Metrics That Matter: A Practical Survey on Tabular Synthetic Data Evaluation

JIM ACHTERBERG[*†], Leiden University Medical Center, The Netherlands and Statistics Netherlands (CBS), The Netherlands
BRAM VAN DIJK[†], Leiden University Medical Center, The Netherlands
SAIF UL ISLAM, WMG, University of Warwick, United Kingdom
GREGORY EPIPHANIOU, WMG, University of Warwick, United Kingdom
CARSTEN MAPLE, WMG, University of Warwick, United Kingdom
MARKUS MUELLER, Erasmus University Rotterdam, The Netherlands
MARCEL HAAS, Leiden University Medical Center, The Netherlands
MARCO SPRUIT, Leiden University Medical Center, The Netherlands and Leiden Institute of Advanced Computer Science, The Netherlands

To determine whether tabular synthetic data (SD) can provide a viable alternative to real data, it is vital to adequately evaluate its quality. A wide variety of metrics exist to assess the three archetypal dimensions of SD evaluation: realism (fidelity), task-specific usefulness (utility), and remaining disclosure risk (privacy). Current work in SD generation often relies on the ad-hoc selection of evaluation metrics without a clear justification, while their suitability strongly depends on the dataset and other contextual factors. This work surveys the field of independent and identically distributed (i.i.d.) tabular SD evaluation, provides a 4-question framework to guide metric selection pertaining to the task, evaluation goal, tabular data characteristics, and domain of SD, and provides general practical recommendations on SD evaluation. Although this work focuses on tabular data, it also considers how the findings may generalize to other data modalities. Experiments on an illustrative dataset of electronic health records show how to bring our recommendations and metric selection framework into practice. Finally, we provide an accessible Python library for tabular SD evaluation which implements the most important metrics discussed in this work, to support researchers and practitioners seeking to evaluate SD holistically across all key dimensions: https://synthyverse.readthedocs.io/.

CCS Concepts: • **General and reference** → **Metrics**; **Evaluation**; **Validation**; • **Mathematics of computing** → *Distribution functions*; *Multivariate statistics*; • **Security and privacy** → *Privacy-preserving protocols*; *Pseudonymity, anonymity and untraceability*.

Additional Key Words and Phrases: Synthetic data, evaluation metrics, generative models, data fidelity, data utility, data privacy, privacy enhancing technology

[*]Corresponding author
[†]Equal contribution

Authors' Contact Information: Jim Achterberg, j.l.achterberg@lumc.nl, Leiden University Medical Center, Leiden, The Netherlands and Statistics Netherlands (CBS), The Hague, The Netherlands; Bram van Dijk, Leiden University Medical Center, Leiden, The Netherlands; Saif Ul Islam, WMG, University of Warwick, Coventry, United Kingdom; Gregory Epiphaniou, WMG, University of Warwick, Coventry, United Kingdom; Carsten Maple, WMG, University of Warwick, Coventry, United Kingdom; Markus Mueller, Erasmus University Rotterdam, Rotterdam, The Netherlands; Marcel Haas, Leiden University Medical Center, Leiden, The Netherlands; Marco Spruit, Leiden University Medical Center, Leiden, The Netherlands and Leiden Institute of Advanced Computer Science, Leiden, The Netherlands.

Manuscript submitted to ACM

ACM Reference Format:
Jim Achterberg, Bram van Dijk, Saif Ul Islam, Gregory Epiphaniou, Carsten Maple, Markus Mueller, Marcel Haas, and Marco Spruit. 2026. Metrics That Matter: A Practical Survey on Tabular Synthetic Data Evaluation. *ACM Comput. Surv.* XXX, XXX, Article XXX (July 2026), 34 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

Synthetic Data (SD) is data generated by an algorithm or mathematical model instead of a real-world process. Its purpose is to replace or augment Real Data (RD) whenever RD is scarce or inaccessible due to privacy or other concerns [60]. SD evaluation typically spans three dimensions: (1) realism (*Fidelity*), (2) task-specific usefulness (*Utility*), and (3) information leakage from RD (*Privacy*) [51, 60]. That is, SD should be similar to RD to ensure similar outcomes across a range of inferences drawn from the SD (high fidelity); should mimic the performance of RD in specific downstream tasks of interest (high utility); should not allow adversaries to infer sensitive information from the SD (high privacy-protection).

Unfortunately, however, SD evaluation metrics are often selected without appropriate reflection or motivation, leading to incorrect interpretation of results, and a lack of alignment in benchmarking methods [9, 62, 84, 118]. For example, statistical distance measures such as Jensen-Shannon Distance (JSD) and Wasserstein Distance (WSD) are often used interchangeably, and two-sample classifier tests use varying underlying classifiers [110], without reflection on which measure or method is most suitable for a given dataset or the drawbacks they might have. Regarding privacy, ad-hoc evaluations based on record-level distances are often conducted, even though SD poses additional privacy risks and should be adequately evaluated [117, 147]. Furthermore, recent work has documented hundreds of evaluation metrics already in use for assessing tabular SD [125], and although various overviews exist that try to reduce this complexity [23, 62, 94], they provide little guidance on choosing appropriate metrics given a particular dataset and context.

This article surveys existing work on SD evaluation, presents a 4-question framework to guide evaluation metric selection across contexts, provides general practical recommendations on SD evaluation, and illustrates SD evaluation in practice with a concrete example. The scope of this survey is cross-sectional, independent and identically distributed (i.i.d.) tabular SD. We therefore do not aim to provide comprehensive guidance for, e.g., longitudinal data, text, or images. Section 5 briefly discusses which high-level principles may transfer to other modalities, but modality-specific evaluation requires dedicated treatment beyond the scope of this work.

We refrain from the idea of a universally valid and applicable set of evaluation metrics, and instead keep a *user perspective* in mind by providing a framework for selecting appropriate metrics for a given dataset and context. We believe that this provides a more practical and helpful guide to SD evaluation. Second, we aim to provide *deeper insight* into which types of metrics are most applicable to various practical scenarios, thereby aiming to improve, homogenise, and enhance the robustness of SD evaluation. This approach makes our framework more robust to newly developed or newly introduced metrics as well, by evaluating their applicability across different scenarios. This is helpful, as we observe that many more statistical metrics are equally suitable for SD evaluation but are rarely used[1], as researchers tend to reproduce metrics they have seen in prior work, or those that have easily available software implementations.

Compared to previous works, this work is broad in its scope, in that it discusses evaluation metrics for all three archetypal dimensions (fidelity, utility, privacy), is not focused on one specific domain, provides explicit guidance on metrics selection through a structured framework, and provides an empirical illustration to showcase how to select and apply different evaluation metrics in practice. In contrast, previous works tend to focus on specific evaluation dimensions (e.g., privacy [94] or utility [72]), or specific domains (e.g., medical [14, 51, 62, 88]). Other works mainly summarize existing literature and do not focus on providing guidance for metric selection [68, 118]. To the best of our knowledge, this work is among the first practical, question-driven frameworks for selecting tabular SD metrics.

[1]A prime example is Jeffrey's divergence: a statistical distance measure which is similar to JSD but rarely used for SD evaluation.

Accompanying this survey, we provide the `synthyverse`[2], a Python library for tabular SD generation and evaluation. It provides an accessible API to implement the most important evaluation metrics discussed in this work, including sensible default parameters as per our findings. The `synthyverse` also wraps some of the most popular tabular SD generators from literature, to allow a single library to be used for both generation and evaluation.

The remainder of the paper is structured as follows: we first provide a categorization of SD evaluation metrics (Section 2), which serves as the foundation for the rest of the paper. Thereafter, we present a structured framework based on four key questions that will help researchers and practitioners to make a better-informed selection of SD evaluation metrics (Section 3). We continue by providing concrete practical recommendations for SD evaluation (Section 4). Then, we dedicate attention to how our findings may generalize to other data modalities (Section 5), and finally, conclude with a concrete illustration of SD generation and evaluation using a dataset of heart failure patients (Section 6).

2 Evaluation Metrics for Tabular Synthetic Data

This section provides a categorization of evaluation metrics for tabular SD. This is not intended to be an exhaustive list; it contains the most common and well-justified metrics from the literature and shows how they can be categorised. Many more similar metrics are currently not discussed but could easily be added.

Notation. We briefly introduce some notation relating to tabular synthetic data generation, which is later used to explain different evaluation dimensions. Let $\mathcal{D} = \{x^{(i)}\}_{i=1}^{n}$ denote a tabular dataset of i.i.d. samples $\mathbf{x} = (x^{(j)})_{j=1}^{d}$ drawn from unknown distribution $p(\mathbf{x})$. The goal of tabular synthetic data generation is to learn a parametric distribution $p_\theta(\mathbf{x})$ – although non-parametric alternatives exist – such that $p_\theta(\mathbf{x}) \sim p(\mathbf{x})$.

2.1 Fidelity

SD fidelity comprises its realism or similarity to RD. It can thereby also be considered a measure of SD general utility; when SD closely resembles RD, it should be possible to perform a range of (related) tasks approximately as successfully with the SD as with the RD. We first distinguish fidelity metrics which evaluate **Domain-Specific Fidelity**. Then, we discuss fidelity metrics which measure either **Univariate Fidelity**, **Bivariate Fidelity**, or **Multivariate Fidelity**. Here, the specific attention to tabular data becomes apparent: univariate and bivariate fidelity measures are especially relevant to structured data, and less so to unstructured data types. Lastly, we dedicate attention to fidelity metrics which disentangle **Realism vs. Diversity**. An overview can be found in Figure 1.

2.1.1 Domain-Specific Fidelity. Domain-specific fidelity assessments evaluate aspects of SD that are specific to the domain of the dataset. This is especially crucial when data validity is required, i.e., SD could (theoretically) belong to a real individual or entity. It can identify failure modes of SD which are missed by more general fidelity metrics that do not account for the domain of the dataset. The content of these assessments naturally vary a lot, but can be roughly classified into i) evaluating whether certain domain-specific relationships or logical rules hold in the SD [82], or ii) having domain-experts directly assess the realism of synthetic samples.

Yan et al. [144] provide insightful examples of logical domain-rule evaluation in the medical domain: systolic blood pressure should always be larger than diastolic blood pressure for a given patient, and pregnancy diagnosis codes should not be associated with genotypically male patients.

[2] https://synthyverse.readthedocs.io/

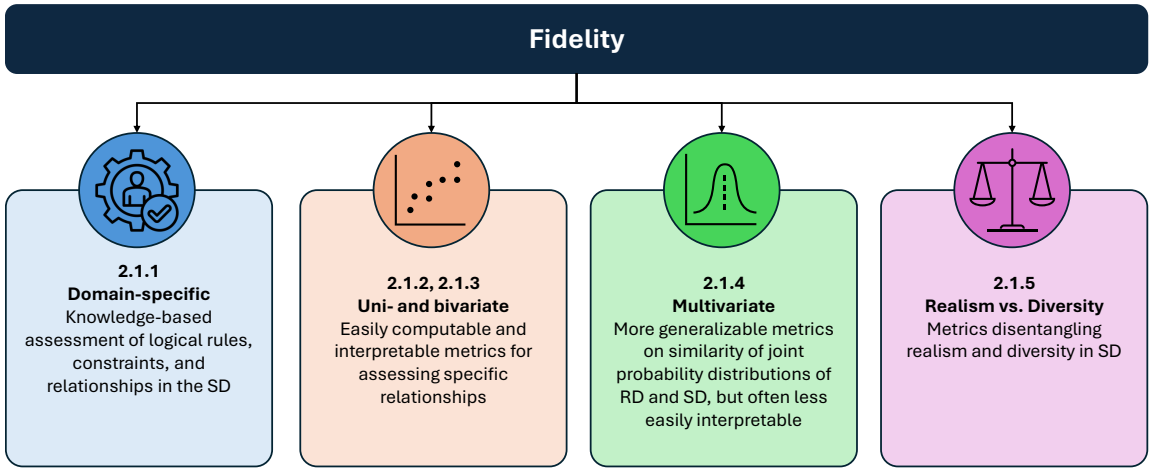

Fig. 1. Overview of categories of fidelity metrics used in this survey.

Choi et al. [20] and Wang et al. [130] provide examples of domain-expert evaluation: both set up an experimental procedure where target audiences blindly score SD realism. If SD is hard to discriminate from RD by domain-experts, this is interpreted as high domain-specific fidelity. However, directly evaluating tabular data becomes cumbersome and opaque in high-dimensions; in many scenarios, domain-expert evaluation instead considers derived statistics, logical rulesets, or otherwise aggregated information.

As the exact contents of these evaluations vary considerably per domain, we do not go deeper into this topic. However, for many domains, a large body of literature exists on exactly such domain-specific fidelity assessments, which need to be considered when evaluating SD. For example, in radiology, previous works provide extensive assessments by domain experts through, e.g., two-alternative forced-choice experiments [33, 81].

2.1.2 Univariate Fidelity. In structured datasets, individual variables are typically directly interpretable and provide important signal for downstream analyses. Evaluating the fidelity of individual variables in tabular SD is therefore crucial, i.e., whether $p_{real}(x_j) = p_{syn}(x_j),\ \forall j \in \{1, \ldots, d\}$. Many works directly compare marginal features of RD and SD through plots and descriptive statistics for an easily interpretable evaluation [1, 70, 109, 144, 154].

More in-depth analyses rely on statistical distance and similarity measures to provide compressed quantitative statistics on the similarity of the distributions of random variables. These can be presented per variable, or they can be aggregated, e.g., by taking the mean or median over feature-wise statistics. Popular choices are statistics with well-established distributions under the null hypothesis, allowing formal statistical testing, such as Kolmogorov-Smirnov statistics (KS) [7, 64, 105, 124, 149]. Other popular choices include, e.g., Wasserstein Distance (WSD) for numerical features and Jensen-Shannon Distance (JSD) for categorical features [70, 86, 158]. The widely popular Shape metric

from the SDMetrics library[3] relies on KS for numerical features and Total Variation Distance (TVD) for categorical features [46, 112, 152].

As statistical distance measures can also be used to assess multivariate fidelity, we go into more detail about the characteristics of individual measures in Section 2.1.4.

2.1.3 Bivariate Fidelity. Associations between individual variables in structured datasets are typically also directly interpretable and well-understood, and are therefore often evaluated to investigate whether SD accurately reflects relationships between individual features, i.e., whether $p_{real}(x_i, x_j) = p_{syn}(x_i, x_j)\ \forall j \in \{1, \ldots, d\},\ j \neq i$. Similar to univariate fidelity evaluation, the simplest and most interpretable methods include bivariate plots and simple statistics. A variety of correlation statistics can be used depending on the type of data considered, for example: Pearson's ρ or Spearman's r for numerical correlations, Cramér's V or Theil's U for categorical correlations, and correlation ratio η^2 for inter-type correlations [70, 86]. The widely popular Trend metric from the SDMetrics library[3] relies on Pearson's ρ for numerical correlations, and contingency similarity measured by TVD for inter-type and categorical correlations [46, 112, 152].

Unfortunately, some of the most commonly used correlation statistics turn out to be relatively uninformative for measuring bivariate fidelity. Scassola et al. [110] show that naive generative models without the capacity to learn any feature associations score very highly on SDMetrics' Trend metric. The underlying Pearson's ρ measures only linear association and is heavily influenced by the shape of the marginal distribution. Spearman's r *is* invariant to the shape of the marginals, but is only sensitive to monotonic dependence; for the trivial case $Y = X^2 \rightarrow\ r \approx 0$, while they are clearly associated. More informative bivariate fidelity metrics are invariant to univariate fidelity *and* sensitive to non-linear associations, such as mutual information-based measures. These had been proposed in Xu and Veeramachaneni [142], but were not widely adopted, as recently reiterated by Scassola et al. [110].

Certain domains tend to rely on more interpretable measures for bivariate associations. For example, association rule mining is often used in healthcare to extract if-then rules between pairs of individual features. Precision/recall analyses between the association rules extracted from SD and RD can then provide a measure for bivariate similarity [7, 8, 64, 146].

2.1.4 Multivariate Fidelity. Multivariate fidelity metrics indicate similarity between the full joint distribution of SD and RD, i.e., whether $p_{real}(\mathbf{x}, y) = p_{syn}(\mathbf{x}, y)$. They are sensitive to univariate, bivariate, ánd higher-order fidelity, and therefore provide a more informative view of fidelity than univariate and bivariate fidelity metrics alone. The latter are used more appropriately to disentangle fidelity failure modes or provide deeper insight, rather than providing conclusions on overall fidelity. Multivariate fidelity metrics are typically also much more generalizable to other data modalities, e.g., text and images, where uni- and bivariate fidelity are not as straightforward as for tabular data, or lack meaning.

Dimensionality Reduction Plots. An initial qualitative assessment of multivariate fidelity can be provided by plotting SD versus RD. This is typically done by first reducing the dimensionality of the data to a reasonable amount for plotting, e.g., 2 or 3. For tabular data, common dimensionality reduction techniques with the aim of plotting SD versus RD are PCA, tSNE, and UMAP [1, 66, 91, 97, 116, 128, 129, 137, 159]. Naturally, the choice of dimensionality reduction technique greatly influences in which respect and to what extent similarity between samples is preserved in the compressed representation, and correspondingly, which conclusions can be derived with respect to SD fidelity. For a more in-depth

[3] https://docs.sdv.dev/sdmetrics

discussion on dimensionality reduction techniques, we refer to Gisbrecht and Hammer [40]. Low-dimensional plots of multivariate data have the advantage of providing an interpretable qualitative view of multivariate fidelity, and of potential failure modes such as mode collapse [1]. However, they can be computationally costly and are prone to lose fine-grained structure (lossy compression), especially in high-dimensional datasets [76].

Statistical Distance and Similarity Measures. These measures tell something about the (dis)similarity between the probability distributions of SD and RD. Their univariate counterparts were also covered in the section on univariate fidelity metrics (Section 2.1.2). Note that we consider probabilistic distance measures rather than sample-wise geometric distance measures, as the latter measure similarity between individual samples, which is typically an indication of privacy or diversity loss instead of high fidelity in SD.

We can distinguish between statistical distance measures which operate on the PDF, CDF, or directly on the metric space. Measures which operate on the PDF require density estimation before they can be computed, e.g., Kullback-Leibler Divergence (KLD), JSD, TVD, and Hellinger distance. Multivariate density estimation remains a difficult problem in tabular data; it is as complex as learning the generative model for the SD itself. Therefore, these statistical distance measures are most popularly used for univariate fidelity assessment only, after simple feature-wise density estimation through, e.g., histogram-based methods. Similarly, measures which operate on the CDF, e.g., KS, are also mainly used for univariate fidelity assessment, since their multivariate extensions are less canonical and harder to interpret than in the one-dimensional case. Statistical distance measures which operate directly on the metric space, e.g., WSD and Maximum Mean Discrepancy (MMD), are used for multivariate fidelity assessment more often [59, 105]. Still, they are not a widely popular choice: WSD can be computationally costly as it requires solving an optimal transport problem (although faster regularized variants based on Sinkhorn divergences exist [19]), and is sensitive to the choice of cost function which defines the metric space. Similarly, MMD depends strongly on the kernel choice, which determines how distributions are embedded in the reproducing kernel Hilbert space, and thus which discrepancies are emphasized [43].

Finally, we note that the various statistical distance measures are suitable to detect different distributional failure modes. For example, in KLD, due to its asymmetric nature, whether we select SD to take the role of P or Q in $D_{KL}(P \,||\, Q)$ determines whether it favours mode seeking or covering behaviour of the SD distribution. JSD is a bounded and symmetrized version of KLD and provides an overall indication of distributional similarity. TVD measures the maximum absolute deviation between distributions, thereby being especially sensitive to distribution shifts. Hellinger distance calculates divergence based on square root probabilities and is especially sensitive to differences in distribution tails. WSD is especially sensitive to distributional geometry. These characteristics can be used to make a selection of statistical distance measures based on which distributional failure modes are especially of interest.

ML-based Similarity Measures. More popular choices for multivariate fidelity assessment in tabular SD are based on (un)supervised Machine Learning (ML) techniques. Most commonly, the Classifier 2-Sample Test (C2ST) [83], also referred to as Detection score [1, 64, 74, 77, 86, 112, 152]. In short, this test trains a binary classifier to distinguish SD samples from RD samples and assesses generalization performance on a holdout set. High accuracy indicates easily discriminated (and therefore dissimilar) joint distributions. The literature on this classifier test dates back over 20 years, and was originally intended as a proxy for a true multivariate goodness-of-fit test [36]. In the current SD literature, however, it is rarely used to apply a statistical test [1], and typically only reports a discriminatory metric (e.g., accuracy, ROCAUC, pMSE [115]) on a holdout set. The main advantages of this measure are the relatively simple procedure through supervised ML models, and its relatively low cost, which is achieved by encoding the samples from the multivariate distributions as single bits (0/1). The measure is, naturally, very sensitive to the chosen classifier. More

flexible/universal classifiers have power against more distributional differences, and are therefore typically more suitable to provide conclusions on overall fidelity. This was already noted by Friedman [36], but recently reiterated by Scassola et al. [110], since many recent works on tabular SD – as well as popular software implementations such as SDMetrics[3] – choose weak learners (logistic regression) to execute the C2ST [112, 152]. More flexible/universal classifiers for tabular data which should be considered can be based on, e.g., boosted decision trees [110]. We do, however, acknowledge that it can be useful to execute the C2ST using ML models with varying inductive biases and degrees of flexibility to provide further insight into structural fidelity. Furthermore, the level of interpretability of the classifier directly translates to the interpretability of the overall fidelity; using inherently interpretable classifiers such as shallow decision trees, or interpretability methods such as feature importances, can provide more insight into causes of fidelity loss.

Finally, we should note that this measure can provide a somewhat disheartening view of absolute SD fidelity in high-dimensional datasets, as the power of this test increases greatly with increasing feature size [36]. In such cases, it does remain an effective *relative* metric to compare different generative models.

Less widely used is the cluster analysis measure [138], which assesses whether SD and RD similarly reside in clusters constructed through unsupervised ML techniques [29, 41, 90, 144]. In short, this measure clusters the pooled SD and RD, and computes the squared error of the cluster-wise SD/RD proportions to the overall SD/RD proportion [138]. It indicates whether SD and RD induce a similar geometric/group structure in a metric space. Similar to the C2ST, it is sensitive to the chosen ML model; it is also sensitive to the choice of metric space and number of clusters. Generally, we do not find this clustering measure to provide a robust measure for SD fidelity. Firstly, it may construct clusters with similar SD/RD proportions, but with substantial within-cluster differences. Secondly, many (especially high-dimensional) datasets are not naturally clusterable, in which case this measure is not informative. Finally, choosing the best or most universal ML model (and corresponding hyperparameters) is much more difficult than for the C2ST, as generalization performance is difficult to evaluate for clustering algorithms.

Finally, some works evaluate the joint distribution via an aggregation of all conditional distributions, i.e., whether $p_{real}(x_j \mid x_{/j}) = p_{syn}(x_j \mid x_{/j}),\ \forall j \in \{1, \ldots, d\}$, through dimension-wise prediction. In short, this metric regresses each feature on all other features in both the SD and RD with an ML model, and evaluates generalization of those models to a real holdout set [7, 20, 65, 146, 157]. Note the overlap with some utility metrics (when the target feature corresponds to the target in downstream predictive analysis, Section 2.2.3) and privacy metrics (when target feature corresponds to target in attribute inference attack, Section 2.3.1).

2.1.5 Realism vs. Diversity. A portion of literature distinguishes between fidelity and diversity, where fidelity indicates whether SD is realistic or typical, and diversity whether SD covers the full range of the RD distribution. Disentangling these properties can be useful to provide further insight into potential failure modes of the generative process that produced the SD. In essence, realism and diversity both fall under our definition of fidelity: overall similarity between SD and RD comprises the full range of their probability distributions, and therefore both notions of realism and diversity.

Sajjadi et al. [107] first introduce precision-recall analyses for generative models, which distinguish whether SD is covered by RD (precision/realism) and whether RD is covered by SD (recall/diversity). Several adaptations have since been made to improve or generalize these measures; Kynkäänniemi et al.'s [71] Precision and Recall, Naeem et al.'s [89] Density and Coverage, and Alaa et al.'s [4] α-Precision and β-Recall are currently most widely used in tabular SD evaluation [45, 47, 67, 70, 112, 152]. These measures differentiate between whether SD samples belong to the RD manifold (realism) and whether RD samples belong to the SD manifold (diversity); they differ slightly in how they estimate the manifold, and whether or not samples belong to it. For example, Kynkäänniemi et al. [71] estimate

binary manifold membership for each SD/RD sample with respect to the opposing dataset's estimated manifold using k-nearest neighbour hyperspheres; typically, $k \in [3, 5]$. Naeem et al. [89] instead count the *total number* of SD samples in each RD hypersphere (Density), and estimate diversity (Coverage) from the perspective of RD hyperspheres instead of SD hyperspheres to limit the influence of SD outliers. Alaa et al. [4] instead estimate whether SD falls inside the hypersphere around the RD mean (or embedding center) containing the most central α fraction of RD (α-Precision), and integrate over all possible values of α to estimate whether SD is concentrated in high-density "typical" regions of RD (realism). β-Recall is calculated similarly from the SD perspective to indicate diversity[4].

As all the measures above rely heavily on sample-wise distances, they can be computationally costly in datasets with many samples or features. Furthermore, choosing an appropriate metric space is vital; we provide further discussion on this in Section 3.

Having discussed various fidelity measures, and how they can be disentangled, we now provide intuitions regarding the suitability of probabilistic distance and ML measures to detect various distributional failure modes. Figure 2 and Table 1 show how, for example, KLD and Coverage are especially sensitive to low diversity/coverage (mode collapse and overfitting), whereas KS and WSD are more sensitive to the shape of the distribution than to, e.g., differences in variance (under- and overfitting).

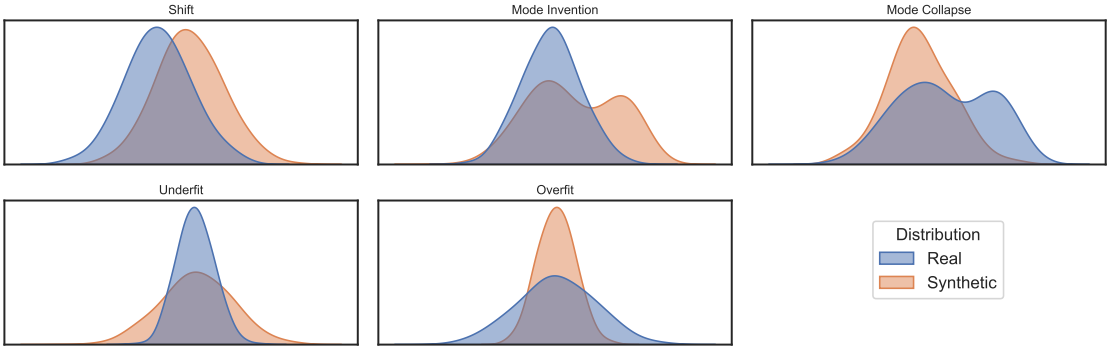

Fig. 2. Examples of different failure modes of SD, keeping overall similarity as measured by JSD relatively constant (see Table 1).

Table 1. Probabilistic distance measures corresponding to Figure 2 ($\mu \pm \sigma$ over 20 seeds). Toy data from (mixtures of) Gaussian distributions ($n = 1000$). Arrows indicate whether higher (↑) or lower (↓) metric values are desirable.

	↓ JSD	↓ KLD	↓ TVD	↓ KS	↓ WSD	↓ C2ST*	↑ Density	↑ Coverage
Shift	$0.360_{\pm 0.018}$	$0.987_{\pm 0.361}$	$0.337_{\pm 0.022}$	$0.346_{\pm 0.023}$	$0.115_{\pm 0.010}$	$0.705_{\pm 0.013}$	$0.997_{\pm 0.025}$	$0.866_{\pm 0.021}$
Mode Invention	$0.363_{\pm 0.013}$	$0.473_{\pm 0.059}$	$0.304_{\pm 0.009}$	$0.308_{\pm 0.007}$	$0.110_{\pm 0.011}$	$0.655_{\pm 0.012}$	$0.977_{\pm 0.040}$	$0.927_{\pm 0.012}$
Mode Collapse	$0.364_{\pm 0.015}$	$2.148_{\pm 1.317}$	$0.305_{\pm 0.012}$	$0.310_{\pm 0.012}$	$0.113_{\pm 0.007}$	$0.656_{\pm 0.009}$	$0.997_{\pm 0.018}$	$0.777_{\pm 0.021}$
Underfit	$0.363_{\pm 0.016}$	$0.453_{\pm 0.041}$	$0.316_{\pm 0.020}$	$0.177_{\pm 0.012}$	$0.061_{\pm 0.005}$	$0.680_{\pm 0.012}$	$0.908_{\pm 0.036}$	$0.912_{\pm 0.019}$
Overfit	$0.359_{\pm 0.015}$	$3.739_{\pm 0.979}$	$0.316_{\pm 0.017}$	$0.177_{\pm 0.008}$	$0.061_{\pm 0.004}$	$0.678_{\pm 0.013}$	$1.002_{\pm 0.010}$	$0.793_{\pm 0.025}$

*Using an XGBoost classifier.

[4] β-Recall also checks whether SD is locally covered by a RD hypersphere; whereas the original paper suggests $k = 5$, most implementations use $k = 2$, which correlates highly with metrics that indicate memorization, as real samples are only considered "recalled" if a synthetic sample is very close.

2.2 Utility

The utility of SD relates to its usefulness *in a specific task* or *set of tasks*. We note that utility in the literature is sometimes also used for metrics that seem to describe fidelity [24, 25, 109]. Since the description by Purdam and Elliot [101, p. 1102] of the loss of utility as the moment when *"a disclosure control method has changed a dataset to the point at which a user reaches a different conclusion from the same analysis"*, utility is more often understood in terms of the *inference* drawn from a specific analysis or task done with SD. Since utility is highly task-specific, we further distinguish utility metrics based on the data analysis-specific goal, i.e., **Explorative Analyses**, **Aetiological Analyses**, **Predictive Analyses**, and **Domain-Specific Analyses**. An overview is given in Figure 3.

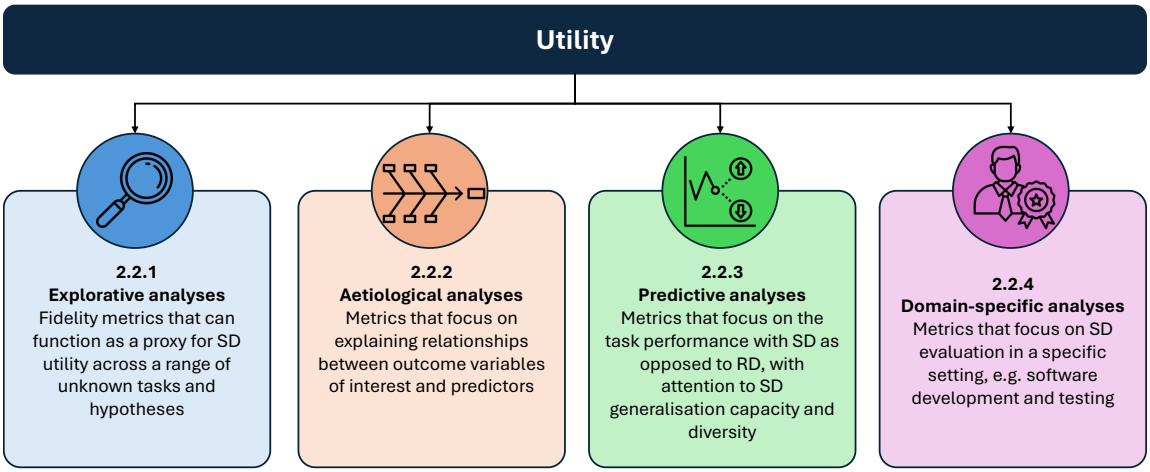

Fig. 3. Overview of categories of utility metrics used in this survey.

2.2.1 Explorative Analyses. Explorative analyses aim to formulate and explore a number of hypotheses or general insights from the SD. Similar to when the task is unknown, there is no specific inference the SD can be evaluated for. This means that no task-specific utility measures can be used directly, and analyses often rely on fidelity metrics as a proxy for usefulness - even though these can be severely limited in that sense [2]. Some common examples of SD generation for unknown or explorative downstream analyses are synthetic census data [78], synthetic Electronic Health Records (EHRs) [1, 144], but also the exploration how well different SD generators perform, i.e., benchmarking generative models according to their ability to accurately model a dataset's joint distribution function [2, 23].

2.2.2 Aetiological Analyses. Aetiological analyses reflect on how, and to what extent, predictions draw on explanatory variables. Here, the relationship between explanatory variables and the outcome is typically of main interest, rather than the accuracy of the predictions. Various utility metrics exist which indicate whether SD and RD similarly relate explanatory variables to outcomes: comparing regression coefficients or their subsequent hypotheses tests in SD and RD,

or computing feature importances (e.g., Shapley values) in SD and RD [22, 24, 26, 56, 63, 87, 92, 144]. Common examples where SD is generated with the aim of performing aetiological analyses are clinical trials [16, 30], epidemiological studies [13, 122], credit card fraud and network intrusion detection [5, 56]. In these contexts it is crucial that human decisions based on patterns in the SD are the same as those based on RD.

2.2.3 Predictive Analyses. Predictive analyses aim to accurately predict an outcome of interest, without necessarily relying on explainability. Predictive utility metrics typically indicate generalization capacity to RD when training a supervised ML model on SD or vice versa, i.e., evaluating whether $p_{real}(y|\mathbf{x}) = p_{syn}(y|\mathbf{x})$. This was formally introduced by Esteban et al. [32] as Train Synthetic Test Real (TSTR) [1, 2, 23, 61, 65, 66, 69, 77, 80, 104, 141, 149, 151] and Train Real Test Synthetic (TRTS) [65, 66, 114, 144] respectively. TSTR is sometimes also referred to as Machine Learning Efficiency/Efficacy (MLE) [112, 152]. Both approaches are usually compared to a baseline performance on purely RD (Train Real Test Real, TRTR). Unlike TSTR, TRTS is not sensitive to the diversity of the synthetic labels. TSTR is therefore often seen as the more important evaluation [32]. However, we acknowledge the potential usefulness of disentangling whether synthetic labels have adequate fidelity ánd diversity by computing both TSTR and TRTS. For example, when TRTS performance is high but TSTR low, this tells us that SD labels are realistic but not diverse enough to generalize well to RD [32]. Additionally, there are specific applications where TRTS more closely aligns with the intended goal, e.g., sharing SD with regulatory bodies to test clinical decision support systems trained on RD [38]. In these applications, reporting TRTS can be valuable, as it directly aligns with SD's intended task.

Some works also include Train Synthetic Test Synthetic (TSTS), which trains predictive algorithms on SD and scores them on a different independent set of SD [34, 114]. However, this is generally not a good measure of SD utility, as it does not indicate how well SD generalizes to real-world scenarios: high TSTS utility may be a result of equally poorly generated train *and* test data.

Choosing the appropriate supervised ML model is vital, as it directly influences the outcome of these metrics. For example, choosing a model that is less robust against overfitting may yield less pronounced differences between different SD generators as opposed to more robust prediction models [108]. On the other hand, using simpler models which are more likely to underfit may positively bias simpler SD generators, e.g., those which preserve only linear associations. Hence, it is best to select a set of models which are appropriate for the task at hand, dependent on, e.g., data type, complexity and domain. Then, reporting utility metrics for all appropriate models can provide additional insight into SD utility, similar to other supervised ML-based metrics such as the C2ST.

2.2.4 Domain-Specific Analyses. SD can also be used for various specific tasks which are not captured under the aforementioned categories (aetiological or predictive analyses). These are typically domain- or application-specific, such that they cannot easily be further categorized. One common example is using SD as test data within software systems [106]. As a general rule, SD utility evaluation should evaluate the discrepancy in results when using RD versus SD. This discrepancy should then be evaluated in light of the domain and task, which informs whether the magnitude of the discrepancy can be deemed acceptable. We provide more discussion on domain- and application-specific evaluation in Section 3.3.

2.3 Privacy

Information from RD can potentially leak through to SD, inducing privacy risks [117]. Common risk-inducing failure modes are memorization of the training samples and overfitting of the generative procedure (see also Section 2.1.5). However, privacy risk may occur even without any memorization or overfitting.

A common distinction regarding privacy risk is between **Attribute Disclosure** and **Membership Disclosure** [62]. In attribute disclosure, an attacker augments an available set of features (quasi-identifiers) with unknown sensitive attributes by mining information from SD. In membership disclosure, an attacker infers which records from an attack dataset were used in training the SD generator, which leaks sensitive information if membership to the SD training set is sensitive in and of itself.

Another type of disclosure risk is sometimes discussed, i.e., identity disclosure, where SD leads to complete re-identification of samples in RD. However, as mentioned by Pilgram et al. [99], identity disclosure in the context of SD is more precisely captured under attribute and membership disclosure, as reidentifying a specific record can lead to identifying additional sensitive attributes (attribute disclosure) *and* that the record was used to generate SD (membership disclosure). Privacy metrics which aim to indicate identity disclosure risk can thus be better thought of as indicating either attribute or membership disclosure or both, depending on the context, as also discussed by Taub et al. [120]. See Figure 4 for an overview of privacy metric categories.

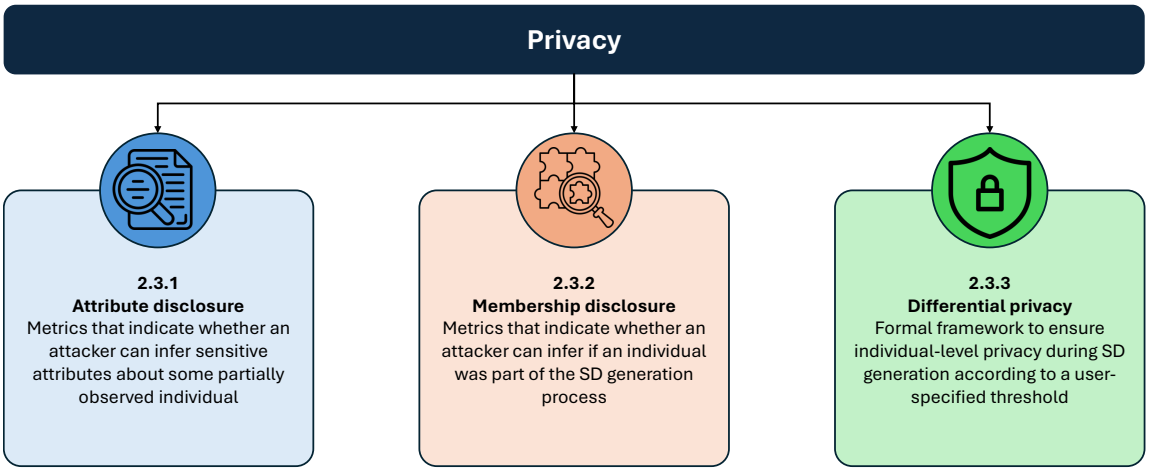

Fig. 4. Overview of categories of privacy metrics used in this survey.

2.3.1 Attribute Disclosure. Attackers can disclose sensitive attributes by employing an Attribute Inference Attack (AIA). AIAs learn the relationship between quasi-identifiers and sensitive attributes from SD, and then infer sensitive attributes in RD using the quasi-identifiers. Attribute inference risk will be considerable when i) there exists some significant association between the quasi-identifiers and sensitive feature and ii) the inference model generalizes well from SD to RD [53]. Therefore, attribute disclosure risk from SD can be considerable even when individual SD samples *do not* leak information (e.g., through memorization). Attribute disclosure risk is often more a property of the dataset and attacker's access to auxiliary data sources, than of the SD generation procedure.

A typical method for attribute inference is training supervised ML models on SD, with quasi-identifiers as the input features and a sensitive feature as the outcome, and performing inference on (a holdout split of) RD [1, 2, 20, 117, 144, 145, 155]. Flexible, inherently regularized or otherwise well-generalizing algorithms such as random forests or boosted trees are often favoured, as there is no way to test generalization capacity of different models due to a lack of ground-truth labels.

Instead of ML, much simpler methods for attribute inference are often also considered, which rely on sample-level similarity between SD and RD. Here, the idea is that when SD and RD can be linked on their quasi-identifiers on a sample-level, the sensitive attributes from SD are potentially good predictors for the sensitive attributes in RD. This relates to the notion of identity disclosure metrics mentioned before: when we can accurately link SD and RD on a sample-level this discloses the full identity of the sample, but more importantly, this discloses membership to the training set (membership disclosure); if the sensitive attributes in SD are close to those in RD, this incurs attribute disclosure as well.

The simplest metrics which implement this notion work with discrete tables and match SD to RD directly on discrete attributes [53, 120]. Other metrics can be based, more generally, on any geometric distance measure suitable to the dataset, e.g., Euclidean, Hamming, or Gower distance for numerical, discrete, or mixed-type data respectively. Many such metrics can be captured under the umbrella of Distance to Closest Record (DCR) metrics, which indicate whether SD samples are “too similar" to RD, or in other words, “memorize" [28, 50, 66, 73, 96, 114, 119, 157]. However, DCR metrics only indicate disclosure risk on a sample-level when compared to distances to other samples in its direct neighbourhood; geometric closeness does not incur disclosure risk when those types of samples are typical [99]. An example of a metric which compares to other samples in its direct neighbourhood is Nearest Neighbour Distance Ratio (NNDR) [93], which normalizes SD-RD distances to the next-nearest RD sample [73, 114]. Extremely low values indicate that these particular SD and RD samples are similar without other RD in its (relative) vicinity; this is a potential case of identity disclosure, and thereby, a potential case of membership as well as attribute disclosure. To assess whether SD truly isolates training points more than what should be expected, we can further normalize against a baseline NNDR between the training data and an independent holdout set (NNDR ratio).

Other similar DCR-based metrics which do not account for relative neighbourhood density are only informative when aggregated across the entire dataset; these are not effective to detect linkability of *specific* samples, from which identity and thus attribute disclosure may occur. Therefore, we discuss these metrics when discussing membership disclosure instead.

2.3.2 Membership Disclosure. Membership disclosure indicates whether an attacker can infer which RD was used in the SD generating procedure [88]. This leaks privacy when membership to the training set exposes sensitive attributes in and of itself. An example is when a health insurer infers that some patient’s data was used to generate SD for a particular illness, thereby exposing that the patient has that illness.

Building on the previous section, sample-level similarity metrics can indicate identity disclosure when compared to other samples in its neighbourhood, in which case membership is disclosed as well. Other than these, some measures are only informative on general disclosure risk across the entire dataset, most typically by providing a general indication of memorization.

Firstly, DCR Share indicates the proportion of RD which is closer to the SD than to a similar-sized holdout set [100]. This effectively executes a paired sign test between the DCR distributions of the RD to the SD, and the RD to the holdout set. Hereby it measures whether RD is more often closer to SD than to the holdout set, which is considered an

indication of memorization. Next to evaluating the location of the DCR distributions, we can also evaluate their tails, e.g., 2% and 5% quantiles [12], to obtain a robust measure.

Somewhat similarly, the Authenticity score [4] indicates the proportion of RD which is closer to RD (from the same set) than to SD [2, 6, 80, 102]. Nearest neighbour adversarial accuracy [143] balances the authenticity score with a second term indicating whether SD is closest to other SD [11, 18, 128, 136, 144, 148]. However, since there is no likely scenario when the first term is high but the second term low, the second term adds little practical value in terms of indicating privacy risk. The identifiability score [149] is calculated similarly to the authenticity score, but has the inverse connotation. It measures the proportion of RD which is closest to SD instead of other RD, and therefore, high scores indicate more instead of less disclosure risk [2, 48, 85, 111]. Also, "closeness" is defined by entropy-weighted distances such that more “identifiable" features (lower entropy) receive a higher score.

The measures above can all be considered to fall under the umbrella of DCR metrics, which have been highly scrutinized for evaluating privacy loss [147]. An important reason is that DCR-based metrics tend to treat all features equally, while Membership Inference Attacks (MIAs, see next paragraph) can learn and leverage which individual features are highly identifiable. Generally, satisfying DCR scores are no substitute for more elaborate evaluation through MIAs when thorough privacy evaluation is required. However, it can be useful as an initial cheap diagnostic: poor DCR Share provides a quick indication that the generative procedure may memorize the RD and will likely result in privacy loss. Finally, DCR metrics are naturally very sensitive to the choice of metric space; see Section 3 for more discussion.

Membership Inference Attacks. MIAs can be constructed to target specific samples and disclose whether they are members of the training set. They typically exploit the notion of local overfitting in SD, i.e., (groups of) samples which are overrepresented in SD and hence likely part of the training set. Firstly, we can distinguish between white-box and black-box attack settings, which indicate whether attackers do (white-box) or do not (black-box) have access to the SD generating model. For example, Hayes et al. [49] employ GAN discriminators to estimate a relative likelihood score that a specific sample is part of the training data or synthetic; in the white-box setting they use the original GAN’s discriminator, while in the black-box setting they train a new GAN from scratch on the SD.

Some efforts construct MIAs using geometric distances with a threshold [144, 156]. However, when these are not calibrated for distances to some auxiliary RD, they do not provide any meaningful notion of membership disclosure, and are not likely to lead to accurate inference attacks.

Many MIAs have been proposed, which typically all share a similar structure: train a model to output scores for samples in an attack dataset which reflect the relative likelihood that a sample belongs to either the synthetic distribution or the real distribution. The synthetic distribution can be directly estimated as the attacker is assumed to have access to the SD. The real distribution can either be estimated from a reference set of RD assumed to be available to the attacker [127], or the reference data is taken from the SD, and the synthetic distribution is estimated from the data generated by an additional generative model trained on the SD [49]. Now, a variety of methods have been proposed to learn the relative likelihood scoring function. As mentioned before, LOGAN [49] uses GAN discriminators, which output a relative likelihood score that a sample is either synthetic (1) or real (0). Similarly, but more generally, we can train *any* binary classifier on the synthetic and reference data to learn the scoring function in a classifier-based MIA [54, 73, 131]. Detecting Overfitting for Membership Inference Attacks against Synthetic Data (DOMIAS) [127] fits density estimators (e.g., KDE or neural autoregressive flows) on the synthetic and reference distributions and output $\frac{p_{syn}}{p_{ref}}$ as a scoring function. Data Plagiarism Index (DPI) [132] forms k-nearest neighbour hyperspheres around attack records and determine the scoring function from the proportion of synthetic to reference points within the

hyperspheres. Generative Likelihood Ratio Attack (Gen-LRA) [133] fits density estimators on the reference data with and without the attack record and evaluate the resulting likelihood ratio over the SD in its local region (k-nearest neighbour hypersphere) as the scoring function.

As shown by Ward et al. [134], there is no one MIA which consistently outperforms others; a conservative privacy risk assessment either evaluates various MIAs separately, or an ensemble combining various MIAs.

Finally, the classification threshold and reported performance measure greatly influence which conclusions can be drawn from an MIA. For example, choosing a low classification threshold may yield high-precision low-recall attacks; in some contexts this may be especially risky, as this discloses information on a few individuals with high confidence. An informative MIA report includes, at least, performance measures on overall discriminatory ability (e.g., ROCAUC) *and* performance at high-confidence thresholds (e.g., Lift@k or TPR@FPR=k). Additionally, calibration measures (e.g., Brier score) may be helpful to indicate whether sample-level predictions provide an accurate indication of membership inference risk.

2.3.3 Differential Privacy. Differential privacy is a formal mathematical framework to ensure a user-specified level of privacy *during* SD generation; it is not a post-hoc metric. In short, it quantifies the contribution of individual samples to the output of an algorithm [27]. In the SD context, this corresponds to the contribution of individual RD points to the generated SD. If this contribution is high, it will be easier to detect which RD was used to generate SD, invoking membership disclosure.

A specific level of differential privacy is typically enforced by injecting noise during the training process [139, 153]. This diminishes the contribution of individual training samples to the output, but also diminishes SD fidelity [139, 153]. Hereby, differential privacy potentially mitigates both membership ánd attribute disclosure.

The parameter $\epsilon \in \mathbb{R}_{\geq 0}$ in differential privacy sets the specified privacy level, with $\delta \in \mathbb{R}_{\geq 0}$ indicating the probability of violating said privacy level for any sample. High values of ϵ indicate low probability of getting the same output with or without a specific sample, meaning the sample is identifiable, and thus high privacy risk. An ϵ of 0 indicates no membership disclosure, but also results in poor fidelity SD due to the fidelity-privacy tradeoff inherent to differential privacy. Typically, $\epsilon \leq 1$ and δ less than some polynomial in the size of the dataset is seen as adequately privacy-preserving [139]. However, setting appropriate levels of differential privacy remains an open question. For any $\epsilon > 0$ some privacy risk may still exist [117], and similar ϵ can result in different privacy guarantees dependent on the domain and characteristics of the data [75]. Therefore, differential privacy parameters can typically not be considered as a sufficient privacy evaluation, and differentially private SD still needs to be evaluated according to the same metrics as SD without differential privacy [99].

All in all, various privacy metrics as discussed in this section exist to inform on privacy risks, which can be utilized to decide whether SD is safe enough to publish. It should be noted, however, that there are currently no clear thresholds or legal precedents which values of privacy metrics deem SD "safe enough", so assessments can, in effect, only be made on a case-by-case basis [10]. This is in part because it is still discussed among legal experts under what circumstances SD could be acknowledged as anonymous data for which, generally, privacy regulations do not apply. This in turn depends on legal definitions of “personally identifiable information" (PII) that may vary per country and sector. In the US context for example, current rule-based conceptions of PII do not take into account the data-generating process and further downstream processing, which do relate to privacy [37]. These issues could help explain the finding that a significant part of research on SD in for example the health domain, lacks a rigorous privacy evaluation [88]. Another factor may be the wrong presumption that SD is “privacy-secure by design" [37, 60, 62].

3 Selecting Evaluation Metrics

Which evaluation metrics are most suitable depends heavily on contextual factors. Here, we survey four of the most important factors by addressing the questions of how dataset characteristics, SD task, evaluation goal, and dataset domain influence the choice of evaluation metrics. We provide an overview of our approach in Figure 5.

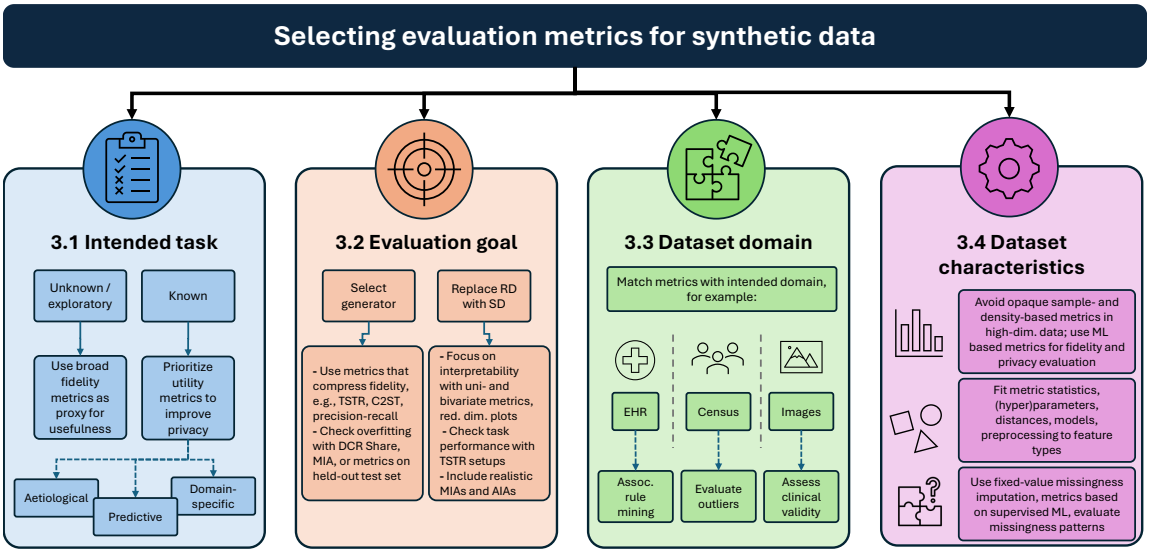

Fig. 5. Overview of the framework for metric selection.

3.1 What Type of Task is Envisioned for Synthetic Data?

Whether or not SD is generated with a specific task in mind, is paramount to how its usefulness should be evaluated. Without a specific task in mind (unknown or explorative task), SD usefulness can only be evaluated through fidelity metrics. Otherwise, for aetiological or predictive tasks, or domain-specific applications, utility measures can be used as well.

Fidelity metrics do not always correlate well with utility in the respective task, as they might not indicate similarity in the dimension of interest [140]. Additionally, reducing fidelity in a manner that maintains task-specific utility can ensure useful SD with stronger privacy guarantees, as a lower degree of similarity typically reduces privacy risk [2]. In task- or domain-specific scenarios, the main focus of SD evaluation should therefore be on relevant utility measures, rather than fidelity measures.

3.2 What is the Goal of Synthetic Data Evaluation?

SD can be evaluated to (i) select the best out of a set of potential data generators, or to (ii) report the quality of a single SD set intended to be published or shared. This essentially determines whether informative, compressed, quantitative metrics (i), or interpretable metrics (ii) have priority in evaluation. Note that these two goals usually do not happen in

isolation, since researchers will often select the best out of a set of potential SD generators – with potentially varying hyperparameters – *before* publishing the "best" SD set, accompanied by an interpretable evaluation report.

3.2.1 Selecting the best data generator. In case of (i), we should emphasise informative quantitative metrics that provide a compressed representation of fidelity, and utility if the intended analysis is known. Statistical fidelity measures (e.g., WSD, C2ST, precision-recall for distributions) provide a numerical representation of SD quality which can easily be compared across generators. These analyses, however, should always be accompanied by privacy metrics and fidelity evaluated on an independent holdout set (see Section 4.1) indicating generalization capacity, to ensure generators do not overfit and harm privacy. AIAs are likely less relevant here, as attribute disclosure is typically not a result of a generator failure (overfitting/memorization), which we aim to detect at this stage.

3.2.2 Replace RD with SD. In case of (ii), more interpretable metrics are often required. Regarding fidelity, absolute values of statistical distance and similarity measures are not well understood by humans. For example, a JSD between SD and RD of 0.25 is not very informative by itself. In order to inform a target audience such as privacy officers, regulators, or fellow researchers of the quality and safety of SD, metrics need to be interpretable or different values and thresholds need to be better understood and documented in the literature. Also note that, instead of selecting a single best SD generator and generating a single SD set in case of ii), it can be beneficial to make inferences on multiple sets of SD from ensembles of SD generators when publishing SD [126].

Structured data fares better than unstructured data regarding interpretability in that it allows separately emphasising univariate and bivariate fidelity. Univariate similarity provides a straightforward indication of what can be expected in the SD: which features are captured well, which values are generally higher, lower, and so on. The same goes for bivariate fidelity: which features correlate similarly as in the RD, and which do not. Dimensionality reduction plots can also be useful, since it provides a graphical representation of fidelity in a single plot, although underlying algorithms can be complex, making it difficult to understand what the differences between SD and RD signify.

If SD is to replace RD in context of a specific task, the task performance obtained with SD should typically match the performance of RD. Scenarios however exist in which the loss in performance could be deemed acceptable, for example because the task is situated in a low-stake context, such as education, training, or model development settings, where no critical decisions are based on SD.

Lastly, when replacing RD with SD, choosing privacy metrics is challenging. Because legal and governance frameworks (e.g., the GDPR) traditionally focus on harm for individuals (e.g., identity disclosure), it makes sense to always include metrics that mirror this focus (e.g., NNDR and MIAs). Still, because there is no obvious mapping between real and synthetic individuals in fully SD [120], as the RD is often unreleased, such metrics are hard to interpret from a legal and governance perspective, leading to a gap between research on SD generation and practice [3]. Furthermore, metrics and potential threshold values depend on the context and problem at hand. Differential privacy is an example that illustrates this tension: it quantifies the influence of an individual on the SD, but its guarantees are global, worst-case, and regulators may not know how to interpret ϵ in compliance terms.

Since MIAs and AIAs are part of the privacy evaluation when replacing RD with SD, we highlight some recommendations for more realistic attacks by Pilgram et al. [99]. For MIAs, attack datasets are often split from the RD and from the same population, but if that population has disease Y, then the assumption is that the attacker already knows this population has disease Y. More realistic MIAs include attack data from different distributions, though this may not always be feasible. In addition, many commonly used metrics (e.g., F1-score) are prevalence-dependent, and need calibration against a baseline MIA, e.g., predicting all attack records to be members.

With respect to AIAs, the information gained about some individual in the SD should be compared to information that can be gained from a naive baseline, e.g., using population data. Only if a sensitive feature can be predicted substantially more reliably with the SD than without, we can say privacy may be violated. Lastly, though intuitive, assuming maximum information for an AIA or MIA (access to all QIs), does not always imply the highest privacy risk [99].

3.3 What is the Domain of the Dataset?

We already mentioned that, in task-specific scenarios, utility measures are more important than fidelity measures (Section 3.1). However, even for more general scenarios where fidelity metrics are more applicable, we can use prior knowledge from the dataset domain to select more relevant metrics; especially when reporting the quality of SD intended to replace or augment RD (see Section 3.2). For example, association rule mining is widely applied in EHRs, whereas census data is mainly used for computing descriptive statistics, and clinical validity as assessed by humans is most important for medical images. It can be useful to place more emphasis on analyses common to the respective data domain to ensure SD is useful to the intended audience.

The dataset domain also influences which distributional characteristics are especially of interest, e.g., whether or not to evaluate outliers, anomalies, and missing values. In some domains, retaining outliers in SD is crucial for useful analyses, e.g., for rare diseases in health data, fraudulent transactions in financial data, or natural disasters in environmental data. In other domains, however, outliers are considered extremely identifiable and are preferably not propagated to SD with high fidelity, e.g., in census data [35], where outliers are either removed before SD generation or perturbed afterwards. Analysis of outlier similarity between SD and RD can be performed through any outlier detection method (e.g., isolation forests), and privacy metrics such as NNDR [93] can indicate whether specific outliers from RD are too closely mimicked in SD, potentially leaking sensitive information.

3.4 What are the Characteristics of the Dataset?

3.4.1 Sample Size and Dimensionality. Sample-wise distances and density estimation become more cumbersome and less meaningful in high-dimensional datasets [121]. Therefore, any metric which relies on these becomes more opaque, e.g., the clustering fidelity metric [138], various probabilistic distance measures, distributional precision-recall metrics, and MIAs such as DOMIAS [127]. High-dimensional (unstructured) data types such as text and images typically solve this by compressing data to a limited dimensionality through feature extraction using, e.g., pre-trained image- or text-encoders [52]. Such pre-trained encoders are not readily available for tabular data, and as such, feature extraction is rarely used as a precursor to further metric computation. However, there are some recent works which do explore tabular SD evaluation in a more compressed or meaningful embedding space, using a variety of methods, e.g., LLM-based embeddings [113], contrastive learning [95], or factor analysis [44].

More straightforward however, we can avoid metrics which rely on sample-wise distances or density estimation in high-dimensional datasets, and focus on, e.g., supervised ML techniques: the C2ST can be used for fidelity evaluation, and a classifier-based MIA for privacy evaluation. Assuming an appropriate and sufficiently regularized model is selected, supervised ML models can handle high-dimensional datasets relatively well.

Similar to high dimensionality, large sample size can put pressure on computational requirements, especially for metrics that depend on sample-wise distances. For other metrics that rely on density estimation or (un)supervised ML models, we can choose an appropriate method based on considerations regarding time complexity and performance

with respect to sample size. For example, deep learning-based methods are likely better considered in larger datasets than small ones.

3.4.2 Feature Types. Tabular data can consist of (ordinal) numerical features, (nominal) categorical features, or a combination of both. The types of features contained in a dataset directly influence which metrics to select, how to choose metric parameters, and how to preprocess the data.

Firstly, some metrics are more appropriate for numerical than categorical features, or vice versa. For example, as a univariate fidelity measure, KS statistics are more appropriate for numerical data, whereas Chi-square statistics are more appropriate for categorical data. Similarly, for bivariate fidelity measures, Pearson/Spearman correlation statistics can be considered for numerical correlations, Cramer/Theil statistics for categorical correlations, or correlation ratio η^2 for inter-type correlations.

Additionally, the feature types influence which *parameters* or *instance* of a metric to select. For example, metrics which directly rely on geometric distances, e.g., the clustering measure [138], Wasserstein distance cost function, MMD kernel, distributional precision-recall [71], DCR, DPI [132], need to choose a geometric distance measure which defines an appropriate metric space. Important factors to consider are the balance between distances of different feature types, and the domain of the dataset. Many previous works simply choose Euclidean distances while one-hot encoding any categorical features [86, 112, 152]. This can be suitable for purely numerical data, however, when categorical data is also included, this greatly overemphasizes differences in categorical features. A better alternative can be a Gower distance [42]; this may still over-emphasize categorical features, but to a much lesser extent [59]. For purely categorical data we can consider, e.g., a Hamming distance.

Similarly, metrics which rely on supervised ML models (e.g., C2ST, TSTR/TRTS, classifier-based MIA) need to choose an appropriate model and corresponding hyperparameters. Typically, we find that tree-based models provide a flexible solution for tabular data – under sensible default hyperparameters – as they handle both numerical and categorical data well, and are efficient to train. This is also shown in popular ML benchmarks such as TabArena [31].

3.4.3 Missingness. Missing values are common in tabular data. In this section we consider how to handle evaluation in the presence of missing values, and how to evaluate similarity in missingness patterns between SD and RD. We do *not* consider how to train a generator in the presence of missing values, or how to synthesize missing values.

There are a variety of methods to handle missing values in SD evaluation. Firstly, we can simply drop any rows containing missing values. This can be acceptable if there are very few unique rows with missing values, but may discard a lot of information. Another common approach is to impute missing values in the SD and RD. For many metrics, imputing with a fixed value (e.g., the RD mean) can be preferable: this ensures that metrics measure differences between the SD and RD, not the imputation strategy. This is different from imputing *before training the generator*. In the latter case, model-based imputation can be preferable, as to avoid artificially simplifying the RD distribution which is to be learned.

When imputing missing values, any further evaluation does not (implicitly) evaluate whether the missingness pattern is similar between SD and RD. However, to that end, we can simply add feature-wise binary missing indicators. Another approach is to rely on metrics which can directly deal with missing values. Notably, we can rely on metrics based on supervised ML models (C2ST, TSTR/TRTS, classifier-based MIA), where we select a model which can inherently handle missing data, such as XGBoost. Since this treats missing values separately, differences in missingness patterns are reflected directly in the evaluation.

Lastly, we can directly evaluate whether missingness patterns themselves are the same in SD and RD. Here, we can construct feature-wise binary missing value indicators, and perform fidelity evaluation on these binary features. This may consist of standard univariate or multivariate fidelity measures.

4 Practical Synthetic Data Evaluation

Having established the existing SD evaluation metrics and the criteria for selecting appropriate ones based on dataset characteristics and contextual factors, we now present some general practical guidelines for SD evaluation.

4.1 Evaluating Generalization and Overfitting

SD generators are trained to mimic a RD training set. However, high similarity between SD and this training set can be the result of overfitting, which can lead to various privacy-related issues (Section 2.3). To assess generalization of SD, fidelity can therefore be evaluated with respect to the training set *and* an independent RD test set [59]. Here, high fidelity with respect to the test set indicates that the SD truly captures the underlying distribution. High fidelity to the training set but relatively low fidelity to the test set may indicate overfitting and poor generalization, given that the test set comes from the same distribution as the training set. Other metrics such as DCR Share and MIAs more directly assess overfitting.

4.2 Uncertainty and Bias

The training process of many SD generators is stochastic and sensitive to random initialization. Secondly, SD generation and evaluation is typically performed on a random train-test split of the RD; results are naturally also sensitive to this random split. Finally, many SD generators base sampling on pseudo-random number generators with a fixed seed; this sampling seed is another potential source of variation and bias.

Combatting all sources of uncertainty and bias calls for a cross-validation procedure where SD generators are trained for multiple initialization and train-test split seeds, and we sample multiple datasets from each trained model using varying sampling seeds. However, we acknowledge that bias can never be fully mitigated, and that such a procedure can incur high computational costs. In practice, which source of bias is more urgent depends on the specific dataset and SD generator. For example, algorithms such as GANs typically incur more variation across initializations than, e.g., VAEs which tend to learn fuzzier distributions [126], whereas in regards to train-test splits, small datasets typically incur more bias than larger datasets due to higher variance across partitions.

Lastly, to quantify uncertainty, we can report the evaluation metric distribution over the cross-validation procedure described above [126].

4.3 Auditing

Next to reporting SD quality, evaluation metrics can be used to improve existing SD post-generation without any changes to the underlying generative model. For example, sample-level metrics can inform an auditing process which removes poor samples and queries additional (potentially good) samples from a generative model [4]. Not all metrics are suited for this, however. Tabular data fidelity can typically only be considered with respect to the full range of the distribution, and not on an individual level, which is in stark contrast to other data modalities such as text and images. Sample-level auditing measures should therefore inform on both realism and diversity, such as sample-level precision-recall type metrics [4]. Another solution is to perform auditing based on larger batches of SD, and iteratively select batches which best match the RD distribution [39].

5 Generalizing to Other Data Modalities

Many of the metrics discussed above can also be computed for other data modalities such as time-series, images, and text. Due to their inherent high-dimensionality compared to tabular data, a common approach is to first compress the dataset to a more compact or meaningful embedding space before computing distance- or density-based metrics. For unstructured data types such as images and text, these embedding spaces typically come from neural network encoders pre-trained on large datasets [52]. For structured data types such as time-series, where such pre-trained encoders are much less prevalent, these can come from self-supervised representation learning on the same dataset [57].

Other metrics provide plug-and-play adaptability to different types of data. Metrics which rely on supervised ML models (e.g., C2ST, TSTR, classifier-based MIA) can simply use an ML model with suitable inductive priors for the considered data type, for example, recurrent neural networks for sequential data (time-series, text) [150], or convolutional neural networks for spatial data (images) [15].

However, in many cases, modality-specific metrics are considered most important, especially if they correlate well with some important property of the data type. A prime example is Fréchet Inception Distance [52], which is often computed for images as it has been found to (somewhat consistently) correlate well with human perceptual quality. As our focus is tabular data, we leave a thorough overview of other modality-specific metrics for future work.

6 Illustrative Experiments: Electronic Health Records of Heart Failure Patients

Now, we show how our survey can inform effective and reliable SD evaluation for an illustrative dataset of tabular EHRs; privacy-preserving SD generation is a widely researched topic in this domain [1, 7, 74, 77, 128, 144, 145, 151]. These experiments illustrate how to leverage the framework to make decisions on metric selection. We therefore consider a somewhat simple EHR dataset, instead of considering *all* scenarios which might be relevant in the framework. Our code for these experiments is available at https://github.com/JimAchterbergLUMC/MetricsThatMatter.

We select a patient cohort from the Medical Information Mart for Intensive Care (MIMIC)-IV (version 3.1) [58]. We include admissions where patients received a heart failure diagnosis (ICD-9 code 428 or ICD-10 code I50). In total, we include 11,194 rows (admissions) with 12 variables, which have a 1:1 numerical:categorical proportion. Rows with missing numerical values are dropped. Variables contain information on demographic characteristics (age, sex, ethnicity, marital status), physiological measurements (bmi, systolic and diastolic blood pressure), and admission (type, location, length of stay, number of diagnoses, and whether patient passed away). We assume that the EHRs are synthesized for a (currently unknown) wide range of use cases.

6.1 Benchmarking Synthetic Data Generators

Our first goal is to select the best SD generator for this scenario by benchmarking several models. For illustrative purposes, we consider SMOTE [17], ARF [135], CTGAN [141], TVAE [141] and TabDDPM [70]. All generators are implemented through the `synthyverse` Python library accompanying this survey. Appendix A provides more details on hyperparameters and implementation.

Plugging the context of this experiment into the framework (Figure 5), it follows that we focus on fidelity metrics and disregard utility metrics as there is no known task (Q1), use compressed quantitative metrics rather than more interpretable but potentially less informative metrics as we benchmark different generators (Q2), omit AIAs from privacy evaluation as these are a property of the dataset's risk rather than the generator (Q2), and consider metrics and distance measures appropriate for mixed-type data (Q4). For now, we disregard further domain-specific metrics

(Q3), since we focus on compressed quantitative fidelity metrics to examine which generator performs best (Q2). Table 2 provides an overview of the selected metrics with their detailed set-up and corresponding justification based on the framework and other insights from this survey. Similar to the generators, all evaluation metrics are implemented through the `synthyverse` Python library accompanying this survey.

We randomly split the data into train:test sets in 8:2 proportion. To mitigate bias from random data splits and model initializations, we utilize 3 different training seeds. To mitigate sampling bias, we utilize 10 different sampling seeds for each trained model when querying synthetic datasets, resulting in 30 synthetic sets per generator type. We report the average results including standard deviations over all training and sampling seeds. The results from the experiments are provided in Table 3.

To evaluate whether differences between models are substantial within each metric, we evaluate 95% confidence intervals of the difference between each model and the best-performing model for that metric[5]. In short, we hierarchically bootstrap the metric results from the considered training and sampling seeds 10^4 times, and evaluate the difference in performance to the best-performing model. This provides a sampling distribution of the difference between the metric results for the considered training/sampling seed combinations. The confidence interval then constitutes the $[\frac{\alpha}{2}, 1-\frac{\alpha}{2}]$ quantile interval with $\alpha = 0.05$ from the sampling distribution. If 0 is contained in the confidence interval, this suggests no substantial difference to the best-performing model for that metric.

The results highlight the importance of using a carefully selected set of informative metrics, since depending on the context of the user, different generators can come out on top. For example, SMOTE provides decent fidelity at very low runtime, but also the highest privacy risk. The latter can be an issue even when privacy-preservation is not the main goal, as poor MIA and DCR Share performance suggests overfitting; this can be an issue even when we do not care about privacy risk. TabDDPM provides strongest overall performance: high fidelity at low privacy risk, and moderate training cost.

6.2 Replacing RD with SD: Evaluation Report for TabDDPM

Next we focus on replacing RD with SD (Q2); we report the quality of a single set of SD generated from TabDDPM to potential stakeholders in the healthcare domain. Plugging this new context into the framework (Figure 5), the main difference with Section 6.1 is that we now focus on *interpretable* and *domain-specific* metrics which are easily understood by stakeholders, and provide an indication of the *absolute* rather than relative fidelity and privacy risks of SD. Also, we include AIAs, as we are now interested in the practical privacy risk when SD is shared or published. We use a 1:1 train:test split to ensure a large independent holdout set in SD evaluation, which is particularly useful to construct a reasonable attack dataset for MIAs. Table 4 provides an overview of the selected metrics for this new scenario.

Firstly, we find that the logical domain constraint (Table 4) holds for 100% of patients in the SD (no violations). The feature-wise plots given in Figure 6 indicate that marginal distributions are generally similar between SD and RD. The correlation matrices given in Figure 7 indicate that the direction and magnitude of pairwise correlations are generally similar as well. However, the SD contains weaker correlation between length of stay (los) and number of diagnoses (n_diagnoses) than the RD. The association rule mining algorithm finds 43 rules in the RD and 45 rules in the SD, with a precision of 0.956 and recall of 1.000. The hallucinated rules found only in SD are: when sex=female → bmi=low, and when n_diagnoses=low → bmi=low.

[5]We choose this approach instead of a statistical test since results from different training and sampling seeds within each generator are highly correlated.

Table 2. Metric selection for Section 6.1 as informed by the framework (Figure 5).

Metric	**Details & Justification**
WSD (marginal)	Average Wasserstein-1 distance for numerical features. Evaluated on both train and test set to guard against overfitting. Quantitative univariate fidelity metric suitable for numerical data.
JSD (marginal)	Average JSD for categorical features. Evaluated on both train and test set to guard against overfitting. Quantitative univariate fidelity metric suitable for categorical data.
MI	Similarity in (weighted and normalized) mutual information between feature pairs [110]. Numerical features are discretized into 20 equal-width bins. Quantitative bivariate fidelity metric which is invariant to marginal shape and sensitive to non-linear associations.
C2ST	ROCAUC of an XGBoost classifier, which is trained for 500 rounds with early stopping (20 rounds) based on a 20% validation set. Quantitative multivariate fidelity metric which is relatively cheap to compute and sensitive to many distributional differences.
α-Precision	α-Precision score [4]. Quantitative fidelity metric which disentangles fidelity and diversity. Uses Gower distance for equal contribution of numerical and categorical features.
β-Recall	β-Recall score using $k = 5$ nearest neighbours for the local coverage hyperspheres [4]. Quantitative diversity metric which disentangles fidelity and diversity. Uses Gower distance for equal contribution of numerical and categorical features.
DCR Share	Proportion of SD closer to the train set than the test set. SD and train set are subsampled to the test set size such that DCR Share = 0.5 indicates that SD is equally close to train and test set. We repeat this 4 times to ensure that (in expectation) all synthetic and train samples are evaluated, and average the results. Cheap general measure for memorization/overfitting, which may in turn indicate membership disclosure. Uses Gower distance for equal contribution of numerical and categorical features.
NNDR Ratio	Ratio between the NNDR to the SD and the NNDR to the test set. Measure of identity disclosure, which can be a cause of attribute and membership disclosure. Also provides the ratio at the 5th quantile (ratio@0.05) as a robust measure of identity disclosure. Uses Gower distance for equal contribution of numerical and categorical features.
MIA	ROCAUC of a rank-averaged ensemble of 3 different MIAs: classifier-based (Random Forest), DOMIAS (KDE on PCA-transformed data), and DPI (using Gower distances). Non-members are taken from the test set for training and evaluation in 1:1 proportion. Members are subsampled from the SD and train set for training and evaluation respectively, in equal size to the non-members to ensure a balanced inference task. We repeat this procedure 5 times and average the results. Simulates membership disclosure risk when the attacker is unaware of which MIA performs best. Also reports lift (enrichment over random baseline) at the 1% and 5% most confident predictions (lift@0.01, lift@0.05) to indicate membership disclosure on small high-risk groups; since the inference task is balanced, the maximum enrichment is 100% (lift@k=2).

After evaluating fidelity, we move to thoroughly investigate privacy risk. As Section 3.2 indicates, we need to evaluate risk of i) identity disclosure, ii) membership disclosure, and iii) attribute disclosure, for which Table 5 provides the results. A DCR Share of 0.525 and an NNDR ratio of 1.064 indicate low risk of memorization and identity disclosure

Table 3. Benchmarking results of SD generators for the MIMIC-IV Heart Failure dataset. Shows $\mu \pm \sigma$ over all training and sampling seeds. Arrows indicate whether higher (↑) or lower (↓) metric values are desirable. Best results per metric are given in **blue**, worst in **orange**; for results in **bold black**, the 95% hierarchical CI of the diff between the **best** model and the respective model over training/sampling seed combinations contains 0, which suggests no substantial difference.

	SMOTE	ARF	CTGAN	TVAE	TabDDPM
↓ WSD (num, train)	$0.006 \pm_{0.000}$	**$0.005 \pm_{0.000}$**	$0.016 \pm_{0.004}$	**$0.017 \pm_{0.005}$**	$0.007 \pm_{0.001}$
↓ WSD (num, test)	$0.008 \pm_{0.001}$	**$0.005 \pm_{0.001}$**	$0.017 \pm_{0.004}$	**$0.018 \pm_{0.005}$**	$0.007 \pm_{0.000}$
↓ JSD (cat, train)	$0.046 \pm_{0.001}$	**$0.008 \pm_{0.001}$**	$0.097 \pm_{0.004}$	**$0.189 \pm_{0.009}$**	$0.013 \pm_{0.002}$
↓ JSD (cat, test)	$0.046 \pm_{0.002}$	**$0.018 \pm_{0.002}$**	$0.100 \pm_{0.004}$	**$0.188 \pm_{0.010}$**	$0.021 \pm_{0.003}$
↑ MI	**$0.993 \pm_{0.001}$**	$0.970 \pm_{0.001}$	$0.965 \pm_{0.002}$	**$0.900 \pm_{0.006}$**	$0.984 \pm_{0.001}$
↓ C2ST (AUC)	$0.898 \pm_{0.005}$	$0.963 \pm_{0.003}$	$0.983 \pm_{0.002}$	**$0.996 \pm_{0.000}$**	**$0.719 \pm_{0.011}$**
↑ α-Precision	$0.869 \pm_{0.004}$	$0.969 \pm_{0.005}$	$0.862 \pm_{0.022}$	**$0.611 \pm_{0.043}$**	**$0.989 \pm_{0.006}$**
↑ β-Recall	**$0.963 \pm_{0.003}$**	$0.895 \pm_{0.006}$	$0.715 \pm_{0.017}$	**$0.419 \pm_{0.007}$**	$0.877 \pm_{0.005}$
↓ DCR Share	**$0.624 \pm_{0.008}$**	$0.522 \pm_{0.006}$	**$0.504 \pm_{0.007}$**	$0.511 \pm_{0.011}$	**$0.502 \pm_{0.005}$**
↑ NNDR (ratio)	**$0.871 \pm_{0.006}$**	$1.041 \pm_{0.004}$	**$1.068 \pm_{0.008}$**	$1.054 \pm_{0.006}$	$1.041 \pm_{0.004}$
↑ NNDR (ratio@0.05)	**$0.326 \pm_{0.020}$**	$1.223 \pm_{0.032}$	$1.298 \pm_{0.031}$	**$1.388 \pm_{0.054}$**	$1.259 \pm_{0.031}$
↓ MIA (AUC)	**$0.537 \pm_{0.004}$**	$0.513 \pm_{0.009}$	**$0.506 \pm_{0.006}$**	**$0.501 \pm_{0.009}$**	**$0.503 \pm_{0.006}$**
↓ MIA (lift@0.01)	**$1.432 \pm_{0.071}$**	**$1.071 \pm_{0.118}$**	**$1.006 \pm_{0.088}$**	**$1.025 \pm_{0.114}$**	**$1.019 \pm_{0.100}$**
↓ MIA (lift@0.05)	**$1.268 \pm_{0.042}$**	$1.047 \pm_{0.050}$	$1.022 \pm_{0.033}$	**$0.964 \pm_{0.060}$**	$1.019 \pm_{0.033}$
↓ Training time (s)	-*	$45.474 \pm_{0.214}$	$50.188 \pm_{12.393}$	$27.825 \pm_{12.076}$	$31.963 \pm_{21.797}$
↓ Sampling time (s)	$1.929 \pm_{0.028}$	$6.364 \pm_{0.043}$	$0.053 \pm_{0.002}$	$0.034 \pm_{0.001}$	$1.576 \pm_{0.054}$

*SMOTE has no learned parameters and therefore no training time.

Table 4. Metric selection for Section 6.2 as informed by the framework (Figure 5).

Metric	**Details & Justification**
Logical domain constraints	Percentage of rows for which systolic blood pressure > diastolic blood pressure in SD. Should hold for all rows in realistic SD. Interpretable domain-specific fidelity metric.
Feature-wise plots	Histograms and barplots for numerical and categorical features respectively. Interpretable univariate fidelity metric.
Correlation matrices	Heatmaps of pairwise feature correlations. Uses Pearson's ρ for numerical correlations, Cramer's V for categorical correlations, and correlation ratio η^2 for mixed correlations. Interpretable bivariate fidelity metric.
Association rule mining	Precision and recall of association rules found in SD and RD through Apriori. Numerical features are discretized into 5 equal-width bins. Apriori uses minimum support of 0.3 and minimum confidence of 0.8. Domain-specific bivariate fidelity metric.
DCR Share	See Table 2.
NNDR Ratio	See Table 2.
MIA	See Table 2.
AIA	ROCAUC and R^2 of an ML-based AIA using XGBoost models.

across the entire SD set, respectively. Even for the 5% most isolated points, SD is not more isolating than an independent holdout set of real data (NNDR ratio@0.05 = 1.444), indicating low risk of identity disclosure. Next, we evaluate an ensemble of MIAs, where we select all features to be available to the attacker; although this isn't necessarily the most risky scenario [99], we observe that it is in this case, after testing various combinations of quasi-identifiers. We find

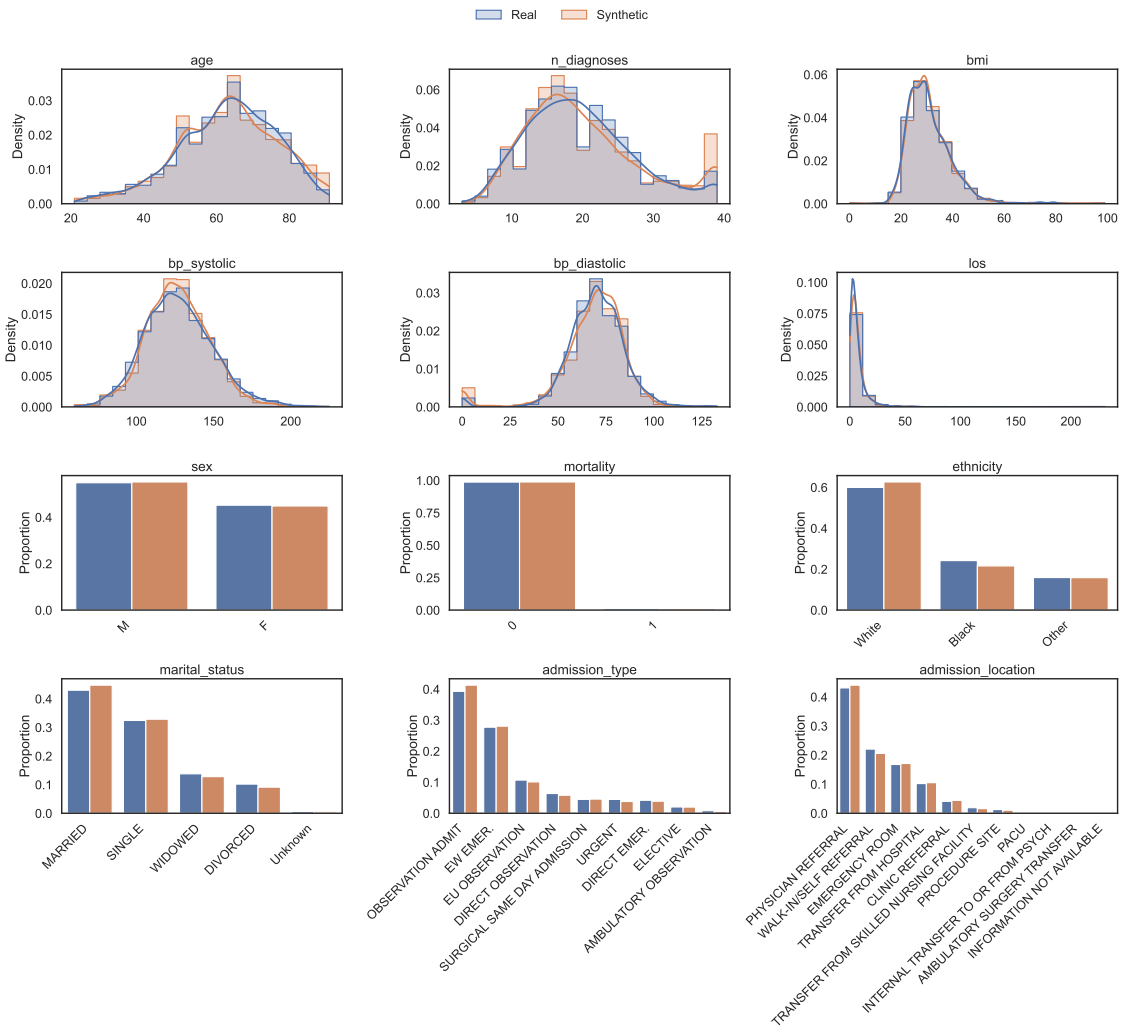

Fig. 6. Marginal distributions of the twelve features of the real and synthetic Heart Failure dataset.

little additional risk of membership disclosure over a random baseline, even for small high-risk groups; lift@0.01 and lift@0.05 are both < 2.5%. Finally, we investigate attribute disclosure risk through an AIA using an XGBoost model as predictor. Here, we assume that attackers might have access to – or are able to make an educated guess on – the quasi-identifiers age, bmi, and sex, and try to infer information on patients previous and current whereabouts through (sensitive) variables on admission location (admission_location) and length of stay (los). Our AIA achieves ROCAUC of 0.869 for admission location and R^2 score of -0.236 for length of stay, indicating that publishing this SD set can potentially inform attackers on patients' admission location, but likely not on length of stay, *for this specific set of quasi-identifiers.*

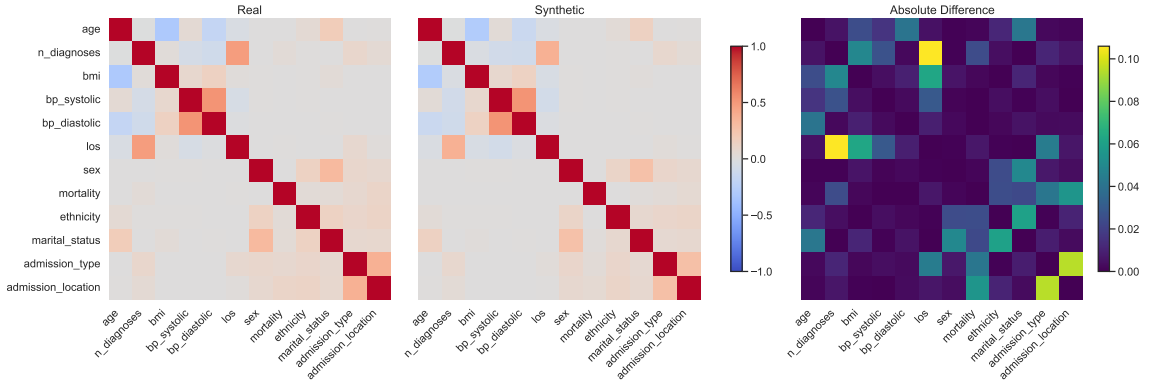

Fig. 7. Correlation matrices of the real and synthetic Heart Failure dataset.

Table 5. Privacy metrics results for replacing Heart Failure dataset with SD from TabDDPM.

	TabDDPM
DCR Share	0.525
NNDR (ratio)	1.064
NNDR (ratio@0.05)	1.444
MIA (AUC)	0.511
MIA (lift@0.01)	1.014
MIA (lift@0.05)	1.023
AIA los (R^2)	−0.236
AIA admission location (AUC)	0.869

6.3 Comparison With Previous Works

The metrics selected in these experiments are not the only valid set of metrics applicable to comparable scenarios. However, according to our suggestions, a valid set of metrics should inform on a variety of aspects which have been underemphasised in previous impactful works which focus on a comparable scenario, i.e., synthesizing mixed-type tabular health records. For example, previous works have placed limited focus on metrics which disentangle fidelity and diversity [7, 64, 123, 128, 144], set-up MIA scenarios without sufficient domain-specific justification or alignment with realistic threat models [144], or only assess a subset of the relevant privacy risk dimensions which should be evaluated (attribute, membership, and identity disclosure) [7, 64, 123, 128].

7 Discussion and Conclusion

Though SD have been widely recognized as offering the next best privacy-preserving technique for data sharing [55], proper evaluation of its fidelity, utility and privacy dimensions is all but straightforward. We mapped the contours of the large set of available evaluation metrics (Section 2) and argued that anyone looking to generate and employ SD should pause on questions about i) the different tasks that may be envisioned, ii) goals that should be obtained, relative to the iii) domain at issue and iv) characteristics of the dataset (Section 3). By setting up an illustrative experiment in line with general practical recommendations on evaluating SD, (Sections 4 and 6), we showed what choices regarding task, goal, domain and data characteristics solicit different kinds of metrics. Here we found, for example, that for a set of

tabular health records there is already quite some nuance in how well different univariate and multivariate distributions are synthesized, that is only visible on closer manual inspection. We have also provided examples of computing relative gains in membership disclosure and AIAs. Here, assumptions on naive baselines and member prevalence are made explicit to provide more context on increases or decreases of privacy risks.

New SD generation methods and evaluation metrics appear at a fast pace [e.g. 21, 98]. This also points towards the limitations of this survey. Recent work in SD evaluation, for example, employs metrics tailored to specific topics in fidelity evaluation, such as the extent to which minorities are represented (fairness) [79]. As we aimed to keep discussion of evaluation metrics general in terms of fidelity, utility and privacy, such more specific metrics were not discussed. Another example that we did not discuss are bounded multi-dimensional metrics designed to summarize fidelity, utility and privacy in a single score [21]. Although such metrics will benefit comparability and benchmarking of SD generators, it is unclear how such scores are viewed by legal and domain experts looking to incorporate SD.

Furthermore, we acknowledge that SD evaluation happens only in part on the side of research and development. The other side concerns regulatory and legal bodies, where important discussions take place which hitherto get little attention in the scientific community. An example is the distinction between generating and further processing SD. Since the former always draws on RD, models generating SD seem fully subject to the GDPR, although the extent to which the GDPR is applicable to downstream processing and use of SD is much less clear. This creates open questions about which parties should be held accountable when generating and using SD, and what metrics can tell a story about the due diligence done by all parties involved.

8 Acknowledgments

This work is co-funded by the HORIZON.2.1 - Health Programme of the European Commission, Grant Agreement number: 101095661 - Innovative applications of assessment and assurance of data and synthetic data for regulatory decision support (INSAFEDARE).

References

[1] Jim Achterberg, Marcel Haas, and Marco Spruit. 2024. On the evaluation of synthetic longitudinal electronic health records. *BMC Medical Research Methodology* 24 (2024), 181. https://doi.org/10.1186/s12874-024-02304-4

[2] Jim Achterberg, Marcel Haas, Bram van Dijk, and Marco Spruit. 2025. Fidelity-agnostic synthetic data generation improves utility while retaining privacy. *Patterns* (2025). https://doi.org/10.1016/j.patter.2025.101287

[3] Jim Achterberg, Bram van Dijk, Saif ul Islam, Hafiz Muhammad Waseem, Parisis Gallos, Gregory Epiphaniou, Carsten Maple, Marcel Haas, and Marco Spruit. 2025. The Data Sharing Paradox of Synthetic Data in Healthcare. In *Studies in Health Technology and Informatics*, Vol. 327. 582–586. https://doi.org/10.3233/SHTI250404

[4] Ahmed Alaa, Boris Van Breugel, Evgeny S Saveliev, and Mihaela van der Schaar. 2022. How faithful is your synthetic data? sample-level metrics for evaluating and auditing generative models. In *International Conference on Machine Learning*. PMLR, 290–306.

[5] Mohammad Ali and Jielun Zhang. 2024. Exploring the Effectiveness of Synthetic Data in Network Intrusion Detection through XAI. In *2024 Cyber Awareness and Research Symposium (CARS)*. IEEE, 1–5.

[6] Marco Aversa, Gabriel Nobis, Miriam Hägele, Kai Standvoss, Mihaela Chirica, Roderick Murray-Smith, Ahmed M. Alaa, Lukas Ruff, Daniela Ivanova, Wojciech Samek, Frederick Klauschen, Bruno Sanguinetti, and Luis Oala. 2023. DiffInfinite: Large Mask-Image Synthesis via Parallel Random Patch Diffusion in Histopathology. In *Advances in Neural Information Processing Systems*, A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (Eds.), Vol. 36. Curran Associates, Inc., 78126–78141. https://proceedings.neurips.cc/paper_files/paper/2023/file/f64927f5de00c47899e6e58c731966b6-Paper-Datasets_and_Benchmarks.pdf

[7] Mrinal Kanti Baowaly, Chia-Ching Lin, Chao-Lin Liu, and Kuan-Ta Chen. 2019. Synthesizing electronic health records using improved generative adversarial networks. *Journal of the American Medical Informatics Association* 26, 3 (2019), 228–241.

[8] Mrinal Kanti Baowaly, Chao-Lin Liu, and Kuan-Ta Chen. 2019. Realistic data synthesis using enhanced generative adversarial networks. In *2019 IEEE Second International Conference on Artificial Intelligence and Knowledge Engineering (AIKE)*. IEEE, 289–292.

[9] André Bauer, Simon Trapp, Michael Stenger, Robert Leppich, Samuel Kounev, Mark Leznik, Kyle Chard, and Ian Foster. 2024. Comprehensive exploration of synthetic data generation: A survey. *arXiv preprint arXiv:2401.02524* (2024).

[10] Steven M Bellovin, Preetam K Dutta, and Nathan Reitinger. 2019. Privacy and synthetic datasets. *Stan. Tech. L. Rev.* 22 (2019), 1.

[11] William W Booker, Dylan D Ray, and Daniel R Schrider. 2023. This population does not exist: learning the distribution of evolutionary histories with generative adversarial networks. *Genetics* 224, 2 (2023), iyad063.

[12] Alexander Theodorus Petrus Boudewijn, Andrea Filippo Ferraris, Daniele Panfilo, Vanessa Cocca, Sabrina Zinutti, Karel De Schepper, and Carlo Rossi Chauvenet. 2023. Privacy Measurements in Tabular Synthetic Data: State of the Art and Future Research Directions. In *NeurIPS 2023 Workshop on Synthetic Data Generation with Generative AI*. https://openreview.net/forum?id=DO8YT1pt4L

[13] Amy Elise Braddon, Suzanne Robinson, Rosa Alati, and Kim S Betts. 2023. 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* 37, 4 (2023), 292–300.

[14] Emmanuella Budu, Kobra Etminani, Amira Soliman, and Thorsteinn Rögnvaldsson. 2024. Evaluation of synthetic electronic health records: A systematic review and experimental assessment. *Neurocomputing* (2024), 128253.

[15] Philippe M Burlina, Neil Joshi, Katia D Pacheco, TY Alvin Liu, and Neil M Bressler. 2019. Assessment of deep generative models for high-resolution synthetic retinal image generation of age-related macular degeneration. *JAMA ophthalmology* 137, 3 (2019), 258–264.

[16] Shantanu Chandra, PKS Prakash, Subhrajit Samanta, and Srinivas Chilukuri. 2024. ClinicalGAN: powering patient monitoring in clinical trials with patient digital twins. *Scientific Reports* 14, 1 (2024), 12236.

[17] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. 2002. SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research* 16 (2002), 321–357.

[18] Yang Chen, Dustin J Kempton, Azim Ahmadzadeh, Junzhi Wen, Anli Ji, and Rafal A Angryk. 2022. CGAN-based synthetic multivariate time-series generation: a solution to data scarcity in solar flare forecasting. *Neural Computing and Applications* 34, 16 (2022), 13339–13353.

[19] Lenaic Chizat, Pierre Roussillon, Flavien Léger, François-Xavier Vialard, and Gabriel Peyré. 2020. Faster Wasserstein distance estimation with the Sinkhorn divergence. *Advances in Neural Information Processing Systems* 33 (2020), 2257–2269.

[20] Edward Choi, Siddharth Biswal, Bradley Malin, Jon Duke, Walter F Stewart, and Jimeng Sun. 2017. Generating multi-label discrete patient records using generative adversarial networks. In *Machine learning for healthcare conference*. PMLR, 286–305.

[21] Vikram S Chundawat, Ayush K Tarun, Murari Mandal, Mukund Lahoti, and Pratik Narang. 2022. A universal metric for robust evaluation of synthetic tabular data. *IEEE Transactions on Artificial Intelligence* 5, 1 (2022), 300–309.

[22] Saverio D'amico, Daniele Dall'Olio, Claudia Sala, Lorenzo Dall'Olio, Elisabetta Sauta, Matteo Zampini, Gianluca Asti, Luca Lanino, Giulia Maggioni, Alessia Campagna, et al. 2023. Synthetic data generation by artificial intelligence to accelerate research and precision medicine in hematology. *JCO Clinical Cancer Informatics* 7 (2023), e2300021.

[23] Fida K Dankar, Mahmoud K Ibrahim, and Leila Ismail. 2022. A multi-dimensional evaluation of synthetic data generators. *IEEE Access* 10 (2022), 11147–11158.

[24] Jorg Drechsler. 2022. Challenges in measuring utility for fully synthetic data. In *International Conference on Privacy in Statistical Databases*. Springer, 220–233.

[25] Jorg Drechsler and Anna-Carolina Haensch. 2024. 30 years of synthetic data. *Statist. Sci.* 39, 2 (2024), 221–242.

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

[27] Cynthia Dwork. 2006. Differential privacy. In *International colloquium on automata, languages, and programming*. Springer, 1–12.

[28] Khaled El Emam, Lucy Mosquera, and Xi Fang. 2022. Validating a membership disclosure metric for synthetic health data. *JAMIA open* 5, 4 (2022), ooac083.

[29] 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.

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

[31] Nick Erickson, Lennart Purucker, Andrej Tschalzev, David Holzmüller, Prateek Desai, David Salinas, and Frank Hutter. 2026. Tabarena: A living benchmark for machine learning on tabular data. *Advances in Neural Information Processing Systems* 38 (2026).

[32] Cristobal Esteban, Stephanie L Hyland, and Gunnar Ratsch. 2017. Real-valued (medical) time series generation with recurrent conditional gans. *arXiv preprint arXiv:1706.02633* (2017).

[33] Roberto Fedrigo, Fereshteh Yousefirizi, Ziping Liu, Abhinav K Jha, Robert V Bergen, Jean-Francois Rajotte, Raymond T Ng, Ingrid Bloise, Sara Harsini, Dan J Kadrmas, et al. 2023. Observer study-based evaluation of TGAN architecture used to generate oncological PET images. In *Medical Imaging 2023: Image Perception, Observer Performance, and Technology Assessment*, Vol. 12467. SPIE, 202–208.

[34] Mohammad Navid Fekri, Ananda Mohon Ghosh, and Katarina Grolinger. 2019. Generating energy data for machine learning with recurrent generative adversarial networks. *Energies* 13, 1 (2019), 130.

[35] Michael Freiman, Amy Lauger, and Jerome Reiter. 2017. Data synthesis and perturbation for the American Community Survey at the US Census Bureau. *US Census Bureau* (2017).

[36] Jerome H Friedman. 2003. On multivariate goodness-of-fit and two-sample testing. *Statistical Problems in Particle Physics, Astrophysics, and Cosmology* 1 (2003), 311.

[37] Michal S Gal and Orla Lynskey. 2023. Synthetic data: legal implications of the data-generation revolution. *Iowa L. Rev.* 109 (2023), 1087.

[38] Parisis Gallos, Nicholas Matragkas, Gregory Epiphaniou, Scott Hansen, Stuart Harrison, Bram van Dijk, Marcel Haas, Giorgos Pappous, Simon Brouwer, Francesco Torlontano, et al. 2024. INSAFEDARE Project: Innovative Applications of Assessment and Assurance of Data and Synthetic Data for Regulatory Decision Support. In *Digital Health and Informatics Innovations for Sustainable Health Care Systems*. IOS Press, 1193–1197.

[39] Daniel Gärber and Lea Demelius. 2025. Iterative Subset Selection for High-fidelity Synthetic Tabular Data. In *EurIPS 2025 Workshop: AI for Tabular Data*. https://openreview.net/forum?id=O3a8P07SmT

[40] Andrej Gisbrecht and Barbara Hammer. 2015. Data visualization by nonlinear dimensionality reduction. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 5, 2 (2015), 51–73.

[41] Andre Goncalves, Priyadip Ray, Braden Soper, Jennifer Stevens, Linda Coyle, and Ana Paula Sales. 2020. Generation and evaluation of synthetic patient data. *BMC medical research methodology* 20 (2020), 1–40.

[42] John C Gower. 1971. A general coefficient of similarity and some of its properties. *Biometrics* (1971), 857–871.

[43] Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. 2012. A kernel two-sample test. *The Journal of Machine Learning Research* 13, 1 (2012), 723–773.

[44] Morgan Guillaudeux, Olivia Rousseau, Julien Petot, Zineb Bennis, Charles-Axel Dein, Thomas Goronflot, Nicolas Vince, Sophie Limou, Matilde Karakachoff, Matthieu Wargny, et al. 2023. Patient-centric synthetic data generation, no reason to risk re-identification in biomedical data analysis. *NPJ Digital Medicine* 6, 1 (2023), 37.

[45] Manbir Gulati and Paul Roysdon. 2023. TabMT: Generating tabular data with masked transformers. In *Advances in Neural Information Processing Systems*, A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (Eds.), Vol. 36. Curran Associates, Inc., 46245–46254. https://proceedings.neurips.cc/paper_files/paper/2023/file/90debc7cedb5cac83145fc8d18378dc5-Paper-Conference.pdf

[46] Andrés Guzmán-Cordero, Floor Eijkelboom, and Jan-Willem van de Meent. 2025. Exponential Family Variational Flow Matching for Tabular Data Generation. In *Forty-second International Conference on Machine Learning*. https://openreview.net/forum?id=kjtvCSkSsy

[47] Jun Han, Zixiang Chen, Yongqian Li, Yiwen Kou, Eran Halperin, Robert E. Tillman, and Quanquan Gu. 2025. Guided Discrete Diffusion for Electronic Health Record Generation. *Transactions on Machine Learning Research* (2025). https://openreview.net/forum?id=N2rWhTgits

[48] Atiye Sadat Hashemi, Kobra Etminani, Amira Soliman, Omar Hamed, and Jens Lundström. 2023. Time-series anonymization of tabular health data using generative adversarial network. In *2023 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 1–8.

[49] Jamie Hayes, Luca Melis, George Danezis, and Emiliano De Cristofaro. 2019. LOGAN: Membership Inference Attacks Against Generative Models. *Proceedings on Privacy Enhancing Technologies* 1 (2019), 133–152.

[50] Mikel Hernadez, 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.

[51] 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.

[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. *Advances in neural information processing systems* 30 (2017).

[53] Markus Hittmeir, Rudolf Mayer, and Andreas Ekelhart. 2020. A baseline for attribute disclosure risk in synthetic data. In *Proceedings of the Tenth ACM Conference on Data and Application Security and Privacy*. 133–143.

[54] Florimond Houssiau, James Jordon, Samuel N Cohen, Owen Daniel, Andrew Elliott, James Geddes, Callum Mole, Camila Rangel-Smith, and Lukasz Szpruch. 2022. Tapas: a toolbox for adversarial privacy auditing of synthetic data. *arXiv preprint arXiv:2211.06550* (2022).

[55] Jiri Hradec, Massimo Craglia, Margherita Di Leo, Sarah De Nigris, Nicole Ostlaender, and Nicholas Nicholson. 2022. Multipurpose synthetic population for policy applications. *No. JRC128595* (2022).

[56] ASIF IQBAL and Biplab Sikdar. 2023. Are Classifiers Trained on Synthetic Data Reliable? An XAI Study. *Authorea Preprints* (2023).

[57] Paul Jeha, Michael Bohlke-Schneider, Pedro Mercado, Shubham Kapoor, Rajbir Singh Nirwan, Valentin Flunkert, Jan Gasthaus, and Tim Januschowski. 2022. PSA-GAN: Progressive self attention GANs for synthetic time series. In *The Tenth International Conference on Learning Representations*.

[58] Alistair Johnson, Lucas Bulgarelli, Tom Pollard, Steven Horng, Leo Anthony Celi, and Roger Mark. 2024. MIMIC-IV (version 3.1). *Physionet* (2024). https://doi.org/10.13026/kpb9-mt58

[59] Alexia Jolicoeur-Martineau, Kilian Fatras, and Tal Kachman. 2024. Generating and imputing tabular data via diffusion and flow-based gradient-boosted trees. In *International conference on artificial intelligence and statistics*. PMLR, 1288–1296.

[60] James Jordon, Lukasz Szpruch, Florimond Houssiau, Mirko Bottarelli, Giovanni Cherubin, Carsten Maple, Samuel N Cohen, and Adrian Weller. 2022. Synthetic Data–what, why and how? *arXiv preprint arXiv:2205.03257* (2022).

[61] James Jordon, Jinsung Yoon, and Mihaela Van Der Schaar. 2018. PATE-GAN: Generating synthetic data with differential privacy guarantees. In *International conference on learning representations*.

[62] Bayrem Kaabachi, Jérémie Despraz, Thierry Meurers, Karen Otte, Mehmed Halilovic, Bogdan Kulynych, Fabian Prasser, and Jean Louis Raisaro. 2025. A scoping review of privacy and utility metrics in medical synthetic data. *npj Digital Medicine* 8, 1 (2025), 60.

[63] Alan F Karr, Christine N Kohnen, Anna Oganian, Jerome P Reiter, and Ashish P Sanil. 2006. A framework for evaluating the utility of data altered to protect confidentiality. *The American Statistician* 60, 3 (2006), 224–232.

[64] 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.

[65] Shahzad Ahmed Khan, Hajra Murtaza, and Musharif Ahmed. 2024. Utility of GAN generated synthetic data for cardiovascular diseases mortality prediction: an experimental study. *Health and Technology* (2024), 1–24.

[66] Jaewon Kim, Hyunwoo Choo, Soo-Yong Shin, and Kyoung Doo Song. 2024. Synthesis and quality assessment of combined time-series and static medical data using a real-world time-series generative adversarial network. *Scientific Reports* 14, 1 (2024), 19064.

[67] Jayoung Kim, Chaejeong Lee, and Noseong Park. 2023. STaSy: Score-based Tabular data Synthesis. In *The Eleventh International Conference on Learning Representations*. https://openreview.net/forum?id=1mNssCWt_v

[68] A Kiran, P Rubini, and S Saravana Kumar. 2025. Comprehensive review of privacy, utility and fairness offered by synthetic data. *IEEE Access* (2025).

[69] Hendrik Klopries and Andreas Schwung. 2024. ITF-GAN: Synthetic time series dataset generation and manipulation by interpretable features. *Knowledge-Based Systems* 283 (2024), 111131.

[70] Akim Kotelnikov, Dmitry Baranchuk, Ivan Rubachev, and Artem Babenko. 2023. Tabddpm: Modelling tabular data with diffusion models. In *International Conference on Machine Learning*. PMLR, 17564–17579.

[71] Tuomas Kynkäänniemi, Tero Karras, Samuli Laine, Jaakko Lehtinen, and Timo Aila. 2019. Improved precision and recall metric for assessing generative models. *Advances in neural information processing systems* 32 (2019).

[72] Anton Danholt Lautrup, Tobias Hyrup, Arthur Zimek, and Peter Schneider-Kamp. 2024. Systematic review of generative modelling tools and utility metrics for fully synthetic tabular data. *Comput. Surveys* 57, 4 (2024), 1–38.

[73] Anton D Lautrup, Tobias Hyrup, Arthur Zimek, and Peter Schneider-Kamp. 2025. Syntheval: a framework for detailed utility and privacy evaluation of tabular synthetic data. *Data Mining and Knowledge Discovery* 39, 1 (2025), 6.

[74] Dongha Lee, Hwanjo Yu, Xiaoqian Jiang, Deevakar Rogith, Meghana Gudala, Mubeen Tejani, Qiuchen Zhang, and Li Xiong. 2020. Generating sequential electronic health records using dual adversarial autoencoder. *Journal of the American Medical Informatics Association* 27, 9 (2020), 1411–1419.

[75] Jaewoo Lee and Chris Clifton. 2011. How Much Is Enough? Choosing ϵ for Differential Privacy. In *Information Security*, Xuejia Lai, Jianying Zhou, and Hui Li (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 325–340.

[76] Joshua Lewis, Laurens Van der Maaten, and Virginia de Sa. 2012. A behavioral investigation of dimensionality reduction. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, Vol. 34.

[77] Jin Li, Benjamin J Cairns, Jingsong Li, and Tingting Zhu. 2023. Generating synthetic mixed-type longitudinal electronic health records for artificial intelligent applications. *NPJ Digital Medicine* 6, 1 (2023), 98.

[78] Claire Little, Richard Allmendinger, and Mark Elliot. 2024. Synthetic census microdata generation: A comparative study of synthesis methods examining the trade-off between disclosure risk and utility. *Journal of Official Statistics* (2024), 0282423X241266523.

[79] Qinyi Liu, Oscar Deho, Farhad Vadiee, Mohammad Khalil, Srecko Joksimovic, and George Siemens. 2025. Can synthetic data be fair and private? A comparative study of synthetic data generation and fairness algorithms. In *Proceedings of the 15th International Learning Analytics and Knowledge Conference*. 591–600.

[80] Tennison Liu, Zhaozhi Qian, Jeroen Berrevoets, and Mihaela van der Schaar. 2022. GOGGLE: Generative modelling for tabular data by learning relational structure. In *The Eleventh International Conference on Learning Representations*.

[81] Ziping Liu, Scott Wolfe, Zitong Yu, Richard Laforest, Joyce C Mhlanga, Tyler J Fraum, Malak Itani, Farrokh Dehdashti, Barry A Siegel, and Abhinav K Jha. 2023. Observer-study-based approaches to quantitatively evaluate the realism of synthetic medical images. *Physics in Medicine & Biology* 68, 7 (2023), 074001.

[82] Yunbo Long, Liming Xu, and Alexandra Brintrup. 2025. Evaluating Inter-Column Logical Relationships in Synthetic Tabular Data Generation. In *Will Synthetic Data Finally Solve the Data Access Problem?* https://openreview.net/forum?id=9FIOO09boS

[83] David Lopez-Paz and Maxime Oquab. 2017. Revisiting Classifier Two-Sample Tests. In *International Conference on Learning Representations*. https://openreview.net/forum?id=SJkXfE5xx

[84] Keith Man and Javaan Chahl. 2022. A review of synthetic image data and its use in computer vision. *Journal of Imaging* 8, 11 (2022), 310.

[85] Marko Miletic and Murat Sariyar. 2024. Assessing the Potentials of LLMs and GANs as State-of-the-Art Tabular Synthetic Data Generation Methods. In *Privacy in Statistical Databases*, Josep Domingo-Ferrer and Melek Önen (Eds.). Springer Nature Switzerland, Cham, 374–389.

[86] Markus Mueller, Kathrin Gruber, and Dennis Fok. 2025. Continuous Diffusion for Mixed-Type Tabular Data. In *The Thirteenth International Conference on Learning Representations*.

[87] Graciela Muniz-Terrera, Ofer Mendelevitch, Rodrigo Barnes, and Michael D Lesh. 2021. Virtual cohorts and synthetic data in dementia: an illustration of their potential to advance research. *Frontiers in Artificial Intelligence* 4 (2021), 613956.

[88] 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 (2023), 100546.

[89] Muhammad Ferjad Naeem, Seong Joon Oh, Youngjung Uh, Yunjey Choi, and Jaejun Yoo. 2020. Reliable Fidelity and Diversity Metrics for Generative Models. In *Proceedings of the 37th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 119)*, Hal Daumé III and Aarti Singh (Eds.). PMLR, 7176–7185. https://proceedings.mlr.press/v119/naeem20a.html

[90] I Nicholas, Hsien Kuo, Federico Garcia, Anders Sönnerborg, Michael Böhm, Rolf Kaiser, Maurizio Zazzi, Mark Polizzotto, Louisa Jorm, Sebastiano Barbieri, et al. 2023. Generating synthetic clinical data that capture class imbalanced distributions with generative adversarial networks: Example using antiretroviral therapy for HIV. *Journal of Biomedical Informatics* 144 (2023), 104436.

[91] Alexander Norcliffe, Bogdan Cebere, Fergus Imrie, Pietro Lio, and Mihaela van der Schaar. 2023. Survivalgan: Generating time-to-event data for survival analysis. In *International Conference on Artificial Intelligence and Statistics*. PMLR, 10279–10304.

[92] Beata Nowok, Gillian M Raab, and Chris Dibben. 2016. synthpop: Bespoke creation of synthetic data in R. *Journal of statistical software* 74 (2016), 1–26.

[93] Samson Otieno Ooko, Didacienne Mukanyiligira, Jean Pierre Munyampundu, and Jimmy Nsenga. 2021. Synthetic Exhaled Breath Data-Based Edge AI Model for the Prediction of Chronic Obstructive Pulmonary Disease. In *2021 International Conference on Computing and Communications Applications and Technologies (I3CAT)*. IEEE, 1–6.

[94] Pablo A Osorio-Marulanda, Gorka Epelde, Mikel Hernandez, Imanol Isasa, Nicolas Moreno Reyes, and Andoni Beristain Iraola. 2024. Privacy mechanisms and evaluation metrics for Synthetic Data Generation: A systematic review. *IEEE Access* (2024).

[95] Milton Nicolás Plasencia Palacios, Sebastiano Saccani, Gabriele Sgroi, Alexander Boudewijn, and Luca Bortolussi. 2025. Contrastive learning-based privacy metrics in tabular synthetic datasets. *arXiv preprint arXiv:2502.13833* (2025).

[96] Noseong Park, Mahmoud Mohammadi, Kshitij Gorde, Sushil Jajodia, Hongkyu Park, and Youngmin Kim. 2018. Data synthesis based on generative adversarial networks. *Proceedings of the VLDB Endowment* 11 (2018), 1071–1083.

[97] Hengzhi Pei, Kan Ren, Yuqing Yang, Chang Liu, Tao Qin, and Dongsheng Li. 2021. Towards generating real-world time series data. In *2021 IEEE International Conference on Data Mining (ICDM)*. IEEE, 469–478.

[98] Vasileios C Pezoulas, Dimitrios I Zaridis, Eugenia Mylona, Christos Androutsos, Kosmas Apostolidis, Nikolaos S Tachos, and Dimitrios I Fotiadis. 2024. Synthetic data generation methods in healthcare: A review on open-source tools and methods. *Computational and structural biotechnology journal* (2024).

[99] Lisa Pilgram, Fida K Dankar, Jorg Drechsler, Mark Elliot, Josep Domingo-Ferrer, Paul Francis, Murat Kantarcioglu, Linglong Kong, Bradley Malin, Krishnamurty Muralidhar, et al. 2025. A Consensus Privacy Metrics Framework for Synthetic Data. *arXiv preprint arXiv:2503.04980* (2025).

[100] Michael Platzer and Thomas Reutterer. 2021. Holdout-based empirical assessment of mixed-type synthetic data. *Frontiers in big Data* 4 (2021), 679939.

[101] Kingsley Purdam and Mark Elliot. 2007. A case study of the impact of statistical disclosure control on data quality in the individual UK Samples of Anonymised Records. *Environment and Planning A* 39, 5 (2007), 1101–1118.

[102] Zhaozhi Qian, Thomas Callender, Bogdan Cebere, Sam M Janes, Neal Navani, and Mihaela van der Schaar. 2024. Synthetic data for privacy-preserving clinical risk prediction. *Scientific Reports* 14, 1 (2024), 25676.

[103] Zhaozhi Qian, Rob Davis, and Mihaela van der Schaar. 2024. Synthcity: a benchmark framework for diverse use cases of tabular synthetic data. *Advances in Neural Information Processing Systems* 36 (2024).

[104] Sina Rashidian, Fusheng Wang, Richard Moffitt, Victor Garcia, Anurag Dutt, Wei Chang, Vishwam Pandya, Janos Hajagos, Mary Saltz, and Joel Saltz. 2020. SMOOTH-GAN: towards sharp and smooth synthetic EHR data generation. In *Artificial Intelligence in Medicine: 18th International Conference on Artificial Intelligence in Medicine, AIME 2020, Minneapolis, MN, USA, August 25–28, 2020, Proceedings 18*. Springer, 37–48.

[105] Antonio J Rodriguez-Almeida, Himar Fabelo, Samuel Ortega, Alejandro Deniz, Francisco J Balea-Fernandez, Eduardo Quevedo, Cristina Soguero-Ruiz, Ana M Wägner, and Gustavo M Callico. 2022. Synthetic patient data generation and evaluation in disease prediction using small and imbalanced datasets. *IEEE Journal of Biomedical and Health Informatics* (2022).

[106] Guru Pramod Rusum and Sunil Anasuri. 2023. Synthetic Test Data Generation Using Generative Models. *International Journal of Emerging Trends in Computer Science and Information Technology* 4, 4 (2023), 96–108.

[107] Mehdi SM Sajjadi, Olivier Bachem, Mario Lucic, Olivier Bousquet, and Sylvain Gelly. 2018. Assessing generative models via precision and recall. *Advances in neural information processing systems* 31 (2018).

[108] Gabriele Santangelo, Giovanna Nicora, Riccardo Bellazzi, Arianna Dagliati, et al. 2024. SynthCheck: A Dashboard for Synthetic Data Quality Assessment.. In *BIOSTEC (2)*. 246–256.

[109] Fatima Jahan Sarmin, Atiquer Rahman Sarkar, Yang Wang, and Noman Mohammed. 2025. Synthetic data: revisiting the privacy-utility trade-off. *International Journal of Information Security* 24, 4 (2025), 156. https://doi.org/10.1007/s10207-025-01072-6

[110] Davide Scassola, Dylan Ponsford, Adrián Javaloy, Sebastiano Saccani, Luca Bortolussi, Henry Gouk, and Antonio Vergari. 2026. A Sobering Look at Tabular Data Generation via Probabilistic Circuits. *arXiv preprint arXiv:2603.23016* (2026).

[111] Jingpu Shi, Dong Wang, Gino Tesei, and Beau Norgeot. 2022. Generating high-fidelity privacy-conscious synthetic patient data for causal effect estimation with multiple treatments. *Frontiers in Artificial Intelligence* 5 (2022), 918813.

[112] Juntong Shi, Minkai Xu, Harper Hua, Hengrui Zhang, Stefano Ermon, and Jure Leskovec. 2025. TabDiff: a Mixed-type Diffusion Model for Tabular Data Generation. In *The Thirteenth International Conference on Learning Representations*. https://openreview.net/forum?id=swvURjrt8z

[113] Andrey Sidorenko, Michael Platzer, Mario Scriminaci, and Paul Tiwald. 2025. Benchmarking synthetic tabular data: A multi-dimensional evaluation framework. *arXiv preprint arXiv:2504.01908* (2025).

[114] Jayanth Sivakumar, Karthik Ramamurthy, Menaka Radhakrishnan, and Daehan Won. 2023. GenerativeMTD: A deep synthetic data generation framework for small datasets. *Knowledge-Based Systems* 280 (2023), 110956.

[115] 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.

[116] Jian Song, Hongruixuan Chen, and Naoto Yokoya. 2024. SyntheWorld: A Large-Scale Synthetic Dataset for Land Cover Mapping and Building Change Detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 8287–8296.

[117] Theresa Stadler, Bristena Oprisanu, and Carmela Troncoso. 2022. Synthetic data–anonymisation groundhog day. In *31st USENIX Security Symposium (USENIX Security 22)*. 1451–1468. https://www.usenix.org/conference/usenixsecurity22/presentation/stadler

[118] Michael Stenger, Robert Leppich, Ian Foster, Samuel Kounev, and André Bauer. 2024. Evaluation is key: a survey on evaluation measures for synthetic time series. *Journal of Big Data* 11, 1 (2024), 66.

[119] Chang Sun, Johan van Soest, and Michel Dumontier. 2023. Generating synthetic personal health data using conditional generative adversarial networks combining with differential privacy. *Journal of Biomedical Informatics* 143 (2023), 104404.

[120] Jennifer Taub, Mark Elliot, Maria Pampaka, and Duncan Smith. 2018. Differential correct attribution probability for synthetic data: an exploration. In *Privacy in Statistical Databases: UNESCO Chair in Data Privacy, International Conference, PSD 2018, Valencia, Spain, September 26–28, 2018, Proceedings*. Springer, 122–137.

[121] Lucas Theis, Aäron van den Oord, and Matthias Bethge. 2015. A note on the evaluation of generative models. *arXiv preprint arXiv:1511.01844* (2015).

[122] Jason A Thomas, Randi E Foraker, Noa Zamstein, Jon D Morrow, Philip RO Payne, and Adam B Wilcox. 2022. Demonstrating an approach for evaluating synthetic geospatial and temporal epidemiologic data utility: results from analyzing> 1.8 million SARS-CoV-2 tests in the United States National COVID Cohort Collaborative (N3C). *Journal of the American Medical Informatics Association* 29, 8 (2022), 1350–1365.

[123] Amirsina Torfi, Edward A Fox, and Chandan K Reddy. 2022. Differentially private synthetic medical data generation using convolutional GANs. *Information Sciences* 586 (2022), 485–500.

[124] 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), 1–13.

[125] Vibeke Binz Vallevik, Aleksandar Babic, Serena Elizabeth Marshall, Elvatun Severin, Helga MB Brøgger, Sharmini Alagaratnam, Bjørn Edwin, Narasimha Raghavan Veeraragavan, Anne Kjersti Befring, and Jan F Nygård. 2024. Can I trust my fake data–A comprehensive quality assessment framework for synthetic tabular data in healthcare. *International Journal of Medical Informatics* (2024), 105413.

[126] Boris Van Breugel, Zhaozhi Qian, and Mihaela Van Der Schaar. 2023. Synthetic data, real errors: how (not) to publish and use synthetic data. In *International Conference on Machine Learning*. PMLR, 34793–34808.

[127] Boris van Breugel, Hao Sun, Zhaozhi Qian, and Mihaela van der Schaar. 2023. Membership Inference Attacks against Synthetic Data through Overfitting Detection. In *International Conference on Artificial Intelligence and Statistics*. PMLR, 3493–3514.

[128] 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.

[129] Yoga Advaith Veturi, William Woof, Teddy Lazebnik, Ismail Moghul, Peter Woodward-Court, Siegfried K Wagner, Thales Antonio Cabral de Guimarães, Malena Daich Varela, Bart Liefers, Praveen J Patel, et al. 2023. SynthEye: investigating the impact of synthetic data on artificial intelligence-assisted gene diagnosis of inherited retinal disease. *Ophthalmology Science* 3, 2 (2023), 100258.

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

[131] Joshua Ward, Xiaofeng Lin, Chi-Hua Wang, and Guang Cheng. 2025. Synth-MIA: A Testbed for Auditing Privacy Leakage in Tabular Data Synthesis. *arXiv preprint arXiv:2509.18014* (2025).

[132] Joshua Ward, Chi-Hua Wang, and Guang Cheng. 2024. Data plagiarism index: Characterizing the privacy risk of data-copying in tabular generative models. *arXiv preprint arXiv:2406.13012* (2024).

[133] Joshua Ward, Chi-Hua Wang, and Guang Cheng. 2025. Privacy Auditing Synthetic Data Release through Local Likelihood Attacks. *arXiv preprint arXiv:2508.21146* (2025).

[134] Joshua Ward, Yuxuan Yang, Chi-Hua Wang, and Guang Cheng. 2025. Ensembling Membership Inference Attacks Against Tabular Generative Models. In *Proceedings of the 18th ACM Workshop on Artificial Intelligence and Security*. 182–193.

[135] David S Watson, Kristin Blesch, Jan Kapar, and Marvin N Wright. 2023. Adversarial random forests for density estimation and generative modeling. In *International Conference on Artificial Intelligence and Statistics*. PMLR, 5357–5375.

[136] Sophie Wharrie, Zhiyu Yang, Vishnu Raj, Remo Monti, Rahul Gupta, Ying Wang, Alicia Martin, Luke J O'Connor, Samuel Kaski, Pekka Marttinen, et al. 2023. HAPNEST: efficient, large-scale generation and evaluation of synthetic datasets for genotypes and phenotypes. *Bioinformatics* 39, 9 (2023), btad535.

[137] Viktor Wolf, Felix Neubürger, and Ralf Lanwehr. 2023. Generating Synthetic Data for Better Prediction Modeling in Skill Demand Forecasting. In *2023 IEEE World Conference on Applied Intelligence and Computing (AIC)*. IEEE, 313–318.

[138] Mi-Ja Woo, Jerome P Reiter, Anna Oganian, and Alan F Karr. 2009. Global measures of data utility for microdata masked for disclosure limitation. *Journal of Privacy and Confidentiality* 1, 1 (2009).

[139] Liyang Xie, Kaixiang Lin, Shu Wang, Fei Wang, and Jiayu Zhou. 2018. Differentially private generative adversarial network. *arXiv preprint arXiv:1802.06739* (2018).

[140] Xiaodan Xing, Federico Felder, Yang Nan, Giorgos Papanastasiou, Simon Walsh, and Guang Yang. 2023. You Don't Have to Be Perfect to Be Amazing: Unveil the Utility of Synthetic Images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 13–22.

[141] Lei Xu, Maria Skoularidou, Alfredo Cuesta-Infante, and Kalyan Veeramachaneni. 2019. Modeling tabular data using conditional gan. *Advances in neural information processing systems* 32 (2019).

[142] Lei Xu and Kalyan Veeramachaneni. 2018. Synthesizing tabular data using generative adversarial networks. *arXiv preprint arXiv:1811.11264* (2018).

[143] Andrew Yale, Saloni Dash, Ritik Dutta, Isabelle Guyon, Adrien Pavao, and Kristin P Bennett. 2019. Privacy Preserving Synthetic Health Data. In *ESANN 2019-European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*.

[144] 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.

[145] Chao Yan, Ziqi Zhang, Steve Nyemba, and Bradley A Malin. 2020. Generating electronic health records with multiple data types and constraints. In *AMIA annual symposium proceedings*, Vol. 2020. American Medical Informatics Association, 1335.

[146] Chao Yan, Ziqi Zhang, Steve Nyemba, and Bradley A Malin. 2021. Generating electronic health records with multiple data types and constraints. In *AMIA annual symposium proceedings*, Vol. 2020. 1335.

[147] Zexi Yao, Nataša Krčo, Georgi Ganev, and Yves-Alexandre de Montjoye. 2025. The DCR delusion: measuring the privacy risk of synthetic data. In *European Symposium on Research in Computer Security*. Springer, 469–487.

[148] Burak Yelmen, Aurélien Decelle, Leila Lea Boulos, Antoine Szatkownik, Cyril Furtlehner, Guillaume Charpiat, and Flora Jay. 2023. Deep convolutional and conditional neural networks for large-scale genomic data generation. *PLoS Computational Biology* 19, 10 (2023), e1011584.

[149] 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.

[150] Jinsung Yoon, Daniel Jarrett, and Mihaela Van der Schaar. 2019. Time-series generative adversarial networks. *Advances in neural information processing systems* 32 (2019).

[151] Jinsung Yoon, Michel Mizrahi, Nahid Farhady Ghalaty, Thomas Jarvinen, Ashwin S Ravi, Peter Brune, Fanyu Kong, Dave Anderson, George Lee, Arie Meir, et al. 2023. EHR-Safe: generating high-fidelity and privacy-preserving synthetic electronic health records. *NPJ Digital Medicine* 6, 1 (2023), 141.

[152] Hengrui Zhang, Jiani Zhang, Zhengyuan Shen, Balasubramaniam Srinivasan, Xiao Qin, Christos Faloutsos, Huzefa Rangwala, and George Karypis. 2024. Mixed-Type Tabular Data Synthesis with Score-based Diffusion in Latent Space. In *The Twelfth International Conference on Learning Representations*. https://openreview.net/forum?id=4Ay23yeuz0

[153] Jun Zhang, Graham Cormode, Cecilia M Procopiuc, Divesh Srivastava, and Xiaokui Xiao. 2017. Privbayes: Private data release via bayesian networks. *ACM Transactions on Database Systems (TODS)* 42, 4 (2017), 1–41.

[154] Yili Zhang, Jia Li Dong, Bai Xue, Yanbao Xiong, Samir Gupta, Maarten Van Segbroeck, Nawar Shara, and Peter McGarvey. 2025. Exploring the Utilization of Synthetic Data in Unsupervised Clustering for Opioid Misuse Analysis. In *AMIA Annual Symposium Proceedings*, Vol. 2024. 1313.

[155] Ziqi Zhang, Chao Yan, Thomas A Lasko, Jimeng Sun, and Bradley A Malin. 2021. SynTEG: a framework for temporal structured electronic health data simulation. *Journal of the American Medical Informatics Association* 28, 3 (2021), 596–604.

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

[157] Ziqi Zhang, Chao Yan, Diego A Mesa, Jimeng Sun, and Bradley A Malin. 2020. Ensuring electronic medical record simulation through better training, modeling, and evaluation. *Journal of the American Medical Informatics Association* 27, 1 (2020), 99–108.

[158] Zilong Zhao, Aditya Kunar, Robert Birke, Hiek Van der Scheer, and Lydia Y Chen. 2024. Ctab-gan+: Enhancing tabular data synthesis. *Frontiers in big Data* 6 (2024), 1296508.

[159] Taiyu Zhu, Kezhi Li, Pau Herrero, and Pantelis Georgiou. 2023. Glugan: Generating personalized glucose time series using generative adversarial networks. *IEEE Journal of Biomedical and Health Informatics* (2023).

A Synthetic Data Generators

All generators are implemented using the `synthyverse` Python library (version 0.2.3) accompanying this survey. The `synthyverse` bases implementations of generators on existing works from literature and/or software libraries[6]. Below we describe the origin of the implementations and more details on hyperparameters.

SMOTE. The implementation of SMOTE is based on that from Mueller et al. [86]. We use 5 nearest-neighbours for interpolation.

ARF. The implementation is based on the `arfpy` Python package (version 0.1.1), and we use the same hyperparameters from the original paper [135].

CTGAN. The implementation uses the `ctgan` Python package (version 0.12.0). We align the networks of all deep generative models – CTGAN, TVAE, and TabDDPM – to 2-hidden layer MLPs with 128 nodes, and the embedding/bottleneck dimension to 128. We also align the batch sizes to 500, and train the models for 300 epochs. Other than that, we use the default parameters from the original paper [141].

TVAE. The implementation uses the `ctgan` Python package (version 0.12.0). The network sizes, bottleneck dimension, batch size and epochs are aligned as described above. Other than that, we use the default parameters from the original paper [141].

TabDDPM. The implementation is based on that from the `synthcity` Python library (version 0.2.12) [103]. The network sizes, bottleneck dimension, batch size and epochs are aligned as described above. We use a learning rate of 1e-3, weight decay of 1e-5, 200 sampling steps, and a linear scheduler.

B Hardware

GPU-based models (CTGAN, TVAE, TabDDPM) were trained using a single MIG instance of an RTX6000 GPU (24GB VRAM) and access to an additional 8 CPU cores (32GB RAM). CPU-based models were trained using 16 CPU cores (64GB RAM).

[6]See the third party notices for more information on licenses and modifications of source implementations.