The DCR Delusion: Measuring the Privacy Risk of Synthetic Data

Zexi Yao[1,*] Nataša Krčo[1,*] Georgi Ganev[2] Yves-Alexandre de Montjoye[1]

[1]Imperial College London
[2]University College London and SAS

Abstract

Synthetic data has become an increasingly popular way to share data without revealing sensitive information. Though Membership Inference Attacks (MIAs) are widely considered the gold standard for empirically assessing the privacy of a synthetic dataset, practitioners and researchers often rely on simpler proxy metrics such as Distance to Closest Record (DCR). These metrics estimate privacy by measuring the similarity between the training data and generated synthetic data. This similarity is also compared against that between the training data and a disjoint holdout set of real records to construct a binary privacy test. If the synthetic data is not more similar to the training data than the holdout set is, it passes the test and is considered private. In this work we show that, while computationally inexpensive, DCR and other distance-based metrics fail to identify privacy leakage. Across multiple datasets and both classical models such as Baynet and CTGAN and more recent diffusion models, we show that datasets deemed private by proxy metrics are highly vulnerable to MIAs. We similarly find both the binary privacy test and the continuous measure based on these metrics to be uninformative of actual membership inference risk. We further show that these failures are consistent across different metric hyperparameter settings and record selection methods. Finally, we argue DCR and other distance-based metrics to be flawed by design and show a example of a simple leakage they miss in practice. With this work, we hope to motivate practitioners to move away from proxy metrics to MIAs as the rigorous, comprehensive standard of evaluating privacy of synthetic data, in particular to make claims of datasets being legally anonymous.

[*]Equal contribution

1 Introduction

Synthetic data is a popular tool for sharing and using sensitive data, used across fields such as medicine [12, 14], finance [57] and public security [48]. Synthetic data generators (SDGs) aim to learn the underlying distribution of a dataset and generate synthetic data that preserves its statistical properties while protecting the privacy of individual records.

Membership inference attacks (MIAs) are widely accepted as the standard method to empirically assess information leakage and the privacy of synthetic data [26, 38, 56, 16, 6, 31] and machine learning models in general [10, 50, 69], both as a direct attack, and as an upper bound on more severe threats such as reconstruction or attribute inference attacks [52, 5]. MIAs evaluate the privacy of synthetic data at an *individual* level by assessing the risk of a particular record's membership in the training dataset being correctly inferred by an attacker.

While MIAs are the state-of-the-art method for evaluating privacy, they typically involve training multiple generative models, leading to high computational costs and motivating the use of simpler distance-based metrics. Distance to Closest Record (DCR) and similar metrics are commonly used as a proxy for MIAs, both in commercial products [41, 58] and for evaluating novel methods such as diffusion models [32, 70, 46]. These metrics assess the privacy of a synthetic dataset as a whole by measuring the similarity between synthetic and training data. Intuitively, the less similar a synthetic dataset is to its training data, the stronger the assumed privacy. The metrics can be used either to construct a binary privacy test τ_{DCR} classifying a dataset as "private" or "non-private," or directly as a continuous measure of privacy μ_{DCR}. The binary test compares the distribution of

distances between synthetic and training records against the distances between a set of real holdout records and the training records, typically at a certain percentile.

Contribution. In this work, we evaluate the effectiveness of DCR and similar metrics as a proxy for membership inference attacks, and show them to be an inadequate measure of both information leakage and the privacy risk of generated synthetic data.

First, we show proxy metrics to provide a misleading measure of privacy risk for well-known classical SDGs: IndHist [49], Baynet [71], and CTGAN [64]. Across 9 experimental setups, we generate more than 10,000 datasets and find the majority of them to be deemed private by proxy metrics as applied in industry. Yet, instantiating MIAs against outlier records in these datasets reveals significant information leakage, with records shown to be highly vulnerable to membership inference attacks (AUC$>$ 0.8). Worse, we show MIAs to perform equally well against datasets deemed "private" and "non-private" by DCR, and an absence of correlation between MIA performance and μ_{DCR}.

Second, we show our empirical results to extend to diffusion models, a more recent popular class of synthetic data generation models. For diffusion models TabDDPM [32] and ClavaDDPM [46], we generate synthetic datasets considered private by τ_{DCR}, and instantiate the state-of-the-art MIAs for tabular diffusion models against them. Similarly to classical models, the MIAs reach high performance (TPR at FPR=0% above 10%) despite passing the binary privacy test τ_{DCR}. μ_{DCR} also shows no correlation with vulnerability to MIAs.

For computational reasons, we previously focused on outlier records which are more likely to be vulnerable to MIAs in the case of classical models. We here study for the Baynet generator and Adult [7] dataset the risk for every record in the target dataset and show that the risk is not limited to outliers. Though the synthetic datasets pass the binary privacy test, an MIA is able to infer the membership of 20% of the training records better than a random guess (AUC$\geq$ 0.6). We also show that our findings hold across different choices of the τ_{DCR} hyperparameter, the comparison percentile. Finally, we study a real-world example of a simple privacy leakage that the proxy metrics are by-design unable to detect.

Taken together, our results show DCR and other distance-based metrics to be poor proxies for measuring privacy risk. They detect only the most severe privacy violations, such as when synthetic data consists mostly of copies of the real data, and potentially leaving more subtle information leakage undetected. They also seem to show no meaningful correlation with actual privacy risk, making them unreliable even as general indicators of privacy. We hope this work will motivate rigorous privacy evaluation using state-of-the-art attacks defined in the literature, and encourage researchers and practitioners to move away from using distance-based metrics.

2 Preliminaries

In this section, we introduce relevant notation, synthetic data generators, and methods for measuring privacy of synthetic data.

Notation. We denote a record consisting of k attributes with $x_i = (x_{i,1}, \ldots, x_{i,k}) \sim \mathcal{D}$, where $\mathcal{D}$ is the distribution over feature space $\mathcal{F} = \mathcal{F}_1 \times \ldots \times \mathcal{F}_k$. $D = \{x_1, \ldots, x_m\}$ denotes a tabular dataset where one record x_i corresponds to one row.

Synthetic Data Generators (SDGs). Let $D_{\text{target}} = \{x_1, \ldots, x_n\}$ be a real tabular dataset. A generative model $M = \phi(D_{\text{target}})$ trained using procedure ϕ estimates the underlying distribution of D_{target}. A synthetic dataset $D_{\text{synthetic}} \sim M$ can then be sampled from the model. In general, we consider synthetic datasets of the same size as the training data, $|D_{\text{synthetic}}| = |D_{\text{target}}|$. We distinguish two main categories of SDGs in this paper, classical and diffusion models. Classical models are older generative models that typically learn distributions of feature values in the training dataset. Diffusion models, introduced more recently, generate data through an iterative denoising process, allowing them to capture more complex data distributions and dependencies.

Measuring privacy of synthetic datasets. The standard approach for evaluating the privacy risk of synthetic data in the literature are membership inference attacks (MIAs). They estimate the risk of a record's membership in the training data of an SDG being correctly inferred by an attacker. However, MIAs are typ-

ically computationally expensive, leading practitioners and researchers to rely on simpler distance-based metrics such as DCR as proxies. In the following sections, we introduce these methods in more detail and evaluate their effectiveness.

3 Privacy Evaluation Techniques for Synthetic Data

In this section, we describe how proxy metrics and membership inference attacks are used in practice to empirically evaluate the privacy risk of synthetic data.

3.1 Distance to Closest Record (DCR) and Other Distance-Based Metrics

Proxy privacy metrics use a notion of distance between a synthetic dataset $D_{\text{synthetic}} \sim M(D_{\text{target}})$ and its corresponding training dataset D_{target}. They then use this metric to either to construct a binary privacy test or as a continuous measure of privacy.

DCR is the one of the most popular metrics used for evaluating privacy risk of synthetic data in both industry [41, 58, 66, 39, 4] and academia [14, 35, 22, 55, 61, 65, 68, 9, 32, 70, 72, 74, 53, 46]. It defines a vector of per-record distances between datasets D_1 and D_2, where each entry is the distance from a record in D_1 to its nearest neighbor in D_2:

$$d_{\text{DCR}}(D_1, D_2) = \{\min_{x_j \in D_2} dist(x_i, x_j)\}_{i=1}^{|D_1|}$$

where $dist(x_i, x_j)$ could be any distance metric but is typically the sum of euclidean distance for continuous features and hamming distances for categorical features between x_i and x_j [41, 4, 53, 32].

Other popular proxy metrics include Nearest Neighbor Distance Ratio (NNDR) and Identical Match Share (IMS). NNDR defines the distance vector $d_{\text{NNDR}}(D_1, D_2)$ by computing, for each record in D_1, the ratio between the distance to its nearest neighbor and the distance to its second-nearest neighbor in D_2. Instead of a distance vector, IMS defines a scalar distance measure $d_{\text{IMS}}(D_1, D_2)$ as the number of records in D_1 with an identical match in D_2.

Privacy test. DCR, NNDR and IMS are often used to construct binary privacy tests to classify a synthetic dataset $D_{\text{synthetic}}$ as "private" or "non-private" [41, 4, 22]. This is done by comparing the distance between $D_{\text{synthetic}}$ and D_{target} to the distance between D_{target} and a holdout set of real records D_{holdout}. If $D_{\text{synthetic}}$ is further away from D_{target} than D_{holdout}, it is considered private.

DCR and NNDR construct privacy tests τ_{DCR} and τ_{NNDR} by comparing the 5th percentile of the respective distance vectors:

$$\tau_{\text{DCR}}(D_{\text{synthetic}}, D_{\text{target}}) = \mathbb{1}\Big[d_{\text{DCR}}(D_{\text{synthetic}}, D_{\text{target}})_{p=0.05} \geq d_{\text{DCR}}(D_{\text{holdout}}, D_{\text{target}})_{p=0.05}\Big]$$

where $p = 0.05$ denotes the 5th percentile. The test is defined analogously for d_{NNDR}.

IMS constructs the privacy test by comparing the number of identical matches:

$$\tau_{\text{IMS}}(D_{\text{synthetic}}, D_{\text{target}}) = \mathbb{1}[d_{\text{IMS}}(D_{\text{synthetic}}, D_{\text{target}}) \leq d_{\text{IMS}}(D_{\text{holdout}}, D_{\text{target}})]$$

Here, the synthetic dataset is considered private if it contains fewer identical matches with the training data than the holdout set does.

Various combinations of these metrics are used in practice, with no agreed-upon, widely used setup. We therefore define a strict privacy test $\tau_{\text{DCR,NNDR,IMS}}$ as a joint privacy test using all three proxy metric. $D_{\text{synthetic}}$ is considered by $\tau_{\text{DCR,NNDR,IMS}}$ to be privacy only if it passes all of them τ_{DCR}, τ_{NNDR}, and τ_{IMS}.

$$\tau_{\text{DCR,NNDR,IMS}}(D_{\text{synthetic}}, D_{\text{target}}) = \mathbb{1}[\tau_{\text{DCR}}(D_{\text{synthetic}}, D_{\text{target}}) = 1 \wedge \tau_{\text{NNDR}}(D_{\text{synthetic}}, D_{\text{target}}) = 1 \wedge \tau_{\text{IMS}}(D_{\text{synthetic}}, D_{\text{target}}) = 1]$$

Continuous privacy measure. DCR is also commonly used as a continuous privacy measure μ_{DCR} for comparing the privacy of different generative models, particularly for diffusion models [53, 32]. Instead of comparing synthetic data to a holdout set,

the continuous measure aggregates the distance vector $d_{\text{DCR}}(D_{\text{synthetic}}, D_{\text{target}})$ to produce a single privacy score. In this work, we follow Kotelnikov et al. [32] and use the mean of the distances in $d_{\text{DCR}}(D_{\text{synthetic}},\ D_{\text{target}})$ as the continuous privacy measure:

$$\mu_{\text{DCR}}(D_{\text{synthetic}}, D_{\text{target}}) = \frac{1}{|D_{\text{synthetic}}|} \sum_{i=1}^{|D_{\text{synthetic}}|} d_{\text{DCR}}(D_{\text{synthetic}}, D_{\text{target}})_i$$

A higher value of $\mu_{\text{DCR}}(D_{\text{synthetic}}, D_{\text{target}})$ indicates that synthetic records are more distant from D_{target} and is thus assumed to imply better privacy protection.

3.2 Membership Inference Attacks (MIAs)

MIAs are the state of the art technique for evaluating privacy risk of synthetic data [26, 62]. They identify privacy leakage using a privacy game where an attacker aims to infer whether a synthetic dataset was generated by a model trained on a specific target record.

Classical models. The state-of-the-art MIA for classical SDGs is extended-TAPAS, a black-box attack introduced by Houssiau et al. [26] and extended by Meeus et al. [38]. Extended-TAPAS models how the inclusion or exclusion of a single target record x impacts the generated synthetic data and trains a meta-classifier to predict membership.

For a target record $x \in D_{\text{target}}$, the attacker trains shadow models on datasets sampled from an auxiliary dataset D_{aux}, which is drawn from the same distribution as the target dataset D_{target} but is disjoint from it. The target record is included in exactly half of the shadow datasets. They then generate a synthetic dataset using each shadow model, and extract query features from each. These features count the number of synthetic records that match the target record across random subsets of attributes. This results in a labeled membership dataset for training a meta-classifier that predicts whether a given synthetic dataset was trained on the target record. In our experiments, we use 1000 shadow models.

The MIA is evaluated across a set of evaluation models, where exactly half are trained on x. In this work, we evaluate in the model-seeded setup of Guépin et al. [21], where half of the evaluation models are trained on D_{target}, and half on D_{target} where x is replaced by a randomly sampled holdout record. The MIA is performed against each dataset, and its ROC AUC score is computed, resulting in a risk estimate of x within D_{target}. In our experiments, we use 1000 evaluation models.

As extended-TAPAS must be developed and evaluated separately for each target record, evaluating the risk of every record in D_{target} across setups is computationally infeasible. Because of this, in our main experiment, we select 100 target records in D_{target} and instantiate the MIAs against them. We use the Achilles vulnerability score introduced by Meeus et al. [38], and select the 100 records with the highest vulnerability score. The final output for each setup is then a set of per-record MIA AUC scores.

Diffusion models. The state-of-the-art MIA for tabular diffusion models was introduced by Wu et al. [62] in the challenge on Membership Inference over Diffusion-models-based Synthetic Tabular data (MIDST) [60]. We consider both the black-box and white-box variant of the attack, and refer to them collectively as the MIDST attacks for simplicity. The black-box attack relies only on data generated by the target model, while the white-box attack has full access to the model and its internal parameters. Both attacks model the model's loss on member versus non-member records.

To train the meta-classifier, the attacker first samples shadow datasets from an auxiliary dataset and trains a shadow diffusion model on each. For each shadow model, they extract features from the initial noise and training loss for both member and non-member records. These features form a labeled dataset used to train a multi-layer perceptron (MLP) classifier that predicts whether a given record was part of the training data. In our experiments, we use 20 shadow datasets to train the MIDST attacks.

As these attacks can be applied to any individual record, and training diffusion models is computationally expensive, evaluation here is typically done on a fixed target model across a set of known members and non-members. The attack is applied to each record, and a single True Positive Rate (TPR) at a fixed low False Positive Rate (FPR) is computed based on the predictions [62, 10]. In our setup, we evaluate the MIDST

attack on 10 diffusion models, each with 200 member and 200 non-member records, resulting in 10 TPR values–one per synthetic dataset.

4 Experimental Setup

In this section, we specify the datasets and models we use in our experiments.

4.1 Models

Classical models. We use three well-known synthetic data generators, using the implementations available in the reprosyn [3] repository.

IndHist [49] is the simplest of our selected models. It uses marginal frequency counts to generate feature values for synthetic records. For each feature, it samples from the distribution of values for that feature among all training records. Different features are sampled independently from each other.

BayNet [71] trains a Bayesian network to learn the relationships between features. Each feature is represented as a node on a network graph, with edges representing relations between two features. The GreedyBayes algorithm introduced by Zhang et al. [71] is then used to estimate the joint probabilities of the features, from which synthetic records can be sampled.

CTGAN [64] trains a generative adversarial network (GAN) consisting of a generator and a discriminator to model feature distribution of records in the training dataset. They are trained jointly with opposing goals: the discriminator attempts to distinguish between real and synthetic records produced by the generator, while the generator aims to produce synthetic data similar enough to real data to fool the discriminator.

Diffusion models. Diffusion models have become increasingly popular in recent years due to their increased utility of generated synthetic data and versatility of applications compared to classical models. We use two tabular diffusion models, with the implementation available in the MIDSTModels repository [60].

TabDDPM [32] is the first diffusion model specifically developed for tabular data. It adapts the diffusion process to account for different feature types by applying Gaussian diffusion to numerical features and multinomial diffusion to categorical and binary features.

ClavaDDPM [46] is a tabular diffusion model designed to generate multi-relational data. It uses latent clustering to model the relationships between the tables defined by foreign keys and enable conditional generation of synthetic tables.

4.2 Datasets

We evaluate the success of privacy measures across the following publicly available datasets, commonly used in literature studying tabular data privacy.

Adult [7] is an anonymized sample of the 1994 US Census data containing 48,842 records. It contains 15 demographic features, 9 of which are categorical.

Bank [40] contains 45,211 records concerning the marketing campaign of a Portuguese banking institution in 2014. Each record contains 17 features of which 4 are demographic and 13 describe the individual's previous interactions with the institution.

UK Census [43] is an anonymized 1% sample of the 2011 Census from Wales and England, published by the UK Office for National Statistics. The dataset is comprised of 569,741 records with 17 categorical demographic features.

Berka [8] is an anonymized database containing information regarding over $5,000$ clients collected in 2000 from a Czech bank. The main dataset, which we refer to as the Berka dataset, contains over $1,000,000$ transactions. Additional tables with account and client information can be linked via foreign keys, e.g. when training ClavaDDPM.

5 Results

5.1 Evaluating DCR and Other Proxy Metrics for Classical Models

We here evaluate the effectiveness of DCR and other proxy metrics for identifying privacy leakage in classical synthetic data generators by comparing them to MIA results across 3 datasets and 3 target models. We develop MIAs against 100 outlier target records per setup, selected using the Achilles score. For each target record, we compute the percentage of evaluation synthetic records

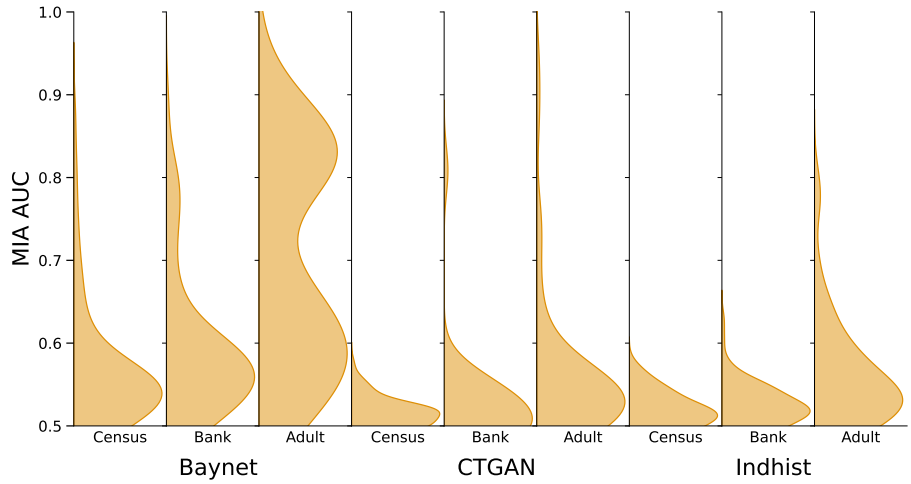

Figure 1: Extended-TAPAS MIA AUC on datasets considered "private" by $\tau_{\text{DCR,NNDR,IMS}}$ across classical SDG setups. Each dataset-SDG setup contains 100 target records selected using the *Achilles* score.

that fail τ_{DCR} and $\tau_{\text{DCR,NNDR,IMS}}$, and the mean μ_{DCR} across all synthetic datasets for that record. We then study the MIA AUC values for the outlier records in datasets deemed "private" by τ_{DCR} and $\tau_{\text{DCR,NNDR,IMS}}$.

In 7 out of the 9 setups, both $\tau_{\text{DCR,NNDR,IMS}}$ and τ_{DCR} consistently classify all 500 synthetic datasets per target record as "private." The only exceptions are observed with the Baynet generator on the Census and Bank datasets. For Census, 12% of the synthetic datasets fail the $\tau_{\text{DCR,NNDR,IMS}}$ and 1.1% fail τ_{DCR} alone. For Bank, 0.4% of the datasets fail the $\tau_{\text{DCR,NNDR,IMS}}$, and 0% fail τ_{DCR}. Fig. 1 shows the datasets to be highly vulnerable to MIAs, despite being considered "private" by the proxy metrics.

Fig. 1 shows the datasets passing the proxy metric privacy tests to leak information about their training data. In the majority of setups, the MIA reaches $AUC \geq 0.6$—shown to indicate information leakage—for a significant fraction of records, and even $AUC \geq 0.8$ for some. This suggests that τ_{DCR} and $\tau_{\text{DCR,NNDR,IMS}}$ often misrepresent synthetic datasets with significant privacy leakage as "private," making them unreliable for verifying privacy of synthetic data for release.

As proxy metrics often fail to flag privacy leakage, we now study whether they can still provide informative

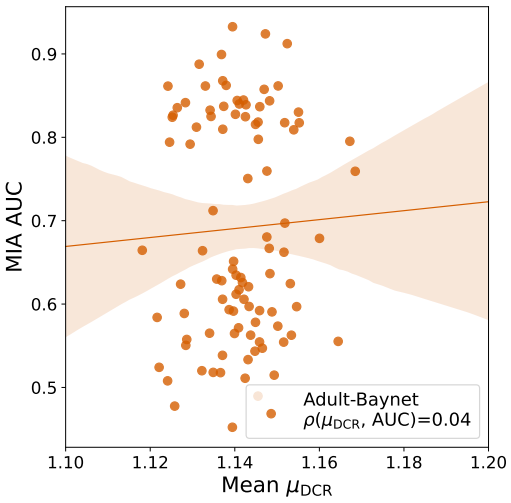

Figure 2: Comparison of mean μ_{DCR} and MIA AUC for the Baynet generator on the Adult dataset. Each point represents a target record's MIA AUC and its mean μ_{DCR} across evaluation datasets.

Table 1: Pearson correlation between TPR@FPR=0% and μ_{DCR}.

Dataset	Black-box attack	White-box attack
TabDDPM	0.47	0.11
ClavaDDPM	0.10	−0.03

signal about the risk of a dataset. We examine whether synthetic datasets considered "non-private" exhibit higher MIA AUC values then those considered "non-private." Then, we evaluate whether the continuous measure μ_{DCR} gives an indication of MIA performance.

For the Census dataset and the Baynet generator, 12% of the synthetic datasets are classified as "non-private" by $\tau_{\text{DCR,NNDR,IMS}}$. We compare MIA performance when instantiated across the full set of synthetic datasets and only the "private" datasets. Fig. 1 shows that there is no meaningful difference in performance between the two sets – MIA AUC remains equally high, regardless of whether evaluation is restricted to the "private" subset or not.

For each target record in the Adult-Baynet setup, we compute the corresponding MIA AUC and the mean μ_{DCR} across the synthetic datasets used for evaluation and study their relationship. Fig. 2 shows that there is no correlation between the two values. AUC values span a wide range (roughly between 0.5 and 1.0), regardless of the value of μ_{DCR}, suggesting that μ_{DCR} is unable to effectively distinguish between datasets with different levels of privacy risk.

5.2 Evaluating DCR for Diffusion Models

We here study the effectiveness of DCR for evaluating the privacy of diffusion models by repeating the analyses done in Section 5.1, and follow the state-of-the-art methods for diffusion models. Specifically, we focus on DCR and measure MIA performance on datasets deemed "private" as TPR at FPR=0% [10]. Using TabDDPM and ClavaDDPM, we train 10 target models for each method and generate one synthetic dataset per model. All generated datasets pass the DCR privacy test τ_{DCR}. We then instantiate both the black-box and white-box MIDST attacks against each of the generated synthetic datasets.

Table 1 shows that for both TabDDPM and ClavaDDPM, the MIDST attacks are able to successfully infer membership in the training data of the target models. The black-box attack achieves TPR@FPR=0% above 5% on the majority of datasets, and exceeds 10% on some datasets for both models. The white-box attack performs even better, reaching TPR@FPR=0% above 20% on all 10 target datasets, and on more than half of the ClavaDDPM datasets. These results indicate clear information leakage that is not detected by τ_{DCR}.

In line with our analysis in Section 5.1, we evaluate whether μ_{DCR} provides any meaningful signal of privacy risk. We compute the μ_{DCR} for all 10 target datasets in each setup, and compare it to the TPR@FPR=0% achieved by the MIAs. Table 1 shows there to be no clear correlation between μ_{DCR} and MIA performance, indicating that μ_{DCR} is not a reliable proxy for privacy risk as identified by MIAs.

5.3 Effect of Adjusting DCR Hyperparameter

τ_{DCR} determines the privacy of a synthetic dataset $D_{\text{synthetic}}$ by comparing the distance vector between $D_{\text{synthetic}}$ and D_{target} to the distance vector between D_{holdout} and D_{target} at the same percentile mark p, typically 5th percentile. This percentile choice is the only hyperparameter of τ_{DCR}. We now study whether tuning this threshold can improve τ_{DCR}'s ability to detect privacy leakage. The condition for passing the privacy test,

$$d_{\text{DCR}}(D_{\text{synthetic}}, D_{\text{target}})_p \geq d_{\text{DCR}}(D_{\text{holdout}}, D_{\text{target}})_p$$

can be rewritten as:

$$d_{\text{DCR}}(D_{\text{synthetic}}, D_{\text{target}})_p - d_{\text{DCR}}(D_{\text{holdout}}, D_{\text{target}})_p \geq 0$$

We examine the effects of adjusting $p \in [0, 0.1]$ for the above condition across all synthetic datasets in the Baynet generator with Adult dataset setup.

Fig. 4 shows an example of adjusting p for a single synthetic dataset trained on a vulnerable record with MIA AUC= 0.84. For this synthetic dataset, the value remains above 0 regardless of the value of p, showing that the dataset passes τ_{DCR} on all thresholds $p \in [0, 0.1]$.

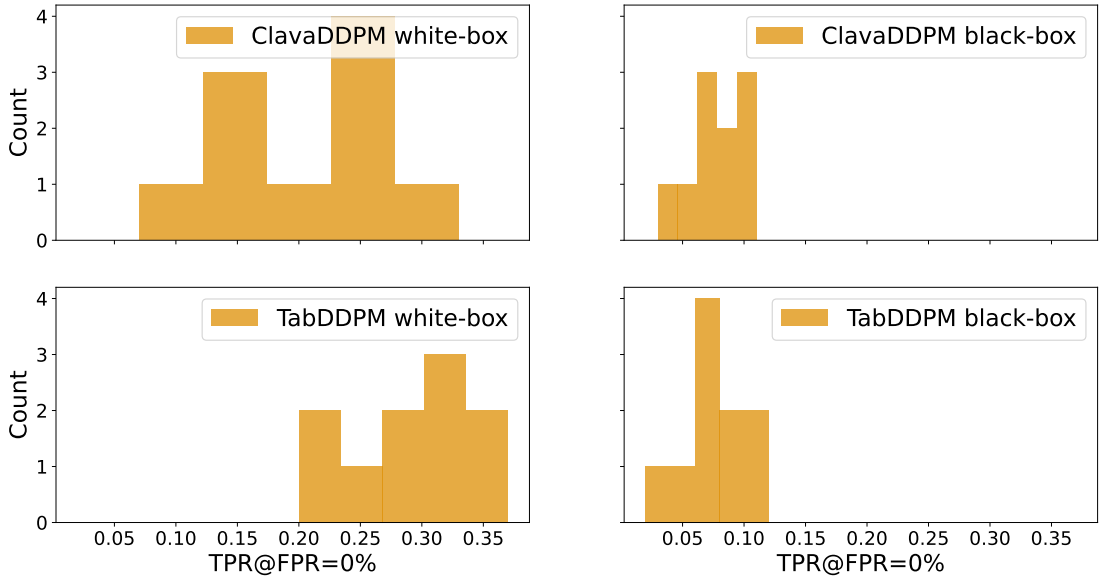

Figure 3: Distribution of TPR@FPR=0% for MIDST attacks against TabDDPM and ClavaDDPM.

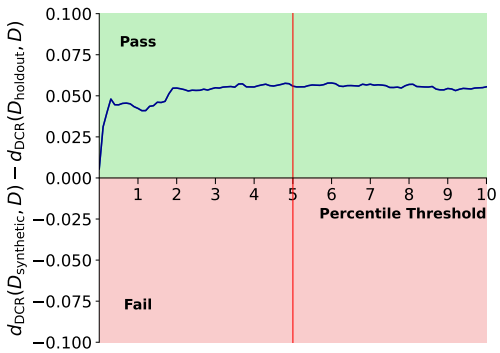

Figure 4: Comparison of $d_{\text{DCR}}(D_{\text{synthetic}}, D_{\text{target}}) - d_{\text{DCR}}(D_{\text{holdout}}, D_{\text{target}})$ across percentile thresholds for a synthetic dataset trained on a vulnerable record with MIA AUC = 0.84.

This result holds across all synthetic datasets in the Baynet-Adult setup – every dataset is deemed "private" by τ_{DCR}, regardless of the choice of threshold.

5.4 Analysis of the Impact of the *Achilles* Vulnerability Score

In our main experiment for classical models in Section 5.1, we reduce computational costs by only performing MIAs against the top 100 records by *Achilles* vulnerability score for each dataset. To eliminate concerns that such sampling may have exaggerated the privacy risk indicated by MIA, we analyze DCR and MIA performance on all 1000 records in D_{target} for one setup (Baynet with Adult). Consistent with our prior results, all synthetic datasets for all 1000 records also pass τ_{DCR} and $\tau_{\text{DCR,NNDR,IMS}}$.

We now compare the distribution of MIA AUC values across all 1000 records to the MIA AUC values of 100 outlier records selected by the *Achilles* vulnerability score in Section 5.1. Fig. 5 shows that while *Achilles* score is more likely to identify vulnerable records than random sampling, a significant proportion of vulnerable records went undetected – the MIA achieves an AUC ≥ 0.8 for 54 records and AUC ≥ 0.6 for 200 records out of 1000 total records. This is still a high percentage of records with information leakage, which indicates significant privacy

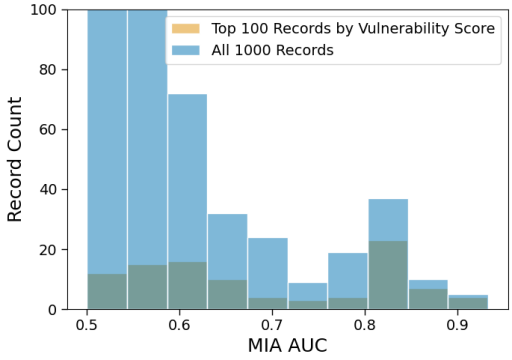

Figure 5: MIA AUCs of all 1000 target records and 100 vulnerable records selected using *Achilles* in the Adult-Baynet setup, in synthetic datasets considered "private" by $\tau_{\text{DCR,NNDR,IMS}}$.

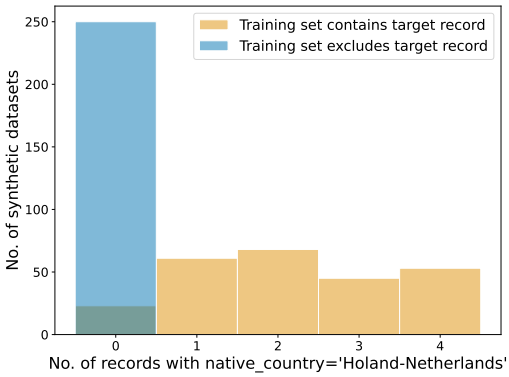

Figure 6: Comparison of synthetic datasets generated by CTGAN-Adult models trained on target record with uniquely identifying native country "Holland-Netherlands."

risk across all synthetic datasets in the setup.

5.5 Detailed Analysis of one Highly Vulnerable Record

We select the record with the highest MIA AUC across all our classical setups for detailed analysis as to why DCR is unable to detect clear privacy violations. This record has an MIA AUC of 0.94 and is from the CTGAN-Adult setup, all synthetic datasets generated by this setup passed both τ_{DCR} and $\tau_{\text{DCR,NNDR,IMS}}$. We start by identifying the cause of high privacy leakage for this record – a distribution shift of generated synthetic datasets between CTGAN models trained on the target record and those that were not.

Notably, as shown in Fig. 6, CTGAN models trained on this record generate synthetic data containing records with a `native-country` value of "Holland-Netherlands" in 92% of cases, while models not trained on it never produce it. We immediately notice the target record is the only record in the entire Adult dataset with the value "Holland-Netherlands" for the `native-country` feature. Thus, the presence of a synthetic record with this feature value would reveal the membership of the target record in D_{target}. While this is a clear privacy concern, DCR instead focuses on distance measurements to the closest record, which is not the cause of the privacy leakage. Furthermore, distance calculations treat all features with the same importance, a single feature has minimal effect on the overall distance metric – synthetic records containing the uniquely identifying value are not even the closest records in $D_{\text{synthetic}}$ to the target record. As a result, τ_{DCR} and $\tau_{\text{DCR,NNDR,IMS}}$ fail to flag this obvious privacy risk.

We believe this to be a core limitation of proxy metrics which by design do not learn and need to make assumptions about what causes privacy leakage. While one could design another proxy metric to check for uniquely identifying feature values, it is just a single example of privacy leakages that distance metrics cannot capture. Prior work has shown that leakage may also arise from more complex feature combinations or dataset-specific characteristics [51], which cannot be identified by DCR and requires designing setup-specific proxy metrics. In contrast, MIAs are able to learn and thus capture a more comprehensive spectrum of privacy risks, including those that were previously unknown and unique to specific setups.

6 Related Work

Synthetic Data Generators (SDGs). Numerous synthetic data generators have been proposed for tabular data [29, 15, 27], spanning approaches from traditional statistical methods like graphical models [71, 36]

and workload/query-based [33, 37], to advanced deep learning techniques, including Variational Autoencoders (VAEs) [2, 1] and Generative Adversarial Networks (GANs) [63, 73, 28, 64]; and more recently, diffusion models [32, 70, 53, 42]. As discussed in Section 4.1, we use a range of best-performing models with reliable and public implementations.

Membership Inference Attacks (MIAs). MIAs were initially introduced as a method to infer the presence of trace amount of DNA from released genomic aggregates [25]. It was then extended to assess privacy leakage in discriminative machine learning models [54, 10, 67, 69]. In the context of generative models on images, Hayes et al. [23] and Hilprecht et al. [24] propose the first MIAs targeting VAEs and GANs, employing a shadow discriminator and Monte Carlo estimation, respectively. Subsequently, Chen et al. [13] introduce a taxonomy of MIAs along with a model-agnostic attack against GANs. More recently, Zhu et al. [75] and Carlini et al. [11] extend MIAs research to diffusion models, demonstrating that these models are more susceptible to memorization than GANs.

For tabular data, Stadler et al. [56] present the first systematic evaluation of MIAs, showing that outliers are particularly vulnerable. Other MIAs against tabular data include TAPAS [26], which relies on running a collection of random queries, and DOMIAS [59], which detects overfitting using a density-based approach. Meeus et al. [38] propose an identification procedure for selecting the most vulnerable records based on distance metrics and an extension of TAPAS [26], called extended-TAPAS. Guépin et al. [20] relax a common assumption in MIAs that the adversary has access to an auxiliary dataset. More recently, researchers have proposed model-specific attacks targeting traditional generative models to audit their privacy properties [6, 18, 19], and diffusion models to mitigate computational overhead [62].

Distance to Closest Record (DCR). DCR is widely adopted to measure and claim privacy in both industry [41, 58, 66, 39, 4] and academia, particularly in the medical domain [14, 35, 22, 55, 61, 65, 68, 30]. Furthermore, a growing number of recently proposed diffusion models – published in top-tier ML and NLP venues such as NeurIPS, ICML, and ICLR – rely exclusively on DCR to support privacy claims, or to demonstrate improvements over prior models [9, 32, 70, 72, 74, 53, 46].

DCR is often used in conjunction with other proxy metrics like Nearest Neighbor Distance Ratio (NNDR) and Identical Match Share (IMS) in real-world synthetic data products to run statistical tests and support privacy claims [41, 58, 4].

This is despite existing research showing that MIAs, such as TAPAS, are more effective at detecting privacy leakage than DCR in traditional models [26, 6]. Additionally, Ganev and De Cristofaro [17] show that relying on DCR to guarantee synthetic data privacy could be dangerous as adversaries operating under strong assumptions – such as repeated black-box access to conditional generation and proxy metric APIs – can successfully perform MIAs and reconstruct entire training records.

7 Discussion & Conclusion

Distance to Closest Record and other proxy privacy metrics are presented both as a statistical test for verifying privacy of synthetic datasets prior to data release and also as a proxy measurement of privacy of synthetic datasets [35, 4, 47, 65, 74, 61, 9, 32, 34, 68, 12, 72, 70, 44, 45, 22, 55, 14, 53, 32, 41].

In this paper, we show across both classical and diffusion models, that DCR and other metric tests consistently fail to identify privacy leakage, including clear privacy violations such as the presence of uniquely identifying feature values. Furthermore, we also show that DCR as a proxy measurement is uninformative for comparing privacy of synthetic datasets for both classical and diffusion models – there is no clear relation between distance of synthetic records to training dataset and MIA vulnerability.

Additionally, we show that privacy violations that are caused by a subset of feature values, such as the case of uniquely identifying feature values in the CTGAN generator with Adult dataset setup, have synthetic records that are distant from the training record. The effect of these important features on synthetic record distance is heavily reduced by the presence of other features, thus making it highly unlikely for DCR to detect such violations.

With DCR and other proxy metrics shown to be unsuitable for use as a privacy test or proxy privacy measurement, it is imperative for both the academic and industry

community to move to Membership Inference Attacks, which is the state-of-the-art for measuring privacy risks of synthetic datasets.

References

[1] Nazmiye Ceren Abay, Yan Zhou, Murat Kantarcioglu, Bhavani Thuraisingham, and Latanya Sweeney. Privacy preserving synthetic data release using deep learning. In *ECML PKDD*, 2019.

[2] Gergely Acs, Luca Melis, Claude Castelluccia, and Emiliano De Cristofaro. Differentially private mixture of generative neural networks. *IEEE TKDE*, 2018.

[3] Alan Turing Institute. Reprosyn. `https://github.com/alan-turing-institute/reprosyn`, 2022.

[4] Amazon AWS. How to Evaluate the Quality of the Synthetic Data – Measuring from the Perspective of Fidelity, Utility, and Privacy, 2022.

[5] Meenatchi Sundaram Muthu Selva Annamalai, Andrea Gadotti, and Luc Rocher. A linear reconstruction approach for attribute inference attacks against synthetic data. In *USENIX Security*, 2024.

[6] Meenatchi Sundaram Muthu Selva Annamalai, Georgi Ganev, and Emiliano De Cristofaro. "What do you want from theory alone?" experimenting with tight auditing of differentially private synthetic data generation. In *USENIX Security*, 2024.

[7] Barry Becker and Ronny Kohavi. Adult. UCI Machine Learning Repository, 1996.

[8] P. Berka et al. Guide to the financial data set. PKDD2000 discovery challenge, 2000.

[9] Vadim Borisov, Kathrin Sessler, Tobias Leemann, Martin Pawelczyk, and Gjergji Kasneci. Language Models are Realistic Tabular Data Generators. In *ICLR*, 2023.

[10] Nicholas Carlini, Steve Chien, Milad Nasr, Shuang Song, Andreas Terzis, and Florian Tramèr. Membership Inference Attacks From First Principles. In *IEEE S&P*, 2022.

[11] Nicolas Carlini, Jamie Hayes, Milad Nasr, Matthew Jagielski, Vikash Sehwag, Florian Tramer, Borja Balle, Daphne Ippolito, and Eric Wallace. Extracting training data from diffusion models. In *USENIX Security*, 2023.

[12] Taha Ceritli, Ghadeer O Ghosheh, Vinod Kumar Chauhan, Tingting Zhu, Andrew P Creagh, and David A Clifton. Synthesizing Mixed-type Electronic Health Records using Diffusion Models. *arXiv:2302.14679*, 2023.

[13] Dingfan Chen, Ning Yu, Yang Zhang, and Mario Fritz. GAN-Leaks: A taxonomy of membership inference attacks against generative models. In *ACM CCS*, 2020.

[14] Saverio D'Amico, Daniele Dall'Olio, Claudia Sala, et al. Synthetic Data Generation by Artificial Intelligence to Accelerate Research and Precision Medicine in Hematology. *JCO Clinical Cancer Informatics*, 2023.

[15] Emiliano De Cristofaro. Synthetic Data: Methods, Use Cases, and Risks. *IEEE S&P Magazine*, 2024.

[16] Jinhao Duan, Fei Kong, Shiqi Wang, Xiaoshuang Shi, and Kaidi Xu. Are diffusion models vulnerable to membership inference attacks? In *ICLR*, 2023.

[17] Georgi Ganev and Emiliano De Cristofaro. The Inadequacy of Similarity-based Privacy Metrics: Privacy Attacks against "Truly Anonymous" Synthetic Datasets. In *IEEE S&P*, 2025.

[18] Georgi Ganev, Meenatchi Sundaram Muthu Selva Annamalai, and Emiliano De Cristofaro. The Elusive Pursuit of Reproducing PATE-GAN: Benchmarking, Auditing, Debugging. *TMLR*, 2025.

[19] Steven Golob, Sikha Pentyala, Anuar Maratkhan, and Martine De Cock. Privacy Vulnerabilities in Marginals-based Synthetic Data. In *SaTML*, 2025.

[20] Florent Guépin, Matthieu Meeus, Ana-Maria Cretu, and Yves-Alexandre de Montjoye. Synthetic is all you need: Removing the auxiliary data assumption for membership inference attacks against synthetic data. In *ESORICS*, 2023.

[21] Florent Guépin, Nataša Krčo, Matthieu Meeus, and Yves-Alexandre de Montjoye. Lost in the averages: A new specific setup to evaluate membership inference attacks against machine learning models. *arXiv:2405.15423*, 2024.

[22] Morgan Guillaudeux, Olivia Rousseau, Julien Petot, Zineb Bennis, Charles-Axel Dein, Thomas Goronflot, Nicolas Vince, Sophie Limou, Matilde Karakachoff, Matthieu Wargny, and Pierre-Antoine Gourraud. Patient-centric synthetic data generation, no reason to risk re-identification in biomedical data analysis. *NPJ Digital Medicine*, 2023.

[23] Jamie Hayes, Luca Melis, George Danezis, and Emiliano De Cristofaro. LOGAN: membership inference attacks against generative models. In *PoPETs*, 2019.

[24] Benjamin Hilprecht, Martin Härterich, and Daniel Bernau. Monte Carlo and reconstruction membership inference attacks against generative models. In *PoPETs*, 2019.

[25] Nils Homer, Szabolcs Szelinger, Margot Redman, David Duggan, Waibhav Tembe, Jill Muehling, John V. Pearson, Dietrich A. Stephan, Stanley F. Nelson, and David W. Craig. Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays. *PLoS Genet*, 2008.

[26] Florimond Houssiau, James Jordon, Samuel N Cohen, Owen Daniel, Andrew Elliott, James Geddes, Callum Mole, Camila Rangel-Smith, and Lukasz Szpruch. Tapas: a toolbox for adversarial privacy auditing of synthetic data. *NeurIPS Workshop on Synthetic Data for Empowering ML Research*, 2022.

[27] Yuzheng Hu, Fan Wu, Qinbin Li, Yunhui Long, Gonzalo Munilla Garrido, Chang Ge, Bolin Ding, David Forsyth, Bo Li, and Dawn Song. Sok: Privacy-preserving data synthesis. In *IEEE S&P*, 2024.

[28] James Jordon, Jinsung Yoon, and Mihaela Van Der Schaar. PATE-GAN: Generating synthetic data with differential privacy guarantees. In *ICLR*, 2018.

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

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

[31] Fei Kong, Jinhao Duan, RuiPeng Ma, Hengtao Shen, Xiaofeng Zhu, Xiaoshuang Shi, and Kaidi Xu. An efficient membership inference attack for the diffusion model by proximal initialization. *arXiv:2305.18355*, 2023.

[32] Akim Kotelnikov, Dmitry Baranchuk, Ivan Rubachev, and Artem Babenko. TabDDPM: Modelling Tabular Data with Diffusion Models. In *ICML*, 2023.

[33] Terrance Liu, Giuseppe Vietri, and Steven Z Wu. Iterative methods for private synthetic data: Unifying framework and new methods. *NeurIPS*, 2021.

[34] Tongyu Liu, Ju Fan, Guoliang Li, Nan Tang, and Xiaoyong Du. Tabular Data Synthesis with Generative Adversarial Networks: Design Space and Optimizations. *VLDBJ*, 2023.

[35] Pei-Hsuan Lu, Pang-Chieh Wang, and Chia-Mu Yu. Empirical Evaluation on Synthetic Data Generation with Generative Adversarial Network. In *WIMS*, 2019.

[36] Ryan McKenna, Gerome Miklau, and Daniel Sheldon. Winning the NIST Contest: a scalable and general approach to differentially private synthetic data. *JPC*, 2021.

[37] Ryan McKenna, Brett Mullins, Daniel Sheldon, and Gerome Miklau. Aim: An adaptive and iterative mechanism for differentially private synthetic data. *PVLDB*, 2022.

[38] Matthieu Meeus, Florent Guepin, Ana-Maria Creţu, and Yves-Alexandre de Montjoye. Achilles' heels: vulnerable record identification in synthetic data publishing. In *ESORICS*, 2023.

[39] Ofer Mendelevitch and Michael D Lesh. Fidelity and privacy of synthetic medical data. *arXiv:2101.08658*, 2021.

[40] S. Moro, P. Rita, and P. Cortez. Bank Marketing. UCI Machine Learning Repository, 2014.

[41] Mostly AI. Truly anonymous synthetic data – evolving legal definitions and technologies (part II). `https://mostly.ai/blog/truly-anonymous-synthetic-data-legal-definitions-part-ii/`, 2020.

[42] Markus Mueller, Kathrin Gruber, and Dennis Fok. Continuous Diffusion for Mixed-Type Tabular Data. In *ICLR*, 2025.

[43] Office for National Statistics. Census microdata teaching files, 2011.

[44] Daniele Panfilo. *Generating Privacy-compliant, Utility-preserving Synthetic Tabular and Relational Datasets through Deep Learning*. University of Trieste, 2022.

[45] Daniele Panfilo, Alexander Boudewijn, Sebastiano Saccani, Andrea Coser, Borut Svara, Carlo Chauvenet, Ciro Mami, and Eric Medvet. A Deep Learning-based Pipeline for the Generation of Synthetic Tabular Data. *IEEE Access*, 2023.

[46] Wei Pang, Masoumeh Shafieinejad, Lucy Liu, Stephanie Hazlewood, and Xi He. Clavaddpm: Multi-relational data synthesis with cluster-guided diffusion models. In *NeurIPS*, 2024.

[47] Noseong Park, Mahmoud Mohammadi, Kshitij Gorde, Sushil Jajodia, Hongkyu Park, and Youngmin Kim. Data Synthesis Based on Generative Adversarial Networks. *PVLDB*, 2018.

[48] Ioannis Pastaltzidis, Nikolaos Dimitriou, Katherine Quezada-Tavarez, Stergios Aidinlis, Thomas Marquenie, Agata Gurzawska, and Dimitrios Tzovaras. Data augmentation for fairness-aware machine learning: Preventing algorithmic bias in law enforcement systems. In *ACM FAccT*, 2022.

[49] Haoyue Ping, Julia Stoyanovich, and Bill Howe. DataSynthesizer: Privacy-Preserving Synthetic Datasets. In *SSDBM*, 2017.

[50] Joseph Pollock, Igor Shilov, Euodia Dodd, and Yves-Alexandre de Montjoye. Free Record-Level Privacy Risk Evaluation Through Artifact-Based Methods. *arXiv:2411.05743*, 2024.

[51] Apostolos Pyrgelis, Carmela Troncoso, and Emiliano De Cristofaro. Knock Knock, Who's There? Membership Inference on Aggregate Location Data. In *NDSS*, 2017.

[52] Ahmed Salem, Giovanni Cherubin, David Evans, Boris Köpf, Andrew Paverd, Anshuman Suri, Shruti Tople, and Santiago Zanella-Béguelin. SoK: Let the privacy games begin! A unified treatment of data inference privacy in machine learning. In *IEEE S&P*, 2023.

[53] Juntong Shi, Minkai Xu, Harper Hua, Hengrui Zhang, Stefano Ermon, and Jure Leskovec. TabDiff: a Mixed-type Diffusion Model for Tabular Data Generation. In *ICLR*, 2025.

[54] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *IEEE S&P*, 2017.

[55] Jayanth Sivakumar, Karthik Ramamurthy, Menaka Radhakrishnan, and Daehan Won. GenerativeMTD: A Deep Synthetic Data Generation Framework for Small Datasets. *KBS*, 2023.

[56] Theresa Stadler, Bristena Oprisanu, and Carmela Troncoso. Synthetic data – anonymization groundhog day. In *USENIX Security*, 2022.

[57] Synthetic Data Expert Group, FCA. Using synthetic data in financial services. `https://www.fca.org.uk/publication/corporate/report-using-synthetic-data-in-financial-services.pdf`, 2024.

[58] Syntho. `https://www.syntho.ai/synthos-quality-assurance-report/`, 2025.

[59] Boris van Breugel, Hao Sun, Zhaozhi Qian, and Mihaela van der Schaar. Membership inference attacks against synthetic data through overfitting detection. *AISTATS*, 2023.

[60] Vector Institute. Midstmodels. `https://github.com/VectorInstitute/MIDST/`, 2025.

[61] Rohit Venugopal, Noman Shafqat, Ishwar Venugopal, Benjamin Mark John Tillbury, Harry Demetrios Stafford, and Aikaterini Bourazeri. Privacy Preserving Generative Adversarial Networks to Model Electronic Health Records. *Neural Networks*, 2022.

[62] Xiaoyu Wu, Yifei Pang, Terrance Liu, and Steven Wu. Winning the MIDST Challenge: New Membership Inference Attacks on Diffusion Models for Tabular Data Synthesis. *arXiv:2503.12008*, 2025.

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

[64] Lei Xu, Maria Skoularidou, Alfredo Cuesta-Infante, and Kalyan Veeramachaneni. Modeling tabular data using conditional gan. In *NeurIPS*, 2019.

[65] Andrew Yale, Saloni Dash, Ritik Dutta, Isabelle Guyon, Adrien Pavao, and Kristin P Bennett. Assessing Privacy and Quality of Synthetic Health Data. In *AIDR*, 2019.

[66] YData. How to evaluate the re-identification risk in Synthetic Data?, 2023.

[67] Jiayuan Ye, Aadyaa Maddi, Sasi Kumar Murakonda, Vincent Bindschaedler, and Reza Shokri. Enhanced membership inference attacks against machine learning models. In *ACM CCS*, 2022.

[68] Jinsung Yoon, Michel Mizrahi, Nahid Farhady Ghalaty, and others. EHR-Safe: Generating High-fidelity and Privacy-preserving Synthetic Electronic Health Records. *NPJ Digital Medicine*, 2023.

[69] Sajjad Zarifzadeh, Philippe Liu, and Reza Shokri. Low-Cost High-Power Membership Inference Attacks. In *ICML*, 2024.

[70] Hengrui Zhang, Jiani Zhang, Zhengyuan Shen, Balasubramaniam Srinivasan, Xiao Qin, Christos Faloutsos, Huzefa Rangwala, and George Karypis. Mixed-Type Tabular Data Synthesis with Score-based Diffusion in Latent Space. In *ICLR*, 2024.

[71] Jun Zhang, Graham Cormode, Cecilia M Procopiuc, Divesh Srivastava, and Xiaokui Xiao. Privbayes: Private data release via bayesian networks. *ACM TODS*, 2017.

[72] Tianping Zhang, Shaowen Wang, Shuicheng Yan, Jian Li, and Qian Liu. Generative Table Pre-training Empowers Models for Tabular Prediction. In *EMNLP*, 2023.

[73] Xinyang Zhang, Shouling Ji, and Ting Wang. Differentially private releasing via deep generative model (technical report). *arXiv:1801.01594*, 2018.

[74] Zilong Zhao, Aditya Kunar, Robert Birke, and Lydia Y Chen. CTAB-GAN: Effective Table Data Synthesizing. In *ACML*, 2021.

[75] Derui Zhu, Dingfan Chen, Jens Grossklags, and Mario Fritz. Data forensics in diffusion models: A systematic analysis of membership privacy. *arXiv:2302.07801*, 2023.