working within specific scientific disciplines often introduce domain-specific metrics to assess the quality, resemblance or usability of the synthetic samples for their intended purpose [1, 8, 21, 92, 119].

Table 6. **Table of identified privacy metrics.** Overview of the privacy metrics used in the retrieved literature. We found two major categories; membership disclosure metrics and metrics that quantified overfitting.

Privacy: measuring disclosure risk of private information
Membership disclosure
Hitting rate (normal: [16, 40, 55, 102, 112, 119], sensitive items: [14, 16, 113]), Attribute disclosure risk (ADR) [37, 66, 112, 116, 119], Membership disclosure risk (MDR) [32, 37, 75, 112, 116, 119].
Overfitting characteristica
Privacy Loss [75, 111, 116, 118, 119], Row-wise square inverse frequency (cloning severity) [14], Distance to closest record (DCR) [40].

Because domain-specific metrics are so descriptive of the intended use of the data, we highly recommend using them wherever appropriate.

In principle, domain- and context-specific metrics can be considered the most practically relevant utility metrics. After all, if synthetic data proves useful for downstream tasks, other considerations, aside from privacy, are of little significance. This dilemma likely contributes to the lack of a universally accepted evaluation framework [51] and supports the division into "analysis-specific" and "global" utility proposed in previous studies [20, 28]. However, to provide a convincing and transparent presentation of results and to verify the statistical validity of the data, especially in the case of data augmentation and amplification to avoid propagating biases [110], it is still important to employ resemblance and utility metrics alongside use-case relevant metrics.

5.2 Privacy evaluation

While this survey is mainly concerned with utility of synthetic data, several metrics for measuring privacy were also uncovered in the literature (see Table 6). This selection is likely not representative of the whole field of privacy evaluation in synthetic data, as privacy was not included in the search strategy. Still, in this part, we summarise our findings since the tradeoff between utility and privacy is fundamental to the application of synthetic data [34, 86, 120].

In the papers selected for this study, most put an effort in addressing privacy: only a small proportion of the retrieved papers has little to no mention of the matter [3, 39, 74, 85, 89, 121, 124] – which is a shame since, those papers have some real interesting contents, and a small discussion of privacy implications would have elevated the work further. An even smaller fraction goes as far as assuming that privacy is a self-evident feature of synthetic data [21, 81, 96]. While non-overfit fully synthetic data has no one-to-one correspondence with real data and has shown to be resistant to adversarial attacks [35, 88, 123], it is inadvisable to assume this holds without solid evidence. Being overenthusiastic about the privacy of synthetic data, without evidence, may dissuade data curators and lawmakers from cooperating with synthetic data researchers. The survey by Murtaza et al. [86], in contrast, found close to half of their included articles assumed that their models protected privacy without any evaluation applied.

The most prominent methods used for estimating privacy are attempts to gauge how much information an adversary could possibly gain from the synthetic dataset by evaluating the risk of membership- or attribute disclosure through different methods. These range from simple hitting rate (or identical match share) and analysis of outliers and rare results, to more complex predictive models trying to infer membership of individuals (membership inference attacks) or recover missing attributes (attribute disclosure risk) from looking at the synthetic data [104, 112]. The second category of privacy metrics we found was concerned with measuring the degree of overfitting. This is often achieved through a "privacy loss" metric which measures some dataset characteristics (such as NNAA) between training and synthetic data, and again between test and synthetic data —

if the data conform better with the training data than the test data, the generative model is likely overfitted to the training data [117].

Another common measure in the synthetic data literature is differential privacy — a well-established privacy enhancing method [33]. Although not an evaluation metric per se, we will still discuss this type of disclosure control, as many papers use the control parameters in conjunction with quantifying the privacy-utility tradeoff. Essentially, differential privacy is a mechanism that can be incorporated into generative models to provide formal privacy guarantees. Given two neighbouring datasets that differ by only one record, ϵ-differential privacy guarantees that any outcome distribution of a corresponding ϵ-DP algorithm cannot differ by more than a factor of $\exp(\epsilon)$ [15, 33, 94]. In other words, the participation of a single individual in a dataset will have a negligible impact on any analysis performed on the data [107]. To achieve this, models inject noise into the model training in different controlled manners [122], and outlier analysis or cleaning may be required prior to training [109]. The noise budget is carefully managed since differential privacy would otherwise require a prohibitive amount of noise compared to the signal in data with many attributes [122]. Close to one-third of the selected papers mention differential privacy as their main implementation of privacy, namely [15, 40, 43, 44, 49, 53, 54, 62, 66, 76, 92, 106, 107]. At the same time, only five emphasise limitations and suggest using empirical privacy evaluation in addition to theoretical guarantees [32, 36, 75, 92, 116].

While the identified metrics provide a starting point for privacy evaluation, the treatment here is by no means complete. Many additional metrics are used outside of the utility-focused papers we reviewed. These include ϵ-identifiability [120], re-identification risk [35, 99], and adversarial accuracy using precision, recall, and F1-scores [66, 88]. In order to provide a comprehensive benchmark of the fidelity of synthetic data, reliable privacy metrics and utility tools are needed to provide a transparent and reproducible analysis. Additionally, further legislative action is needed in order to guide the field in a more constructive direction [61, 79].

6 RQ3: Methods for comparing generative models

The final research question on how generative models can be compared objectively and universally, is answered in two parts. First, we discuss the approaches considered in the retrieved literature and related works. Second, since very few of the considered approaches were general or robust enough, we will illustrate how a larger multifaceted benchmark of different models can be conducted through an example.

6.1 Approaches for comparing generative models

Comparing and ranking generative models is a generalisation of comparing and ranking synthetic datasets. In particular, for one model to be generally recommendable over another, it has to surpass the other in a comparison or ranking of some measurable performance dimension (e.g., privacy and utility) not only on one dataset but across a diverse selection of datasets. In the general case, the datasets should have varied characteristics, and for context-specific models, they should cover enough examples to demonstrate domain efficacy. Either way, it can be difficult to be certain of a model's excellence, but with more (diverse) datasets we can improve our confidence.

Among the retrieved papers in Table 3, are several approaches to comparing synthetic datasets and models. Many primarily focus on selecting the best dataset for a particular downstream task and, as a result, use only a limited selection of metrics and/or datasets in their comparisons (e.g., [16, 78, 81]). Others propose new generative models or attempt to compare the models in general to provide recommendations on which models to use. However, in many of these cases, too few metrics and datasets are tested (e.g., [3, 76, 121]).

On the other hand, some papers include many metrics and datasets and thoroughly motivate the choice of metrics (e.g., [28, 36]). One drawback to many of these papers is the tendency to look for models which performed the best across a lot of different datasets with different compositions of categorical and numerical datatypes.

Table 7. **Benchmark datasets.** Table shows the datasets we use in the experiment part. We divided them into 6 subgroups to test model capabilities at different scales. The number of entries is reduced, compared to the values listed on the dataset documentation, because we remove entries with missing values. Free text fields and unique keys were removed from the columns.

scale	name	#ents.	#atts. total	#atts. cat.	source
small, few atts.	[D1] diabetes	768	9	1	kaggle
	[D2] penguins	344	7	3	kaggle
	[D3] titanic	712	8	5	kaggle
small, some atts.	[D4] dermatology	358	34	33	UCI
	[D5] cervical cancer	668	34	24	UCI
	[D6] spect	267	45	1	UCI
small, many atts.	[D7] spectrometer	531	100	1	OpenML
	[D8] diabetic mellitus	281	97	92	OpenML
	[D9] mice protein	552	81	4	UCI
large, few atts.	[D10] stroke	4909	11	8	kaggle
	[D11] spaceship titanic	6923	11	5	kaggle
	[D12] white whine quality	4898	12	1	UCI
large, some atts.	[D13] cardiotocography	2126	36	15	OpenML
	[D14] steel plates faults	1941	34	10	OpenML
	[D15] one hundred plants	1600	65	1	OpenML
large, many atts.	[D16] speed dating	8242	121	62	OpenML
	[D17] Taiwanese bankruptcy	6819	96	2	UCI
	[D18] yeast-ml8	2417	117	14	OpenML

While this, in principle, allows for identifying generalist models, this also forfeits the opportunity for having hyper-specialised models, i.e., models that may do better on small datasets or on purely categorical data.

As seen in the previous section, a wide range of metrics are used to check the quality of synthetic data. In particular, machine learning accuracy, pMSE, and Hellinger distance for utility and matching risk, adversarial attacks, and distance-based approaches for privacy are used in the papers comparing models and are deemed descriptive for this task. Some papers also experiment with making aggregate metrics [15, 36, 39, 43, 112, 118, 119], to make comparisons at a glance easy. However, this also comes with the threat of suppressing important information [31, 80].

Some evaluation frameworks were also uncovered: DAISYnt framework [112] and synthetic EHR benchmarking framework [119]. DAISYnt only have a demo on PyPI[7] and EHR benchmarking is a hardcoded paper supplement[8], though the main article has some interesting ideas. Another evaluation methodology that we are aware of is STDG evaluation metrics by Hernandez et al. [51]. Again the ideas are interesting, but the supplied code[9] is mainly a paper supplement. Existing synthetic data evaluation frameworks we are aware of, with practically usable code, are SynthCity [91], SynthEval [73], SDMetrics [29], and Table Evaluator [17], each with their strengths and limitations.

6.2 Experiments with a multifaceted model benchmark

Now that we have highlighted models, metrics, and comparison methodologies, we are ready to investigate how a thorough multifaceted model benchmark might be conducted as objectively and universally as possible. In this

[7] https://pypi.org/project/daisynt/

[8] https://github.com/yy6linda/synthetic-ehr-benchmarking

[9] https://github.com/Vicomtech/STDG-evaluation-metrics

Table 8. **Summary scores of model benchmarks.** Shows the combined utility (see Eq. 11) and privacy scores (see Eq. 12) of three specific implementations of BN, GAN and CART models averaged at the different scales of datasets. The propagated error is reported to two significant figures.

	model	small, few atts.	small, some atts.	small, many atts.	large, few atts.	large, some atts.	large, many atts.
utility	CTGAN (GAN)	0.717(50)	0.625(62)	0.724(40)	0.719(31)	0.606(26)	0.528(62)
	DataSynthesizer (BN)	0.864(18)	0.814(23)	0.818(31)	0.774(61)	0.735(96)	0.613(59)
	synthpop (CART)	**0.917**(13)	**0.837**(30)	**0.886**(21)	**0.8718**(47)	**0.823**(30)	**0.743**(69)
privacy	CTGAN (GAN)	**0.717**(77)	**0.7312**(53)	**0.754**(61)	**0.703**(65)	**0.774**(45)	0.824(60)
	DataSynthesizer (BN)	0.604(46)	0.684(19)	0.72(12)	0.667(72)	0.748(39)	**0.865**(72)
	synthpop (CART)	0.515(70)	0.729(11)	0.605(32)	0.55(11)	0.704(24)	0.6858(27)

example, we compare three freely available models — DataSynthesizer [90] (BN), synthpop [87] (CART), and CTGAN [115] (GAN) — since such a comparison may reveal interesting insights about the three most popular model families. Endres et al. [39] and El Emam et al. [36] previously investigated this combination of model families, but the former uses only a few datasets, and the latter focuses more on evaluating metrics rather than the models themselves.

For our benchmark, we select 18 datasets of varying sizes and numbers of attributes, as well as with different compositions of numerical and categorical data (as seen in Table 7). This allows for subcategorisation of the results, to see if a model performs better in certain regimes. Next, we define the comparison paradigm, by selecting 14 metrics among the most descriptive and popular metrics for both utility and privacy. Our goal is not efficiency, but to make as fair of a scoring as possible. For utility, we pick correlation- and mutual information matrix differences, KS statistic and fraction of significant tests, CIO, NNAA, pMSE, Hellinger distance, and the average classification F1 difference across four different types of classifiers on both training and test set. For privacy, we measure privacy loss as the difference in NNAA value between training and test set, hitting rate, epsilon identifiability risk, and membership inference risk. All the metrics are implemented in our open-source evaluation framework, SynthEval[10] [73], where we also direct the reader for implementation details. A Codebook for recreating all results and figures in this section is available in the supplementary repository at https://github.com/schneiderkamplab/syntheval-model-benchmark-example.

We use the official implementations of DataSynthesizer in Python and synthpop in R, and CTGAN through the SynthCity [91] library[11]. In all three cases, we run the model implementations with default hyperparameter settings on a machine with an NVIDIA Tesla V100-PCIE-32Gb GPU and a 2.4GHz AMD EPYC 7501 32-Core Processor CPU.

We run the full selection of metrics on the synthetic datasets produced by each of the three synthetic data generators. The detailed results are left in the data file in the supplement repository, but an overview of high-level results in the form of summary utility and privacy scores are presented in Table 8. The score is calculated as an average of the metrics we test; accordingly, we map the different metrics to the zero-one interval where zero is the worst performance and one is the best. The correlation and mutual information matrix differences were somewhat of a challenge since they are not bounded from above. To solve this problem we apply the hyperbolic tangent function. We also considered the logistic function and error function. However, we ultimately chose tanh since we preferred its behaviour for small values.

[10] https://github.com/schneiderkamplab/syntheval

[11] Available from: https://github.com/DataResponsibly/DataSynthesizer, https://synthpop.org.uk, and https://github.com/vanderschaarlab/synthcity

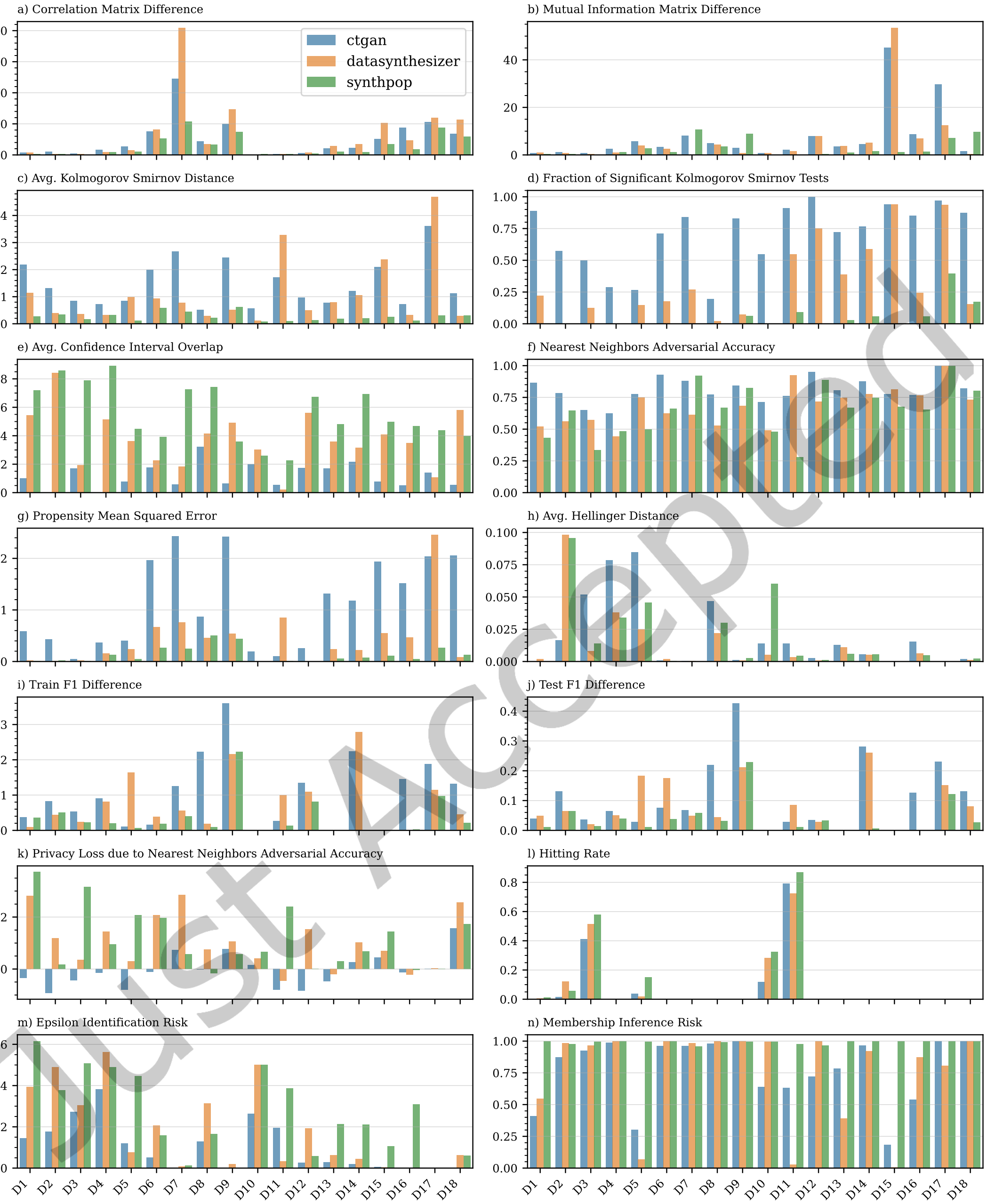


Fig. 10. **Histograms of experiment results.** The results of applying the fourteen metrics described in the text to synthetic data modelled on 18 benchmark datasets, generated by three different generative processes. All metrics save for CIO (e) should be as close to zero as possible, CIO on the other hand should ideally be close to 1. We note that the synthpop CART model excessively outperforms the other two on nearly every dataset. Histograms (k-n) show privacy metrics.

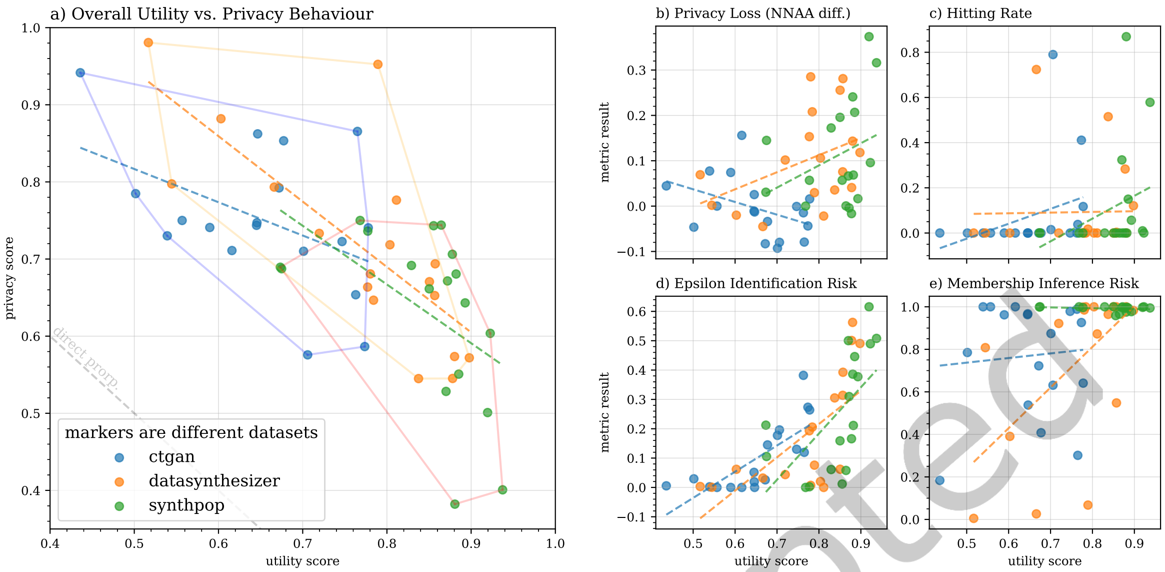


Fig. 11. **Privacy-utility trade-off behaviour.** a) Scatter plot showing all the 54 pairs of utility and privacy scores from across all the synthetic datasets, coloured by the model used to create it, illustrating the privacy-utility tradeoff. Convex hulls show an indication of different behaviour of the models. A grey line in the bottom left illustrates where a direct proportional trade-off would lie. b-e) Minor scatter plots showing how the constituent privacy metric results behave when plotted against their utility ranking. Primitive linear trendlines are plotted to help discern inclination.

$$\begin{aligned}\text{UTIL} = \frac{1}{10}\Big[&(1-\tanh\text{corr. diff.}) + (1-\tanh\text{MI diff.}) + (1-\text{KS dist.}) + (1-\text{KS sig.frac.}) + \text{CIO} \\ &+ (1-\text{H dist.}) + \left(1-\frac{\text{pMSE}}{0.25}\right) + (1-\text{NNAA}) + (1-\text{train F1 diff.}) + (1-\text{test F1 diff.})\Big] \quad (11)\end{aligned}$$

$$\text{PRIV} = \frac{1}{4}\left[(1-|\text{priv. loss}|) + (1-\text{hit rate}) + (1-\text{eps. risk}) + (1-\text{memb. inf. risk})\right] \quad (12)$$

The overall scores show good performances of all three models in terms of utility across all subdivisions of the data considered. The CART model places in the top in terms of the overall scores, and also performs the best across most metrics and datasets when looking at the detailed results in Figure 10. The GAN model underperforms on the default settings, underlining the issues with GAN training being finicky, i.e., requiring hyperparameter tuning and lots of patience. Especially on D6, D7, and D9, datasets with few categorical features, the results stand out as much worse than those produced by the other models, e.g., see (g) and (i-j) in Figure 10. The BN model follows close to the CART model on the small datasets and falls more in-between on the larger ones. For datasets with mainly numerical values like D7 and D15 the BN model fails at reproducing some of the pairwise relationships faithfully (a-b in 10).

On privacy, the GAN model performs better and quite consistently, the BN model performs best on the large datasets with many attributes. The CART model suffers on the datasets with few attributes (e.g., D1, D3, D10, and D11) but seems to do better when more attributes are available. On the individual metrics, (k-n) in Figure 10, the GAN model almost always does better than the other two models. Largely, the tradeoff between privacy

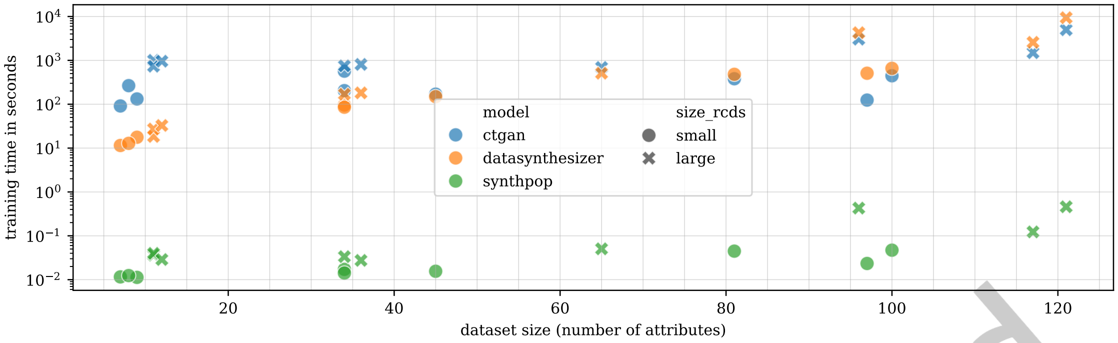


Fig. 12. **Time to train vs. the number of attributes.** In the figure, each dot indicates the training time of a model being fit to a dataset of some number of attributes. The circles are the smaller datasets while the crosses are the larger ones. The trend seems to be that the CART model is by far the most efficient, by several orders of magnitude compared with BN and the GAN model. Note how the BN model seems to scale poorer at higher dimensions than the GAN.

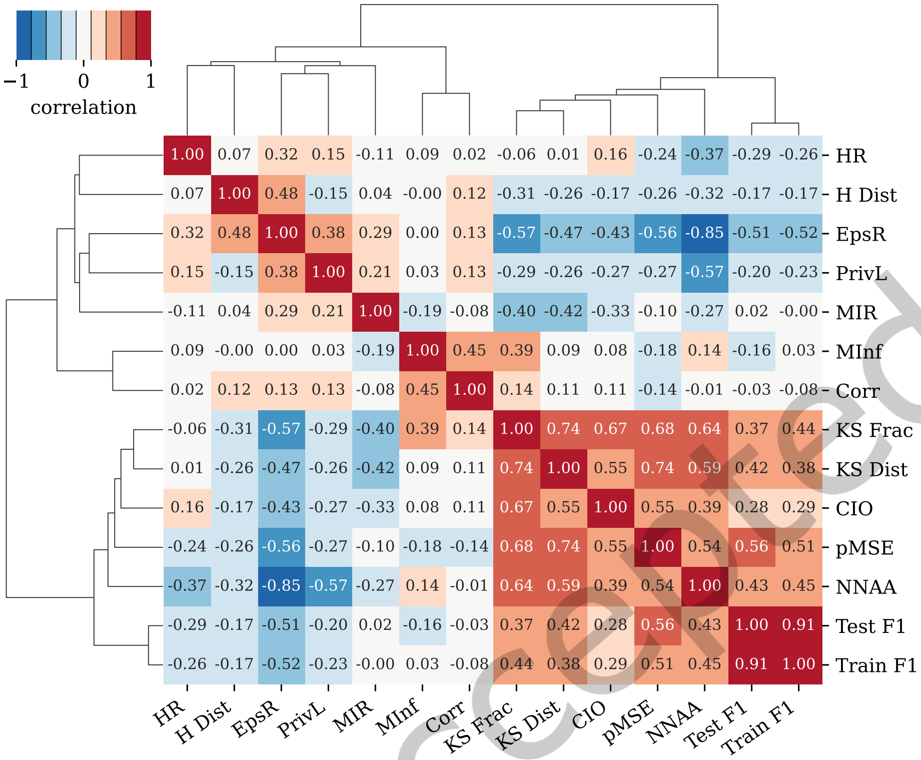
and utility is as expected; with increasing utility — privacy is negatively impacted. In Figure 11 we show all the privacy and utility scores collected in the benchmark plotted against one another. According to this empirical presentation, no datasets are placed below a direct proportionality line, and the population bulge toward the upper right. This is perhaps one of the most uplifting observations of this comparison since it suggests that up to a point an increase in privacy results in a negligible decrease in utility. In addition, different generative models may have different tradeoffs as suggested by primitive trendlines and convex hulls on Figure 11a. The synthpop (CART) generative model, for example, has lower variation in terms of utility than in privacy score (convex hull is an elongated shape in the lower right). On the other hand, CTGAN produces datasets that are closer in privacy than in utility. DataSynthesizer (BN) achieve a more balanced collection of results. The individual metric results in Figure 11 (b-e) also indicate a difference in model behaviours. Privacy loss (b)[12] indicates a tendency for overfitting in datasets with higher utility. Hitting rate and epsilon identifiability both check for synthetic data points that are too close to real records. Compared with Figure 10, it seems that larger datasets with more attributes are safer, still, these metrics generally worsen with growing utility. Finally, membership inference risk reveals that even in large datasets it is possible to determine if a record was used for training if we use machine learning rather than distance considerations.

In terms of efficiency, we mainly look at how the number of attributes influences the training time (see Figure 12). Here, the trends seem to be the same; fitting the GAN model on low-dimensional data is consistently inefficient compared with the other models, the BN model does well on low-dimensional data, but efficiency drops faster than that of the GAN as we add more attributes to fit. Finally, the CART model trains fastest on every dataset by several orders of magnitude. However, considering the limitations of such experiments [71], knowing how the models behave outside of the chosen range is difficult, but extrapolating the trends beyond 120 attributes indicates that all models scale at least exponentially, if not superexponentially, with the BN model being the worst on many attributes.

6.3 Analysis of the connectedness of the metrics

After the initial experiment, the data we gathered allowed us to run some basic data mining operations to assess the evaluation metrics we included. In Figure 13, the correlations obtained for the metrics across the 54 table

[12] Seemingly Yale et al. [116] do not suggest taking the absolute value, so privacy loss can also be negative if the NNAA on test samples is smaller than the NNAA of the training set.



	HR	H Dist	EpsR	PrivL	MIR	MInf	Corr	KS Frac	KS Dist	CIO	pMSE	NNAA	Test F1	Train F1
HR	1.00	0.07	0.32	0.15	-0.11	0.09	0.02	-0.06	0.01	0.16	-0.24	-0.37	-0.29	-0.26
H Dist	0.07	1.00	0.48	-0.15	0.04	-0.00	0.12	-0.31	-0.26	-0.17	-0.26	-0.32	-0.17	-0.17
EpsR	0.32	0.48	1.00	0.38	0.29	0.00	0.13	-0.57	-0.47	-0.43	-0.56	-0.85	-0.51	-0.52
PrivL	0.15	-0.15	0.38	1.00	0.21	0.03	0.13	-0.29	-0.26	-0.27	-0.27	-0.57	-0.20	-0.23
MIR	-0.11	0.04	0.29	0.21	1.00	-0.19	-0.08	-0.40	-0.42	-0.33	-0.10	-0.27	0.02	-0.00
MInf	0.09	-0.00	0.00	0.03	-0.19	1.00	0.45	0.39	0.09	0.08	-0.18	0.14	-0.16	0.03
Corr	0.02	0.12	0.13	0.13	-0.08	0.45	1.00	0.14	0.11	0.11	-0.14	-0.01	-0.03	-0.08
KS Frac	-0.06	-0.31	-0.57	-0.29	-0.40	0.39	0.14	1.00	0.74	0.67	0.68	0.64	0.37	0.44
KS Dist	0.01	-0.26	-0.47	-0.26	-0.42	0.09	0.11	0.74	1.00	0.55	0.74	0.59	0.42	0.38
CIO	0.16	-0.17	-0.43	-0.27	-0.33	0.08	0.11	0.67	0.55	1.00	0.55	0.39	0.28	0.29
pMSE	-0.24	-0.26	-0.56	-0.27	-0.10	-0.18	-0.14	0.68	0.74	0.55	1.00	0.54	0.56	0.51
NNAA	-0.37	-0.32	-0.85	-0.57	-0.27	0.14	-0.01	0.64	0.59	0.39	0.54	1.00	0.43	0.45
Test F1	-0.29	-0.17	-0.51	-0.20	0.02	-0.16	-0.03	0.37	0.42	0.28	0.56	0.43	1.00	0.91
Train F1	-0.26	-0.17	-0.52	-0.23	-0.00	0.03	-0.08	0.44	0.38	0.29	0.51	0.45	0.91	1.00

Fig. 13. **Matrix of metrics correlations.** The heatmap shows the correlations between the values produced by the metrics across the full experiment. The tree hierarchy shows which metrics are most similar. The metrics have been normalised in such a way that a higher value is good. Shorthands: Corr – Correlation Matrix Difference, MInf – Mutual Information Matrix Difference, KS – Kolmogorov Smirnov, CIO – Confidence Interval Overlap, NNAA – Nearest Neighbour Adversarial Accuracy, pMSE – propensity Mean Squared Error, H – Hellinger, PrivL – Privacy Loss (NNAA diff.), HR – Hitting Rate, EpsR – Epsilon Identification Risk, MIR – Membership Inference Risk.

rows of the full data are shown, with hierarchical clustering to indicate similarity. While many of the results are more or less expected, there are some interesting observations in there.

Probably the most noteworthy observation is that the test and train F1-difference are the two most closely associated metrics we tried, i.e., having both does not necessarily reveal much new information. Thus, saving data for testing may not be merited to the degree often portrayed in the literature, which can be a crucial observation for those projects where data are scarce. Similarly, the fraction of significant KS-test in the data has very similar correlations to the average KS distance — though the former seems slightly stronger correlated with metrics outside of its own local neighbourhood. The next addition to the group is the confidence interval overlap (note that the scales have been normalised to make them read the same way, i.e., a higher value is better). The relationship between CIO and the KS statistic seems obvious: the further apart the confidence intervals, the bigger the gap between the empirical cumulative distribution functions. Propensity MSE and NNAA are added next, which is perhaps more difficult to explain. However, both measure populations overlap in some sense. If the real and the synthetic datasets occupy different regions of the joint data space, both metrics give off bad values (no real records for nearest neighbours or easily detectable differences from the pMSE model). The same dataset behaviours would in turn admit a big difference in cumulative distribution functions which is likely why they attach to the KS and CIO grouping.

In a neighbouring group, we find the correlation and mutual information matrix difference coefficients. They are not too closely associated, yet they are each other's closest neighbours, and we can see that they share many similar correlation relationships. Of the two, correlation difference stands out the most, with many weak correlations, indicating that we learn slightly more by including it than by the mutual information difference metric.

In the leftmost grouping, we have the privacy metrics together with the Hellinger distance. This grouping consists of the metrics that were furthest away from the main group and each other, in particular, Hellinger distance may be the most unique utility metric, indicating that it is quite a valuable addition to the benchmark, i.e., the knowledge it contributes is not available from other sources (this result is also seen in Dankar et al. [28]). The privacy metrics, as we have already established above, connect oppositely to good utility values, so it is no surprise that they would have a significant number of negative correlations and that they would group together. Epsilon identifiability couples the strongest to the utility metrics, in particular NNAA which is also based on nearest neighbour comparisons.

Using a correlation map such as Figure 13 can help make efficient selections of metrics that complement and support each other. In the future, we hope to continue gathering data on metrics and models so the clustering of metrics can become more precise and insightful.

7 Discussion

In this systematic review, we analysed the recent advances in synthetic data generation techniques and their corresponding utility evaluation. The review showed that many different approaches are still worth considering in the generation of high-fidelity synthetic data and that new ideas are challenging to promote in the absence of a universal evaluation framework. Our findings should provide researchers new to the field an indication of noteworthy directions in generation approaches while informing newcomers and veterans alike about evaluation tools that may prove useful for their research. This paper set out to answer three key research questions in a reproducible manner. Above, we explained our process to do so. In the following, we discuss our findings for each of the three research questions.

7.1 RQ1: *What are the most reliable solutions to generating high-fidelity fully synthetic tabular data?*

In the review of the 47 different papers, we identified a total of 13 different larger model families. Generative Adversarial Networks, Bayesian Networks, and Classification And Regression Trees were found to be the most successful and well-documented methods for generating synthetic tabular data. Contrary to previous findings that have been rather GAN-centric, CART methods and BNs were found to be on par with GANs, even outperforming them in many cases. In general, our findings suggest that CART models should offer a strong start for most tasks, being both efficient and able to handle heterogeneous data. BN models also seem like a decent choice with an aptitude for acceptable utility-privacy balance, however, efficiency decreases with an increase in dimensionality. For large high-dimensional datasets, GANs may yet be a strong suit, knowing that they require lots of computation and fine-tuning to contend, but with the added benefit of more control on the privacy side. Our recommendation based on our review findings is presented in Table 9. The table was created by examining the results in Table 3 and comparing the data scales to the winning models. However, most authors aim to find models that perform well across multiple dataset scales instead of identifying which models work better at individual scales. Below, we discuss our empirical findings (from Section 6) where we conducted a non-exhaustive benchmark specifically for each of these scales.

Table 9. **Most likely effective models overview.** The table is based on the generalised dataset sizes, seen in Table 3, and is also in agreement with our non-exhaustive experiment of Section 6.

#atts.	few	some	many
small dataset	BN	BN / CART	CART
large dataset	BN / CART	CART	CART / GAN

Apart from the scales where the models perform well, other noteworthy observations can be made regarding the models. One benefit of BN and CART models is their transparency compared to deep learning approaches like GANs. While interpretability decreases with the increase of attributes, it is still possible to follow the process, such as how certain attributes are decided based on priors, or how conditional probabilities look for different classes [66, 72, 113]. Additionally, CART models are versatile in accepting various datatypes, inherently enabling switching between classification and regression tree models for categorical and numerical data respectively. In contrast, BNs and GANs need some tweaking to adjust them to different datatypes, potentially compromising the utility of the data unless the model is specifically adapted for the particular dataset. Moreover, BNs and GANs show to be less efficient than basic CART models, at the scales we investigate.

Another significant aspect in contrasting generative models is how well they preserve privacy. GAN models are the most flexible model on this part with numerous possible alterations that allow improved privacy control such as ADS-GAN [120] which changes the cost function to include a privacy term or DPGAN [114] which adds differential privacy constraints. Bayesian networks can also be made to satisfy differential privacy (e.g. [122]) but the concept seems less explored for CART models. Some papers suggest that BN and CART models are naturally more private [30, 37, 105], but it is uncertain whether this is a sufficient assurance for the sharing of sensitive data.

7.2 RQ2: *What methods are used in the evaluation of synthetic data utility?*

The most striking finding of the literature review is the wealth of different evaluation metrics that are used in evaluating the utility of synthetic data. This is in and of itself not a new discovery [41, 50, 82], but an important one nonetheless. Why no universal metrics are used remains an open question, and several factors may contribute to this, including a lack of agreement on the best evaluation methods, challenges in accommodating mixed data types, the need for metrics specific to particular applications, and the tendency in research to publish new and novel approaches. In particular, pressure to publish results that surpass or tie the state-of-the-art may drive authors to focus on developing new metrics that highlight the strengths of their proposed model rather than using established metrics that may be more appropriate.

This lack of direction is muddling the waters of this active research field. While we should never shy away from a diversity of metrics, researchers should at least agree on a selection of metrics (and how to use them) that should be present throughout the literature, so that findings are somewhat comparable across different studies[13] [36, 79]. A recognised tool or benchmark would better enable new model architectures to demonstrate their competitiveness on an equal footing and, thereby, increase their trustworthiness and adoption rate. We should be careful with summary scores which may show a skewed representation of utility (or privacy), and may result in loss of important information [31, 80]. Inspecting granular results can help alleviate this issue, and work like Dankar et al. [28] and the analysis of the correlation matrix presented above (Figure 13) provide some foundation, with beginnings of indications of redundant metrics, as well as metrics that provide most unique information.

[13] Similar to how the Fréchet inception distance (FID) [52] is widely used within image synthesis [65].

Another aspect that can improve transparency and trust, but which was remarkably absent in most of the retrieved literature, was having baselines/controls to compare with. Arguably, some papers used other generative models as their baselines (e.g., [66, 107, 111]), but this becomes a circular argument in favour of using synthetic data, overlooking how other data augmentation and/or anonymisation techniques perform on the same task. Since specialised uses of generating synthetic data may be unsuitable for employing general benchmarks, using proper baselines become even more important. Some works that did use baselines which can be reviewed for inspiration are [39, 89, 118].

All evaluation of synthetic data should aim to include general evaluation metrics that are widely used, as well as domain-specific metrics recognised within the particular area of research in order to maximise trust. Novel or obscure metrics should be avoided or only used with careful consideration and proper justification. Additionally, validation should include basic privacy metrics, regardless of provable privacy guarantees of the generative model such as differential privacy. This will help convey transparency and provide values to compare with, improving credibility and reproducibility. Our evaluation framework, SynthEval[14], integrates many of the metrics discussed above, making them easy to apply in evaluating synthetic data.

7.3 RQ3: *How can generative models be compared in an objective and universal manner?*

Of the papers retrieved in this survey, more than half attempted to compare synthetic datasets or the models that generate them in some way. Many attempted to show that a proposed model could outperform the state-of-the-art, at some task, and some tested existing models to identify the most effective generation method. In the former case, objectivity and universality may remain unconsidered, few metrics and datasets may be used, and positive results may be optimistically extended without proper analysis of the degree to which results are generalisable. In the latter case, papers are more objective, and some propose quite solid methodologies including many datasets and metrics. Building on these previous works, we show how a model benchmark study can be conducted using the identified metrics from above, for a total of 14 points of comparisons across 18 datasets. We subdivide the results based on the size and number of attributes of the analysed datasets, which reveal granular results that are overlooked in previous research. Notably, this treatment is non-exhaustive, and future works could look into mixture-fractions of numerical and categorical attributes or, for example, investigate efficacy on domains such as bioinformatics or finance data.

Since comparing all the synthetic data generation methods uncovered in this review quantitatively is beyond the scope of this publication, we selected one representative for each of the three most promising model architecture families: GAN, CART, and BN. The most exciting finding of this experiment was that the Synthpop CART model [87], used by the majority of the CART papers, is highly efficient and performs well on utility even without fine-tuning. We also found that the DataSynthesizer BN model, used in slightly less than half of the BN papers, worked impressively, although its efficiency dropped the fastest with increasing problem size. Finally, the GAN model that we used, CTGAN, also showed good results for utility albeit the worst of the three. Fine-tuning of the hyper-parameters for each dataset is likely to increase performance, however, in our experiments, we kept to the default settings. On privacy the rankings were inverted, the CART model did worse, especially on the datasets with few attributes, the BN, on the other hand, performed well on datasets with many attributes, and the GAN model did the best overall. By illustrating the utility-privacy tradeoff in Figure 11, we observe that a decent balance of utility and privacy could be realistically achievable.

7.4 Limitations

While we believe to have provided a solid foundation for new researchers, through the narrative of this systematic review, we have to acknowledge certain limitations. In particular, our focus on the major model architectures

[14] Available at: https://github.com/schneiderkamplab/syntheval

excluded more recent works from the discussion (e.g., diffusion and graph neural networks [70, 77]). Novelty is important, but this work has illustrated a problem with evaluation that makes appraising recent works difficult. Additionally, this survey focused only on tabular data, and not other modalities often contained in patient, citizen, or customer records. During this survey, we did not find any works considering multimodal generation with tabular data, and models and metrics for time series, image, and text data, would be sufficiently different to warrant separate treatment. Moreover, our literature selection process was focused on the utility of synthetic data, which likely provided a skewed perspective on privacy-preserving technologies.

Additionally, considering that the methodology employed in our example model benchmark can be foundational for future work, it is worth pointing out some limitations that need to be addressed in future iterations. First, one might argue, that a comparison should be based on fine-tuned models that achieve their best possible performance on each dataset. This might have provided a more optimistic assessment of the model performances. Parameter optimisation would result in considerably more complex experiments that would have been less controlled and reproducible as they would have heavily depended on the fine-tuning choices made by the authors. That is, objective criteria for fine-tuning might not be available in all cases and, more importantly, this would obfuscate the major focus on the evaluation measures. Second, more datasets could lead to more insightful results in the dataset-size/scale discussion, and it would have been interesting to include more metrics although this would also easily render the discussion too long and complex. Finally, the scores produced by the metrics should also be viewed with a certain scepticism, and not taken as a definitive assessment outside of context. In particular, they are biased by the metrics chosen and the implementation, e.g., including both test and train F1 differences likely inflates the average even for a mediocre result. Many metrics are correlated, as evidenced in Figure 13, and future benchmarks could choose a more compressed set of metrics, with a smaller descriptive overlap.

8 Conclusion

Generating analytically valid substitutes for sensitive datasets is a crucial step towards preserving privacy and confidentiality while still enabling research, analysis, and model development in a multitude of domains. This anticipated benefit is the main motivation for the recent explosion in synthetic data research. Enhancing the accessibility of data can enable a wider range of organizations and researchers to access and benefit from data-driven insights, ultimately leading to better decision-making and improved outcomes. Therefore, ensuring that synthetic data conforms to privacy constraints while it also simultaneously models the real data accurately is of paramount importance.

The primary objective of this systematic review was to explore the currently used generative modelling tools and evaluation methods used to create high-fidelity fully synthetic tabular data. Many methods for generating synthetic data have been proposed. We found that BNs and CART models that were previously dismissed as obsolete remain very relevant alongside deep learning models such as GANs. Many other models were also considered, but it is these three that have the largest presence in the reviewed literature.

Validation of synthetic data is probably the most important open problem in the field. Too many tools and metrics are used haphazardly across the literature; researchers are hesitant to unify behind a framework or benchmark. In the review, we identified relevant metrics, some of which seem more flexible and less contentious than others, and compiled them into a Python library for anyone interested to use. The solution to the evaluation crisis is likely more nuanced, but our proposal should provide a sturdier onset for making the next generation of synthetic data evaluation tools.

If nothing else, we hope to have provided the reader with a starting point for delving into the important field of synthetic data generation. While we have put a great emphasis on the lack of standardisation of validation apparent in the literature, we also believe this to be the most important open question, blocking the way for novel generative model designs to gain traction. If this challenge can be overcome, it will be an important milestone for synthetic data.

Acknowledgement. This study was funded by Innovation Fund Denmark in the project "PREPARE: Personalized Risk Estimation and Prevention of Cardiovascular Disease".

References

[1] Masoud Abedi, Lars Hempel, Sina Sadeghi, and Toralf Kirsten. 2022. GAN-Based Approaches for Generating Structured Data in the Medical Domain. *Applied Sciences* 12, 14 (2022), 7075.

[2] M. Jahangir Alam, Benoit Dostie, Jörg Drechsler, and Lars Vilhuber. 2020. Applying data synthesis for longitudinal business data across three countries. *Statistics in Transition New Series* 21, 4 (2020), 212–236.

[3] Hanan Hammad Alharbi and Masaomi Kimura. 2020. Missing Data Imputation Using Data Generated By GAN. In *ICCBD 2020: The 3rd International Conference on Computing and Big Data*. ACM, Taichung, Taiwan, 73–77.

[4] Frank J. Anscombe. 1973. Graphs in Statistical Analysis. *The American Statistician* 27 (1973), 17–21. Issue 1.

[5] Arno Appenzeller, Moritz Leitner, Patrick Philipp, Erik Krempel, and Jürgen Beyerer. 2022. Privacy and Utility of Private Synthetic Data for Medical Data Analyses. *Applied Sciences* 12, 23 (2022), 12320.

[6] Martín Arjovsky, Soumith Chintala, and Léon Bottou. 2017. Wasserstein Generative Adversarial Networks. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017*, Vol. 70. PMLR, Sydney, NSW, Australia, 214–223.

[7] Samuel A. Assefa, Danial Dervovic, Mahmoud Mahfouz, Robert E. Tillman, Prashant Reddy, and Manuela Veloso. 2020. Generating synthetic data in finance: opportunities, challenges and pitfalls. In *ICAIF '20: The First ACM International Conference on AI in Finance*. ACM, New York, NY, USA, 44:1–44:8.

[8] Zahra Azizi, Chaoyi Zheng, Lucy Mosquera, Louise Pilote, and Khaled El Emam. 2021. Can synthetic data be a proxy for real clinical trial data? A validation study. *BMJ Open* 11, 4 (2021), e043497.

[9] Ludwig Baringhaus and Carsten Franz. 2004. On a new multivariate two-sample test. *Journal of Multivariate Analysis* 88, 1 (2004), 190–206.

[10] Karan Bhanot, Miao Qi, John S. Erickson, Isabelle Guyon, and Kristin P. Bennett. 2021. The Problem of Fairness in Synthetic Healthcare Data. *Entropy* 23, 9 (2021), 1165.

[11] Yuemin Bian and Xiang-Qun Xie. 2021. Generative chemistry: drug discovery with deep learning generative models. *Journal of Molecular Modeling* 27, 3 (2021), 71.

[12] Peter J. Bickel. 1969. A Distribution Free Version of the Smirnov Two Sample Test in the p-Variate Case. *The Annals of Mathematical Statistics* 40, 1 (2 1969), 1–23.

[13] Sam Bond-Taylor, Adam Leach, Yang Long, and Chris G. Willcocks. 2022. Deep Generative Modelling: A Comparative Review of VAEs, GANs, Normalizing Flows, Energy-Based and Autoregressive Models. *IEEE Trans. Pattern Anal. Mach. Intell.* 44, 11 (2022), 7327–7347.

[14] Claire McKay Bowen, Victoria Bryant, Leonard Burman, Surachai Khitatrakun, Robert McClelland, Philip Stallworth, Kyle Ueyama, and Aaron R. Williams. 2020. A Synthetic Supplemental Public Use File of Low-Income Information Return Data: Methodology, Utility, and Privacy Implications. In *Privacy in Statistical Databases - UNESCO Chair in Data Privacy, International Conference, PSD 2020 (Lecture Notes in Computer Science, Vol. 12276)*. Springer, Tarragona, Spain, 257–270.

[15] Claire McKay Bowen and Joshua Snoke. 2021. Comparative Study of Differentially Private Synthetic Data Algorithms from the NIST PSCR Differential Privacy Synthetic Data Challenge. *Journal of Privacy and Confidentiality* 11, 1 (2021), 32 pages.

[16] Amy Elise Braddon, Suzanne Robinson, Rosa Alati, and Kim S. Betts. 2022. Exploring the utility of synthetic data to extract more value from sensitive health data assets: A focused example in perinatal epidemiology. *Paediatric and Perinatal Epidemiology* 00 (12 2022), 1–9.

[17] Bauke Brenninkmeijer. 2021. Table Evaluator. GitHub, code repository. https://github.com/Baukebrenninkmeijer/table-evaluator/ Version 1.6.1.

[18] Erik Buhmann, Sascha Diefenbacher, Engin Eren, Frank Gaede, Gregor Kasieczka, Anatolii Korol, and Katja Krüger. 2021. Getting High: High Fidelity Simulation of High Granularity Calorimeters with High Speed. *Comput. Softw. Big Sci.* 5, 1 (2021), 13.

[19] Brandon Buncher, Awshesh N. Sharma, and Matias Carrasco-Kind. 2021. Survey2Survey: A deep learning generative model approach for cross-survey image mapping. *Monthly Notices of the Royal Astronomical Society* 503, 1 (2021), 777–796.

[20] Gunjan Chandra, Pekka Siirtola, Satu Tamminen, Mikael Knip, Riitta Veijola, and Juha Röning. 2022. Impacts of Data Synthesis: A Metric for Quantifiable Data Standards and Performances. *Data* 7, 12 (2022), 178.

[21] Junjie Chen, Mohammad Erfan Mowlaei, and Xinghua Shi. 2020. Population-scale Genomic Data Augmentation Based on Conditional Generative Adversarial Networks. In *BCB '20: 11th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*. ACM, Virtual Event, USA, 26:1–26:6.

[22] Richard J. Chen, Ming Y. Lu, Tiffany Y. Chen, Drew F.K. Williamson, and Faisal Mahmood. 2021. Synthetic data in machine learning for medicine and healthcare. *Nature Biomedical Engineering* 5, 6 (2021), 493–497.

[23] Jake R. Conway, Alexander Lex, and Nils Gehlenborg. 2017. UpSetR: an R package for the visualization of intersecting sets and their properties. *Bioinformatics* 33, 18 (2017), 2938–2940.

[24] Paulo Cortez, António Cerdeira, Fernando Almeida, Telmo Matos, and José Reis. 2009. Modeling wine preferences by data mining from physicochemical properties. *Decis. Support Syst.* 47, 4 (2009), 547–553.

[25] João Coutinho-Almeida, Pedro Pereira Rodrigues, and Ricardo João Cruz Correia. 2021. GANs for Tabular Healthcare Data Generation: A Review on Utility and Privacy. In *Discovery Science - 24th International Conference, DS 2021 (Lecture Notes in Computer Science, Vol. 12986)*. Springer, Halifax, NS, Canada, 282–291.

[26] Thomas M. Cover and Joy A. Thomas. 2006. *Elements of Information Theory* (2 ed.). John Wiley & Sons, New York, NY, USA.

[27] Fida K. Dankar and Mahmoud Ibrahim. 2021. Fake It Till You Make It: Guidelines for Effective Synthetic Data Generation. *Applied Sciences* 11, 5 (Feb. 2021), 2158.

[28] Fida K. Dankar, Mahmoud K. Ibrahim, and Leila Ismail. 2022. A Multi-Dimensional Evaluation of Synthetic Data Generators. *IEEE Access* 10 (2022), 11147–11158.

[29] DataCebo, Inc. 2023. *Synthetic Data Metrics*. DataCebo, Inc. https://docs.sdv.dev/sdmetrics/ Version 0.14.1.

[30] Irina Deeva, Petr D. Andriushchenko, Anna V. Kalyuzhnaya, and Alexander V. Boukhanovsky. 2020. Bayesian Networks-based personal data synthesis. In *GoodTechs '20: 6th EAI International Conference on Smart Objects and Technologies for Social Good*. ACM, Antwerp, Belgium, 6–11.

[31] Jörg Drechsler and Jingchen Hu. 2020. Synthesizing Geocodes to Facilitate Access to Detailed Geographical Information in Large-Scale Administrative Data. *Journal of Survey Statistics and Methodology* 9, 3 (Dec. 2020), 523–548.

[32] Shaoming Duan, Chuanyi Liu, Peiyi Han, Xiaopeng Jin, Xinyi Zhang, Tianyu He, Hezhong Pan, and Xiayu Xiang. 2023. HT-Fed-GAN: Federated Generative Model for Decentralized Tabular Data Synthesis. *Entropy* 25, 1 (2023), 88.

[33] Cynthia Dwork and Aaron Roth. 2013. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science* 9, 3-4 (2013), 211–487.

[34] Khaled El Emam. 2023. Status of Synthetic Data Generation for Structured Health Data. *JCO Clinical Cancer Informatics* 7 (6 2023), e2300071. Issue 7.

[35] Khaled El Emam, Lucy Mosquera, and Jason Bass. 2020. Evaluating Identity Disclosure Risk in Fully Synthetic Health Data: Model Development and Validation. *Journal of Medical Internet Research* 22, 11 (2020), e23139.

[36] Khaled El Emam, Lucy Mosquera, Xi Fang, and Alaa El-Hussuna. 2022. Utility Metrics for Evaluating Synthetic Health Data Generation Methods: Validation Study. *JMIR Medical Informatics* 10, 4 (2022), e35734.

[37] Khaled El Emam, Lucy Mosquera, Elizabeth Jonker, and Harpreet Sood. 2021. Evaluating the utility of synthetic COVID-19 case data. *JAMIA Open* 4, 1 (2021), ooab012.

[38] Khaled El Emam, Lucy Mosquera, and Chaoyi Zheng. 2021. Optimizing the synthesis of clinical trial data using sequential trees. *J. Am. Medical Informatics Assoc.* 28, 1 (2021), 3–13.

[39] Markus Endres, Asha Mannarapotta Venugopal, and Tung Son Tran. 2022. Synthetic Data Generation: A Comparative Study. In *IDEAS'22: International Database Engineered Applications Symposium*. ACM, Budapest, Hungary, 94–102.

[40] Ju Fan, Tongyu Liu, Guoliang Li, Junyou Chen, Yuwei Shen, and Xiaoyong Du. 2020. Relational Data Synthesis using Generative Adversarial Networks: A Design Space Exploration. *Proc. VLDB Endow.* 13, 11 (2020), 1962–1975.

[41] Alvaro Figueira and Bruno Vaz. 2022. Survey on Synthetic Data Generation, Evaluation Methods and GANs. *Mathematics* 10, 15 (8 2022), 2733.

[42] David J. Gagne, Hannah M. Christensen, Aneesh C. Subramanian, and Adam H. Monahan. 2020. Machine Learning for Stochastic Parameterization: Generative Adversarial Networks in the Lorenz '96 Model. *Journal of Advances in Modeling Earth Systems* 12, 3 (2020), e2019MS001896.

[43] Andrea Galloni, Imre Lendák, and Tomás Horváth. 2020. A Novel Evaluation Metric for Synthetic Data Generation. In *IDEAL 2020 - 21st International Conference (Lecture Notes in Computer Science, Vol. 12490)*. Springer, Guimaraes, Portugal, 25–34.

[44] Chang Ge, Shubhankar Mohapatra, Xi He, and Ihab F. Ilyas. 2021. Kamino: Constraint-Aware Differentially Private Data Synthesis. *Proc. VLDB Endow.* 14, 10 (2021), 1886–1899.

[45] Ian J. Goodfellow. 2016. NIPS 2016 Tutorial: Generative Adversarial Networks. arXiv, preprint. https://doi.org/10.48550/arXiv.1701.00160

[46] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative Adversarial Networks. arXiv, preprint. https://doi.org/10.48550/arXiv.1406.2661

[47] Shijie Guo and Jingchen Hu. 2022. Data Privacy Protection and Utility Preservation through Bayesian Data Synthesis: A Case Study on Airbnb Listings. *The American Statistician* 77, 2 (2022), 192–200.

[48] Alon Halevy, Peter Norvig, and Fernando Pereira. 2009. The Unreasonable Effectiveness of Data. *IEEE Intelligent Systems* 24, 2 (2009), 8–12.

[49] Frederik Harder, Kamil Adamczewski, and Mijung Park. 2021. DP-MERF: Differentially Private Mean Embeddings with RandomFeatures for Practical Privacy-preserving Data Generation. In *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics, AISTATS 2021 (Proceedings of Machine Learning Research, Vol. 130)*. PMLR, San Diego, California, USA, 1819–1827.

[50] Mikel Hernandez, Gorka Epelde, Ane Alberdi, Rodrigo Cilla, and Debbie Rankin. 2022. Synthetic data generation for tabular health records: A systematic review. *Neurocomputing* 493 (2022), 28–45.

[51] Mikel Hernandez, Gorka Epelde, Ane Alberdi, Rodrigo Cilla, and Debbie Rankin. 2023. Synthetic Tabular Data Evaluation in the Health Domain Covering Resemblance, Utility, and Privacy Dimensions. *Methods of Information in Medicine* 62, S 01 (2023), e19–e38.

[52] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. arXiv, preprint. https://doi.org/10.48550/arXiv.1706.08500

[53] Stella Ho, Youyang Qu, Longxiang Gao, Jianxin Li, and Yong Xiang. 2019. Generative Adversarial Nets Enhanced Continual Data Release Using Differential Privacy. In *Algorithms and Architectures for Parallel Processing - 19th International Conference, ICA3PP 2019 (Lecture Notes in Computer Science, Vol. 11945)*. Springer, Melbourne, VIC, Australia, 418–426.

[54] Michael Holmes and George Theodorakopoulos. 2020. Towards using differentially private synthetic data for machine learning in collaborative data science projects. In *ARES 2020: The 15th International Conference on Availability, Reliability and Security*. ACM, Virtual Event, Ireland, 28:1–28:6.

[55] Ryan Hornby and Jingchen Hu. 2021. Identification Risks Evaluation of Partially Synthetic Data with the IdentificationRiskCalculation R Package. *Trans. Data Priv.* 14, 1 (2021), 37–52.

[56] Allison Marie Horst, Alison Presmanes Hill, and Kristen B Gorman. 2020. palmerpenguins: Palmer Archipelago (Antarctica) penguin data. https://doi.org/10.5281/zenodo.3960218 R package version 0.1.0.

[57] Torsten Hothorn, Kurt Hornik, and Achim Zeileis. 2006. Unbiased Recursive Partitioning: A Conditional Inference Framework. *Journal of Computational and Graphical Statistics* 15, 3 (2006), 651–674.

[58] Bill Howe, Julia Stoyanovich, Haoyue Ping, Bernease Herman, and Matt Gee. 2017. Synthetic Data for Social Good. arXiv, preprint. https://doi.org/10.48550/arXiv.1710.08874

[59] Jingchen Hu, Terrance D. Savitsky, and Matthew R. Williams. 2021. Risk-Efficient Bayesian Data Synthesis for Privacy Protection. *Journal of Survey Statistics and Methodology* 10, 5 (2021), 1370–1399.

[60] Mingze Huang, Christian L. Müller, and Irina Gaynanova. 2021. latentcor: An R Package for estimating latent correlations from mixed data types. *J. Open Source Softw.* 6, 65 (2021), 3634.

[61] Tobias Hyrup, Anton Danholt Lautrup, Arthur Zimek, and Peter Schneider-Kamp. 2023. Sharing is CAIRing: Characterizing Principles and Assessing Properties of Universal Privacy Evaluation for Synthetic Tabular Data. arXiv, preprint. https://doi.org/10.48550/arXiv.2312.12216

[62] James Jackson, Robin Mitra, Brian Francis, and Iain Dove. 2022. Using Saturated Count Models for User-Friendly Synthesis of Large Confidential Administrative Databases. *Journal of the Royal Statistical Society Series A: Statistics in Society* 185, 4 (2022), 1613–1643.

[63] Anuj Karpatne, Gowtham Atluri, James H. Faghmous, Michael Steinbach, Arindam Banerjee, Auroop Ganguly, Shashi Shekhar, Nagiza Samatova, and Vipin Kumar. 2017. Theory-Guided Data Science: A New Paradigm for Scientific Discovery from Data. *IEEE Transactions on Knowledge and Data Engineering* 29, 10 (2017), 2318–2331.

[64] Alan F. Karr, 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, 3 (2006), 224–232.

[65] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. 2020. Analyzing and Improving the Image Quality of StyleGAN. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020*. Computer Vision Foundation / IEEE, Seattle, WA, USA, 8107–8116.

[66] Dhamanpreet Kaur, Matthew Sobiesk, Shubham Patil, Jin Liu, Puran Bhagat, Amar Gupta, and Natasha Markuzon. 2021. Application of Bayesian networks to generate synthetic health data. *Journal of the American Medical Informatics Association* 28, 4 (2021), 801–811.

[67] Khalid S. Khan, Regina Kunz, Jos Kleijnen, and Gerd Antes. 2003. Five Steps to Conducting a Systematic Review. *Journal of the Royal Society of Medicine* 96, 3 (2003), 118–121.

[68] Diederik P. Kingma, Shakir Mohamed, Danilo Jimenez Rezende, and Max Welling. 2014. Semi-supervised Learning with Deep Generative Models. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*. MIT Press, Cambridge, MA, USA, 3581–3589.

[69] Barbara Kitchenham, O. Pearl Brereton, David Budgen, Mark Turner, John Bailey, and Stephen Linkman. 2009. Systematic literature reviews in software engineering – A systematic literature review. *Information and Software Technology* 51, 1 (2009), 7–15.

[70] Akim Kotelnikov, Dmitry Baranchuk, Ivan Rubachev, and Artem Babenko. 2023. TabDDPM: Modelling Tabular Data with Diffusion Models. In *International Conference on Machine Learning, ICML 2023 (Proceedings of Machine Learning Research, Vol. 202)*. PMLR, Honolulu, Hawaii, USA, 17564–17579.

[71] Hans-Peter Kriegel, Erich Schubert, and Arthur Zimek. 2017. The (black) art of runtime evaluation: Are we comparing algorithms or implementations? *Knowl. Inf. Syst.* 52, 2 (2017), 341–378.

[72] Carmen Lacave and Francisco J. Díez. 2002. A review of explanation methods for Bayesian networks. *The Knowledge Engineering Review* 17, 2 (2002), 107–127.

[73] Anton D Lautrup, Tobias Hyrup, Arthur Zimek, and Peter Schneider-Kamp. 2024. SynthEval: A Framework for Detailed Utility and Privacy Evaluation of Tabular Synthetic Data. arXiv, preprint. https://doi.org/10.48550/arXiv.2404.15821 Code available on GitHub

v1.4.1.

[74] Marta Lenatti, Alessia Paglialonga, Vanessa Orani, Melissa Ferretti, and Maurizio Mongelli. 2023. Characterization of Synthetic Health Data Using Rule-Based Artificial Intelligence Models. *IEEE Journal of Biomedical and Health Informatics* PP (2023), 1–9.

[75] Stefan Lenz, Moritz Hess, and Harald Binder. 2021. Deep generative models in DataSHIELD. *BMC Medical Research Methodology* 21 (2021), 16 pages.

[76] Mingchen Li, Di Zhuang, and J. Morris Chang. 2023. MC-GEN: Multi-level clustering for private synthetic data generation. *Knowl. Based Syst.* 264 (2023), 110239.

[77] Tennison Liu, Zhaozhi Qian, Jeroen Berrevoets, and Mihaela van der Schaar. 2023. GOGGLE: Generative Modelling for Tabular Data by Learning Relational Structure. In *The Eleventh International Conference on Learning Representations, ICLR 2023*. OpenReview.net, Kigali, Rwanda, 22 pages.

[78] Majlinda Llugiqi and Rudolf Mayer. 2022. An Empirical Analysis of Synthetic-Data-Based Anomaly Detection. In *Machine Learning and Knowledge Extraction - 6th IFIP TC 5, TC 12, WG 8.4, WG 8.9, WG 12.9 International Cross-Domain Conference, CD-MAKE 2022 (Lecture Notes in Computer Science, Vol. 13480)*. Springer, Vienna, Austria, 306–327.

[79] Tshilidzi Marwala, Eleonore Fournier-Tombs, and Serge Stinckwich. 2023. The Use of Synthetic Data to Train AI Models: Opportunities and Risks for Sustainable Development. arXiv, preprint. https://doi.org/10.48550/arXiv.2309.00652

[80] Justin Matejka and George W. Fitzmaurice. 2017. Same Stats, Different Graphs: Generating Datasets with Varied Appearance and Identical Statistics through Simulated Annealing. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. ACM, Denver, CO, USA, 1290–1294.

[81] Rudolf Mayer, Markus Hittmeir, and Andreas Ekelhart. 2020. Privacy-Preserving Anomaly Detection Using Synthetic Data. In *Data and Applications Security and Privacy XXXIV - 34th Annual IFIP WG 11.3 Conference, DBSec 2020 (Lecture Notes in Computer Science, Vol. 12122)*. Springer, Regensburg, Germany, 195–207.

[82] Daniel McDuff, Theodore Curran, and Achuta Kadambi. 2023. Synthetic Data in Healthcare. arXiv, preprint. https://doi.org/10.48550/arXiv.2304.03243

[83] Mehdi Mirza and Simon Osindero. 2014. Conditional Generative Adversarial Nets. arXiv, preprint. https://doi.org/10.48550/arxiv.1411.1784

[84] José Arnaldo Barra Montevechi, Afonso Teberga Campos, Gustavo Teodoro Gabriel, and Carlos Henrique dos Santos. 2021. Input Data Modeling: An Approach Using Generative Adversarial Networks. In *Winter Simulation Conference, WSC 2021*. IEEE, Phoenix, AZ, USA, 1–12.

[85] José Arnaldo Barra Montevechi, Gustavo Teodoro Gabriel, Afonso Teberga Campos, Carlos Henrique dos Santos, Fabiano Leal, and Michael E. F. H. S. Machado. 2022. Using Generative Adversarial Networks to Validate Discrete Event Simulation Models. In *Winter Simulation Conference, WSC 2022*. IEEE, Singapore, 2772–2783.

[86] Hajra Murtaza, Musharif Ahmed, Naurin Farooq Khan, Ghulam Murtaza, Saad Zafar, and Ambreen Bano. 2023. Synthetic data generation: State of the art in health care domain. *Computer Science Review* 48 (May 2023), 100546.

[87] Beata Nowok, Gillian M. Raab, and Chris Dibben. 2016. synthpop: Bespoke Creation of Synthetic Data in R. *Journal of Statistical Software* 74, 11 (2016), 1–26.

[88] Nari Park, Yeong Hyeon Gu, and Seong Joon Yoo. 2021. Synthesizing Individual Consumers′ Credit Historical Data Using Generative Adversarial Networks. *Applied Sciences* 11, 3 (2021), 1126.

[89] Vasileios C. Pezoulas, Nikolaos S. Tachos, George Gkois, Iacopo Olivotto, Fausto Barlocco, and Dimitrios I. Fotiadis. 2022. Bayesian Inference-Based Gaussian Mixture Models With Optimal Components Estimation Towards Large-Scale Synthetic Data Generation for In Silico Clinical Trials. *IEEE Open Journal of Engineering in Medicine and Biology* 3 (2022), 108–114.

[90] Haoyue Ping, Julia Stoyanovich, and Bill Howe. 2017. DataSynthesizer: Privacy-Preserving Synthetic Datasets. In *Proceedings of the 29th International Conference on Scientific and Statistical Database Management*. ACM, Chicago, IL, USA, 42:1–42:5.

[91] Zhaozhi Qian, Bogdan-Constantin Cebere, and Mihaela van der Schaar. 2023. Synthcity: facilitating innovative use cases of synthetic data in different data modalities. arXiv, preprint. https://doi.org/10.48550/arXiv.2301.07573

[92] Harrison Quick. 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 (2022), 596–617.

[93] Gillian M. Raab, Beata Nowok, and Chris Dibben. 2017. Guidelines for Producing Useful Synthetic Data. arXiv, preprint. https://doi.org/10.48550/arXiv.1712.04078

[94] Trivellore E. Raghunathan. 2021. Synthetic Data. *Annual Review of Statistics and Its Application* 8, 1 (2021), 129–140.

[95] Trivellore E Raghunathan, James M Lepkowski, John Van Hoewyk, Peter Solenberger, and John van Hoewyk. 2001. A multivariate technique for multiply imputing missing values using a sequence of regression models. *Survey methodology* 27, 1 (2001), 85–96.

[96] Debbie Rankin, Michaela Black, Raymond Bond, Jonathan Wallace, Maurice Mulvenna, and Gorka Epelde. 2020. Reliability of Supervised Machine Learning Using Synthetic Data in Health Care: Model to Preserve Privacy for Data Sharing. *JMIR Medical Informatics* 8, 7 (2020), e18910.

[97] Jerome P. Reiter. 2004. New Approaches to Data Dissemination: A Glimpse into the Future? *CHANCE* 17, 3 (2004), 11–15.

[98] Jerome P. Reiter. 2005. Using CART to Generate Partially Synthetic Public Use Microdata. *Journal of Official Statistics* 21, 3 (2005), 441–462.

[99] Luc Rocher, Julien M. Hendrickx, and Yves-Alexandre de Montjoye. 2019. Estimating the success of re-identifications in incomplete datasets using generative models. *Nature Communications* 10, 1 (2019), 3069.

[100] Natsuki Sano. 2022. Utility and Risk Evaluation of Synthetic Data by Orthogonal Transformation. *The Review of Socionetwork Strategies* 16 (2022), 71–79.

[101] Hoo-Chang Shin, Neil A. Tenenholtz, Jameson K. Rogers, Christopher G. Schwarz, Matthew L. Senjem, Jeffrey L. Gunter, Katherine P. Andriole, and Mark Michalski. 2018. Medical Image Synthesis for Data Augmentation and Anonymization Using Generative Adversarial Networks. In *Simulation and Synthesis in Medical Imaging - Third International Workshop, SASHIMI 2018, Held in Conjunction with MICCAI 2018 (Lecture Notes in Computer Science, Vol. 11037)*. Springer, Granada, Spain, 1–11.

[102] Duncan Smith, Mark Elliot, and Joseph W. Sakshaug. 2023. To Link or Synthesize? An Approach to Data Quality Comparison. *Journal of Data and Information Quality* 15, 2 (2023), 14:1–20.

[103] Joshua Snoke, Gillian M. Raab, Beata Nowok, Chris Dibben, and Aleksandra Slavkovic. 2018. General and specific utility measures for synthetic data. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 181, 3 (2018), 663–688.

[104] Hui Sun, Tianqing Zhu, Zhiqiu Zhang, Dawei Jin, Ping Xiong, and Wanlei Zhou. 2023. Adversarial Attacks Against Deep Generative Models on Data: A Survey. *IEEE Trans. Knowl. Data Eng.* 35, 4 (2023), 3367–3388.

[105] Lijun Sun and Alexander Erath. 2015. A Bayesian network approach for population synthesis. *Transportation Research Part C: Emerging Technologies* 61 (2015), 49–62.

[106] Bo-Chen Tai, Szu-Chuang Li, Yennun Huang, and Pang-Chieh Wang. 2022. Examining the Utility of Differentially Private Synthetic Data Generated using Variational Autoencoder with TensorFlow Privacy. In *27th IEEE Pacific Rim International Symposium on Dependable Computing, PRDC 2022*. IEEE, Beijing, China, 236–241.

[107] Shun Takagi, Tsubasa Takahashi, Yang Cao, and Masatoshi Yoshikawa. 2021. P3GM: Private High-Dimensional Data Release via Privacy Preserving Phased Generative Model. In *37th IEEE International Conference on Data Engineering, ICDE 2021*. IEEE, Chania, Greece, 169–180.

[108] Jennifer Taub, Mark Elliot, and Joseph W. Sakshaug. 2020. The Impact of Synthetic Data Generation on Data Utility with Application to the 1991 UK Samples of Anonymised Records. *Trans. Data Priv.* 13, 1 (2020), 1 – 23.

[109] Allan Tucker, Zhenchen Wang, Ylenia Rotalinti, and Puja Myles. 2020. Generating high-fidelity synthetic patient data for assessing machine learning healthcare software. *npj Digital Medicine* 3, 1 (2020), 147.

[110] Boris van Breugel, Trent Kyono, Jeroen Berrevoets, and Mihaela van der Schaar. 2021. DECAF: Generating Fair Synthetic Data Using Causally-Aware Generative Networks. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021*. Curran Associates, Inc., Virtual Event, 22221–22233.

[111] Rohit Venugopal, Noman Shafqat, Ishwar Venugopal, Benjamin Mark John Tillbury, Harry Demetrios Stafford, and Aikaterini Bourazeri. 2022. Privacy preserving Generative Adversarial Networks to model Electronic Health Records. *Neural Networks* 153 (2022), 339–348.

[112] Giorgio Visani, Giacomo Graffi, Mattia Alfero, Enrico Bagli, Federico Chesani, and Davide Capuzzo. 2022. Enabling Synthetic Data adoption in regulated domains. In *9th IEEE International Conference on Data Science and Advanced Analytics, DSAA 2022*. IEEE, Shenzhen, China, 1–10.

[113] Zhenchen Wang, Puja Myles, and Allan Tucker. 2021. Generating and evaluating cross-sectional synthetic electronic healthcare data: Preserving data utility and patient privacy. *Comput. Intell.* 37, 2 (2021), 819–851.

[114] Liyang Xie, Kaixiang Lin, Shu Wang, Fei Wang, and Jiayu Zhou. 2018. Differentially Private Generative Adversarial Network. arXiv, preprint. https://doi.org/10.48550/arxiv.1802.06739

[115] Lei Xu, Maria Skoularidou, Alfredo Cuesta-Infante, and Kalyan Veeramachaneni. 2019. Modeling Tabular data using Conditional GAN. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019*. Curran Associates, Inc., Vancouver, BC, Canada, 7333–7343.

[116] Andrew Yale, Saloni Dash, Karan Bhanot, Isabelle Guyon, John S. Erickson, and Kristin P. Bennett. 2020. Synthesizing Quality Open Data Assets from Private Health Research Studies. In *Business Information Systems Workshops - BIS 2020 International Workshops (Lecture Notes in Business Information Processing, Vol. 394)*. Springer, Colorado Springs, CO, USA, 324–335.

[117] Andrew Yale, Saloni Dash, Ritik Dutta, Isabelle Guyon, Adrien Pavao, and Kristin P. Bennett. 2019. Privacy Preserving Synthetic Health Data. In *27th European Symposium on Artificial Neural Networks, ESANN 2019*. i6doc.com, Bruges, Belgium, 10 pages.

[118] Andrew Yale, Saloni Dash, Ritik Dutta, Isabelle Guyon, Adrien Pavao, and Kristin P. Bennett. 2020. Generation and evaluation of privacy preserving synthetic health data. *Neurocomputing* 416 (2020), 244–255.

[119] Chao Yan, Yao Yan, Zhiyu Wan, Ziqi Zhang, Larsson Omberg, Justin Guinney, Sean D. Mooney, and Bradley A. Malin. 2022. A Multifaceted benchmarking of synthetic electronic health record generation models. *Nature Communications* 13, 1 (2022), 7609.

[120] Jinsung Yoon, Lydia N. Drumright, and Mihaela van der Schaar. 2020. Anonymization Through Data Synthesis Using Generative Adversarial Networks (ADS-GAN). *IEEE Journal of Biomedical and Health Informatics* 24, 8 (2020), 2378–2388.

[121] Mandi Yu, Yulei He, and Trivellore E Raghunathan. 2022. A Semiparametric Multiple Imputation Approach to Fully Synthetic Data for Complex Surveys. *Journal of Survey Statistics and Methodology* 10, 3 (2022), 618–641.

[122] Jun Zhang, Graham Cormode, Cecilia M. Procopiuc, Divesh Srivastava, and Xiaokui Xiao. 2017. PrivBayes. *ACM Transactions on Database Systems* 42, 4 (2017), 1–41.

[123] Ziqi Zhang, Chao Yan, and Bradley A. Malin. 2022. Membership inference attacks against synthetic health data. *Journal of Biomedical Informatics* 125 (2022), 103977.

[124] Yujin Zhu, Zilong Zhao, Robert Birke, and Lydia Y. Chen. 2022. Permutation-Invariant Tabular Data Synthesis. In *IEEE International Conference on Big Data, Big Data 2022.* IEEE, Osaka, Japan, 5855–5864.

Received 4 May 2023; revised 13 August 2024; accepted 4 November 2024