Metrics That Matter: A Practical Survey on 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
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

Assessing the quality of synthetic data (SD) is vital to determine whether it can provide a viable alternative to real data. A wide variety of metrics exist to examine 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 the suitability of metrics strongly depend on the dataset and other contextual factors. This paper surveys the field of SD evaluation, provides guidance regarding metric selection based on four key questions pertaining to the task, goal, data type, and domain of SD, and provides general practical recommendations on SD evaluation. Finally, experiments on an illustrative dataset of electronic health records show how researchers can bring our insights and recommendations for SD evaluation into practice. By doing so, we aim to support researchers and practitioners seeking to generate and evaluate SD.

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

ACM Reference Format:
Jim Achterberg, Bram van Dijk, Saif Ul Islam, Gregory Epiphaniou, Carsten Maple, Marcel Haas, and Marco Spruit. 2025. Metrics That Matter: A Practical Survey on Synthetic Data Evaluation. *ACM Comput. Surv.* 37, 4, Article 111 (August 2025), 36 pages. https://doi.org/XXXXXXX.XXXXXXX

*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; 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

1 Introduction

Synthetic Data (SD), which is data generated by an algorithm or mathematical model instead of a real-world process, can replace or augment Real Data (RD) whenever RD is scarce or inaccessible due to privacy or other concerns [77]. The evaluation of SD typically spans assessing its statistical similarity to RD (*Fidelity*), task-specific usefulness (*Utility*), and safety (*Privacy*) [65, 77], which is crucial to determine whether SD can indeed satisfactorily replace or augment RD. That is, realistic SD should be similar to RD to ensure similar outcomes across a range of inferences drawn from the SD (high fidelity); should show good performance in specific downstream tasks of interest (high utility); and should not allow adversaries to disclose sensitive information from the SD (high privacy).

In practice, however, the metrics for evaluating fidelity, utility and privacy are often selected without appropriate reflection or motivation, leading to issues such as incorrect interpretation of results, as well as a lack of alignment in benchmarking methods [11, 80, 109, 151]. For example, Fréchet Inception Distance (FID) as a fidelity metric [66] is fit for natural images, but often misleading when applied to other domains [117]. Additionally, probabilistic fidelity metrics such as KL divergence, Jensen-Shannon (JS) divergence, and Wasserstein distance are often used interchangeably, without reflection on which is most suitable for a given dataset, or the drawbacks they might have. Regarding privacy, ad-hoc evaluations based on similarity of individual synthetic and real records are typically provided, while it is known that SD exhibits further privacy risk and should be adequately evaluated [150]. Furthermore, recent work has documented hundreds of evaluation metrics already in use for assessing synthetic tabular data only [162], and though many different overviews exist that try to reduce this complexity (e.g. [124], [30], [80]), they provide little guidance on choosing appropriate metrics given a particular dataset and context.

This article surveys existing work on SD evaluation, provides guidelines for choosing appropriate evaluation metrics in different contexts, provides general practical recommendations on SD evaluation, and illustrates SD evaluation in practice with a concrete example. First of all, in our discussions, we will keep a *user perspective* in mind, that is, specific use cases and accompanying dataset characteristics, and refrain from the idea of a universally valid and applicable set of evaluation metrics. 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 new metrics. These metrics may come from other fields or may be newly developed. The framework can help evaluate the applicability of such metrics across different scenarios. This is helpful, as we observe that many more statistical metrics are equally suitable for SD evaluation but are rarely used, as researchers tend to reproduce metrics they have already seen in prior work.

Table 1 contains an overview of other surveys and review articles that dedicate substantial attention to SD evaluation. Our work is broad in its scope, in that it discusses evaluation metrics for the three archetypal dimensions (fidelity, utility, privacy), is not focused on one specific domain, discusses multiple data types (tabular, time-series, images, text), provides explicit guidance on which metrics to select in specific scenarios, and provides an empirical illustration to showcase how to select and apply different evaluation metrics in practice. In contrast, Kaabachi et al. [80], Murtaza et al. [115], Hernandez et al. [65], and Budu et al. [18] focus only on structured medical data. Stenger et al. [151] focuses only on time-series data. Osorio-Marulanda et al. [124] focuses only on privacy metrics. Kiran et al. [87] are not explicitly focused only on structured data, but do not mention evaluation metrics for unstructured data types. Lautrup et al. [92] focuses only on tabular data and provides only a minor discussion on privacy metrics. To our knowledge, ours is the only work that explicitly provides guidance on which evaluation metrics are most appropriate in different scenarios.

Table 1. Summary of contributions of this work versus previous works.

	Ours	[80]	[45]	[115]	[151]	[124]	[87]	[65]	[92]	[18]
Summarizes Evaluation Metrics	×	×	×	×	×	×	×	×	×	×
Agnostic to Evaluation Dimension	×	×	×	×	×		×	×	×	×
Agnostic to Data Domain	×		×		×	×	×		×	
Agnostic to Data Type	×		×			×				
Guidance on Metric Selection	×									
Empirical Illustration	×								×	×

As this article is a survey rather than a systematic review, we do not employ a systematic methodology for literature collection. Accordingly, we assemble the bibliography through iterative, ad-hoc searches and snowballing across the major scholarly repositories, e.g., ACM Digital Library, IEEE Xplore, PubMed, Scopus, and Web of Science. Search queries combined terms for SD with synonyms for evaluation, utility, fidelity, and privacy. We include papers which introduce, adapt, or empirically apply at least one evaluation metric for SD ánd appear in a peer-reviewed venue; influential pre-prints are only included when they accumulated over 50 citations by July 2025. No other constraints were imposed. The resulting references are therefore representative rather than exhaustive - sufficient to populate and stress-test our taxonomy (Section 2) and framework for guidance on metric selection (Section 3), while leaving readers room to map additional metrics onto the same taxonomy and framework.

The remainder of the paper is structured as follows: we first provide a taxonomy of SD evaluation metrics (Section 2), which serves as the foundation for the rest of the paper. Thereafter, we address four key questions that will help researchers and practitioners to make a better informed selection of SD evaluation metrics (Section 3). We continue with providing some general, practical recommendations on SD evaluation (Section 4) and conclude with a concrete illustration of SD generation and evaluation with a dataset on heart failure patients (Section 5). Overall, this article has structured data as its vantage point, although at several points we also indicate and discuss evaluation metrics appropriate for unstructured SD, such as images and text.

2 Synthetic Data Evaluation Metrics

Evaluating SD requires assessing its fidelity, utility, and privacy to ensure its realism, usability and safety. This section outlines key metrics used to assess these aspects and provides a taxonomical classification as shown in Figure 1. Note that this taxonomy is not intended to be exhaustive, but rather indicates how these metrics can be categorised. Many more similar evaluation metrics could be easily inserted, for example, probabilistic distance measures including Jeffrey's divergence [74] and projection similarity algorithms such as Isomap [156].

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 categorize fidelity metrics as falling under **Human Evaluation**, **Descriptive Statistics**, **Projection Similarity**, **Feature Association**, and **Statistical Distance and Similarity Measures** as shown in Figure 1.

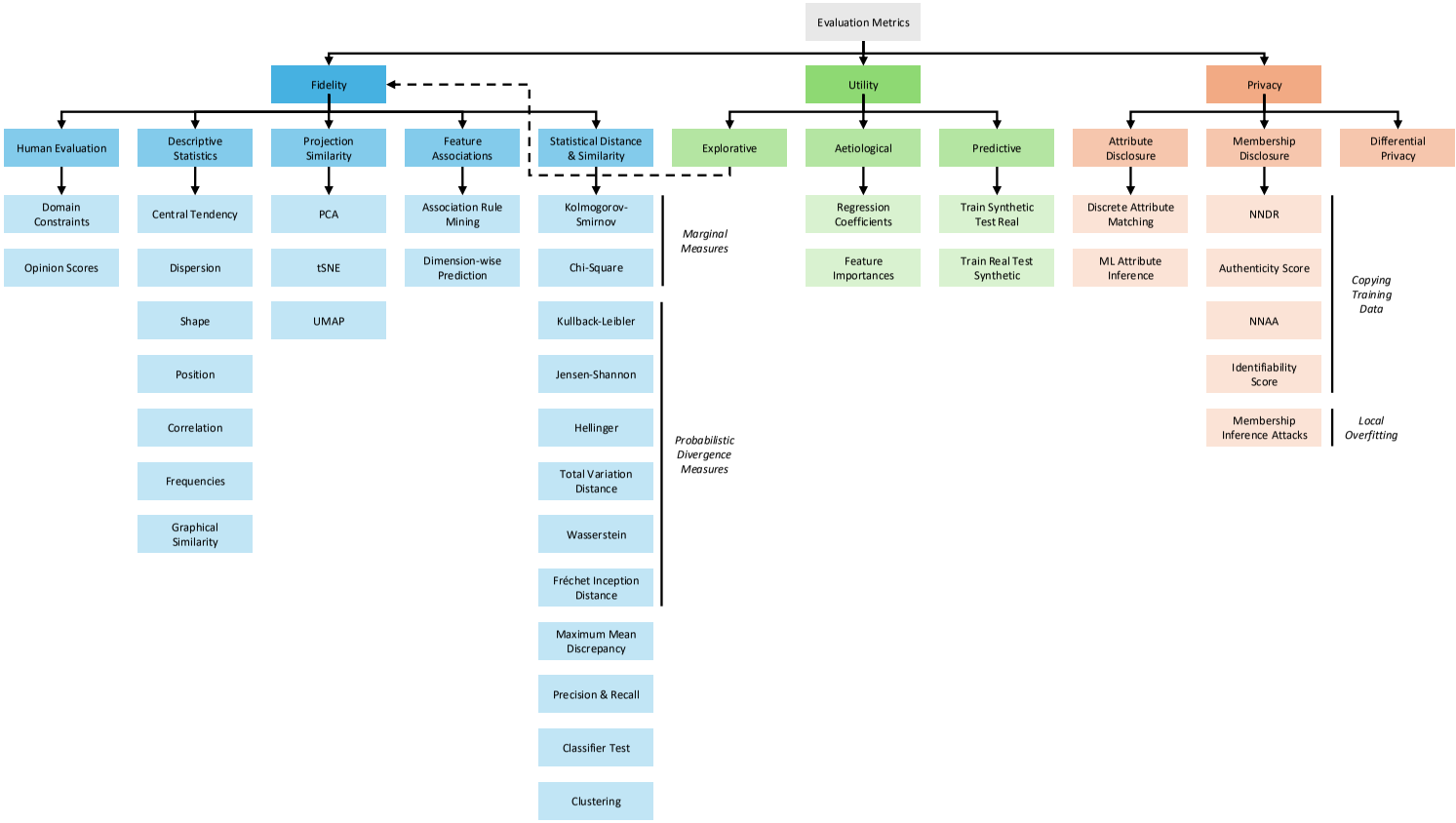

Fig. 1. Taxonomical Classification of Evaluation Metrics for Privacy-Preserving Synthetic Data. The dashed lines indicate that fidelity metrics are preferred when evaluating utility for explorative analyses. In these cases, SD should have high general utility (Section 2.2).

2.1.1 Human Evaluation. Human evaluation provides manual checks on SD fidelity without relying on extensive statistical methods. For example, assessing whether SD similarly follows specific domain constraints or business rules found in RD [184], or setting up experimental procedures allowing target audiences to blindly score SD realism [19, 26, 171].

Human evaluation is especially crucial when data validity is required, i.e., SD could (theoretically) belong to a real individual or entity. It is also an essential part of assessing the realism of some unstructured data types such as speech [37], for which humans generally have a good intuition of what is realistic or not. Also, for some domains, expert knowledge is often required to assess desired and specific features which general fidelity metrics may miss, e.g., co-occurring local and global pathological features in radiological images [89].

Furthermore, SD may demonstrate high statistical quality yet still exhibit unrealistic artifacts. As a result, such SD is often considered unusable, especially in high-stakes domains such as healthcare, which poses challenges for its adoption and reduces practitioners' confidence in its reliability within the respective field.

2.1.2 Descriptive Statistics. Descriptive statistics can provide intuitive checks on statistical fidelity. Here, we limit ourselves to univariate and bivariate statistics that are commonly used because they are easy to compute and interpret. Univariate metrics typically concern measuring central tendency, dispersion, shape, position, and frequency of a specific synthetic feature distribution against its real counterpart [1, 63, 144, 184, 188, 193]. Bivariate metrics such as covariance matrices, contingency tables, and mutual information, express the relationship between two features [30]. Graphical plots such as histograms, kernel density plots, and scatterplots can be used to get an accompanying visual intuition for how much univariate or bivariate distributions overlap [110].

2.1.3 Projection Similarity. Projection similarity provides graphical evaluation of high-dimensional data by plotting compressed versions of SD and RD against each other [1, 86, 119, 130, 149, 167, 169, 175, 199]. Dimensionality reduction algorithms compress data down to a manageable number of dimensions for plotting, e.g., two or three, while aiming to maintain similar statistical properties. These plots provide a qualitative rather than quantitative evaluation of SD fidelity (even when difficult in original high-dimensional data) and can inform on issues such as overfitting and mode collapse in SD, e.g., when SD is clustered into dense clouds *within* clusters of RD [1]. However, they can also be costly to compute, lossy, and contingent on subjectivity [96]. Choosing the appropriate dimensionality reduction algorithm is key, paying attention to the tradeoff between quality and computational complexity. For example, PCA is generally fast, non-stochastic, and insensitive to hyperparameters, but its linear orthogonal projections only retain global data structure [69]. Contrarily, tSNE and UMAP are computationally more expensive, stochastic, and sensitive to hyperparameters, but project non-linearly and aim to retain both local and global data structure [111, 165]. See Gisbrecht and Hammer [53] for a more extensive discussion on selecting dimensionality reduction algorithms. To limit computational overhead, one possibility is to investigate more complex projections (e.g., tSNE or UMAP) only when simple projections (e.g., PCA) are already deemed similar.

2.1.4 Feature Association. Feature association measures indicate further multivariate dependencies beyond simple bivariate correlation statistics. Association rule mining and dimension-wise prediction are common methods. Association rule mining extracts interpretable if-then rules between features in a dataset; similar rules in SD and RD (measured through, e.g., precision-recall) indicate similar feature dependencies [9, 10, 83, 186]. Although association rule mining provides for an interpretable metric, it relies on discrete concepts, which may oversimplify evaluation.

Dimension-wise prediction assesses similarity in the generalization capacity of predictive algorithms trained on SD and RD for predicting each feature from other features to a real test set [9, 26, 84, 186, 196]. Note the overlap with utility metrics (when the target feature corresponds to 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). Utilizing different predictive algorithms ranging from less to more flexible, e.g., linear regression to boosted trees, provides additional insight into the structural fidelity of SD.

2.1.5 Statistical Distance and Similarity Measures. Statistical distance and similarity metrics tell something about the (dis)similarity between the probability distributions of SD and RD. The selection of these metrics should cover overall similarity but ideally also be sensitive to general failure modes such as under- and overfitting, mode collapse, and mode invention. In addition, it should also cover specific failure modes if applicable, such as retaining outliers or tail events. Note that metrics that measure similarity between individual samples are an indication of privacy risk rather than high fidelity in privacy-preserving SD, as they can provide insight as to whether RD is (partially) copied, so we exclude them from discussion here.

Marginal statistics can provide insights into whether individual feature distributions have the same underlying distribution. Common choices are statistics with well-established distributions under the null hypothesis ($H_0 : F_{RD}(x) = F_{SD}(x),\ \forall x$), allowing formal statistical testing, such as Kolmogorov-Smirnov (numerical features) and Chi-square (discrete features) [9, 83, 139, 160, 188].

Probabilistic distance measures indicate distance between multivariate SD and RD distributions, typically requiring density estimation beforehand. KL divergence, JS, total variation distance, Hellinger distance and Wasserstein distance are common choices [30, 32, 40, 40, 81, 86, 103, 119, 133, 139, 178, 188, 198]. Different probabilistic distance measures are suited to indicate different failure modes in SD. For example, due to its asymmetric nature, forward KL divergence favours “likely" SD (SD outside RD support is penalised more heavily) whereas backward KL divergence favours “diverse" SD (RD outside SD support is penalised more heavily). JS divergence is a bounded and symmetrized version of KL divergence and provides an overall indication of distributional similarity. Total variation distance 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; this can be useful when focus on retaining tails is warranted, e.g., when modelling rare diseases, credit risk, or natural disasters. Wasserstein distance indicates geometric similarity of distributions, penalising mode collapsed (or inventing) distributions heavily; this has made it a popular cost function for training generative models while preventing mode collapse [7]. Wasserstein distance does not require density estimation beforehand, but relies on linear programming techniques, which can make it computationally costly in large datasets - although faster approximations exist [25]. Precision-recall analyses distinguish whether SD is covered by RD (precision) and whether RD is covered by SD (recall) [140]; several adaptations to the precision and recall metrics of Sajjadi et al. [140] have been made to improve or generalize these measures [5, 91, 117]. For other measures, choosing an appropriate density estimator is key, e.g., kernel- or deep learning-based methods based on the size and dimensionality of the RD and the available computational resources. Other metrics exist that circumvent direct density estimation, such as maximum mean discrepancy, by comparing means of high-dimensional embeddings between SD and RD directly, making them especially suited to high-dimensional data [56].

More implicit procedures using machine learning algorithms can also be used to indicate SD fidelity. For example, assessing whether SD and RD similarly reside in point clusters [40, 55, 118, 176, 184], or more commonly, whether SD

and RD can be accurately distinguished using a classifier [1, 83, 94, 97]. For the clustering metric, selecting the number of clusters is key; we recommend selecting this on RD only (e.g., through silhouette score), since this should not change when integrating SD. For the classification test, we recommend selecting an accuracy metric which is insensitive to class imbalance and classification threshold, e.g., AUC or (standardised) propensity Mean Squared Error (pMSE) [148]. The classification procedure can also be used for two-sample testing of the SD versus RD distribution [47, 62, 85], although conclusions should be drawn with care, since failing to reject the null hypothesis could simply be the result of an underpowered classifier [1]. Both the clustering and classification measures are highly influenced by which algorithm is selected, and assessing multiple algorithms may provide additional insight into structural fidelity. They both provide an indication of overall fidelity, and fail to disentangle failure modes, as poor clustering and classification scores can be due to a number of other reasons, such as poorly chosen hyperparameters or distance metrics.

Figure 2 and Table 2 provide intuitions regarding the suitability of probabilistic distance and machine learning measures to detect various failure modes in SD. They show how, for example, Wasserstein distance is especially sensitive to geometric similarity of distributions (mode collapse and invention), precision and recall to under- and overfitting (either generally or through mode collapse/invention), while overall similarity (as given by JS divergence and a classifier test) remains similar.

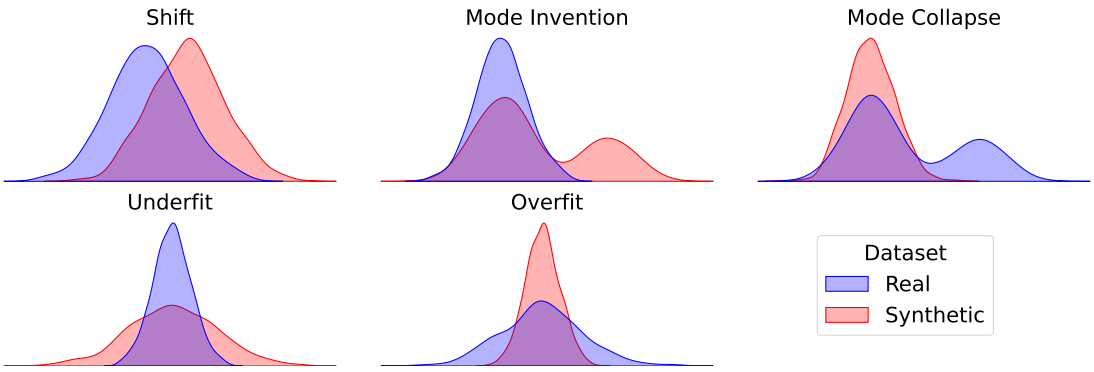

Fig. 2. Examples of different failure modes of SD, keeping overall similarity relatively constant (see Table 2). Toy data from (mixtures of) Gaussians (n = 1000 for SD and RD).

Table 2. Probabilistic distance measures corresponding to Figure 2.

	Jensen-Shannon	Wasserstein	Precision	Recall	Classifier test*
Shift	0.350	0.559	0.982	0.991	0.554
Mode Invention	0.350	1.756	0.710	1.000	0.525
Mode Collapse	0.351	2.127	0.999	0.878	0.535
Underfit	0.352	0.826	0.840	1.000	0.562
Overfit	0.348	0.884	1.000	0.878	0.562

*SD distinguishability from RD as given by the AUC of an XGBoost classifier.

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 [33, 34, 144]. Since the description by Purdam and Elliot [135] 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 task-specific, we further distinguish utility metrics based on the data analysis-specific goal, i.e., **Explorative**, **Aetiological**, or **Predictive**, as shown in Figure 1.

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 (micro-level) census data [100], synthetic electronic health records [1, 184], but also the exploration how well different SD generators perform, i.e., benchmarking generative models according to their ability to accurately model a datasets' joint distribution function [2, 30].

2.2.2 Aetiological Analyses. Aetiological analyses aim to explain causes of outcomes by reflecting on how predictions relate to explanatory variables. Here, we do not consider strict causality, but also correlational studies, as they serve a similar goal. Various utility metrics exist which indicate whether SD and RD similarly relate explanatory variables to outcomes, e.g., by comparing regression coefficients (or their subsequent hypotheses tests) or feature importances, e.g., through Shapley values or feature effect plots [29, 33, 35, 73, 82, 114, 120, 184]. Some common examples of SD generation for aetiological analyses are clinical trials [22, 42], epidemiological studies [16, 158], credit card fraud and network intrusion detection [6, 73]. 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 indicate, e.g., generalisation capacity to RD when training on SD or vice versa - formally introduced by Esteban et al. [43] as Train Synthetic Test Real (TSTR) [1, 2, 30, 78, 84, 86, 88, 97, 102, 138, 182, 188, 190] and Train Real Test Synthetic (TRTS) [84, 86, 146, 184]. Both approaches are usually compared to performance on purely RD (Train Real Test Real, TRTR). TRTS indicates whether synthetic labels are "likely", i.e., similar examples were seen when training on RD, whereas TSTR indicates whether synthetic labels are "diverse", i.e., SD contains similar labels as in the RD test set. Note the similarity to Sajjadi et al. [140]'s disentanglement of precision-recall for distributions. TSTR is often seen the more important evaluation [43]. However, we acknowledge the usefulness of disentangling whether synthetic labels are adequately likely ánd diverse by computing both TSTR and TRTS. For example, when TRTS is high but TSTR low, this tells us that SD labels are realistic but not diverse enough to generalize well to RD [43]. 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 [50]. In these applications, reporting TRTS can be valuable, as it directly aligns with SD's intended task.

Some efforts also include Train Synthetic Test Synthetic (TSTS), which trains predictive algorithms on SD and scores them on a synthetic holdout set [44, 146]. 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 predictive algorithm is vital, as it directly influences the outcome of these metrics. For example, choosing a prediction model that is less robust against overfitting may yield less pronounced differences between different SD generators as opposed to more robust prediction models [143]. Hence, it is best to select a set of prediction 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 algorithms can provide additional insight into SD utility - similar to some fidelity metrics such as classifier distinguishability.

2.3 Privacy

Although we consider privacy-preserving SD, information from RD can potentially leak through to SD, inducing privacy risks [150]. A common distinction regarding privacy risk is between **Attribute Disclosure** and **Membership Disclosure** [80], as shown in Figure 1. 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. [132], 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/or* 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. [155].

2.3.1 Attribute Disclosure. Attackers can disclose sensitive information by employing an Attribute Inference Attack (AIA). 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 generalises well from SD to RD [68]. Therefore, attribute disclosure risk from SD can be considerable even when individual SD samples do not leak information, and attribute disclosure risk is thus often more a property of the dataset and attacker's access to auxiliary data sources, than of the SD generation process.

A typical method for attribute inference is training predictive machine learning algorithms on SD and performing inference on (a holdout split) of RD [1, 1, 2, 26, 150, 184, 185, 194]. 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. Highly parameterized models which require careful architectural tuning based on a validation set containing ground-truth labels, with neural networks as prime example, are thus seldom the best choice.

Instead of machine learning, 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), and 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 [68, 155]. 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 [39, 64, 86, 93, 127, 146, 152, 196]. SD should not artificially distance itself from RD however, e.g., by some fixed scaling factor, as this can be leveraged by attackers to reidentify samples. DCR can account for this by separately normalizing SD and RD to the same scale [113]. In this case, distances no longer reflect distances in the original data space, but they do account for SD which artificially distances itself from RD. Furthermore, DCR metrics only indicate disclosure risk when compared to distances to other samples in its direct neighbourhood, since geometric closeness does not incur disclosure risk when those types of samples are typical [132]. An example of a metric which compares to other samples in its direct neighbourhood is Nearest Neighbour Distance Ratio (NNDR), which normalizes SD-RD distances to the next-nearest RD sample [93, 121, 146]. Extremely low values indicate that these particular SD and RD samples are similar without other RD in the vicinity; this is a potential case of identity disclosure, and thereby, a potential case of membership as well as attribute disclosure. NNDR is computed with respect to the training set, but can also be normalized by NNDRs with respect to a holdout set (NNDR ratio). When NNDRs are low with respect to both the train and holdout set, similar “outliers" exist in the holdout set as in the training set, in which case those points are less identifiable than initially thought by assessing only the training set.

Other such metrics exist which assess sample-level similarity, but these 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 training the SD generating algorithm [115]. This leaks privacy when membership to the training set exposes sensitive attributes in 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 distance-based metrics which can imply specific samples to be at risk, e.g., NNDR [121], some measures are only informative on general disclosure risk across the entire dataset. For example, the authenticity score from Alaa et al. [5] indicates the proportion of samples from RD which are closest in proximity to RD instead of SD [2, 8, 102, 136]. Low scores (compared to the overall SD-RD proportion) indicate that SD is at risk of copying samples from the training data, in which case membership disclosure risk is high - although it does not provide any information on *which* samples are at risk. Also, high scores cannot easily be compared, as we cannot conclude whether a slightly higher score indicates less privacy risk, as this may only be the result of reduced fidelity, and both datasets may pose little disclosure risk. Other similar metrics are nearest neighbour adversarial accuracy [183] and the identifiability score [188]. Nearest neighbour adversarial accuracy balances the authenticity score with a second term indicating whether SD is closest to other SD [15, 24, 167, 174, 184, 187]. 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 is calculated similarly to the authenticity score, but has the inverse connotation. It measures the proportion of RD which is closest to SD, and therefore, high scores indicate more instead of less disclosure risk [2, 59, 112, 145]. Also, "closeness" is not

defined through standard geometric distance, but rather, distances are weighted by the discrete entropy of features, such that low entropy features, for which different values are more identifiable, receive a higher score.

Instead of these measures which provide a general indication of membership disclosure across the entire dataset, a more specific Membership Inference Attack (MIA) can be constructed to target specific samples and disclose whether they are members of the training set. These 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. Various efforts have been made to construct such MIAs, e.g., Hayes et al. [60], which employ discriminators of GANs to directly estimate a relative score relating to the likelihood of some dataset belonging to the SD generator training data. In the white-box setting, i.e., an attacker has access to the GAN SD generator, the original discriminator can be used directly. Otherwise, for VAE generators, the reconstruction error can be used [67]. In the black-box setting, i.e., no access to the SD generator, a generator is trained on SD by the attacker and a classifier used to distinguish training data from non-training data. Otherwise, Zhang et al. [194] estimate the likelihood of training data and some test data under a density estimator fitted on SD - which has a similar architecture as the SD generator - and compare their perplexity. Similar perplexity indicates that no unreasonably high likelihood is placed on (portions of) training data compared to non-training data. However, they do not provide any formal attack accuracy. Next, van Breugel et al. [164] provide a more formal framework for MIAs to detect local overfitting. Firstly, they assume attackers have access to some auxiliary RD which they can use to estimate the RD density. They separately estimate the SD and (auxiliary) RD density using neural autoregressive flows (or Gaussian kernel density estimators). Then, they scale the SD density by the RD density to compute a *relative* likelihood score for RD samples of interest to infer membership: relatively high likelihood, according to some threshold such as the median, indicates local overfitting and thus membership.

Other efforts construct MIAs using geometric distances with a threshold [184, 195]. 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.

2.3.3 Differential Privacy. Instead of a post-hoc metric, differential privacy is a formal mathematical framework to ensure a user-specified level of privacy *during* SD generation. In essence, differential privacy quantifies the contribution of individual samples to the output of an algorithm [36]. 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 [180, 191]. This diminishes the contribution of individual training samples to the output, but also diminishes SD fidelity [180, 191]. Hereby, differential privacy potentially mitigates both membership ánd attribute disclosure. This is also why differential privacy is shown as a separate category of privacy measures in Figure 1.

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 [180]. However, setting appropriate levels of differential privacy remains an open question. For any $\epsilon > 0$ some privacy risk may still exist [150], and similar ϵ can result in different privacy guarantees dependent on the domain and characteristics of the data [95]. Therefore, differential privacy parameters can typically not be considered as a

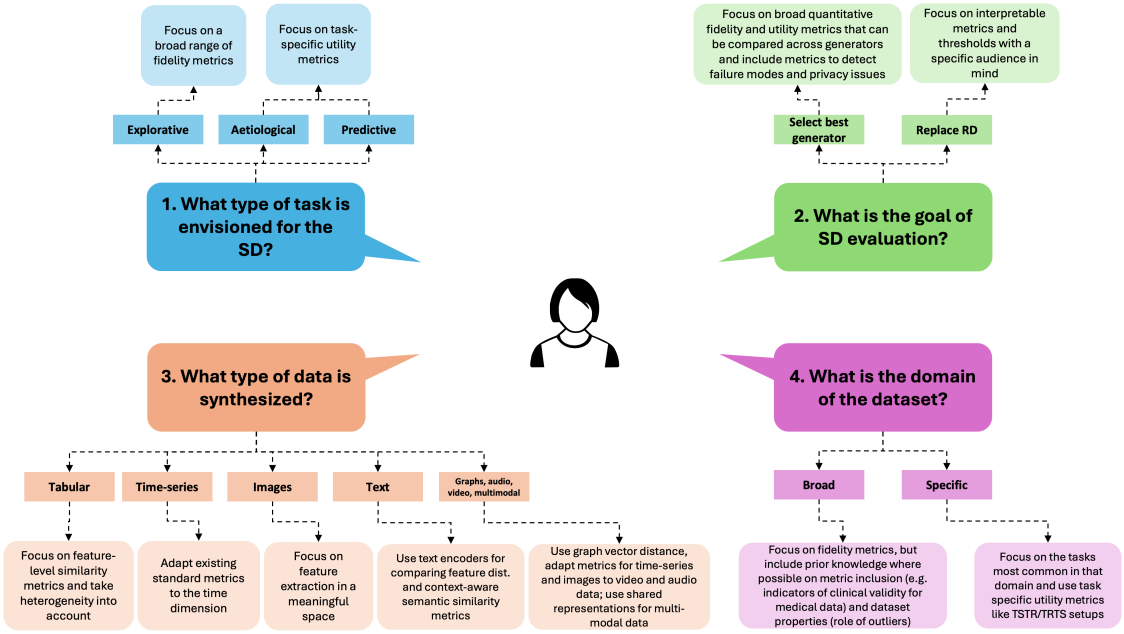

Fig. 3. A framework consisting of four essential questions to address when selecting evaluation metrics, and main takeaways of the answers we propose.

sufficient privacy evaluation, and differentially private SD still needs to be evaluated according to the same metrics as SD without differential privacy [132].

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 [12]. 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 [49]. 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 [115]. Another factor may be the wrong presumption that SD is “privacy-secure by design" [49, 77, 80].

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 SD task, data type, evaluation goal, and dataset domain influence the choice of evaluation metrics. Given that for privacy evaluation, no shared standards can be given, as discussed in the previous section, most of the discussion will revolve around fidelity and utility evaluation.

3.1 What Type of Task is Envisioned for Synthetic Data?

Whether SD is generated to perform a specific task or is generated for explorative purposes is paramount to how its usefulness should be evaluated. For explorative (or unknown) tasks, SD usefulness can only be evaluated through fidelity metrics. For aetiological or predictive tasks, task-specific utility measures can be used as well.

For aetiological and predictive tasks, fidelity metrics do not always correlate well with utility in the respective task. Fidelity metrics might not indicate similarity in the dimension of interest [181]. Additionally, reducing fidelity whilst maintaining task-specific utility can ensure useful SD with stronger privacy guarantees, as a lower degree of similarity typically reduces privacy risk - effectively providing a pareto improvement on the utility-privacy trade-off in SD [2]. In task-specific scenarios, such as aetiological and predictive analyses, the main focus of SD evaluation should thus be on task-specific utility 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 replace or augment RD. These usually do not happen in isolation, since researchers will often perform an SD generator selection process before publishing the “best" SD set, accompanied by an 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. Probabilistic fidelity measures or utility measures (e.g., TSTR, Wasserstein distance, 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 indicating generalization capacity, to ensure generators do not overfit and harm privacy, e.g., authenticity score [5] and DOMIAS [164].

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 JS divergence to 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 either 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 [163].

Structured data fares better regarding interpretability in that it allows separately emphasising marginal distributions and covariate dependencies. Feature-level 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 covariate dependencies: which features correlate similarly as in the RD, and which do not. Projection similarity 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. These methods, however, work less well for complex datasets with many feature dependencies and interactions. For unstructured data, however, the task of reporting interpretable metrics is more difficult. Many efforts focus on metrics that correlate well with human judgment [48, 66, 108, 177], showing that (proxies for) opinion scores from humans are considered important.

If SD is to replace RD, one key criterion regarding utility in predictive use cases is that for a specific task, task performance obtained with SD should match the performance of RD. Still, the loss in performance could also 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 models trained on SD. For aetiological use cases, it is important that the SD relates explanatory variables to outcomes in similar ways. For example, when the intended goal of SD is to support faster assurance of some medical device, a regulator may need to validate a manufacturer's claims about the device's performance for minority populations, which should be borne out by both RD and SD.

Lastly, concerning privacy, interpretable privacy evaluations are still challenging because of a gap between metrics and practice [3]. For example, mathematical frameworks such as differential privacy may lack practical application, as current legal frameworks provide less guidance with respect to the use of this type of metrics [12]. Still, the benefits of SD are arguably strongest in the cases where it can actually replace RD, hence, can be shared. Here, regulations and guidelines on privacy protection, such as the GDPR and the European Data Protection Board, come into play that identify key privacy risks *from a regulatory perspective* that must be addressed. Such guidelines focus on the degree to which a dataset renders individuals anonymous, and the quality and robustness of this anonymisation process, in other words the risk of 1) "singling out" or disclosing identity; 2) "linkability" or membership disclosure; and 3) attribute disclosure [52, 128]. Note that this corresponds to our discussion of identity, attribute and membership disclosure in Section 2.3.

In guiding privacy evaluation for these three risk types, we focus on tabular data, which often has sensitive features that are easy to interpret. First, for singling out, in fully SD there is no obvious mapping between real and synthetic records [155] – even in the case of outliers in the RD, such as rare diseases or financial transactions, an attacker may infer membership but no link with RD since it is typically inaccessible. Still, a distance to the closest record metric such as NNDR (Section 2.3) should always be reported so that it becomes clear whether SD records are close to RD without other RD nearby.

For general recommendations regarding membership disclosure, we draw on work by Pilgram et al. [132]. A good estimation of membership disclosure risk depends on a convincing MIA. An attack dataset is often a split from the RD, hence from the same distribution. This practice is convenient in many setups, as it is hard to reason about the resources of many possible attackers, but its assumptions may not always be realistic: if an attacker had data from the same distribution, not much would be learned from inferring membership besides sample-level information. A good MIA also requires a good performance metric. The F1 score is often used out-of-the-box, but it should be adjusted relative to a naive baseline F1 that gauges successful membership prediction without access to SD. In addition, it should be adjusted for the prevalence of membership in the attack dataset, and its precision and recall weights should be adjusted to the type of RD and vulnerabilities potentially exposed. Lastly, assuming maximum information for an adversary (i.e., access to many quasi-identifiers), however intuitive this may be, does not imply the largest privacy risk [132].

For recommendations on attribute disclosure, it is best to resort to an AIA with a prediction model, and compare the information gained about some individual in the SD to information that can be gained from a naive baseline, such as from population data. Only if a sensitive feature can be predicted substantially more reliably with the SD than without (in contrast to, say, an AUC of about .5), we can say privacy might be violated [132].

3.3 What Type of Data is Synthesized?

What type of data synthesized is the most crucial determinant of which evaluation metrics are applicable, and how they need to be adapted (if necessary)? We focus on the most common modalities for privacy-preserving SD generation below, i.e., tabular, time-series, images, and text, and provide only a short discussion on other modalities (graphs, audio, video, and multi-modal).

3.3.1 Tabular. Tabular data is a structured data format consisting of rows and columns. Rows (samples) are assumed to be independent. Tabular data often exhibits domain constraints or business rules between columns, which need to be enforced in SD [184]. Assessing this is especially crucial when SD generators use machine learning to generate data, as these might fail to learn vital rules in the data; it is less relevant for rule-based SD generators, which inherently enforce constraints. Due to its structured nature, humans typically have difficulty with providing valid opinion scores, especially for high-dimensional datasets - but domain experts may still provide useful insight [26, 171].

Since individual features in tabular data typically relate to well-understood concepts, SD fidelity evaluation should attend to feature-level similarity through, e.g., descriptive statistics and marginal distance measures. One of the main difficulties in tabular data, however is potential heterogeneity across features. For descriptive statistics, they need to be tailored to the relevant data types, e.g., use adequate bivariate correlation metrics such as Pearson or Spearman for numerical correlations, Cramérs V for categorical correlations, and correlation ratio η^2 for categorical-numerical correlations. In projection algorithms, it is key to select the the adequate extension (e.g., PCA, MCA, or FAMD), or for distance-based algorithms, the adequate distance measure (e.g., Euclidean, Hamming, or Gower distance in tSNE and UMAP). For measures relying on probability distributions (e.g., probabilistic distance measures, MIAs [164]), select an adequate density estimator, e.g., histogram- or kernel-based methods, or deep generative models for more complex heterogeneous tables.

Differential privacy has been enabled for a variety of tabular SD generators such as GANs [78, 180], VAEs [4], and Bayesian networks [191]. However, previous research has shown that, for relatively complex heterogeneous tabular datasets, enforcing even weak differential privacy can severely affect utility [51].

3.3.2 Time-Series. Time-series data is a structured data format also consisting of rows and columns. However, rows are generally *not* considered to be independent, as some sequential dependency exists. We can distinguish single time-series (t, k) and multiple time-series (n, t, k), for n samples, t timesteps, and k features.

Time-series adds another layer of complexity on top of the aforementioned practical considerations regarding tabular structured data. This may increase complexity for human evaluators. For other metrics, this necessitates adapting standard metrics, e.g., descriptive statistics need to be computed over the time-dimension, and additional statistics such as autocorrelations may need to be added [99]. Otherwise, metrics which rely on algorithms such as association rule mining, density estimation (probabilistic distance measures, MIAs), distance computation (projection similarity, precision-recall, membership disclosure attacks), clustering, classification and regression (classifier test, TSTR, TRTS, AIAs), can often utilize time-series algorithms as drop-in replacement. For example, time-series association rule mining [134], autoregressive flows with sequential layers for density estimation, dynamic time warping for distance computation [1], and time-series classifiers, regressors, and clustering [105]. Another option is to embed time-series as a single row per sample, allowing to use standard evaluation metrics. For example, Jeha et al. [75] compute a FID-like metric from unsupervised time-series embeddings. The embedding process can be time-consuming and lossy, but it does allow for a unified framework for evaluating SD.

Some other metrics have a more natural extension to time-series. For example, for maximum mean discrepancy, we can regard each sample as a matrix instead of a vector (for multiple time-series), embed matrices using an appropriate kernel, and compute Frobenius- instead of $l2$-norm to calculate the metric [43]. For other standard probabilistic distance measures such as JS divergence and Wasserstein distance, approximations for time-series exist [151].

Since time-series predictors are typically more complex, we do see that some metrics lose their applicability. For example, in multiple time-series, we often need to rely on feature importances instead of regression coefficients to

assess aetiological utility, due to the complex nature of popular predictive models (e.g., recurrent neural networks). Additionally, time-series analyses may yield additional use cases such as forecasting, such that predictive analyses need to predict next-step temporal vectors instead of a single target feature, using the TSTR or TRTS framework [189].

3.3.3 Images. Images are an unstructured data format ubiquitous in domains where privacy-preservation is vital, e.g., healthcare (MRIs, CT scans) and biometrics (facial recognition). Human evaluation is more widely applicable to images than to structured data formats; especially human opinion scores can be valuable to assess SD realism and alignment with human preference [19].

Individual features in pixels or patches are generally uninformative, motivating adaptation of common and new evaluation metrics [48]. Image-specific descriptive statistics relating to, e.g., brightness (mean pixel value), contrast (range of pixel values), and sharpness (Laplacian variance), can provide an indication whether SD accurately captures high-level image properties. However, this is no guarantee that SD is even remotely realistic, unlike in structured data formats, where similarity in a wide range of descriptive statistics is typically a fairly good predictor for overall fidelity. For images, descriptive statistics should rather be seen as a supplement to other evaluations, instead of providing an indication of overall fidelity.

Other metrics can be computed in a more meaningful feature space constructed through image models. For example, for association rule mining on images, we can first extract image labels using, e.g., object detection models, and perform association rule mining on extracted labels [122]. Otherwise, for projection similarity and statistical distance and similarity measures, we can compute metrics in an embedding space constructed through (pre-trained) image encoders. This is often more useful, since random artifacts produced in the SD generation process can lead to poor similarity scores in pixel-space even when samples are perceptually appealing [106, 140, 142, 157]. Tanfoni et al. [154] show that, for example, the classifier test can mainly draw on background artifacts to distinguish SD from RD in facial images.

The importance of perceptual appeal in images has also given rise to new image-specific probabilistic measures designed to align with human preference. FID encodes images using an Inception model [153] and computes Fréchet distance between Gaussians fitted on SD and RD image embeddings. Heusel et al. [66] mention that by using Fréchet distance, the metric correlates well with human judgement: it is more sensitive to types of disturbances such as implanted black rectangles, insertion of external images, and salt and pepper noise than low levels of Gaussian noise or blur. Also, the metric correlates well with the amount of disturbance added [66]. In the medical domain, lower Fréchet distance has been found to overlap with the indistinguishability of synthetic and real images by experts on synthetic dermatological images [108]. However, Bińkowski et al. [13] rightly state that Gaussians are unsuitable priors for Inception embeddings since they rely on ReLU activations and are thus skewed, and propose squared maximum mean discrepancy instead of Fréchet distance, coining the metric Kernel Inception Distance. Current work on metrics that align well with human preferences aim to strike a balance between low-level patch-based metrics and high-level concepts in an image such as object categories [48].

Pre-trained image feature extractors need to be appropriate to the dataset domain. When the distribution of target data moves away from the training data (of the image model), the extracted features become unreliable and may either miss problematic artifacts, or emphasise irrelevant differences. For example, Naeem et al. [117] observe that random embeddings are more suitable than Inception embeddings (trained on natural images) for evaluating handwritten digits and spectrograms. Other studies adapt metrics such as FID by using domain-specific image encoders instead of the Inception network [125]. There is some pushback against this idea, however, since domain-specific encoders do not always align better with human judgment than Inception networks [177]. Whether or not to adapt feature extractors to

the dataset domain depends on the specific dataset and its distance to the training set (e.g., natural images), and the availability of large and diverse domain-specific datasets for pre-training.

We omit some very common image metrics found in image augmentation and domain translation studies, since they either do not measure similarity to the RD (e.g., Inception Score), or measure sample-level similarity, thereby indicating privacy loss rather than fidelity (e.g., SSIM [170], LPIPS [192]).

Aetiological and predictive utility measures for images can employ deep learning methods for prediction (e.g., convolutional neural networks, vision transformers). For aetiological analyses, however, typical feature importance measures (saliency maps, SHAP) provide image-level rather than feature-level explanations, complicating the assessment of overall feature importance similarity. Rather, analyses rely on manually investigating whether predictions adequately draw on objects or cues found in synthetic images [58].

Privacy measures for synthetic images similarly adapt to (partly) operate in a meaningful feature space rather than (solely) pixel space. High feature-level similarity can still leak privacy even when pixel-level similarity is low, e.g., when rotating an image. For example, Chen et al. [23] measure distance-based membership disclosure risk using a weighted measure including distance in pixel- and feature-space.

3.3.4 Text. Natural language is regarded as unstructured data since its features of interest typically have to be extracted by some further qualitative or computational method. Social media posts and newspaper articles are examples of text data that provide relevant information to social science, e.g. regarding sentiment about policy [172], but also to healthcare, for example in forums that discuss adverse drug effects [126]. The Transformer network architecture [166] in combination with large swathes of web text has mainly driven the ascent of Large Language Models (LLMs) [17], which made it much easier to generate domain-specific texts of high quality via API calls at scale [14] .

The fidelity of synthetic text can be evaluated with approaches similar to FID, where instead of an image encoder a text encoder such as BERT [31] is used to embed synthetic and real texts and evaluate to what extent high-level features co-occur in representation space [179]. Fidelity evaluation that draws on assessing projection similarity of extracted features is sometimes also used [72], though this is not very common as features may be hard to interpret – they could be the presence of specific words but also less obvious combinations of tokens or grammatical structures. The training distribution of common text encoders such as BERT is often different from the real-world distribution of interest [168]. This means that using such models for embedding e.g. clinical notes may not capture features of interest well, so using a specific model such as ClinicalBERT [71] is often preferred. Relatedly, as many easily accessible LLMs are also general-purpose, human evaluation of LLM-generated synthetic text for specific domains will often reveal unrealistic properties, such as too verbose or articulate clinical notes or tweets [107, 168].

As human evaluation is costly, proxies for this kind of fidelity include n-gram overlap, for example ROUGE or Self-BLEU scores between real and synthetic texts [27, 72]. Such scores gauge sample-level similarity and could also be used to indicate generator failure modes and privacy risks such as overfitting or memorisation. These are relatively simple metrics and are best complemented with sample similarity based on cosine similarity which is more sensitive to semantics and context [72, 104], or the BERT-score mentioned above for overall similarity. Classifier test setups are another common and good choice to assess the realism of synthetic texts – if a classifier cannot reliably distinguish real from synthetic texts, this is an indication that the synthetic text is of high quality [116]. Another probabilistic fidelity metric is perplexity, which indicates the average log-likelihood of some synthetic text sequence from the perspective of some reference model [116]. Often the generator is used to calculate how dissimilar the synthetic text is compared to the learned distribution, which can be problematic as it is insensitive to failure modes such as overfitting. Using

another reference model is not a solution, as perplexity is ill-defined for models that have different tokenizers hence different vocabularies [79].

Regarding utility evaluation, synthetic text can be employed in TSTR and TRTS setups on downstream tasks such as text classification, entity recognition, and sentiment analysis, and evaluated with common metrics such as AUC, precision and recall, and F1-score. Synthetic text can also be evaluated against open benchmarks to evaluate specific aspects such as its factuality [104].

Concerning privacy, given that many frequently used LLMs such as GPT-4o or Llama are trained on private web datasets, it is unclear what sensitive information may surface when generating SD. This is different in scenarios where a researcher trains a generator with full control over the RD. LLMs can leak sensitive information even when they are not overfit, and successful retrieval of sensitive information is often prompt-dependent hence variable and difficult to quantify [20]. Synthetic texts are frequently used for supervised fine-tuning to further optimise LLMs for comprehension and reasoning, and here MIAs and PII leakage are good strategies to assess privacy [61]. Combinations of rule-based and manual inspection for sensitive information are also still used [27] and may be advisable in high-stakes contexts.

Synthetic text is predominantly used for fine-tuning, instruction-tuning and aligning LLMs, or augmenting existing NLP-datasets, which is why evaluation is often in terms of utility [129]. LLMs have capacities that other data generators lack, such as self-reflection, which introduces non-standard evaluation metrics such as the number of necessary iterations over its own synthetic text to further improve it [72]. In addition, given the increasing size and training complexity of LLMs, 'meta' evaluation metrics that target efficiency such as SD generation time and power consumption will likely become more important in evaluation [147].

3.3.5 Graphs, Audio, Video, and Multi-Modal. Graph data consists of nodes connected by edges, and is commonly used to express connected networks in a structured format. For graph SD, fidelity is assessed through graph-specific properties such as spectrum, distance distribution, betweennes centrality, likelihood, degree distribution, and clustering [98]. These graph features are often put into a feature vector, whereby we can measure similarity (i.e., fidelity) between graphs by measuring distance between feature vectors, e.g., L2 norm [141]. Which features are most important, and which distance metric is appropriate, remains an open question, however [98]. Utility can be more straightforwardly assessed by measuring prediction accuracy from SD through a TSTR or TRTS setup [57]. Privacy metrics for synthetic graphs similarly consider whether individual graphs (or queries from graphs) are not "too similar" to the RD (especially regarding sensitive information), or whether adversaries can construct attacks to infer information (e.g., reidentify nodes) [197]. Differential privacy can also be incorporated in synthetic graph generation, allowing a more controlled tradeoff between graph fidelity and privacy [41].

Privacy-preserving synthetic audio generation is still in its infancy; most audio generation studies focus on data augmentation or domain translation (e.g., text-to-speech). Privacy-preserving approaches for audio focus mostly on perturbing real signals such that they are unidentifiable [123]. Audio SD can be seen as single or multiple continuous time-series (dependent on whether a single or multiple fragments are processed) in 1 dimension (raw waveforms) or 2 dimensions (spectrograms). For audio in stereo, another dimension is added. Correspondingly, waveform data is often processed similarly to time-series, while spectrograms are processed similarly to images, and (adaptations to) evaluation metrics for SD correspond to those found for these two categories, see Section 3.3.2 and Section 3.3.3. Due to the humanly interpretable nature of audio, human opinion scores are considered much more valuable than for regular structured time-series, however [38, 54].

Video data consists essentially of series of images. Video SD quality can be assessed by adapting image metrics to use feature extractors for videos instead of images [161], e.g., I3D models which use 3D instead of 2D convolutions to capture the temporal dimension [21].

Multi-modal data consists of combinations of different types of data, e.g., tabular and time-series, image and text, or video and audio. Deep learning provides methods to learn shared representations of multiple modalities, which can be useful for, e.g., density estimation, distance or similarity computation, and prediction, allowing the aforementioned evaluation metrics to be computed (Section 2). For example, synthetic electronic health records often contain both tabular and time-series data, for which we can use neural networks with separate recurrent and feedforward layers for, e.g., TSTR and the classifier test [1].

3.4 What is the Domain of the Dataset?

Domain-specific feature extractors can be useful to compute evaluation metrics in a feature space relevant to the dataset domain (see e.g. Section 3.3.3). Additionally, since privacy-preserving SD replaces RD, it will be used in analyses common to that domain, making these specific analyses more important during evaluation. 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 a SD set intended to replace or augment RD (see Section 3.2). For example, association rule mining is widely applied in electronic health records, 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 common analyses to ensure SD is useful to the intended audience.

The dataset domain also influences whether outliers or anomalies (and corresponding metrics) require special attention. 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 [46], 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 [121] can indicate whether outliers from RD are too closely mimicked in SD, potentially leaking sensitive information.

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

SD generators are trained to mimic a RD training set. However, high similarity between SD and this RD training set can be the result of overfitting, which can lead to various privacy-related issues (Section 2.3). To assess generalization of SD it should, therefore, be evaluated with respect to an independent RD test set whenever possible. The training set should, typically, only be used in privacy evaluations, to investigate whether indeed any overfitting to the training data has occurred.

4.2 Uncertainty and Bias

The training process of many SD generators is stochastic, and therefore, different random initializations lead to different evaluation metrics. To quantify uncertainty in evaluation metrics we can train SD generators across different random initializations and report the metric distribution over folds [163].

Evaluation metrics can be biased in multiple ways. As mentioned above, random initialization of SD generators affect downstream evaluation metrics, thus, biased initialization leads to biased evaluation. One solution is to infer a point estimate (e.g., mean or median) from a distribution over metrics from SD generators with varying initializations.

Secondly, metrics are usually computed with respect to an independent RD test set to assess generalization capacity of the SD generator (Section 4.1). Random train-test splits can be biased, which is propagated to evaluation. To combat this, we can compute metrics for different train-test partitions, and infer a point estimate (if necessary).

Combatting both these sources of bias calls for a sort of nested cross validation procedure, where SD generators are trained for multiple initializations and multiple train-test splits. 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 [163], whereas in regards to train-test splits, small datasets typically incur more bias than larger datasets due to higher variance across partitions.

4.3 Auditing

Next to reporting SD quality, evaluation metrics can be used to improve existing SD post-generation. Sample-level metrics can inform an auditing process which removes poor samples and queries additional (potentially good) samples from a generative model [5]. This process can improve SD quality without changes to the underlying generative model or training process. Not all metrics are suited for this, however. For example, when aiming to improve SD fidelity, the metric should inform on both likelihood ánd diversity of SD (and potentially generalization), such that auditing does not incur mode collapse. Good choices are, therefore, precision-recall type metrics [5].

5 Illustrative Experiments: Electronic Health Records of Heart Failure Patients

Now, for an illustrative tabular dataset, we show how our survey can inform effective and reliable SD evaluation. To ensure impactful and realistic experiments, we select a dataset of Electronic Health Records; privacy-preserving SD generation is a widely researched topic in this domain [1, 9, 94, 97, 167, 184, 185, 190]. We select a patient cohort from the Medical Information Mart for Intensive Care (MIMIC)-IV (version 3.1) [76]. 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. Table 3 summarizes all included variables. Our code for these experiments is available at https://github.com/JimAchterbergLUMC/EvaluationMetricsSD/.

Firstly, we consider the 4 questions (denoted as Q1, Q2, Q3, Q4, respectively) from the framework displayed in Figure 3 to assess which metrics are relevant for this dataset. With respect to Q1, this dataset can be used for a variety of analyses, e.g., explorative (clustering for patient risk stratification), aetiological (analyzing regression coefficients to find which features affect mortality), and predictive (predicting mortality). We assume a currently unknown downstream task. This means we focus mainly on fidelity metrics rather than task-specific utility metrics. With respect to Q2, we first consider selecting the best SD generator out of a variety of generators, after which we also consider the goal of replacing RD with a set of SD. The former requires compressed quantitative metrics, whereas the latter requires metrics

Table 3. Summary of Heart Failure Dataset from MIMIC-IV.

Variable	Type	Description
age	Numerical	Patient age at admission. This is an approximation, as the exact year patients received care has been deidentified by grouping several years together.
sex	Categorical	Genotypical sex of the patient: Male ('M') or Female ('F').
ethnicity	Categorical	Patient ethnicity. Grouped into 'White', 'Black', or 'Other'.
marital_status	Categorical	Patient marital status: 'single', 'married', 'widowed', or 'divorced'.
bmi	Numerical	Patient Body Mass Index as measured closest to admission time.
bp_systolic	Numerical	Patient systolic blood pressure as measured closest to admission time.
bp_diastolic	Numerical	Patient diastolic blood pressure as measured closest to admission time.
n_diagnoses	Numerical	Total number of diagnoses during admission.
admission_type	Categorical	Admission characteristics useful for classifying admission urgency. There are 9 possibilities: 'AMBULATORY OBSERVATION', 'DIRECT EMER.', 'DIRECT OBSERVATION', 'ELECTIVE', 'EU OBSERVATION', 'EW EMER.', 'OBSERVATION ADMIT', 'SURGICAL SAME DAY ADMISSION', 'URGENT'.
admission_location	Categorical	Location of the patient prior to admission at the Emergency Department. There are 11 possibilities: 'PHYSICIAN REFERRAL', 'WALK-IN/SELF REFERRAL', 'EMERGENCY ROOM', 'TRANSFER FROM HOSPITAL', 'CLINIC REFERRAL', 'TRANSFER FROM SKILLED NURSING FACILITY', 'PROCEDURE SITE', 'PACU', 'AMBULATORY SURGERY TRANSFER', 'INTERNAL TRANSFER TO OR FROM PSYCH', 'INFORMATION NOT AVAILABLE'
los	Numerical	Length of admission stay in days.
mortality	Categorical	Whether patient passed away during admission.

which are more interpretable by our target audience. Regarding Q3, we consider a tabular dataset, which means we incorporate metrics on feature-level similarity and take data heterogeneity into account. Finally, regarding Q4, we consider a dataset in the healthcare domain. This allows us to incorporate analyses and measures which are commonly found in this domain, such as certain physiological constraints, and association rule mining.

5.1 Selecting Synthetic Data Generator

We first focus on the goal of comparing SD generators and consider a variety of SD algorithms with different architectural styles, using their implementation from the Synthcity library [137]: ARF [173], Bayesian Network (BN), CTGAN [182], TVAE [182] and TabDDPM [90]. Details on hyperparameters can be found in the Appendix B.2.

When benchmarking SD generators, we focus on easily comparable quantitative metrics (see Q2 in Figure 3). Furthermore, as Section 3.1 describes, we need to utilize metrics to get insight into various aspects of SD fidelity, e.g., overall similarity (classifier test), geometric similarity (Wasserstein distance), and general over- and underfitting (precision-recall). We use the classifier test rather than, e.g., JS divergence, to circumvent direct density estimation. Additionally, we include the authenticity score and DOMIAS to indicate failure modes and privacy issues caused by the training process (i.e., overfitting). Other, more practical privacy risks, e.g., attribute disclosure or singling out, are more of a property of the dataset and information available to an attacker than a failure during the training process, and thus less relevant during benchmarking than for reporting privacy risk of a specific SD set to stakeholders.

For the classifier test we use an XGBoost classifier and report the AUC. For precision and recall we use Kynkäänniemi et al. [91]'s method for estimating coverage. For DOMIAS we use an attack dataset consisting of members (training set) and non-members (half of the test set), and a reference set (other half of the test set) which attackers can use to

estimate RD density. We estimate density using Gaussian KDE; as this requires weakly correlated input features, we apply PCA first. Since PCA can over-smooth data when retaining fewer components which makes it more difficult to estimate local overfitting, and since the dataset has relatively few features, we retain a large amount of principal components, i.e., preserving at least 99% of the variance.

To mitigate bias from random initializations and train-test splits (Section 4.2), we perform 3-fold cross validation and use 3 different initializations within each split, effectively training each generator 9 times.

Table 4. Benchmarking results of SD generators for the Heart Failure dataset. Best results per metric in bold, second best underlined. Arrows indicate whether higher (↑) or lower (↓) metric values are desirable.

	Fidelity				Privacy		Training
	Classifier test ↓	Wasserstein ↓	Precision ↑	Recall ↑	Authenticity ↑	DOMIAS ↓	Time (m)* ↓
ARF	0.910 ± 0.091	1.731 ± 0.036	0.744 ± 0.013	**0.973 ± 0.003**	0.628 ± 0.007	0.501 ± 0.008	<u>0.541 ± 0.006</u>
BN	<u>0.905 ± 0.101</u>	<u>1.646 ± 0.014</u>	<u>0.805 ± 0.011</u>	<u>0.960 ± 0.004</u>	0.431 ± 0.007	0.651 ± 0.004	**0.046 ± 0.002**
CTGAN	0.972 ± 0.036	1.863 ± 0.189	0.773 ± 0.077	0.860 ± 0.054	<u>0.684 ± 0.078</u>	0.495 ± 0.005	1.774 ± 0.044
TVAE	0.956 ± 0.052	1.920 ± 0.194	0.745 ± 0.058	0.849 ± 0.068	**0.704 ± 0.067**	<u>0.492 ± 0.006</u>	0.933 ± 0.027
TabDDPM	**0.677 ± 0.026**	**1.484 ± 0.055**	**0.917 ± 0.006**	0.933 ± 0.006	0.577 ± 0.003	**0.488 ± 0.007**	2.662 ± 0.017

*Average training time in minutes - CTGAN, TVAE and TabDDPM on GPU, ARF and BN on CPU (see Appendix C for more details).

The results in Table 4 highlight the importance of using a carefully selected set of informative metrics, since dependent on the context of the user, different generators can come out on top. For a strong overall tradeoff between fidelity and privacy, the TabDDPM generator seems the most obvious choice. However, when computational costs are also of concern, the BN can be a strong choice, as it scores second in terms of fidelity at a fraction of the training time of TabDDPM. When strong privacy guarantees are also required, the BN is not the strongest choice, and the ARF might bring a better tradeoff between fidelity, privacy, and computational costs. For now, we assume adequate access to resources and care mainly about a strong tradeoff between fidelity and privacy, therefore selecting TabDDPM as our final generator.

5.2 Replacing RD with SD: Evaluation Report for TabDDPM

Next, we focus on the goal of replacing RD with SD and assume we need to report the quality of SD generated from TabDDPM (intended to replace RD) to a set of stakeholders in the healthcare domain. We generate a single set of SD from TabDDPM from a random 50% training split of RD - we use a relatively small training set to ensure a larger independent validation set can be used to evaluate SD. This is especially useful to construct a reasonable attack dataset for MIAs, for which usually it is assumed that the attacker has access to data from the same population as the RD. It is, however, debated whether an adversary can really learn new information in this setup [132].

The framework (Figure 3) indicates that, for this scenario, we now focus on interpretable metrics (Q2) with a strong focus on feature-level similarity (Q3), and incorporate analyses and domain knowledge relevant to healthcare (Q4). For the latter, we can identify various physiological constraints that should hold in patient records, e.g., systolic blood pressure should exceed diastolic blood pressure. Additionally, we can incorporate association rule mining as a common data mining analysis in healthcare [9, 10, 83, 186], to investigate whether SD exhibits similar feature associations. Regarding interpretable feature-level similarity measures we include feature-wise plots and bivariate correlation statistics.

Systolic blood pressure should always exceed diastolic blood pressure in patient records (bp_systolic > bp_diastolic). For RD this constraint holds for all patients, whereas for SD this holds for 99.8% of patients (13 violations in the test

set). This shows that even even generators which seem to generate high fidelity SD according to statistical metrics (Table 4), can exhibit unrealistic artifacts which are easily picked out by domain experts.

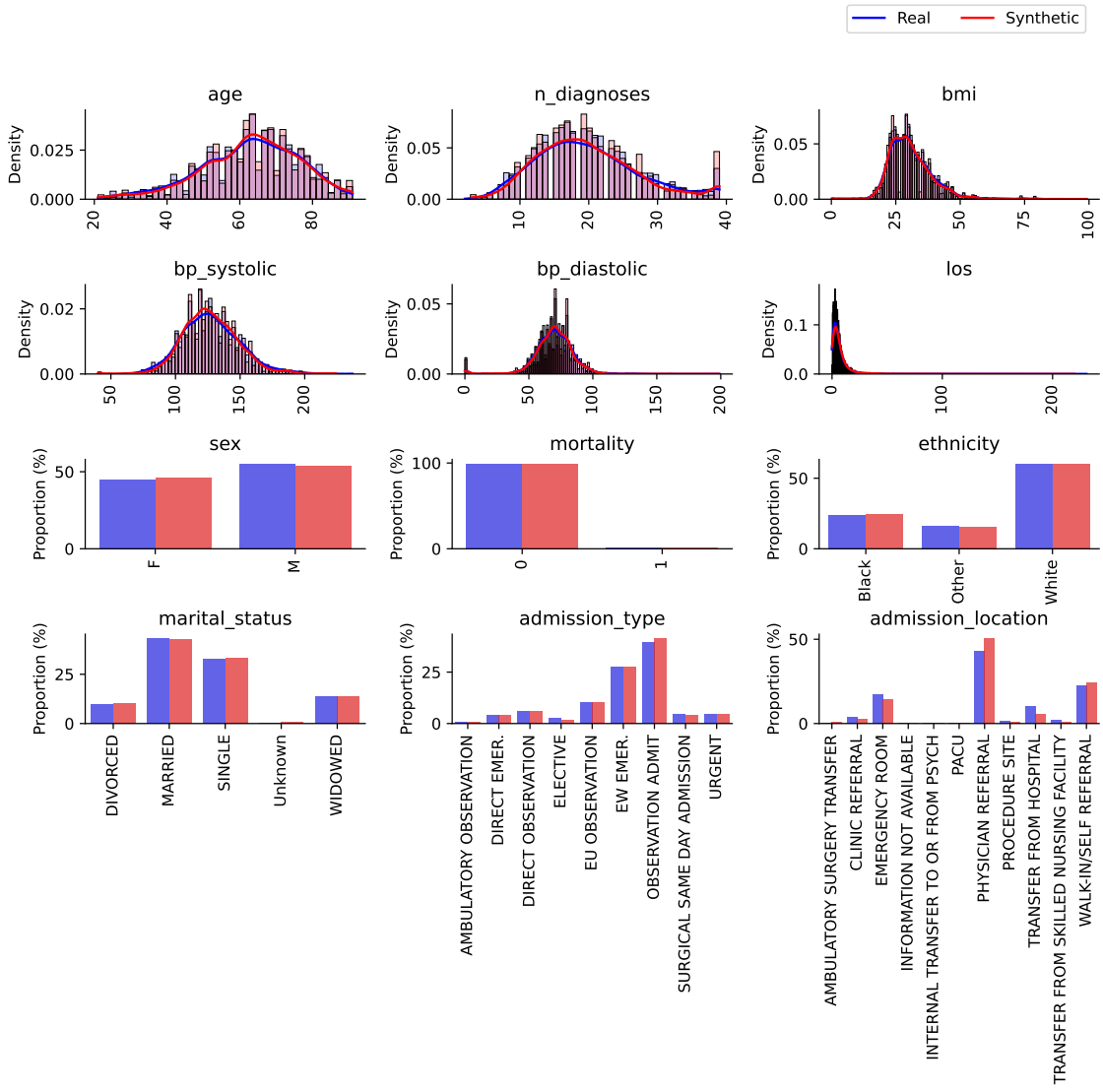

Fig. 4. Marginal distributions of the twelve features of the real and synthetic Heart Failure dataset. See Table 3 for feature descriptions.

Figure 4 shows each feature from SD and RD for the Heart Failure dataset. TabDDPM captures most marginal distributions well. It only seems to generate lower quality SD for the admission_location feature: it overestimates frequencies of the two most occurring categories, and underestimates frequencies for others. Figure 5 shows correlation matrices of SD and RD, i.e., Spearman's r for numerical features, Cramérs V for categorical features, and correlation

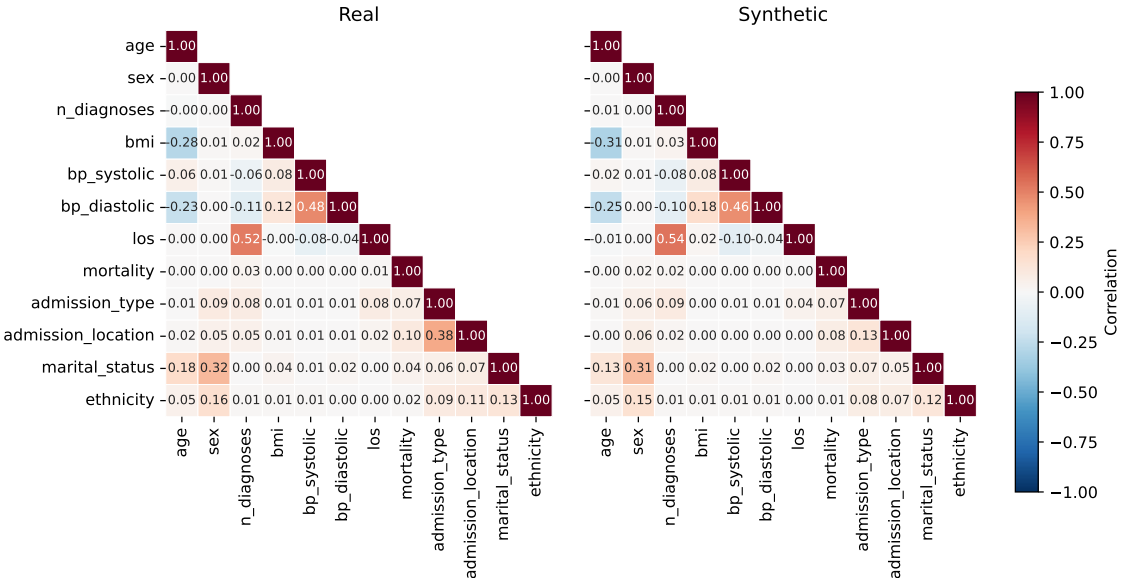

Fig. 5. Correlation matrices of the real and synthetic Heart Failure dataset.

ratio η^2 for numerical-categorical features. Overall, SD seems to contain mostly similar bivariate correlation statistics in terms of their direction and order of magnitude.

From an association rule mining algorithm (Apriori, see Appendix B.3 for more details) we find 7 association rules in RD and 21 rules in SD, with precision and recall of 0.33 and 1.000 respectively. This indicates that the SD contains all rules rules found in RD, but also "hallucinates" other rules which were not found in the RD. An example of a rule found in both SD and RD is: when BMI (bmi) is low, length of stay (los) is also low. An example of a hallucinated rule found in SD is: when patients were referred by a physician (admission_location), length of stay (los) is low.

After evaluating fidelity, we move to thoroughly investigate privacy risk of SD generated from TabDDPM. As Section 3.2 indicates, we need to evaluate risk of i) identity disclosure, ii) membership disclosure, and iii) attribute disclosure. Table 5 provides results for metrics on identity, membership, and attribute disclosure.

To identify records at risk of singling out, we can compute NNDRs and investigate samples with notably low values. For these samples, SD is close to RD without other RD in the vicinity (relatively speaking), which could be an indication that these samples are unique/identifiable and present in SD - a potential risk for singling out. Identity disclosure risk still depends on context, given that the RD or generator will often not be available to an attacker, so NNDR can only provide guidance to select and further evaluate datapoints at risk of identity disclosure. For the TabDDPM SD we find NNDR ratio ≈ 1 for the first few percentiles (Table 5), indicating low risk of identity disclosure even for those samples which are closest to training data without other data in the vicinity.

An authenticity score of 0.566 indicates that there is no significant risk that training data is copied across the entire SD set. To further investigate membership disclosure risk, we analyse local overfitting in SD. To this end, we perform the DOMIAS MIA [164], using the same set-up as in Section 5.1. Here, we select all features to be available to the attacker; although this isn't *necessarily* the most risky scenario [132], we observe that it is in this case, after testing various combinations of quasi-identifiers.

This time, we use a different accuracy metric than AUC, as we are now interested in *how many* and *to what extent* RD is at risk, which AUC cannot discern. As Pilgram et al. [132] indicate and we discuss in Section 3.2.2, metrics such as F1 score are commonly used, where we should adjust the score for i) a naive baseline MIA, ii) prevalence of membership in the attack dataset, and iii) the type of RD and vulnerabilities potentially exposed. As a naive baseline, we predict all samples to be members in the dataset. The attack dataset's members:non-members ratio is equal to 2:1. To provide more insight into the attack, we report precision and recall instead of only its weighted F1. The naive baseline's precision and recall are 0.667 and 1.000 respectively. Note that the naive baseline always achieves perfect recall. Thus, in this scenario, our attack can only improve upon precision. This can be considered a realistic threat scenario, since high precision attacks can often be considered more risky than, e.g., moderate precision and recall attacks. Since, in high precision attacks, attackers are highly accurate at disclosing membership of a select group of individuals, which can be more risky than attackers being moderately accurate at disclosing membership for a broader population.

To simulate a high precision attack, we change the standard classification threshold in DOMIAS from the median to the top 5th percentile. This way, we only aim to disclose membership for the top 5% most likely individuals - according to DOMIAS' relative probability scores. Note that recall will be harmed significantly. Appendix A.1 provides more details and evidence on the potential gain which can be achieved through this high precision attack set-up.

Our attack achieves precision and recall of 0.676 and 0.051 respectively, indicating a 1.3% increase in precision. Releasing this SD likely only slightly increases the confidence of attackers who aim to disclose membership of a (select) group of RD to the training set used for SD generation. The extent to which this increase is problematic depends, however, on the broader context, which should also involve privacy officers and law experts, that place this number in a broader assessment of how long SD will be available, how likely it is that other data possibly relevant to an attacker becomes available, advances in data mining techniques, and so on.

To investigate attribute disclosure risk, we simulate an AIA using an XGBoost model as predictor. Here, we need to make some assumptions on which quasi-identifiers might be available to an attacker, and which sensitive features might be at risk of reidentification. In this experiment we assume that attackers might have access to (or are able to make an educated guess on) the features age, bmi, and sex, and try to infer information on patients previous and current whereabouts through variables on admission location (admission_location) and length of stay (los). Similar as for membership disclosure, attribute inference accuracy metrics need to indicate whether publishing SD poses *additional* risk over some naive baseline. To this end, we use the AUC for admission location and R^2 score for length of stay. R^2 score indicates whether predictions are better than the best unbiased constant prediction (the average), in which case $R^2 > 0$. Our AIA achieves AUC of 0.892 for admission location and R^2 score of -0.028 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.*

5.3 Comparison With Previous Works

We do not imply that the metrics we use in these experiments are 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 inform on failure modes such as mode collapse/invention and overfitting/underfitting (Figure 2) [9, 83, 159, 167, 184], set-up MIA scenarios without sufficient domain-specific justification or alignment with realistic threat models [184], or only

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

Disclosure Risk	Metric	Measure	Value
Identity Disclosure	NNDR Ratio	Percentile 1	0.982
	NNDR Ratio	Percentile 2	0.986
	NNDR Ratio	Percentile 5	0.994
Membership Disclosure	Authenticity	Score	0.566
	DOMIAS (naive)	Precision	0.667
	DOMIAS (naive)	Recall	1.000
	DOMIAS (KDE)	Precision	0.676
	DOMIAS (KDE)	Recall	0.051
Attribute Disclosure	admission_location (XGBoost)	AUC	0.892
	los (XGBoost)	R^2	-0.028

assess a subset of the relevant privacy risk dimensions which should be evaluated (attribute, membership, and identity disclosure) [9, 83, 159, 167].

6 Discussion and Conclusion

Though SD have been widely recognized as offering the next best privacy-preserving technique for providing data in increasingly data-driven societies [70], 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) specific type of data and iv) domain at issue (Section 3). By setting up an illustrative experiment in line with general practical recommendations on evaluating SD, (Sections 4 and 5), we showed what choices regarding task, goal, data type and domain 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. 28, 131]. 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) [101]. 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 [28]. 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.

7 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] Gergely Acs, Luca Melis, Claude Castelluccia, and Emiliano De Cristofaro. 2018. Differentially private mixture of generative neural networks. *IEEE Transactions on Knowledge and Data Engineering* 31, 6 (2018), 1109–1121.

[5] 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.

[6] 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.

[7] Martin Arjovsky, Soumith Chintala, and Léon Bottou. 2017. Wasserstein generative adversarial networks. In *International conference on machine learning*. PMLR, 214–223.

[8] 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

[9] 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.

[10] 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.

[11] 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).

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

[13] Mikołaj Bińkowski, Dougal J. Sutherland, Michael Arbel, and Arthur Gretton. 2018. Demystifying MMD GANs. In *International Conference on Learning Representations*. https://openreview.net/forum?id=r1lUOzWCW

[14] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258* (2021).

[15] 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.

[16] 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.

[17] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems* 33 (2020), 1877–1901.

[18] 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.

[19] Francesco Calimeri, Aldo Marzullo, Claudio Stamile, and Giorgio Terracina. 2017. Biomedical data augmentation using generative adversarial neural networks. In *International conference on artificial neural networks*. Springer, 626–634.

[20] Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, et al. 2021. Extracting training data from large language models. In *30th USENIX security symposium (USENIX Security 21)*. 2633–2650.

[21] Joao Carreira and Andrew Zisserman. 2017. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 6299–6308.

[22] 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.

[23] Dingfan Chen, Ning Yu, Yang Zhang, and Mario Fritz. 2020. Gan-leaks: A taxonomy of membership inference attacks against generative models. In *Proceedings of the 2020 ACM SIGSAC conference on computer and communications security*. 343–362.

[24] 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.

[25] 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.

[26] 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.

[27] Yao-Shun Chuang, Atiquer Rahman Sarkar, Yu-Chun Hsu, Noman Mohammed, and Xiaoqian Jiang. 2025. Robust privacy amidst innovation with large language models through a critical assessment of the risks. *Journal of the American Medical Informatics Association* (2025), ocaf037.

[28] 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.

[29] 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.

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

[31] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*. 4171–4186.

[32] Mihai Dogariu, Liviu-Daniel Ştefan, Bogdan Andrei Boteanu, Claudiu Lamba, Bomi Kim, and Bogdan Ionescu. 2022. Generation of realistic synthetic financial time-series. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 18, 4 (2022), 1–27.

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

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

[35] 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.

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

[37] Peter Eigenschink, Thomas Reutterer, Stefan Vamosi, Ralf Vamosi, Chang Sun, and Klaudius Kalcher. 2023. Deep Generative Models for Synthetic Data: A Survey. *IEEE Access* 11 (2023), 47304–47320. https://doi.org/10.1109/ACCESS.2023.3275134

[38] Peter Eigenschink, Thomas Reutterer, Stefan Vamosi, Ralf Vamosi, Chang Sun, and Klaudius Kalcher. 2023. Deep Generative Models for Synthetic Sequential Data: A Survey. *IEEE Access* (2023).

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

[40] 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.

[41] Marek Eliáš, Michael Kapralov, Janardhan Kulkarni, and Yin Tat Lee. 2020. Differentially private release of synthetic graphs. In *Proceedings of the Fourteenth Annual ACM-SIAM Symposium on Discrete Algorithms*. SIAM, 560–578.

[42] 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.

[43] 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).

[44] 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.

[45] Alvaro Figueira and Bruno Vaz. 2022. Survey on synthetic data generation, evaluation methods and GANs. *Mathematics* 10, 15 (2022), 2733.

[46] 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).

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

[48] Stephanie Fu, Netanel Tamir, Shobhita Sundaram, Lucy Chai, Richard Zhang, Tali Dekel, and Phillip Isola. 2023. DreamSim: Learning New Dimensions of Human Visual Similarity using Synthetic Data. 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., 50742–50768. https://proceedings.neurips.cc/paper_files/paper/2023/file/9f09f316a3eaf59d9ced5ffaefe97e0f-Paper-Conference.pdf

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

[50] 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.

[51] Georgi Ganev, Kai Xu, and Emiliano De Cristofaro. 2024. Graphical vs. Deep Generative Models: Measuring the Impact of Differentially Private Mechanisms and Budgets on Utility. In *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security*. 1596–1610.

[52] EU GDPR. 2018. General data protection regulation (gdpr).

[53] 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.

[54] Karan Goel, Albert Gu, Chris Donahue, and Christopher Ré. 2022. It's raw! audio generation with state-space models. In *International conference on machine learning*. PMLR, 7616–7633.

[55] 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.

[56] 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.

[57] Xiaotian Han, Zhimeng Jiang, Ninghao Liu, and Xia Hu. 2022. G-mixup: Graph data augmentation for graph classification. In *International Conference on Machine Learning*. PMLR, 8230–8248.

[58] Romain Hardy, Joe Klepich, Ryan Mitchell, Steve Hall, Jericho Villareal, and Cornelia Ilin. 2023. Improving nonalcoholic fatty liver disease classification performance with latent diffusion models. *Scientific Reports* 13, 1 (2023), 21619.

[59] 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.

[60] 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.

[61] Yu He, Boheng Li, Liu Liu, Zhongjie Ba, Wei Dong, Yiming Li, Zhan Qin, Kui Ren, and Chun Chen. 2025. Towards label-only membership inference attack against pre-trained large language models. In *USENIX Security*.

[62] Simon Hediger, Loris Michel, and Jeffrey Näf. 2022. On the use of random forest for two-sample testing. *Computational Statistics & Data Analysis* 170 (2022), 107435.

[63] John Heine, Erin EE Fowler, Anders Berglund, Michael J Schell, and Steven Eschrich. 2023. Techniques to produce and evaluate realistic multivariate synthetic data. *Scientific Reports* 13, 1 (2023), 12266.

[64] 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.

[65] 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.

[66] 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).

[67] Benjamin Hilprecht, Martin Härterich, and Daniel Bernau. 2019. Monte carlo and reconstruction membership inference attacks against generative models. *Proceedings on Privacy Enhancing Technologies* (2019).

[68] 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.

[69] Harold Hotelling. 1933. Analysis of a complex of statistical variables into principal components. *Journal of educational psychology* 24, 6 (1933), 417.

[70] 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).

[71] Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. 2019. Clinicalbert: Modeling clinical notes and predicting hospital readmission. *arXiv preprint arXiv:1904.05342* (2019).

[72] Yue Huang, Siyuan Wu, Chujie Gao, Dongping Chen, Qihui Zhang, Yao Wan, Tianyi Zhou, Chaowei Xiao, Jianfeng Gao, Lichao Sun, and Xiangliang Zhang. 2025. DataGen: Unified Synthetic Dataset Generation via Large Language Models. In *The Thirteenth International Conference on Learning Representations*. https://openreview.net/forum?id=F5R0lG74Tu

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

[74] Harold Jeffreys. 1946. An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences* 186, 1007 (1946), 453–461.

[75] 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*.

[76] 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

[77] 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).

[78] 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*.

[79] Daniel Jurafsky and James H. Martin. 2025. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models, 3rd edition*. Online manuscript released January 12. https://web.stanford.edu/ jurafsky/slp3.

[80] 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.

[81] Ha Ye Jin Kang, Erdenebileg Batbaatar, Dong-Woo Choi, Kui Son Choi, Minsam Ko, and Kwang Sun Ryu. 2023. Synthetic tabular data based on generative adversarial networks in health care: generation and validation using the divide-and-conquer strategy. *JMIR Medical Informatics* 11 (2023), e47859.

[82] 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.

[83] 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.

[84] 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.

[85] Ilmun Kim, Aaditya Ramdas, Aarti Singh, and Larry Wasserman. 2021. Classification accuracy as a proxy for two-sample testing. *The Annals of Statistics* 49 (2021), 411–434.

[86] 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.

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

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

[89] Lennart R Koetzier, Jie Wu, Domenico Mastrodicasa, Aline Lutz, Matthew Chung, W Adam Koszek, Jayanth Pratap, Akshay S Chaudhari, Pranav Rajpurkar, Matthew P Lungren, et al. 2024. Generating synthetic data for medical imaging. *Radiology* 312, 3 (2024), e232471.

[90] 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.

[91] 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).

[92] 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.

[93] 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.

[94] 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.

[95] 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.

[96] 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.

[97] 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.

[98] Seung-Hwan Lim, Sangkeun Lee, Sarah S Powers, Mallikarjun Shankar, and Neena Imam. 2016. *Survey of Approaches to Generate Realistic Synthetic Graphs.* Technical Report. Oak Ridge National Laboratory.

[99] Zinan Lin, Alankar Jain, Chen Wang, Giulia Fanti, and Vyas Sekar. 2020. Using gans for sharing networked time series data: Challenges, initial promise, and open questions. In *Proceedings of the ACM Internet Measurement Conference*. 464–483.

[100] 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.

[101] 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.

[102] 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*.

[103] Richard K Lomotey, Sandra Kumi, Madhurima Ray, and Ralph Deters. 2024. Synthetic Data Digital Twins and Data Trusts Control for Privacy in Health Data Sharing. In *Proceedings of the 2024 ACM Workshop on Secure and Trustworthy Cyber-Physical Systems*. 1–10.

[104] Lin Long, Rui Wang, Ruixuan Xiao, Junbo Zhao, Xiao Ding, Gang Chen, and Haobo Wang. 2024. On LLMs-Driven Synthetic Data Generation, Curation, and Evaluation: A Survey. In *Findings of the Association for Computational Linguistics: ACL 2024*, Lun-Wei Ku, Andre Martins, and Vivek Srikumar (Eds.). Association for Computational Linguistics, Bangkok, Thailand, 11065–11082. https://doi.org/10.18653/v1/2024.findings-acl.658

[105] Markus Löning, Anthony Bagnall, Sajaysurya Ganesh, Viktor Kazakov, Jason Lines, and Franz J Király. 2019. sktime: A unified interface for machine learning with time series. *arXiv preprint arXiv:1909.07872* (2019).

[106] David Lopez-Paz and Maxime Oquab. 2016. Revisiting classifier two-sample tests. *arXiv preprint arXiv:1610.06545* (2016).

[107] Isabelle Lorge, Dan W Joyce, Niall Taylor, Alejo Nevado-Holgado, Andrea Cipriani, and Andrey Kormilitzin. 2025. Detecting the clinical features of difficult-to-treat depression using synthetic data from large language models. *Computers in Biology and Medicine* 194 (2025), 110246.

[108] Alessio Luschi, Linda Tognetti, Alessandra Cartocci, Gabriele Cevenini, Pietro Rubegni, and Ernesto Iadanza. 2025. Advancing synthetic data for dermatology: GAN comparison with multi-metric and expert validation approach. *Health and Technology* 15, 3 (2025), 553–562.

[109] 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.

[110] Miro Mannino and Azza Abouzied. 2019. Is this real? generating synthetic data that looks real. In *Proceedings of the 32nd Annual ACM Symposium on User Interface Software and Technology*. 549–561.

[111] Leland McInnes, John Healy, and James Melville. 2018. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426* (2018).

[112] 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.

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

[114] 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.

[115] 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.

[116] Mihai Nadăș, Laura Dioșan, and Andreea Tomescu. 2025. Synthetic Data Generation Using Large Language Models: Advances in Text and Code. *IEEE Access* (2025).

[117] 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

[118] 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.

[119] 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.

[120] 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.

[121] 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.

[122] Carlos Ordonez and Edward Omiecinski. 1999. Discovering association rules based on image content. In *Proceedings IEEE Forum on Research and Technology Advances in Digital Libraries*. IEEE, 38–49.

[123] Patrick O'Reilly, Andreas Bugler, Keshav Bhandari, Max Morrison, and Bryan Pardo. 2022. Voiceblock: Privacy through real-time adversarial attacks with audio-to-audio models. *Advances in Neural Information Processing Systems* 35 (2022), 30058–30070.

[124] 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).

[125] Richard Osuala, Grzegorz Skorupko, Noussair Lazrak, Lidia Garrucho, Eloy García, Smriti Joshi, Socayna Jouide, Michael Rutherford, Fred Prior, Kaisar Kushibar, et al. 2023. medigan: a Python library of pretrained generative models for medical image synthesis. *Journal of Medical Imaging* 10, 6 (2023), 061403–061403.

[126] Oladapo Oyebode and Rita Orji. 2023. Identifying adverse drug reactions from patient reviews on social media using natural language processing. *Health informatics journal* 29, 1 (2023), 14604582221136712.

[127] 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.

[128] WORKING PARTY. 2014. European Commission. Article 29 - Data Protection Working Party. (2014).

[129] Ajay Patel, Colin Raffel, and Chris Callison-Burch. 2024. DataDreamer: A Tool for Synthetic Data Generation and Reproducible LLM Workflows. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Lun-Wei Ku, Andre Martins, and Vivek Srikumar (Eds.). Association for Computational Linguistics, Bangkok, Thailand, 3781–3799. https://doi.org/10.18653/v1/2024.acl-long.208

[130] 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.

[131] 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).

[132] 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).

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

[134] Gaurav N Pradhan and B Prabhakaran. 2017. Association rule mining in multiple, multidimensional time series medical data. *Journal of Healthcare Informatics Research* 1 (2017), 92–118.

[135] 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.

[136] 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.

[137] 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).

[138] 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.

[139] 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).

[140] 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).

[141] Alessandra Sala, Lili Cao, Christo Wilson, Robert Zablit, Haitao Zheng, and Ben Y Zhao. 2010. Measurement-calibrated graph models for social network experiments. In *Proceedings of the 19th international conference on World wide web*. 861–870.

[142] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. 2016. Improved techniques for training gans. *Advances in neural information processing systems* 29 (2016).

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

[144] 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

[145] 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.

[146] 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.

[147] Daniel Smolyak, Margrét V Bjarnadóttir, Kenyon Crowley, and Ritu Agarwal. 2024. Large language models and synthetic health data: progress and prospects. *JAMIA open* 7, 4 (2024), ooae114.

[148] 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.

[149] 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.

[150] 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

[151] 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.

[152] 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.

[153] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. 2015. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1–9.

[154] Marco Tanfoni, Elia Giuseppe Ceroni, Sara Marziali, Niccolò Pancino, Marco Maggini, and Monica Bianchini. 2024. Generated or Not Generated (GNG): The Importance of Background in the Detection of Fake Images. *Electronics* 13, 16 (2024), 3161.

[155] 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.

[156] Joshua B Tenenbaum, Vin de Silva, and John C Langford. 2000. A global geometric framework for nonlinear dimensionality reduction. *science* 290, 5500 (2000), 2319–2323.

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

[158] 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.

[159] 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.

[160] 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.

[161] Thomas Unterthiner, Sjoerd van Steenkiste, Karol Kurach, Raphaël Marinier, Marcin Michalski, and Sylvain Gelly. 2019. FVD: A new Metric for Video Generation. https://openreview.net/forum?id=rylgEULtdN

[162] 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.

[163] 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.

[164] 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.

[165] Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of machine learning research* 9, 11 (2008).

[166] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).

[167] 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.

[168] Veniamin Veselovsky, Manoel Horta Ribeiro, Akhil Arora, Martin Josifoski, Ashton Anderson, and Robert West. 2023. Generating faithful synthetic data with large language models: A case study in computational social science. *arXiv preprint arXiv:2305.15041* (2023).

[169] 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.

[170] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing* 13, 4 (2004), 600–612.

[171] 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.

[172] Sandra Wankmüller. 2024. Introduction to neural transfer learning with transformers for social science text analysis. *Sociological Methods & Research* 53, 4 (2024), 1676–1752.

[173] 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.

[174] 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.

[175] 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.

[176] 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).

[177] McKell Woodland, Austin Castelo, Mais Al Taie, Jessica Albuquerque Marques Silva, Mohamed Eltaher, Frank Mohn, Alexander Shieh, Suprateek Kundu, Joshua P Yung, Ankit B Patel, et al. 2024. Feature extraction for generative medical imaging evaluation: New evidence against an evolving trend. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 87–97.

[178] Hao Wu, Yue Ning, Prithwish Chakraborty, Jilles Vreeken, Nikolaj Tatti, and Naren Ramakrishnan. 2018. Generating realistic synthetic population datasets. *ACM Transactions on Knowledge Discovery from Data (TKDD)* 12, 4 (2018), 1–22.

[179] Jiannan Xiang, Yahui Liu, Deng Cai, Huayang Li, Defu Lian, and Lemao Liu. 2021. Assessing Dialogue Systems with Distribution Distances. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (Eds.). Association for Computational Linguistics, Online, 2192–2198. https://doi.org/10.18653/v1/2021.findings-acl.193

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

[181] 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.

[182] 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).

[183] 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*.

[184] 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.

[185] 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.

[186] 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.

[187] 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.

[188] 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.

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

[190] 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.

[191] 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.

[192] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 586–595.

[193] 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.

[194] 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.

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

[196] 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.

[197] Yuchen Zhao and Isabel Wagner. 2020. Using metrics suites to improve the measurement of privacy in graphs. *IEEE Transactions on Dependable and Secure Computing* 19, 1 (2020), 259–274.

[198] Zilong Zhao, Aditya Kunar, Robert Birke, and Lydia Y Chen. 2021. Ctab-gan: Effective table data synthesizing. In *Asian Conference on Machine Learning*. PMLR, 97–112.

[199] 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 Additional Results

A.1 ARF: Membership Inference Attack

We provide results for an MIA on SD from an ARF, using the same attack set-up as in Section 5.2. The goal is to show the major gain in precision which can be achieved when attempting to disclose membership for a select group of individuals.

As Table 6 shows, different thresholds in the DOMIAS set-up lead to varying precision and recall scores. Lowering the threshold corresponds to attempting to increase the *confidence* in disclosing membership (precision), whereas increasing the threshold corresponds to increasing the *amount* of members disclosed. The former can often be considered more risky, as attackers likely do not know the real member ratio in their attack dataset, and can act with more confidence on their member predictions when using a low threshold.

Table 6 shows that lowering the DOMIAS threshold from the default median [164], or even from a threshold informed by the true membership proportion, can increase the precision gain from 2-4% to 11% over the naive baseline.

B Hyperparameters

B.1 Prediction Models

XGBoost classifiers or regressors always use default parameters with a maximum depth of 3 in our experiments.

Table 6. Membership Inference Attack (DOMIAS) results for SD from ARF for various thresholds.

Percentile Threshold	Precision		Recall	
	Score	% ↑ naive	Score	% ↑ naive
100*	0.667	NA	1.000	NA
50*	0.683	2%	0.512	-49%
33*	0.691	4%	0.342	-66%
25	0.701	5%	0.263	-74%
15	0.711	7%	0.160	-84%
10	0.740	11%	0.111	-89%
5	0.738	11%	0.055	-95%

*100=naive baseline, 50=default threshold, 33=true member proportion.

B.2 Synthetic Data Generators

B.2.1 ARF. We use the default parameters of Synthcity for the ARF generator.

B.2.2 Bayesian Network. We use the default parameters of Synthcity for the BN generator, except for the maximum number of parents per node, which we set to 2.

B.2.3 CTGAN. We use the same hyperparameters for CTGAN as mentioned in [182].

B.2.4 TVAE. We use the same hyperparameters for TVAE as mentioned in [182].

B.2.5 TabDDPM. The denoising model in TabDDPM consists of an MLP of 2 hidden layers of 128 nodes, with an embedding dimension of 128 as well. The Adam optimizer uses a learning rate of 1e-3 with weight decay of 1e-5. We use mean squared error for Gaussian loss, and a linear scheduler. The generator is trained for 1000 epochs, 200 timesteps, on batches of 500 samples.

B.3 Association Rule Mining

We use the Apriori algorithm to extract association rules. Since this requires discrete inputs, we divide continuous features into 3 equal-width bins - which can be understood as low, medium, or high levels for that feature. In the Apriori algorithm, we use a minimum support of 0.500 and minimum confidence of 0.750 to find association rules.

C Hardware

All experiments were performed using a single NVIDIA TITAN Xp GPU (12GB RAM) and an Intel Platinum 8160 CPU.