Systematic Assessment of Tabular Data Synthesis Algorithms

Anonymous Author(s)

ABSTRACT

Data synthesis has been advocated as an important approach for utilizing data while protecting data privacy. A large number of tabular data synthesis algorithms (which we call synthesizers) have been proposed. Some synthesizers satisfy Differential Privacy, while others aim to provide privacy in a heuristic fashion. A comprehensive understanding of the strengths and weaknesses of these synthesizers remains elusive due to drawbacks in evaluation metrics and missing head-to-head comparisons of newly developed synthesizers that take advantage of diffusion models and large language models with state-of-the-art marginal-based synthesizers.

In this paper, we present a systematic evaluation framework for assessing tabular data synthesis algorithms. Specifically, we examine and critique existing evaluation metrics, and introduce a set of new metrics in terms of fidelity, privacy, and utility to address their limitations. Based on the proposed metrics, we also devise a unified objective for tuning, which can consistently improve the quality of synthetic data for all methods. We conducted extensive evaluations of 8 different types of synthesizers on 12 real-world datasets and identified some interesting findings, which offer new directions for privacy-preserving data synthesis.

CCS CONCEPTS

• **Security and privacy**;

KEYWORDS

Privacy-preserving data synthesis, evaluation protocols

ACM Reference Format:

Anonymous Author(s). 2025. Systematic Assessment of Tabular Data Synthesis Algorithms. In *Proceedings of International Conference on Management of Data (SIGMOD '25).* ACM, New York, NY, USA, 20 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Data-driven decision-making has emerged as the prevailing approach to advance science, industrial applications, and governance, creating the necessity to share and publish tabular data. At the same time, growing concerns about the privacy breaches caused by data disclosure call for data publishing approaches that preserve privacy. One increasingly advocated and adopted approach to reduce privacy risks while sharing data is to release synthetic data. Ideally, synthetic data can effectively fit any data processing workflow designed for the original data without privacy concerns. Data synthesis initiatives have been promoted not only by the research community [30, 62] but also among non-profit organizations [47, 65] and government agencies [5, 45, 46].

In this paper, we study data synthesis algorithms for tabular data, which we call **synthesizers**. In recent years, a plethora of synthesizers have been proposed, which can be roughly categorized into two groups: statistical and deep generative. Statistical synthesizers [41, 43, 73, 74] use low-order marginals to create synthetic datasets that match real distributions. They were the best-performing algorithms in NIST competitions [45, 46]. Deep generative synthesizers [7, 31, 33, 51, 69], on the other hand, learn the data distribution from real data and generate synthetic instances by sampling from the distribution. With the recent development in deep generative models (*e.g.,* diffusion models [27] and large language models [57, 67]), new synthesizers are proposed [7, 32, 33, 72] to extend these successes to the realm of tabular data synthesis.

While recent state-of-the-art approaches achieve compelling results in synthesizing authentic data, a comprehensive understanding of the strengths and weaknesses of different synthesizers remains elusive. We are unaware of any head-to-head comparison of the best statistical synthesizers [41, 43, 74] against the best deep generative synthesizers [7, 32, 33].

In addition, there is a lack of principled and widely accepted evaluation metrics. It is known that evaluating synthesizers is inherently difficult [2, 13, 30, 63], and qualitative evaluation of tabular data through visual inspection is also infeasible. Table 1 shows that existing papers use different evaluation metrics, resulting in substantially disparate conclusions.

Since synthetic data is proposed to protect the privacy of real data, privacy should be an indispensable part of the evaluation. While many *differentially private (DP)* synthesizers [41, 42, 73, 74] have been proposed with provable privacy guarantees, they often suffer from significant utility loss [22, 28, 62]. Therefore, a parallel line of research focuses on what we call *heuristically private (HP)* synthesizers [7, 32, 33, 51, 72, 75], which do not integrate DP for model learning. These methods typically rely on similarity-based metrics to empirically evaluate privacy risks[1]. The underlying rationale is that the closer the synthetic data is to real data, the higher the risk of information leakage. However, the meaning of these metrics has not been sufficiently examined. Ironically, recent attacks [21] utilize these similarity-based metrics to successfully recover a large proportion of outliers in real data, casting doubt on the effectiveness and reliability of current privacy measurements.

Another problem of existing evaluations is the prevalent reliance on default hyperparameter settings for training synthesizers [7, 32, 61, 72]. This approach can result in suboptimal performance of synthesizers and skewed comparisons, leading to misleading conclusions[2]. The root of this problem lies in the fact that

SIGMOD '25, June 22-27, 2025, Berlin, Germany

ACM ISBN 978-1-4503-XXXX-X/18/06
https://doi.org/XXXXXXX.XXXXXXX

[1]These metrics are widely used both in academia and industry as default privacy evaluation protocols, see [21] for detailed analysis.

[2]We discuss these misleading conclusions in our technical report version.

Table 1: Main evaluation protocols for synthesizers in the literature. "-" means that the synthesizer does not consider this property. " ✓ " denotes that the model is viewed as the state-of-the-art method by many related studies [22, 42, 49, 62]. " ✗ " denotes that the model is frequently used in recent papers[32, 33, 35] for comparison as a state-of-the-art method.

Algorithm	Privacy Type	Architecture	Fidelity	Utility	State-of-the-art
PGM [41, 43]	Differentially Private	Probabilistic Graphical Model	Total Variation Distance	Range Query	✓
PrivSyn [74]	Differentially Private	Non-parametric Model	Total Variation Distance	Machine Learning Efficacy + Range Query	✓
TVAE [69]	Heuristically Private	Variational Autoencoder	Likelihood Fitness	Machine Learning Efficacy	✗
CTGAN [69]	Heuristically Private	Generative Adversarial Network	Likelihood Fitness	Machine Learning Efficacy	✗
PATE-GAN [31]	Differentially Private	Generative Adversarial Network	-	Machine Learning Efficacy	✗
TabDDPM [33]	Differentially Private	Diffusion Model	Correlations	Machine Learning Efficacy	✓
TableDiffusion [64]	Differentially Private	Diffusion Model	Kolmogorov-Smirnov Test	Machine Learning Efficacy	✓
GReaT [7]	Heuristically Private	Large Language Model	-	Machine Learning Efficacy	✓

many synthesizers do not provide details about hyperparameter selections, posing a challenge in conducting fair comparisons among various synthesizers on an equal footing.

The above concerns motivate us to design a systematic evaluation framework for data synthesis to elucidate the current advancements in this field. We believe that principled and systematic assessments are crucial: not only can they help practitioners choose from a large set of synthesizers, but they can also aid in understanding the strengths and weaknesses of each type of method, making progress toward developing better synthesis algorithms.

1.1 Our Contributions

In this paper, we examine, characterize, and critique the commonly used evaluation metrics, and propose a systematic framework as well as a set of new metrics for data synthesis evaluation. Our assessments unfold along three main axes:

- *Fidelity.* To address the heterogeneity and high dimensionality of tabular data, we introduce a new fidelity metric based on Wasserstein distance, providing a unified way to evaluate the numerical, discrete, and mixture marginals under the same criteria.
- *Privacy.* We identify the inadequacy of existing similarity-based metrics and membership inference attacks and propose a novel privacy metric called membership disclosure score to directly gauge the disclosure risk of synthesis algorithms. Our measurement aligns with the principles of differential privacy and only requires black-box access to synthesizers.
- *Utility.* We advocate two tasks for assessing the utility of synthesizers: machine learning prediction and range (point) query. To eliminate the inconsistent performance caused by the choice of different machine learning models, we introduce a new and robust metric to quantify the distribution shift of synthetic data.

SynMeter. We implement a systematic evaluation framework called SynMeter to support the assessment of data synthesis algorithms with the proposed metrics. Differing from the existing evaluations, SynMeter incorporates the model tuning phase, which eases hyperparameter selection and consistently improves the performance of synthesizers for fair comparison. With a modular design, SynMeter can easily integrate additional synthesizers and datasets by implementing new functional codes to the relevant modules. Our code is publicly available[3], facilitating researchers to leverage existing pipelines to tune, assess, or benchmark (new) synthesis algorithms.

[3]https://anonymous.4open.science/r/SynMeter

Main Findings. Extensive experiments have been conducted on a wide range of state-of-the-art heuristically private (HP) and differentially private (DP) synthesizers (list in Table 1) over 12 real-world datasets. Our evaluation finds that no single synthesis algorithm emerges as the definitive leader. Synthesizers struggle to balance competitive performance with robust privacy measures, exhibiting a pronounced performance disparity between HP and DP synthesis algorithms. Specifically, we have the following interesting findings:

- Diffusion models are surprisingly good at synthesizing tabular data. Although they were originally introduced for image generation, our evaluation indicates their impressive capability of synthesizing highly authentic tabular data. For instance, the HP diffusion-based synthesizer (*i.e.,* TabDDPM) surpasses its counterparts and nearly reaches the empirical upper bound in terms of both fidelity and utility. However, it suffers from significant membership privacy risks, and directly applying differential privacy to it would negate its advantages, highlighting a critical challenge for further exploration.
- Statistical methods are still competitive synthesizers. State-of-the-art statistical methods consistently outperform all deep generative synthesizers when DP is required, especially when the privacy budget is small (*e.g.,* $\epsilon = 0.5$). Impressively, even in heuristically private settings, these statistical approaches still maintain competitive performance and empirical privacy protection over many sophisticated synthesizers like TVAE and GReaT.
- Large language models (LLM) are semantic-aware synthesizers. LLM-based method (*i.e.,* GReaT) excels at generating realistic tabular data when the input dataset consists of rich semantic attributes, establishing a new paradigm in data synthesis.
- Despite CTGAN being widely recognized as the strong baseline in data synthesis, a closer examination using proposed metrics reveals that it struggles to learn marginal distributions, which leads to unsatisfactory results on complex tabular data.

Related Studies. Several studies [3, 9, 28, 62] have been conducted to benchmark tabular synthesis algorithms. However, they all focus on DP synthesizers, neglecting extensively studied HP synthesizers. In particular, they did not include the newly proposed deep generative synthesizers using diffusion models or LLM [7, 32, 33], which we found to be very promising in our studies. Furthermore, they directly use existing metrics for comparisons, while we identify the limitations of these metrics and propose a new set of metrics as well as the systematic assessment for tabular data synthesis.

Roadmap. We first propose a systematic framework for tabular data synthesis evaluation in Section 2. Then from Section 3 to Section 5, we discuss existing evaluation methods and propose our metrics from fidelity, privacy, and utility. In Section 6 we introduce a unified tuning objective and our framework implementation. The experimental results are shown in Section 7. We summarize our findings and offer some suggestions for practitioners in Section 8.

2 A FRAMEWORK FOR DATA SYNTHESIS EVALUATION

2.1 Evaluation Criteria

A synthesizer A takes as input a dataset $\mathcal{D}$, which is assumed to be sampled i.i.d. from some (often unknown) underlying data distribution $\mathbb{D}$, and outputs a generator G. Given an integer n, G outputs a dataset $\mathcal{S}$ with n data instances. We consider three classes of desirable properties for synthesizers: Fidelity, Privacy, and Utility.

Fidelity. As the substitute for real data, the distribution for the synthetic data instances should be close to $\mathbb{D}$. Since $\mathbb{D}$ is often unknown, fidelity is measured by the distance between the input dataset $\mathcal{D}$ and the synthetic dataset $\mathcal{S}$. If one partitions the input dataset $\mathcal{D}$ into a training set $\mathcal{D}_{\text{train}}$ and a testing set $\mathcal{D}_{\text{test}}$, one can measure fidelity as closeness to either $\mathcal{D}_{\text{train}}$ or $\mathcal{D}_{\text{test}}$.

Privacy. Using synthetic data is usually motivated by the desire to protect the input dataset. Without this consideration, one may share the input dataset to achieve the best fidelity and utility. For privacy, differential privacy [17] is generally considered to be the golden standard. Some synthesizers are designed to satisfy DP, either by directly perturbing data statistics or adding noises during training. However, satisfying DP under reasonable parameters may result in poor performance. Heuristically private (HP) synthesizers do not satisfy DP, and aim to protect privacy empirically. Therefore, privacy metrics are necessary to evaluate HP synthesis algorithms. In addition, privacy metrics are also useful for evaluating DP synthesizers, as the privacy analysis may not be tight and the degrees of tightness vary based on the algorithms. Two algorithms with the same DP parameters ϵ, δ may offer different levels of privacy protection (analyzed in Section 7.5).

Utility. Synthetic data often replaces real datasets to perform specific downstream tasks. In these cases, the synthetic data may not necessarily need to preserve high fidelity, as long as it achieves good utility for these tasks. Hence, utility evaluation is useful to measure the effectiveness of synthesizers for common tasks.

2.2 Evaluation Pipelines

We propose SynMeter, a systematic evaluation framework for synthesizers. The SynMeter pipeline consists of four phases: data preparation, model tuning, model training, and model evaluation.

The *data preparation* phase preprocesses data for learning algorithms. (Here we assume no missing values in the original data. The missing values problem has been extensively studied [54], which is orthogonal to data synthesis.) In this phase, statistical methods [41, 73, 74] select low-dimensional marginals to serve as compact representations for capturing data distributions. Deep generative models [27, 69, 72, 75] apply standard data processing techniques like data encoding and normalization.

The goal of *model tuning* phase is to select the optimal hyperparameters for data synthesizers. While this phase is often neglected in many data synthesis models, we show in Section 7.5 that a proper tuning objective can boost the performance of synthesizers.

The *model training* phase focuses on model learning with tuned hyperparameters. Various generative models implement different architectures and optimization objectives.

In the *model evaluation* phase, the trained model generates some synthetic data, which are used for evaluation.

3 FIDELITY EVALUATION

In this section, we first review existing fidelity measurements, identify their limitations, and introduce a new fidelity metric that addresses these limitations.

3.1 Existing Metrics and Limitations

Low-order Statistics. Marginals are the workhorses of statistical data analysis and well-established statistics for one(two)-way marginals have been used to assess the quality of synthetic data.

Distribution Measurements. Total Variation Distance (TVD) [33, 41, 74] and Kolmogorov-Smirnov Test (KST) [72] are used to measure the univariate distribution similarity for categorical and numerical attributes, respectively. The main problem with this approach is the lack of versatility. Each type of marginal requires a distinct statistical measure, which complicates the ability to perform a comprehensive comparison across various attribute types.

Correlation Statistics. Some researchers use correlation difference, *i.e.,* the difference of correlation scores on synthetic and real data, to measure the pairwise distribution similarity. Popular correlation statistics like Theil's uncertainty coefficient [75], Pearson correlation [72], and the correlation ratio [33] are applied for different types of two-way marginals (categorical, continuous, and mixed). In addition to the lack of universality, this approach also suffers from the problem that correlation scores capture only limited information about the data distribution. Two attributes may have the same correlation score both in the real data and in the synthetic data, yet their underlying distributions diverge significantly—a phenomenon known as the scale invariance of correlation statistics [68].

Likelihood Fitness. Some studies [69] assume the input data are generated from some known probabilistic models (*e.g.,* Bayesian networks), thus the likelihood of synthetic data can be derived by fitting them to the priors. While likelihood fitness can naturally reflect the closeness of synthetic data to the assumed prior distribution, it is only feasible for data whose priors are known, which is inaccessible for most real-world complex datasets.

Other Approaches. α-Precision [2] defines fidelity as the proportion that the synthetic samples are covered by real data. A synthetic record is covered if its distance to the nearest real point is within a predefined threshold. Probabilistic mean squared error (pMSE) [60] employs a logistic regression discriminator to distinguish between synthetic and real data, using relative prediction confidence as the fidelity metric. The effectiveness of these heuristic metrics highly relies on the choice of auxiliary evaluators (*e.g.,* threshold and discriminator), which requires careful calibration to ensure meaningful comparisons across different datasets and synthesizers.

3.2 Proposed Metric: Wasserstein Distance

We opt for Wasserstein distance to measure the distribution discrepancies between synthetic data and real data. Originating from optimal transport theory [53], the Wasserstein distance provides a structure-aware measure of the minimal amount of work required to transform one distribution into another. Formally, Let $\mathbf{P} = (p_1, p_2, \ldots p_n)$ and $\mathbf{Q} = (q_1, q_2, \ldots q_n)$ be the two probability distributions, and $\mathbf{C}$ be a matrix of size $n \times n$ in which $\mathbf{C}_{ij} \geq 0$ is the cost of moving an element i of $\mathbf{P}$ to the element j of $\mathbf{Q}$ ($\mathbf{C}_{ii} = 0$ for all element i). The optimal transport plan $\mathbf{A}$ is defined as below:

$$\begin{aligned} &\min_{\mathbf{A}} \quad \langle \mathbf{C}, \mathbf{A} \rangle \\ &\text{s.t. } \mathbf{A}\mathbb{1} = \mathbf{P}, \quad \mathbf{A}^\top \mathbb{1} = \mathbf{Q}, \end{aligned} \tag{1}$$

where $\langle \cdot, \cdot \rangle$ is inner product between two matrices, $\mathbb{1}$ denotes a vector of all ones. Let $\mathbf{A}^*$ denote the solution to the above optimization problem, the Wasserstein distance is defined as:

$$\mathcal{W}(\mathbf{P}, \mathbf{Q}) = \langle \mathbf{C}, \mathbf{A}^* \rangle. \tag{2}$$

Now we can use Wasserstein distance to define the fidelity:

Definition 1. *(Wasserstein-based Fidelity Metric) Let* $\mathbf{P}$ *and* $\mathbf{Q}$ *be the marginal distributions derived from synthetic data* $\mathcal{S}$ *and real data* $\mathcal{D}$*, the fidelity of synthesis algorithm* A *is:*

$$\text{Fidelity}(\mathrm{A}) \coloneqq \mathop{\mathbb{E}}_{\substack{\mathbf{P} \sim \mathcal{S} \\ \mathbf{Q} \sim \mathcal{D}}} \left[\mathcal{W}(\mathbf{P}, \mathbf{Q}) \right], \tag{3}$$

where $\mathcal{W}(\mathbf{P}, \mathbf{Q})$ *is the Wasserstein distance between synthetic and real marginal distributions.*

Determining Cost Matrix. The Wasserstein distance requires the predefined cost matrix $\mathbf{C}$, which encapsulates the "cost" of transitioning from one distribution element to another. For k-way marginal distributions $\mathbf{P}$ and $\mathbf{Q}$, the cost matrix is formulated by summing the pairwise distances between corresponding elements:

$$\mathbf{C}_{ij} = \sum_r d(v_i^r, v_j^r). \tag{4}$$

Here, $v_i, v_j \in \mathbb{R}^k$ are the element located in i and j in k-way probability distributions. The distance $d(\cdot, \cdot)$ is tailored to the nature of the attributes, differing for numerical and categorical values:

$$d(v_i^r, v_j^r) = \begin{cases} \|v_i^r - v_j^r\|_1 & \text{if numerical} \\ \infty, & \forall v_i^r \neq v_j^r, \text{if categorical.} \end{cases} \tag{5}$$

This means we use l_1 distance for numerical values and consider the strict match for categorical values (*i.e.,* cost for any different categorical values is infinity). Note that it is also feasible to assign semantic distance for categorical attributes [38], we omit it because it requires specific context for optimization and most synthesizers do not model the semantics in tabular data.

Implementations. In practice, the real dataset $\mathcal{D}$ can be designated as either $\mathcal{D}_{train}$ or $\mathcal{D}_{test}$, depending on the specific requirements of analysis. The computation of Wasserstein distance involves solving the optimization problem in Equation 1, which is an instance of linear programming problems (LP) [53]. There are many open-source libraries like CVXPY [15] and OPT [20] that can be used to solve LP reasonably fast[4]. In our evaluation, we use OPT package [20] to compute all the one-way and two-way marginals and use the mean as the final fidelity score.

Merits of Wasserstein-based Fidelity Metric. Wasserstein distance offers several distinct advantages for fidelity evaluations: (i) Faithfulness. It serves as a natural and structure-aware statistic measurement for analyzing distribution discrepancies, which generalizes existing metrics such as TVD and contingency similarity [52, 72]. (ii) Universality. It can accommodate both numerical and categorical attributes and extend to any multivariate marginals under the same criterion, facilitating the evaluation of heterogeneous types of marginals.

4 PRIVACY EVALUATION

In this section, we examine the popular privacy evaluation methodologies employed for synthesizers and propose a novel and effective privacy metric: membership disclosure score.

4.1 Existing Metrics and Limitations

Similarity-based Privacy Metrics. Similarity-based approaches measure the empirical privacy risk of synthetic data by comparing the closeness with real data, which have become the conventional privacy evaluation metrics for HP synthesizers [21, 30, 40]. Among all heuristics, the Distance to Closest Records (DCR) is the most widely used both in academia [33, 70, 72, 75] and industry [4, 25, 44]. DCR looks at the distribution of the distances from each synthetic data point to its nearest real one, and uses the 5th percentile (or the mean) of this distribution as the privacy score. A small score indicates that the synthetic dataset is too close (similar) to real data, signaling a high risk of information leakage.

We point out two drawbacks of the DCR metrics. First, DCR measures privacy loss *averaged over the data points*, whereas it is crucial to protect the privacy of *every* individual. When measuring the privacy leakage across different individuals, one needs to ensure that the worst-case leakage is bounded. It is unacceptable to use a mechanism that sacrifices the privacy of some individuals even though the protection averaged over the population is good. This point is illustrated by the fact that the re-identification of one or a few individuals is commonly accepted as privacy breaches [17, 37, 47, 65]. DCR uses the 5th percentile (or mean) proximity to real data as a measurement, and fails to provide worst-case protection.

Second, DCR overestimates the privacy risks when data points are naturally clustered close together. As illustrated by discussions about differential privacy [18, 36], leaking information regarding an individual should not be considered a privacy violation if the leakage can occur even if the individual's data is not used. Analogously, having some synthetic data very close to real ones does *not* mean worse privacy if this situation can occur even if each data point is removed. Consider, for example, a dataset that is a mixture of two Gaussians with small standard deviations. A good synthetic dataset is likely to follow the same distribution, and has many data points very close to the real ones. DCR interprets this closeness as a high privacy risk, overlooking the fact that the influence of any individual training instance on synthetic data is insignificant.

Another concern of the similarity-based privacy metrics is the risk of additional information leakage when the adversary gets access to these metric values. Since these metrics directly measure the

[4] If the computation is still too slow, Sinkhorn distance [14], Sliced-Wasserstein distance [6] and random sampling can be used to approximate the Wasserstein distance. We implement these methods in our framework and detailed in technical report version.

distance between synthetic data points and real ones, these metric values provide direct information about the real data points. Recent research [21] has demonstrated the feasibility of reconstructing significant portions of training data when given black-box access to a synthesizer and an oracle for computing the DCR. This revelation is a stark warning to practitioners relying on such metrics.

Membership Inference Attacks. Another way to empirically verify the privacy of machine learning models is membership inference (MI) attacks [58], which have been extensively studied on discriminative models. Some studies [61] try to perform the attack on data synthesis by utilizing handcrafted features extracted from synthetic data distribution to train shadow models. Another work [29] approximates the likelihood of target records with density estimation and uses it as the confidence score for membership inference. However, as pointed out in [23] and [56], current MI attacks on synthetic data either depend on unrealistic settings or are too weak to differentiate the membership disclosure risks among various synthesizers [16]. Therefore, existing synthesis algorithms rarely use MI attacks for privacy evaluation [56].

4.2 Proposed Metric: Membership Disclosure Score (MDS)

We propose a new privacy metric, termed membership disclosure score (MDS), to estimate the membership disclosure risks of data synthesizers. MDS follows the same intuition behind DCR, namely, including a data point x in the data synthesizing process may result in the generation of data points that are close to x, and the degree of closeness is related to the degree of privacy leakage. However, we avoid the two drawbacks of DCR identified above. First, given an input dataset $\mathcal{D}$ and synthesizer A, we quantify the disclosure risk of one record $x \in \mathcal{D}$ as follows:

$$\mathrm{DS}(x, \mathrm{A}, \mathcal{D}) := \mathop{\mathbb{E}}_{\substack{\mathcal{H},\mathcal{H}' \subset \mathcal{D} \\ \mathcal{H}\cap\mathcal{H}'=\{x\}}} \left[||\mathrm{NN}(\mathcal{S}, x) - \mathrm{NN}(\mathcal{S}', x)||_1 \right], \quad (6)$$

where the $\mathcal{H}$ and $\mathcal{H}'$ are subsets of training samples sampled from the dataset $\mathcal{D}$, and they only differ in x. $\mathcal{S}$ and $\mathcal{S}'$ represent the synthetic datasets when the synthesizer A is trained on $\mathcal{H}$ and $\mathcal{H}'$. The term $\mathrm{NN}(\mathcal{S}, x)$ denotes the nearest neighbor (under l_1 distance) for record x within the synthetic dataset $\mathcal{S}$. We also use l_1 distance (empirically we find that the difference between using l_1 and l_2 distance is negligible) to measure the discrepancy resulting from training the model with or without the record x.

$\mathrm{DS}(x, \mathrm{A}, \mathcal{D})$ addresses DCR's second drawback identified above (over-estimating leakage) by quantifying how much including x *changes* the distance between x and the closest synthetic data. If including x results in records much closer to x to be generated, then the disclosure risk is high. If records close to x are generated whether or not x is included, then the disclosure risk for x is low.

We then define the membership disclosure risk of synthesizer A on dataset $\mathcal{D}$ to be the maximum disclosure risk among all records. This addresses DCR's first drawback identified above.

Definition 2. *(Membership Disclosure Score) Let* $\mathrm{DS}(x, \mathrm{A}, \mathcal{D})$ *be the disclosure risk of record* x *in real dataset* $\mathcal{D}$ *for synthesis algorithm* A. *The membership disclosure score of* A *is given by:*

$$\mathrm{MDS}(\mathrm{A}) := \max_{x \in \mathcal{D}} \mathrm{DS}(x, \mathrm{A}, \mathcal{D}). \quad (7)$$

Here we omit the notion of dataset $\mathcal{D}$ for simplicity. This metric signifies the maximum disclosure risk across the entire training dataset, which indicates that MDS is a *worst-case* privacy measurement. Intuitively, for a privacy-preserving synthesizer, the change of output (synthetic data) should remain insignificant by any individual training samples, resulting in a low MDS. It is worth mentioning that the disclosure score is calculated based on the difference between neighboring synthetic data, which ensures that real data information remains undisclosed (unlike DCR).

Implementation. Directly computing this score entails retraining many models for each record, which is computationally prohibitive. Alternatively, we train m models once on random subsets of $\mathcal{D}$, and compute the disclosure score for all instances using the same set of m synthesis models (statistically each record is trained on half of the models, while the other half are not). This approach provides an efficient and practical approach to estimating MDS in real-world scenarios. Empirically we find that $m = 80$ is sufficient to accurately estimate MDS for all synthesizers.

Limitations of MDS. In Section 7.5, we show that MDS can differentiate the privacy risks of different synthesizers where DCR cannot. However, MDS has its own limitations and is not designed to replace metrics based on differential privacy or MI attacks. Specifically, MDS applies only to algorithms that learn the distribution of the input records and then synthesize from the distribution. There exist pathological synthesizers for which MDS is inappropriate. One example of such a synthesizer is the one that simply adds the same large constant to each record in the input dataset. In datasets where the records are far apart, this results in a large MDS (meaning low disclosure risks) yet completely reveals the dataset, assuming that the mechanism of the synthesizer is known. However, most (if not all) practical synthesizers are not such pathological ones, *i.e.,* they learn the distribution and then generate records from it. In summary, whenever DCR is used as a heuristic privacy metric, we believe that it is better to use MDS instead. More discussion about MDS is included in our technical report version.

5 UTILITY EVALUATION

In this section, we examine the existing utility measurements and propose two robust evaluation metrics for data synthesis.

5.1 Existing Metrics and Limitations

Machine Learning Efficacy. Machine learning efficacy has emerged as the predominant utility metric for data synthesis [32, 33, 69, 72]. Specifically, it first forms a classification (regression) problem by treating the target attribute as labels. Then, it computes the average performance of certain classifiers or regressors in a "train on synthetic data, test on real data" (TSTR) framework [31]. This metric, ideally, should reflect the synthetic data's ability to preserve the real data distribution through test performance.

Shortfalls of Machine Learning Efficacy. The computation of this metric involves the selection of machine learning models (we also call "evaluators"). It is known that tabular data performs differently when evaluated with various evaluators [33], and there is no dominant model that can achieve the best performance on all tabular datasets [24] (shown in Section 7.5). Consequently, selecting different evaluators for comparison can lead to divergent outcomes, making fair assessment difficult. Furthermore, it is expected that high-quality synthetic data should closely mimic the performance behavior of real data across a spectrum of machine learning models, which is pivotal for tasks like model selection. However, directly using the accuracy of certain evaluators as the utility metric fails to reflect the performance behaviors of synthetic data.

5.2 Proposed Metrics: MLA and Query Error

Machine Learning Affinity (MLA). The inconsistent performance of different machine learning models is caused by the distribution shift of synthetic data. It is known that machine learning models vary in their sensitivity to such shifts, leading to divergent levels of performance degradation [39]. To accurately reflect these performance fluctuations on synthetic data, we introduce an interpretable, easy-to-compute metric called machine learning affinity. It measures the relative accuracy difference between synthetic and real data under the TSTR framework. More formally we define:

Definition 3. *(Machine Learning Affinity) Let $\mathcal{E}$ be a set of candidate machine learning models (evaluators), let $e_{\mathcal{D}_{train}}$ and $e_{\mathcal{D}_{\mathcal{S}}}$ be evaluators trained on real training data $\mathcal{D}_{train}$ and synthetic data $\mathcal{S}$, $\mathcal{A}(e, \mathcal{D}_{test})$ denotes the evaluator's accuracy (F1 score or RMSE) when performed on test dataset $\mathcal{D}_{test}$. The MLA of synthesizer* A *is:*

$$\mathrm{MLA}(\mathsf{A}) \coloneqq \mathop{\mathbb{E}}_{e\in\mathcal{E}}\Big[\frac{\mathcal{A}(e_{\mathcal{D}_{train}}, \mathcal{D}_{test}) - \mathcal{A}(e_{\mathcal{S}}, \mathcal{D}_{test})}{\mathcal{A}(e_{\mathcal{D}_{train}}, \mathcal{D}_{test})}\Big]. \tag{8}$$

Instead of directly averaging the test performance, MLA measures the *relative* performance discrepancy over various machine learning models. Therefore, the low MLA not only indicates the effectiveness of the synthetic data for machine learning tasks but also represents similar performance behaviors with real ones.

Query Error. Range query and point query are fundamental tasks for tabular data analysis [11]. However, they are often overlooked in the evaluation of most state-of-the-art synthesizers [7, 64, 69, 72]. To accommodate both range query and point query, we follow [43] to define the query error as below:

Definition 4. *(Query Error) Consider a subset of k attributes $a = \{a_1, a_2, ..., a_k\}$ sampled from data $\mathcal{D}$. For each attribute, if a_i is categorical, a value v_i is randomly chosen from its domain $\mathbb{R}(a_i)$, which forms the basis for a point query condition; for numerical attributes, two values s_i and d_i from $\mathbb{R}(a_i)$ are randomly sampled as the start and end points, to construct a range query condition. The final query $c \in C$ combines k sub-queries and is executed on both real and synthetic data to obtain query frequency ratios $\mu_c^{\mathcal{D}_{test}}$ and $\mu_c^{\mathcal{S}}$. The query error of synthesis algorithm* A *is defined as:*

$$\mathrm{QueryError}(\mathsf{A}) \coloneqq \mathop{\mathbb{E}}_{c\in C}\big[||\mu_c^{\mathcal{D}_{test}} - \mu_c^{\mathcal{S}}||_1\big]. \tag{9}$$

Implementations. We utilize eight machine learning models to compute MLA: SVM, Logistic Regression (or Ridge Regression), Decision Tree, Random Forest, Multilayer Perceptron (MLP), XGBoost [12], CatBoost [55], and Transformers [24]. Each model is extensively tuned on real training data to ensure optimal performance (refer to the technical report version for details). Performance on classification and regression is evaluated by the F1 score and RMSE, respectively. For query error, we randomly construct 1,000 3-way query conditions and conduct range (point) queries for both synthetic and real data.

6 SynMeter

This section introduces SynMeter, a modular toolkit tailored for assessing data synthesis algorithms across three dimensions: fidelity, privacy, and utility.

6.1 Tuning Objective

Most synthesizers do not provide details about how to tune their hyperparameters [35, 43, 64, 69]. Instead, default settings are used for evaluations, which may not always yield optimal results, leading to subpar performance and skewed comparisons. To address this issue, we introduce the model tuning phase within our evaluation framework to facilitate the selection of hyperparameters for synthesizers. More precisely, we propose a simple and unified tuning objective, $\mathcal{L}$, to steer the tuning process:

$$\mathcal{L}(\mathsf{A}) = \alpha_1\mathrm{Fidelity}(\mathsf{A}) + \alpha_2\mathrm{MLA}(\mathsf{A}) + \alpha_3\mathrm{QueryError}(\mathsf{A}).$$

Here, α_1, α_2, and α_3 are coefficients assigned to fidelity (*i.e.,* Wasserstein distance), machine learning affinity, and query error. In our evaluation, we set all coefficients to 1, since we find that it can significantly improve the quality of the synthetic data compared to default configurations. We utilize Optuna [1] for hyperparameter optimizations by minimizing $\mathcal{L}$.

Discussion. Despite the simplicity of the setup described, we observe consistent improvement when applying this tuning strategy. For instance, this approach can boost the fidelity of TabDDPM by 13% and utility by at least 11% (refer to Section 7.5 for detailed comparisons). In practice, the values of fidelity and utility metrics fall within the same scale (although they use different measurements), which alleviates the need to assign different weights for each aspect. Practitioners can adjust these coefficients based on the specific requirements of applications (*e.g.,* higher α_2 for model selection [59]). We leave other assignment strategies for future work.

It is worth mentioning that we do not include the privacy metric for model tuning. The reason is two-fold: (1) DP synthesizers already offer provable privacy guarantees, eliminating the need for privacy-oriented tuning. By excluding the privacy metric, we establish a more generalized tuning objective that can apply to both HP and DP synthesizers. (2) Empirically we find that incorporating MDS into the tuning objective $\mathcal{L}$ results in negligible improvements in the privacy of synthetic data (this may be because some synthesizers are inherently private, discussed in Section 7.5). Moreover, computing MDS requires training multiple models for each tuning iteration, which can significantly hinder the efficiency of the tuning phase.

6.2 Implementation

SynMeter's design is inherently modular, offering significant flexibility to incorporate new synthesizers and datasets. As depicted in Figure 1, SynMeter comprises four modules, each module is implemented with an abstract interface for any synthesis algorithm. We envisage that SynMeter can be used for the following purposes:

- With eight state-of-the-art synthesis algorithms implemented, SynMeter facilitates data owners to tune, train, and select different synthesizers for data publishing.

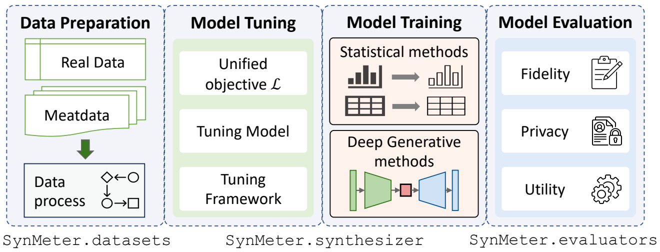


Figure 1: Overview of SynMeter**.**

- It serves as a benchmark for data synthesis, providing systematic evaluation metrics for comparative studies.
- The modular architecture, connected via abstract interfaces, allows for easy integration of any synthesizers and enables adaptation into other domains (*e.g.,* genomic data synthesis [48]), by re-implementing module-specific processing functions.

7 EXPERIMENTS

Based on the proposed evaluation framework and metrics, in this section, we present a series of comprehensive experiments to answer the following question:

- **RQ1:** How do the various synthesizers perform under the systematic assessment? What are the new findings?
- **RQ2:** Why do these methods work well (or not so well) on certain properties? How can our metrics help for in-depth analysis?
- **RQ3:** How effective are the proposed metrics and evaluation framework?

7.1 Data Synthesis Models

We study a wide range of state-of-the-art data synthesizers, from statistical methods to deep generative models. We select them as they are either generally considered to perform best in practice [45, 46], widely used [50, 69], or recently emerged [7, 33, 64]. These synthesizers can be categorized into two groups: heuristic private (HP) synthesizers and differentially private (DP) synthesizers.

HP Synthesizers. Synthesizers in this category are developed without integrating DP:

- **CTGAN** [69] is one of the most widely used HP synthesis algorithms. It utilizes generative adversarial networks to learn the tabular data distributions. Training techniques like conditional generation and Wasserstein loss [26] are used.
- **TVAE** [69] is the state-of-the-art variational autoencoder for tabular data synthesizer, which uses mode-specific normalization to tackle the non-Gaussian problems of continuous distributions.
- **TabDDPM** [33] is the state-of-the-art diffusion model for data synthesis. It leverages the Gaussian diffusion process and the multinomial diffusion process to model continuous and discrete distributions respectively.
- **GReaT** [7] utilizes the large language model (LLM) for data synthesis. It converts records to textual representations for LLM and generates synthetic data with prompts.

DP Synthesizers. These methods are either inherently designed with DP or are adaptations of HP models with additional mechanisms to offer provable privacy guarantees:

Table 2: List of datasets used for evaluation.

Name	# Train	# Val	# Test	# Num	# Cat	Task type
Adult	20838	5210	6513	6	9	Binclass
Shoppers	7891	1973	2466	10	8	Binclass
Phishing	7075	1769	2211	0	31	Binclass
Magic	12172	3044	3804	10	1	Binclass
Faults	1241	311	389	24	4	Multiclass(7)
Bean	8710	2178	2723	16	1	Multiclass(7)
Obesity	1350	338	423	8	9	Multiclass(7)
Robot	3491	873	1092	24	1	Multiclass(4)
Abalone	2672	668	836	8	1	Regression
News	25372	6343	7929	46	14	Regression
Insurance	856	214	268	3	4	Regression
Wine	3134	784	980	12	0	Regression

- **PGM** [41] is the state-of-the-art DP synthesizer, which uses probabilistic graphical models [43] to learn the dependence of low-dimensional marginals for data synthesis. It won the NIST Differential Privacy Synthetic Data Challenge [45].
- **PrivSyn** [74] is a non-parametric DP synthesizer, which iteratively updates the synthetic dataset to make it match the target noise marginals. This method also shows strong performance in NIST competitions [45, 46].
- **PATE-GAN** [31] shares a similar architecture with CTGAN, but leverages the Private Aggregation of Teacher Ensembles (PATE) mechanism [50] to offer DP guarantees.
- **TableDiffusion** [64] is a newly proposed diffusion model for tabular data synthesis, which uses Differentially Private Stochastic Gradient Descent (DP-SGD) to enforce privacy.

Note that all DP synthesizers can be adapted to the HP scenario by either using their HP equivalents (*i.e.,* CTGAN as the HP model for PATE-GAN, TabDDPM for TableDiffusion) or setting the privacy budget to infinity (*i.e.,* PGM and PrivSyn). However, some HP synthesizers, such as TVAE and GReaT, do not have corresponding DP variants. Thus, we only assess their performance within the context of HP models.

Our goal is not to benchmark all possible synthesizers but to focus on the best-known and broad spectrum of state-of-the-art synthesizers. Hence, some synthesizers [32, 35, 42, 75] are omitted as they are considered the variants of the algorithms we assess. Some approaches (*e.g.,* TabDDPM and GReaT) are nearly simultaneously introduced, limiting opportunities for an extensive comparison. Our evaluation fills this gap by providing the first comprehensive evaluation of all state-of-the-art in a standardized setting.

7.2 Experimental Setups

Datasets. For systematic investigation of the performance of synthesis algorithms, we consider a diverse set of 12 real-world public datasets. These datasets have various sizes, nature, attributes, and distributions. We explicitly divide datasets into training and test with a ratio of 8:2, then split 20% of the training dataset as the validation set, which is used for model tuning. All datasets are publicly available and widely used, and the full list is detailed in Table 2.

Evaluation Details. We use implementations from the original papers and make some small modifications to integrate them into our evaluation framework. For PGM and PrivSyn, we discretize

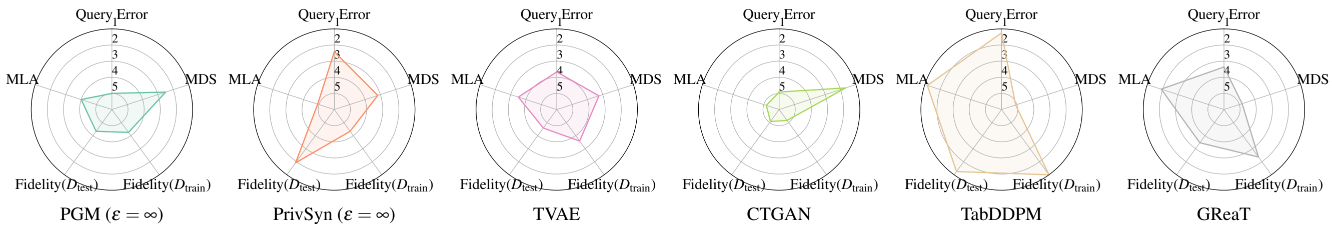


Figure 2: Average ranking comparison for six HP synthesizers (outer means higher rank and better performance). Each vertex is the average performance rank of the method across 12 datasets, and each axis is the evaluation metric. "Fidelity($\mathcal{D}_{\text{train}}/\mathcal{D}_{\text{test}}$)" denotes the fidelity (*i.e.,* Wasserstein distance) is evaluated on the training/test dataset. "MDS" is the proposed membership disclosure score (detailed in Section 4.2), and "MLA" is machine learning affinity (detailed in Section 5.2).

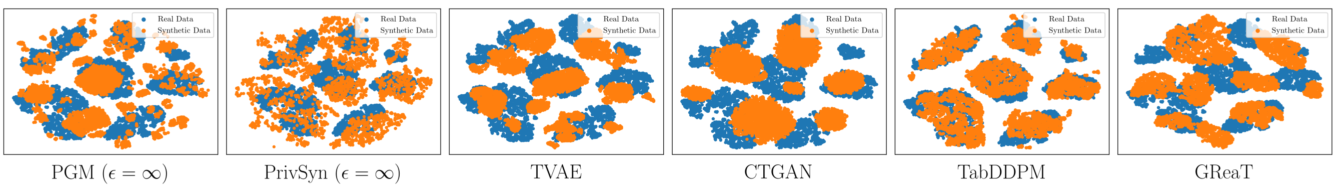


Figure 3: Visualization comparison of HP synthesizers on Bean dataset. Real data are in blue and synthetic data are in orange.

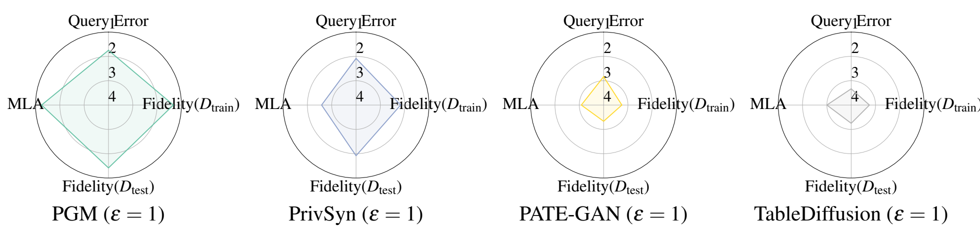


Figure 4: Average ranking comparison for four DP synthesizers. All methods offer provable privacy guarantees so we remove the privacy axis for comparison (the empirical privacy protections of these synthesizers are evaluated in Section 7.4).

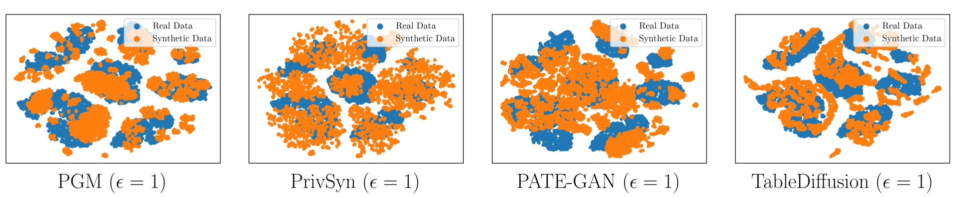


Figure 5: Visualization comparison of DP synthesizers on Bean dataset. Real data are in blue and synthetic data are in orange.

numerical attributes into discrete ones by binning and setting the privacy budget to infinity when compared with HP synthesis algorithms. During the evaluation, we first tune the synthesizers with our proposed tuning objective using training and validation data and train them with optimal performance. Then, synthetic data are generated by trained generator G for evaluation, where we test 10 times and report the mean and standard deviation as the final score. The random seeds are fixed for each model to improve reproducibility. The complete hyperparameter search space of each synthesis algorithm is detailed in our technical report version.

Baselines. We introduce additional baselines to better understand the performance of synthesizers on the proposed evaluation metrics. For fidelity, the `HALF` baseline randomly divides the real data into two equal parts: one serving as the training dataset $\mathcal{D}$, and the other as the synthetic data $\mathcal{S}$. Since the two are drawn from the same distribution, this serves as an upper bound on the fidelity scores. For `HISTOGRAM`, the synthetic dataset is generated by PGM with only one-way marginals (*i.e.,* histograms). Thus in the synthetic dataset, all attributes are independent from each other. We use `HISTOGRAM` as the empirical lower bounds of fidelity.

Table 3: Fidelity evaluation (*i.e.,* Wasserstein distance, the lower score indicates the synthetic data is of higher quality) of synthesis algorithms on train data $\mathcal{D}_{\text{train}}$. The privacy budget ϵ of HP synthesizers is ∞ (the top part), and the budget for DP synthesizers is 1 (the middle part). `HALF` and `HISTOGRAM` are the baselines that serve as the empirical upper bound and lower bound of the fidelity for HP synthesizers (detailed in Section 7.2). The best result is in bold.

	Adult	Shoppers	Phishing	Magic	Faults	Bean	Obesity	Robot	Abalone	News	Insurance	Wine
PGM	0.186±.010	0.092±.002	0.019±.001	0.037±.002	0.056±.002	0.040±.002	0.041±.001	0.050±.002	0.037±.002	0.060±.001	0.038±.005	0.066±.001
PrivSyn	0.024±.001	0.030±.001	**0.010±.001**	0.015±.003	0.064±.006	0.035±.002	0.034±.002	0.065±.012	0.024±.004	**0.018±.001**	0.033±.002	0.017±.000
TVAE	0.085±.002	0.156±.001	0.024±.001	0.021±.003	0.055±.007	0.047±.006	0.055±.004	0.053±.001	0.048±.003	0.081±.001	0.078±.007	0.039±.000
CTGAN	0.059±.001	0.062±.001	0.062±.002	0.157±.006	0.133±.004	0.139±.005	0.072±.002	0.106±.003	0.049±.004	0.040±.001	0.090±.004	0.033±.001
TabDDPM	**0.020±.001**	**0.022±.001**	0.015±.001	**0.011±.003**	**0.026±.002**	**0.015±.002**	**0.017±.001**	**0.015±.002**	**0.015±.004**	0.034±.001	**0.028±.005**	**0.011±.000**
GReaT[1]	0.050±.002	0.049±.003	0.076±.002	0.037±.003	0.050±.006	0.020±.001	0.055±.005	0.055±.003	0.022±.005	-	0.094±.004	0.019±.001
PGM (ϵ = 1)	0.198±.013	0.103±.002	**0.023±.001**	**0.042±.003**	**0.086±.003**	**0.048±.004**	**0.063±.001**	**0.065±.001**	**0.052±.003**	**0.062±.002**	**0.071±.002**	**0.068±.001**
PrivSyn (ϵ = 1)	**0.045±.002**	**0.077±.005**	0.033±.002	0.052±.003	0.228±.007	0.142±.007	0.167±.009	0.169±.021	0.127±.009	0.070±.002	0.124±.006	0.156±.004
PATE-GAN (ϵ = 1)	0.139±.001	0.176±.002	0.173±.002	0.153±.005	0.204±.003	0.520±.006	0.086±.003	0.477±.002	0.331±.005	0.065±.002	0.385±.003	0.251±.000
TableDiffusion (ϵ = 1)	0.180±.002	0.209±.002	0.123±.002	0.132±.003	0.369±.002	0.148±.005	0.347±.003	0.203±.001	0.232±.005	0.135±.001	0.343±.002	0.108±.001
`HALF` (upper bound)	0.020±.002	0.018±.001	0.010±.002	0.011±.004	0.017±.002	0.015±.004	0.017±.003	0.010±.001	0.012±.004	0.009±.001	0.026±.004	0.006±.000
`HISTOGRAM` (lower bound)	0.213±.013	0.101±.003	0.027±.001	0.051±.003	0.081±.002	0.087±.002	0.051±.001	0.061±.002	0.069±.001	0.063±.002	0.046±.002	0.068±.000

[1] GReaT cannot be applied to the News dataset because of the maximum length limit of large language models.

Table 4: Privacy evaluation (*i.e.,* membership disclosure score, lower value means better empirical privacy protection) of HP synthesizers ($\epsilon = \infty$). `SELF` is the baseline that serves as the empirical lower bound of MDS (the upper bound of MDS is 0 by definition). The best result is in bold.

	Adult	Shoppers	Phishing	Magic	Faults	Bean	Obesity	Robot	Abalone	News	Insurance	Wine
PGM	0.131±.001	**0.133±.002**	**0.108±.003**	**0.028±.001**	0.125±.002	0.032±.003	0.138±.001	**0.031±.002**	0.113±.003	0.041±.001	0.242±.002	0.120±.003
PrivSyn	0.187±.002	0.164±.001	0.139±.003	0.033±.002	0.083±.001	0.037±.003	0.132±.002	0.078±.001	0.113±.003	0.059±.002	0.240±.001	0.071±.003
TVAE	0.199±.003	0.186±.002	0.172±.001	0.030±.005	0.093±.002	0.042±.001	0.155±.003	0.082±.002	**0.045±.001**	0.055±.003	0.217±.002	0.046±.001
CTGAN	**0.072±.002**	0.171±.003	0.149±.001	0.030±.003	**0.075±.003**	**0.031±.001**	**0.112±.001**	0.052±.003	0.045±.002	0.051±.005	**0.176±.003**	**0.024±.001**
TabDDPM	0.550±.001	0.230±.002	0.702±.003	0.031±.001	0.130±.002	0.064±.003	0.543±.001	0.164±.002	0.141±.003	**0.030±.001**	0.383±.002	0.120±.003
GReaT	0.495±.002	0.236±.003	0.568±.001	0.082±.002	0.125±.003	0.036±.004	0.357±.002	0.181±.003	0.065±.001	-	0.534±.002	0.072±.003
`SELF`	0.732±.000	0.498±.000	0.718±.000	0.178±.000	0.426±.000	0.257±.000	0.659±.000	0.318±.000	0.284±.000	0.267±.000	0.537±.000	0.342±.000

For privacy, we introduce the `SELF` baseline, which uses a direct copy of the real data as the synthetic data, establishing the worst privacy protection (lower bound). By the definition of MDS, the perfect privacy-preserving synthesizer would achieve a score of 0, which is the upper bound of privacy.

Having these lower and upper bounds for fidelity and privacy enables us to better interpret the scores of different synthesizers. We do not include baselines for utility metrics (*i.e.,* machine learning affinity and query errors), since their scores are ratios that can naturally reflect the relative deviation from the ground truth.

7.3 Overall Evaluation (RQ1)

Overview. Figure 2 and Figure 4 report the overview ranking results for HP and DP synthesizers, respectively. For HP synthesizers, TabDDPM exhibits superior fidelity and utility, albeit at the expense of compromising privacy. Statistical methods like PrivSyn also achieve impressive fidelity, surpassing complex deep generative models like TVAE and GReaT. Conversely, CTGAN, the most popular HP synthesizer, shows the least satisfactory results in synthetic data quality but offers better privacy protection.

In terms of DP synthesizers, statistical methods remain effective in both fidelity and utility. The performance of deep generative models drops significantly to ensure differential privacy. Even the strongest model (*i.e.,* TableDiffusion) underperforms statistical approaches by a large margin, which starkly contrasts with its performance in the HP context, indicating a pronounced impact of privacy constraints on deep generative models.

We also visualize the synthetic and real data by using the t-SNE technique [66], as depicted in Figure 3 and Figure 5. The visualizations generally reveal a considerable overlap between synthetic and real data points, demonstrating the synthesizers' ability to generate data resembling the original. One interesting observation is that PrivSyn is the only method that does not have clear cluster patterns. This is because PrivSyn is the only non-parametric approach that directly modifies the synthetic records to match the marginal distribution, potentially resulting in less realistic records.

Fidelity Evaluation. We assess the data synthesis fidelity by applying the Wasserstein distance to both training $\mathcal{D}_{\text{train}}$ (shown in Table 3) and test data $\mathcal{D}_{\text{test}}$ (included in technical report version due to similar results). The results demonstrate that TabDDPM leads among HP synthesizers (nearly reaches the empirical upper bound), while statistical methods (*e.g.,* PGM) excel within the DP synthesizers category. To achieve differential privacy, all deep generative models experience a significant fidelity loss, whereas statistical methods maintain stable performance. This stability is further highlighted when comparing performance across training and test data, where statistical methods exhibit a lesser performance

Table 5: Machine learning affinity of data synthesis. The lower value means the performance of synthetic data is more similar to real data. The privacy budget ϵ of HP synthesizers is set as ∞ (the top part), and the budget for DP synthesizers is set as 1 (the bottom part). The best result of each category is in bold.

	Adult	Shoppers	Phishing	Magic	Faults	Bean	Obesity	Robot	Abalone	News	Insurance	Wine
PGM	$0.086_{\pm.001}$	$0.193_{\pm.002}$	$0.037_{\pm.003}$	$0.073_{\pm.001}$	$0.255_{\pm.002}$	$0.035_{\pm.003}$	$0.332_{\pm.001}$	$0.146_{\pm.002}$	$0.096_{\pm.003}$	$0.498_{\pm.001}$	$0.270_{\pm.002}$	$0.347_{\pm.003}$
PrivSyn	$0.120_{\pm.003}$	$0.040_{\pm.001}$	$0.057_{\pm.002}$	$0.085_{\pm.003}$	$0.532_{\pm.001}$	$0.039_{\pm.002}$	$0.604_{\pm.003}$	$0.406_{\pm.001}$	$0.210_{\pm.002}$	$1.992_{\pm.003}$	$0.518_{\pm.001}$	$0.201_{\pm.002}$
TVAE	$0.035_{\pm.002}$	$0.011_{\pm.003}$	$0.031_{\pm.001}$	$0.075_{\pm.002}$	$0.217_{\pm.003}$	$0.059_{\pm.001}$	$0.294_{\pm.002}$	$0.128_{\pm.003}$	$0.245_{\pm.001}$	$0.147_{\pm.002}$	$0.336_{\pm.003}$	$0.091_{\pm.001}$
CTGAN	$0.039_{\pm.003}$	$0.031_{\pm.002}$	$0.068_{\pm.001}$	$0.154_{\pm.003}$	$0.525_{\pm.002}$	$0.103_{\pm.001}$	$0.893_{\pm.003}$	$0.434_{\pm.002}$	$0.282_{\pm.001}$	$0.104_{\pm.003}$	$1.700_{\pm.002}$	$0.222_{\pm.001}$
TabDDPM	$0.014_{\pm.001}$	$\mathbf{0.006_{\pm.002}}$	$\mathbf{0.007_{\pm.003}}$	$\mathbf{0.007_{\pm.001}}$	$\mathbf{0.085_{\pm.002}}$	$\mathbf{0.003_{\pm.003}}$	$\mathbf{0.021_{\pm.001}}$	$\mathbf{0.011_{\pm.002}}$	$\mathbf{0.043_{\pm.003}}$	$\mathbf{0.047_{\pm.001}}$	$\mathbf{0.140_{\pm.002}}$	$\mathbf{0.047_{\pm.003}}$
GReaT	$\mathbf{0.009_{\pm.002}}$	$0.009_{\pm.003}$	$0.020_{\pm.001}$	$0.033_{\pm.002}$	$0.183_{\pm.003}$	$0.017_{\pm.001}$	$0.117_{\pm.002}$	$0.050_{\pm.003}$	$0.038_{\pm.001}$	-	$0.292_{\pm.002}$	$0.083_{\pm.003}$
PGM ($\epsilon = 1$)	$\mathbf{0.101_{\pm.003}}$	$\mathbf{0.048_{\pm.001}}$	$\mathbf{0.041_{\pm.002}}$	$\mathbf{0.093_{\pm.003}}$	$\mathbf{0.489_{\pm.001}}$	$\mathbf{0.054_{\pm.002}}$	$\mathbf{0.531_{\pm.003}}$	$\mathbf{0.245_{\pm.001}}$	$\mathbf{0.241_{\pm.002}}$	$1.072_{\pm.003}$	$\mathbf{1.366_{\pm.001}}$	$0.340_{\pm.002}$
PrivSyn ($\epsilon = 1$)	$0.120_{\pm.002}$	$0.177_{\pm.003}$	$0.085_{\pm.001}$	$0.217_{\pm.002}$	$0.753_{\pm.003}$	$0.466_{\pm.001}$	$0.821_{\pm.002}$	$0.608_{\pm.003}$	$0.624_{\pm.001}$	$4.538_{\pm.002}$	$1.878_{\pm.003}$	$\mathbf{0.302_{\pm.001}}$
PATE-GAN ($\epsilon = 1$)	$0.126_{\pm.001}$	$0.135_{\pm.002}$	$0.530_{\pm.003}$	$0.394_{\pm.001}$	$0.781_{\pm.002}$	$0.781_{\pm.003}$	$0.877_{\pm.001}$	$0.755_{\pm.002}$	$2.119_{\pm.003}$	$0.259_{\pm.001}$	$2.325_{\pm.002}$	$0.405_{\pm.003}$
TableDiffusion ($\epsilon = 1$)	$0.198_{\pm.002}$	$0.135_{\pm.003}$	$0.074_{\pm.001}$	$0.133_{\pm.002}$	$0.904_{\pm.003}$	$0.981_{\pm.001}$	$0.968_{\pm.002}$	$0.439_{\pm.003}$	$0.287_{\pm.001}$	$\mathbf{0.781_{\pm.002}}$	$2.503_{\pm.003}$	$0.489_{\pm.001}$

Table 6: Query error of data synthesis. A lower value means a smaller query error. The privacy budget ϵ of HP synthesizers is set as ∞ (the top part), and the budget for DP synthesizers is set as 1 (the bottom part). The best result of each category is in bold.

	Adult	Shoppers	Phishing	Magic	Faults	Bean	Obesity	Robot	Abalone	News	Insurance	Wine
PGM	$0.056_{\pm.018}$	$0.044_{\pm.005}$	$\mathbf{0.009_{\pm.001}}$	$0.035_{\pm.004}$	$0.041_{\pm.003}$	$0.036_{\pm.003}$	$0.035_{\pm.007}$	$0.049_{\pm.005}$	$0.040_{\pm.004}$	$0.033_{\pm.005}$	$0.039_{\pm.004}$	$0.042_{\pm.005}$
PrivSyn	$0.009_{\pm.002}$	$0.011_{\pm.006}$	$0.011_{\pm.002}$	$0.011_{\pm.002}$	$0.027_{\pm.004}$	$0.034_{\pm.002}$	$0.027_{\pm.006}$	$0.029_{\pm.003}$	$0.014_{\pm.001}$	$\mathbf{0.010_{\pm.005}}$	$0.035_{\pm.007}$	$0.013_{\pm.002}$
TVAE	$0.025_{\pm.005}$	$0.034_{\pm.006}$	$0.018_{\pm.000}$	$0.014_{\pm.002}$	$0.026_{\pm.003}$	$0.019_{\pm.001}$	$0.027_{\pm.003}$	$0.020_{\pm.001}$	$0.016_{\pm.002}$	$0.030_{\pm.006}$	$0.050_{\pm.009}$	$0.028_{\pm.004}$
CTGAN	$0.015_{\pm.001}$	$0.017_{\pm.001}$	$0.051_{\pm.002}$	$0.037_{\pm.002}$	$0.047_{\pm.006}$	$0.030_{\pm.003}$	$0.037_{\pm.004}$	$0.033_{\pm.004}$	$0.036_{\pm.005}$	$0.018_{\pm.003}$	$0.055_{\pm.006}$	$0.016_{\pm.003}$
TabDDPM	$\mathbf{0.006_{\pm.001}}$	$\mathbf{0.008_{\pm.001}}$	$0.012_{\pm.001}$	$\mathbf{0.006_{\pm.001}}$	$\mathbf{0.021_{\pm.002}}$	$\mathbf{0.006_{\pm.001}}$	$\mathbf{0.017_{\pm.003}}$	$\mathbf{0.008_{\pm.001}}$	$\mathbf{0.011_{\pm.003}}$	$0.017_{\pm.002}$	$\mathbf{0.027_{\pm.007}}$	$\mathbf{0.010_{\pm.001}}$
GReaT	$0.014_{\pm.002}$	$0.014_{\pm.004}$	$0.049_{\pm.002}$	$0.029_{\pm.003}$	$0.028_{\pm.003}$	$0.011_{\pm.001}$	$0.030_{\pm.003}$	$0.014_{\pm.001}$	$0.019_{\pm.003}$	-	$0.041_{\pm.007}$	$0.013_{\pm.001}$
PGM ($\epsilon = 1$)	$0.071_{\pm.014}$	$0.052_{\pm.017}$	$\mathbf{0.012_{\pm.001}}$	$0.036_{\pm.003}$	$\mathbf{0.045_{\pm.002}}$	$\mathbf{0.037_{\pm.002}}$	$0.043_{\pm.005}$	$\mathbf{0.050_{\pm.008}}$	$\mathbf{0.041_{\pm.003}}$	$0.043_{\pm.004}$	$\mathbf{0.033_{\pm.004}}$	$\mathbf{0.045_{\pm.004}}$
PrivSyn ($\epsilon = 1$)	$\mathbf{0.010_{\pm.001}}$	$0.027_{\pm.007}$	$0.016_{\pm.002}$	$\mathbf{0.025_{\pm.003}}$	$0.100_{\pm.006}$	$0.048_{\pm.004}$	$0.060_{\pm.004}$	$0.095_{\pm.002}$	$0.051_{\pm.003}$	$\mathbf{0.027_{\pm.005}}$	$0.062_{\pm.007}$	$0.064_{\pm.008}$
PATE-GAN ($\epsilon = 1$)	$0.028_{\pm.004}$	$\mathbf{0.024_{\pm.002}}$	$0.117_{\pm.009}$	$0.058_{\pm.005}$	$0.088_{\pm.009}$	$0.191_{\pm.017}$	$\mathbf{0.037_{\pm.001}}$	$0.150_{\pm.023}$	$0.223_{\pm.032}$	$0.029_{\pm.003}$	$0.138_{\pm.011}$	$0.158_{\pm.013}$
TableDiffusion ($\epsilon = 1$)	$0.057_{\pm.006}$	$0.054_{\pm.005}$	$0.071_{\pm.007}$	$0.074_{\pm.011}$	$0.119_{\pm.009}$	$0.052_{\pm.007}$	$0.108_{\pm.010}$	$0.071_{\pm.012}$	$0.085_{\pm.010}$	$0.050_{\pm.003}$	$0.195_{\pm.011}$	$0.048_{\pm.006}$

drop compared to deep generative models. A closer examination of the fidelity of each type of attribute (detailed in technical report version) reveals that statistical methods generally achieve better fidelity performance on categorical attributes but fall short with numerical ones. This discrepancy can be attributed to the inherent design of statistical methods for discrete spaces, making them less effective for modeling numerical attributes.

Privacy Evaluation. Table 4 presents the privacy assessment for HP synthesizers. Contrary to the fidelity evaluation findings, CTGAN, which exhibited the lowest fidelity performance, provides the best privacy protections against membership disclosure. Statistical methods, such as PGM, also demonstrate commendable empirical privacy protections. Nonetheless, the unsatisfied results of strong synthesis algorithms like TabDDPM and GReaT reveal their vulnerability to membership disclosure.

Utility Evaluation. The utility of data synthesis is assessed by performing downstream tasks on the synthetic datasets and measuring with the proposed metrics (*i.e.,* machine learning affinity and query errors), as shown in Table 5 and Table 6. In the context of machine learning tasks, TabDDPM excels among HP synthesizers, which contributes to its class-conditional framework that learns label dependencies during its training process. However, this advantage diminishes when adding random noise to ensure privacy, where PGM takes the lead with its robust and superior performance. The outcomes for range (point) query tasks echo the results of fidelity evaluation, where TabDDPM shows superior performance in HP settings, and statistical methods (*i.e.,* PGM) can surpass other methods under DP constraints.

7.4 In-depth Analysis (RQ2)

In this section, we delve into the underlying reasons for the observed performances. Specifically, we employ the proposed fidelity metrics as tools for auditing the synthesizers' learning process and explore the impact of differing privacy budgets on DP synthesizers.

Why Does CTGAN Perform Poorly? CTGAN is widely recognized as a strong synthesizer in many studies [7, 33, 75]. Despite its popularity, our evaluation reveals that it produces the lowest quality synthetic data. This discrepancy prompts an investigation into the reasons behind CTGAN's apparent underachievement.

To address this, we scrutinize CTGAN's learning trajectory, particularly evaluating the fidelity across different marginal types throughout training. Our findings illustrated in Figure 7(a), reveal an unexpected stagnation in the improvement of Wasserstein distances for both numerical and categorical marginals throughout the entire training period. As such, the quality of synthetic data completely relies on data preprocessing, which employs a variational Gaussian mixture model to approximate numerical data and conditional sampling for categorical attributes. The closer the data distribution to Gaussianity, the better the fidelity. Yet, most tabular data display complexities beyond simple Gaussian distributions [24], leading to CTGAN's subpar performance. Furthermore, this phenomenon sheds light on CTGAN's strong empirical privacy

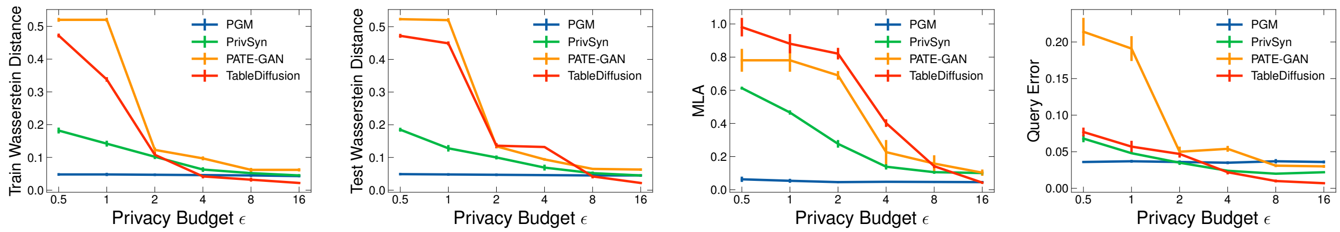
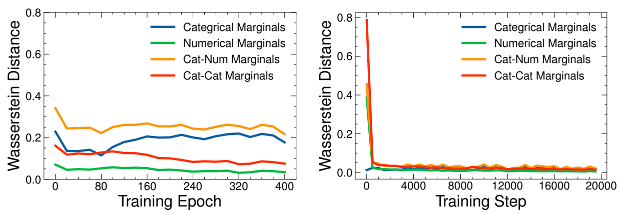


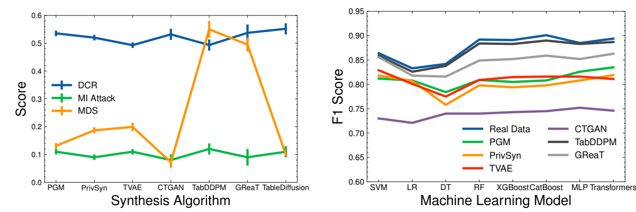
Figure 6: Impact of privacy budget ϵ on Bean dataset. The lower score indicates higher fidelity/utility. "Train (test) Wasserstein Distance" denotes the fidelity is evaluated on train (test) dataset $\mathcal{D}_{\text{train}}$ ($\mathcal{D}_{\text{test}}$). "MLA" is the proposed utility metric.



(a) CTGAN

(b) TabDDPM

Figure 7: Auditing the learning process of CTGAN and TabDDPM with fidelity metrics on the Bean dataset. The fidelity of numerical-numerical marginals is omitted since the Bean dataset only has one numerical attribute.



(a) Privacy Metric Comparison

(b) Machine Learning Performance

Figure 8: (a) Comparing different privacy metrics on Adult dataset. Existing metrics are uninformative and only MDS can accurately detect the privacy risks of different synthesizers. (b) Performance behaviors of HP synthesizers on Magic dataset. "LR" denotes Linear Regression, "DT" is Decision Tree, and "RF" means Random Forest.

protections. The nearly random (and unlearnable) nature of the output does not depend on the presence of any individual training data. This aspect, while beneficial for privacy, underscores the challenges in generating high-fidelity synthetic data using CTGAN.

Why Does TabDDPM Excel? A notable finding from our evaluations is TabDDPM's capability to achieve high fidelity. This contradicts with previous statements that deep generative models generally fail for tabular data synthesis [42, 62]. To better illustrate this, we utilized fidelity metrics to audit TabDDPM's learning process. From the results depicted in Figure 7(b) we observe a rapid decrease in the Wasserstein distance for all types of marginals, indicating the ability to model both numerical and categorical attributes. We contribute its success to the design of the model architecture. Diffusion models are proved to minimize the Wasserstein distance between synthetic and real data [34]. This represents a methodological superiority compared to other generative models, which primarily aim at reducing the Kullback–Leibler divergence. Moreover, integrating the multinomial diffusion process addresses the challenge of modeling discrete spaces for conventional diffusion models. However, TabDDPM faces serious privacy disclosure risks which are ignored by previous work [33], and directly applying DP (*i.e.,* TableDiffusion) would greatly impact the quality of synthetic data. Despite these challenges, we deem that diffusion-based methods represent a promising frontier for data synthesis.

Large Language Models Are Semantic-aware Synthesizers. We also notice that the recently emerged LLM-based synthesizer (*i.e.,* GReaT) also shows competitive performance, especially on datasets that consist of rich semantic attributes. For example, the Adult dataset comprises detailed personal information (*e.g.,* age, income, and relationship), and GReaT achieves the best performance on machine learning predictions among HP synthesizers. Nevertheless, this advantage diminishes in cases where the attributes' semantics are unclear. This is evident in the Phishing dataset, where all attributes' values have been preprocessed into integers and GReaT's performance is notably weaker. Given the rapid development of LLM and the inherent rich semantics of most tabular data, it may become a new paradigm for realistic data synthesis.

The Impact of Privacy Budget. We conduct experiments on different privacy budgets to further analyze the impact of differential privacy for data synthesis. Specifically, we vary ϵ from 0.5 to 16, and evaluate the fidelity and utility of synthetic data, as shown in Figure 6. It is observed that statistical methods (*e.g.,* PGM) demonstrate robust performance even with the small privacy budget (*e.g.,* $\epsilon = 0.5$), whereas deep generative models generally require a much larger privacy budget (*e.g.,* $\epsilon = 8$) to yield similar outcomes. This is reasonable since statistical methods only rely on the estimation of a small set of marginals. In contrast, deep generative models aim to directly capture the entire joint distribution, making them inherently more susceptible to random perturbations.

7.5 Effectiveness of proposed Metrics (RQ3)

Effectiveness of Membership Disclosure Score (MDS). Figure 8(a) shows the performance of different privacy metrics (*i.e.,* DCR [75], MI attack [61] and MDS, detailed implementations are

Table 7: Average performance improvements (%) on fidelity and utility when tuning with the proposed tuning objective.

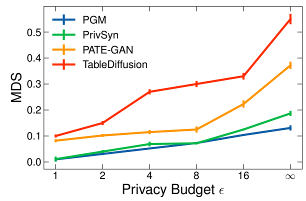
Models	Fidelity ↑		Utility ↑	
	$\mathcal{D}_{\text{Train}}$	$\mathcal{D}_{\text{Test}}$	MLA	Query Error
PGM	0.33	0.34	17.35	3.39
PrivSyn	1.60	2.92	12.08	1.12
TVAE	1.06	0.67	5.29	2.67
CTGAN	9.87	9.60	0.57	8.63
PATE-GAN	6.27	8.48	0.75	7.04
TabDDPM	13.62	13.65	13.67	11.95
TableDiffusion	11.34	10.95	8.32	7.86
GReaT	3.84	9.21	1.14	1.77

included in technical report version). To evaluate the privacy protection, we add a DP synthesizer (*i.e.,* TableDiffusion with $\epsilon = 1$) for reference. A good privacy metric should differentiate the privacy risks between TableDiffusion and its HP equivalent (*i.e.,* TabDDPM). However, the results of DCR indicate all synthesizers have the same level of privacy protection, and the attack accuracy of MI attacks is rather low for all synthesizers. On the contrary, MDS can accurately detect the privacy risks of different synthesizers. For example, it indicates the membership disclosure risks of TabDDPM and GReaT are remarkably high, and applying DP to the synthesizer (*i.e.,* TableDiffusion) can largely alleviate the disclosure risks.

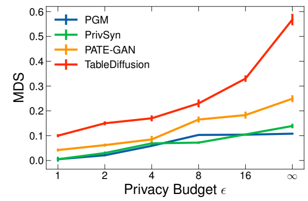
Effectiveness of Machine Learning Affinity. The performance of various data synthesizers fluctuates significantly across different machine learning models, as illustrated in Figure 8(b). Such variations in performance underscore the impact of the choice of evaluation models. For instance, while PrivSyn is ranked third when evaluated using linear regression, it falls to fifth when assessed with decision trees. The proposed MLA addresses this challenge by quantifying relative discrepancies in performance trends, thereby facilitating a more nuanced and consistent utility evaluation.

Impact of Model Tuning Phase. We showcase the average performance improvements by incorporating the tuning phase in Section 6.1. Generally, the tuning objective can consistently improve synthetic data's fidelity and utility for all synthesizers. For statistical methods, the improvement in utility surpasses those in fidelity, while deep generative models exhibit the reverse trend. In addition, the tuning phrase is exceptionally beneficial for TabDDPM, yielding significant improvements in both fidelity and utility.

Empirical Privacy Analysis with MDS. We further investigate the empirical privacy protections of DP synthesizers by applying the proposed privacy metric (*i.e.,* MDS). Specifically, we set the privacy budget ϵ from 1 to infinity and calculate the corresponding MDS. The results on the Adult and Phishing dataset are shown in Figure 9. It is observed that a larger privacy budget would lead to higher membership disclosure risks (thus larger MDS). Interestingly, we find that statistical methods can provide reasonable empirical privacy protections, even without DP guarantees (*i.e.,* $\epsilon = \infty$). On the contrary, deep generative models like TableDiffusion need a smaller privacy budget for good empirical privacy protection. This may be because statistical methods are inherently more private



(a) Adult Dataset



(b) Phishing Dataset

Figure 9: Empirical privacy evaluation with membership disclosure score (MDS) on DP synthesizers.

since they only rely on statistics for synthesis. We leave it as an open question for future research.

8 DISCUSSION AND KEY TAKEAWAYS

Data synthesis has been advocated as an important approach for utilizing data while protecting data privacy. Despite the plethora of available data synthesizers, a comprehensive understanding of their strengths and weaknesses remains elusive, due to the limitations of existing evaluation metrics. In this paper, we examine and critique existing metrics, and introduce a systematic framework as well as a new suite of evaluation criteria for assessing data synthesizers. We also provide a unified tuning objective that can consistently improve the performance of synthesizers and ensure that evaluation results are less affected by accidental choices of hyperparameters.

Our results identify several guidelines for practitioners seeking the best performance for their task and facing the daunting task of selecting and configuring appropriate synthesizers.

- *Model tuning is indispensable.* Tuning hyperparameters can significantly improve synthetic data quality, especially for deep generative models.
- *Statistical methods should be preferred for applications where privacy is paramount.* PRM and PrivSyn achieve the best fidelity among DP synthesizers, and they also offer good empirical privacy protection even in HP settings.
- *Diffusion models provide the best fidelity and utility.* Practitioners are suggested to use diffusion models (*e.g.,* TabDDPM) for tabular synthesis when the quality of synthetic data is the priority over privacy due to their impressive ability to generate highly authentic synthetic data.
- *Deep generative models can be tailored for specific tasks.* The flexible design spaces of deep generative models make them suitable for scenarios where the applications of the synthetic data are known in advance (*e.g.,* machine learning prediction). In addition, the LLM-based synthesizer, GReaT, is particularly effective at preserving semantic information in synthetic data.

Our systematic assessment leads us to conclude that recently emerged generative models achieve impressive performance on tabular data synthesis and offer new directions in this field. At the same time, several critical challenges are also revealed (*e.g.,* privacy issues of diffusion models, performance gaps between DP and HP synthesizers, *etc.*). Our evaluation metrics and framework play a crucial role in elucidating the advancements in data synthesis as well as serve as a benchmark for future works.

REFERENCES

[1] Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. 2019. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*. 2623–2631.

[2] 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*. 290–306.

[3] Christian Arnold and Marcel Neunhoeffer. 2020. Really Useful Synthetic Data–A Framework to Evaluate the Quality of Differentially Private Synthetic Data. *arXiv preprint arXiv:2004.07740* (2020).

[4] AWS. 2022. How to evaluate the quality of the synthetic data: measuring from the perspective of fidelity, utility, and privacy. https://aws.amazon.com/blogs/machine-learning/how-to-evaluate-the-quality-of-the-synthetic-data-measuring-from-the-perspective-of-fidelity-utility-and-privacy/.

[5] Gary Benedetto, Martha Stinson, and John M Abowd. 2018. The creation and use of the SIPP Synthetic. (2018).

[6] Nicolas Bonneel, Julien Rabin, Gabriel Peyré, and Hanspeter Pfister. 2015. Sliced and radon wasserstein barycenters of measures. *Journal of Mathematical Imaging and Vision* 51 (2015), 22–45.

[7] Vadim Borisov, Kathrin Sessler, Tobias Leemann, Martin Pawelczyk, and Gjergji Kasneci. 2023. Language Models are Realistic Tabular Data Generators. In *International Conference on Learning Representations*.

[8] Olivier Bousquet and André Elisseeff. 2002. Stability and generalization. *The Journal of Machine Learning Research* 2 (2002), 499–526.

[9] Claire McKay Bowen and Joshua Snoke. 2019. Comparative study of differentially private synthetic data algorithms from the NIST PSCR differential privacy synthetic data challenge. *arXiv preprint arXiv:1911.12704* (2019).

[10] Nicholas Carlini, Steve Chien, Milad Nasr, Shuang Song, Andreas Terzis, and Florian Tramer. 2022. Membership inference attacks from first principles. In *2022 IEEE Symposium on Security and Privacy (SP)*. 1897–1914.

[11] C Chatfied and AJ Collins. 2013. *Introduction to Multivariate Analysis*. Springer.

[12] Tianqi Chen and Carlos Guestrin. 2016. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*. 785–794.

[13] 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* (2022), 1–11.

[14] Marco Cuturi. 2013. Sinkhorn distances: Lightspeed computation of optimal transport. In *Proceedings of the 26 th International Conference on Neural Information Processing Systems*.

[15] Steven Diamond and Stephen Boyd. 2016. CVXPY: A Python-embedded modeling language for convex optimization. *Journal of Machine Learning Research* 17, 83 (2016), 1–5.

[16] Antreas Dionysiou and Elias Athanasopoulos. 2023. SoK: Membership Inference is Harder Than Previously Thought. *Proceedings on Privacy Enhancing Technologies* 3 (2023), 286–306.

[17] Cynthia Dwork. 2006. Differential Privacy. In *Automata, Languages and Programming*. 1–12.

[18] Cynthia Dwork, Aaron Roth, et al. 2014. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science* 9, 3–4 (2014), 211–407.

[19] Vitaly Feldman. 2020. Does learning require memorization? a short tale about a long tail. In *Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing*. 954–959.

[20] Rémi Flamary, Nicolas Courty, Alexandre Gramfort, Mokhtar Z Alaya, Aurélie Boisbunon, Stanislas Chambon, Laetitia Chapel, Adrien Corenflos, Kilian Fatras, Nemo Fournier, et al. 2021. Pot: Python optimal transport. *The Journal of Machine Learning Research* 22, 1 (2021), 3571–3578.

[21] Georgi Ganev and Emiliano De Cristofaro. 2023. On the Inadequacy of Similarity-based Privacy Metrics: Reconstruction Attacks against" Truly Anonymous Synthetic Data". *arXiv preprint arXiv:2312.05114* (2023).

[22] Georgi Ganev, Bristena Oprisanu, and Emiliano De Cristofaro. 2022. Robin hood and matthew effects: Differential privacy has disparate impact on synthetic data. In *International Conference on Machine Learning*. 6944–6959.

[23] Matteo Giomi, Franziska Boenisch, Christoph Wehmeyer, and Borbála Tasnádi. 2023. A Unified Framework for Quantifying Privacy Risk in Synthetic Data. *Proceedings on Privacy Enhancing Technologies* 2 (2023), 312–328.

[24] Yury Gorishniy, Ivan Rubachev, Valentin Khrulkov, and Artem Babenko. 2021. Revisiting deep learning models for tabular data. In *Proceedings of the 35th International Conference on Neural Information Processing Systems*. 18932–18943.

[25] Gretel. 2023. Build smarter with the right data. fast. safe. accurate. https://gretel.ai/synthetics.

[26] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. 2017. Improved training of wasserstein gans. In *Proceedings of the 31th International Conference on Neural Information Processing Systems*.

[27] Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*. 6840–6851.

[28] Yuzheng Hu, Fan Wu, Qinbin Li, Yunhui Long, Gonzalo Garrido, Chang Ge, Bolin Ding, David Forsyth, Bo Li, and Dawn Song. 2024. SoK: Privacy-Preserving Data Synthesis. In *2024 IEEE Symposium on Security and Privacy (SP)*. 2–2.

[29] Jihyeon Hyeong, Jayoung Kim, Noseong Park, and Sushil Jajodia. 2022. An empirical study on the membership inference attack against tabular data synthesis models. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*. 4064–4068.

[30] James Jordon, Lukasz Szpruch, Florimond Houssiau, Mirko Bottarelli, Giovanni Cherubin, Carsten Maple, Samuel N. Cohen, and Adrian Weller. 2021. Synthetic Data – what, why and how? *arXiv preprint arXiv:2205.03257* (2021).

[31] 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*.

[32] Jayoung Kim, Chaejeong Lee, and Noseong Park. 2022. STaSy: Score-based Tabular data Synthesis. In *International Conference on Learning Representations*.

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

[34] Dohyun Kwon, Ying Fan, and Kangwook Lee. 2022. Score-based Generative Modeling Secretly Minimizes the Wasserstein Distance. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*.

[35] Chaejeong Lee, Jayoung Kim, and Noseong Park. 2023. CoDi: Co-Evolving Contrastive Diffusion Models for Mixed-Type Tabular Synthesis. In *International Conference on Machine Learning*. 18940–18956.

[36] Ninghui Li, Min Lyu, Dong Su, and Weining Yang. 2017. *Differential privacy: From theory to practice*.

[37] Ninghui Li, Wahbeh Qardaji, Dong Su, Yi Wu, and Weining Yang. 2013. Membership privacy: A unifying framework for privacy definitions. In *Proceedings of the 2013 ACM SIGSAC conference on Computer & communications security*. 889–900.

[38] Zitao Li, Trung Dang, Tianhao Wang, and Ninghui Li. 2021. MGD: a utility metric for private data publication. In *Proceedings of the 8th International Conference on Networking, Systems and Security*. 106–119.

[39] Raphael Gontijo Lopes, Sylvia J Smullin, Ekin Dogus Cubuk, and Ethan Dyer. 2021. Tradeoffs in Data Augmentation: An Empirical Study.. In *International Conference on Learning Representations*.

[40] Pei-Hsuan Lu, Pang-Chieh Wang, and Chia-Mu Yu. 2019. Empirical evaluation on synthetic data generation with generative adversarial network. In *Proceedings of the 9th International Conference on Web Intelligence, Mining and Semantics*. 1–6.

[41] Ryan McKenna, Gerome Miklau, and Daniel Sheldon. 2021. Winning the NIST Contest: A scalable and general approach to differentially private synthetic data. *arXiv preprint arXiv:2108.04978* (2021).

[42] Ryan McKenna, Brett Mullins, Daniel Sheldon, and Gerome Miklau. 2022. AIM: an adaptive and iterative mechanism for differentially private synthetic data. *Proceedings of the VLDB Endowment* 15, 11 (2022), 2599–2612.

[43] Ryan McKenna, Daniel Sheldon, and Gerome Miklau. 2019. Graphical-model based estimation and inference for differential privacy. In *International Conference on Machine Learning*. 4435–4444.

[44] Ofer Mendelevitch and Michael D Lesh. 2021. Fidelity and privacy of synthetic medical data. *arXiv preprint arXiv:2101.08658* (2021).

[45] NIST. 2018. Differential privacy synthetic data challenge. https://www.nist.gov/ctl/pscr/open-innovation-prize-challenges/past-prize-challenges/2018-differential-privacy-synthetic.

[46] NIST. 2020. Differential privacy temporal map challenge. https://www.nist.gov/ctl/pscr/open-innovation-prize-challenges/past-prize-challenges/2020-differential-privacy-temporal.

[47] OECD. 2023. Emerging privacy-enhancing technologies. https://www.oecd-ilibrary.org/content/paper/bf121be4-en.

[48] Bristena Oprisanu, Georgi Ganev, and Emiliano De Cristofaro. 2022. On Utility and Privacy in Synthetic Genomic Data. In *29th Annual Network and Distributed System Security Symposium*.

[49] Yidong Ouyang, Liyan Xie, Chongxuan Li, and Guang Cheng. 2023. Missdiff: Training diffusion models on tabular data with missing values. *arXiv preprint arXiv:2307.00467* (2023).

[50] Nicolas Papernot, Shuang Song, Ilya Mironov, Ananth Raghunathan, Kunal Talwar, and Ulfar Erlingsson. 2018. Scalable Private Learning with PATE. In *International Conference on Learning Representations*.

[51] 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, 10 (2018).

[52] Neha Patki, Roy Wedge, and Kalyan Veeramachaneni. 2016. The Synthetic data vault. In *IEEE International Conference on Data Science and Advanced Analytics (DSAA)*. 399–410.

[53] Gabriel Peyré, Marco Cuturi, et al. 2019. Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning* 11, 5-6 (2019), 355–607.

[54] Therese D Pigott. 2001. A review of methods for missing data. *Educational research and evaluation* 7, 4 (2001), 353–383.
[55] Liudmila Prokhorenkova, Gleb Gusev, Aleksandr Vorobev, Anna Veronika Dorogush, and Andrey Gulin. 2018. CatBoost: unbiased boosting with categorical features. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*.
[56] Zhaozhi Qian, Bogdan-Constantin Cebere, and Mihaela van der Schaar. 2023. Synthcity: facilitating innovative use cases of synthetic data in different data modalities. *arXiv preprint arXiv:2301.07573* (2023).
[57] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog* 1, 8 (2019), 9.
[58] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. 2017. Membership inference attacks against machine learning models. In *2017 IEEE symposium on security and privacy (SP)*. 3–18.
[59] Alon Shoshan, Nadav Bhonker, Igor Kviatkovsky, Matan Fintz, and Gérard Medioni. 2023. Synthetic data for model selection. In *International Conference on Machine Learning*. 31633–31656.
[60] 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.
[61] Theresa Stadler, Bristena Oprisanu, and Carmela Troncoso. 2022. Synthetic data–anonymisation groundhog day. In *31st USENIX Security Symposium (USENIX Security 22)*. 1451–1468.
[62] Yuchao Tao, Ryan McKenna, Michael Hay, Ashwin Machanavajjhala, and Gerome Miklau. 2021. Benchmarking differentially private synthetic data generation algorithms. *arXiv preprint arXiv:2112.09238* (2021).
[63] L. Theis, A. van den Oord, and M. Bethge. 2016. A note on the evaluation of generative models. In *ICLR*.
[64] Gianluca Truda. 2023. Generating tabular datasets under differential privacy. *arXiv preprint arXiv:2308.14784* (2023).
[65] UN. 2023. The United Nations Guide on privacy-enhancing technologies for official statistics. https://unstats.un.org/bigdata/task-teams/privacy/guide/2023_UN%20PET%20Guide.pdf.
[66] Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of machine learning research* 9, 11 (2008).
[67] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31th International Conference on Neural Information Processing Systems*.
[68] Wikipedia contributors. 2024. https://en.wikipedia.org/wiki/Pearson_correlation_coefficient#Mathematical_properties
[69] Lei Xu, Maria Skoularidou, Alfredo Cuesta-Infante, and Kalyan Veeramachaneni. 2019. Modeling tabular data using conditional GAN. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*. 7335–7345.
[70] Andrew Yale, Saloni Dash, Ritik Dutta, Isabelle Guyon, Adrien Pavao, and Kristin P Bennett. 2019. Assessing privacy and quality of synthetic health data. In *Proceedings of the Conference on Artificial Intelligence for Data Discovery and Reuse*. 1–4.
[71] Chiyuan Zhang, Daphne Ippolito, Katherine Lee, Matthew Jagielski, Florian Tramèr, and Nicholas Carlini. 2023. Counterfactual memorization in neural language models. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*.
[72] Hengrui Zhang, Jiani Zhang, Balasubramaniam Srinivasan, Zhengyuan Shen, Xiao Qin, Christos Faloutsos, Huzefa Rangwala, and George Karypis. 2023. Mixed-Type Tabular Data Synthesis with Score-based Diffusion in Latent Space. *arXiv preprint arXiv:2310.09656* (2023).
[73] 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.
[74] Zhikun Zhang, Tianhao Wang, Ninghui Li, Jean Honorio, Michael Backes, Shibo He, Jiming Chen, and Yang Zhang. 2021. PrivSyn: Differentially Private Data Synthesis. In *30th USENIX Security Symposium (USENIX Security 21)*. 929–946.
[75] Zilong Zhao, Aditya Kunar, Robert Birke, and Lydia Y. Chen. 2021. CTAB-GAN: Effective Table Data Synthesizing. In *Proceedings of The 13th Asian Conference on Machine Learning*. 97–112.

APPENDIX

A HYPERPARAMETERS SEARCH SPACES OF SYNTHESIS ALGORITHMS

In this paper, we tune and use the following synthesizers:

- PGM [43] in Table A1.
- PrivSyn [74] in Table A2.
- TVAE [69] in Table A4.
- CTGAN [69] in Table A3.
- TabDDPM [33] in Table A5.
- GReaT [7] in Table A6.
- PATE-GAN [31] in Table A7.
- TableDiffusion [64] in Table A8.

We set the privacy budget to infinity ($\epsilon = 100,000,000$) when using DP synthesizers for HP evaluations.

Table A1: PGM [43] hyperparameters search space.

Parameter	Distribution
Number of two-way marginals	Int[10, 50]
Number of three-way marginals	Int[5, 20]
Number of bins	Int[5, 20]
Maximum number of iterations	Int[3000, 5000]
Number of tuning trials	50

Table A2: PrivSyn [74] hyperparameters search space.

Parameter	Distribution
Number of bins	Int[5, 20]
Maximum number of iterations	Int[10, 100]
Number of tuning trials	50

Table A3: TVAE [69] hyperparameters search space.

Parameter	Distribution
Number of epochs	Int[100, 500]
Batch size	Int[500, 5000]
Loss factor	Float[1, 5]
Embedding dimension	Int[128, 512]
Compression dimension	Int[128, 512]
Decompression dimension	Int[128, 512]
L_2 regularization	LogUniform[$1e$-6, $1e$-3]
Number of tuning trials	50

B DATASETS DESCRIPTIONS

We use 12 real-world datasets for evaluations, varying in size, nature, and distribution. These datasets are all publicly available:

- **Adult** is to predict whether income exceeds 50K/yr based on census data. (https://archive.ics.uci.edu/dataset/2/adult)

Table A4: CTGAN [69] hyperparameters search space.

Parameter	Distribution
Number of epochs	Int[100, 500]
Batch size	Int[500, 5000]
Embedding dimension	Int[128, 512]
Generator dimension	Int[128, 512]
Discriminator dimension	Int[128, 512]
Learning rate of generator	LogUniform[1*e*-5, 1*e*-3]
Learning rate of discriminator	LogUniform[1*e*-5, 1*e*-3]
Number of tuning trials	50

Table A5: TabDDPM [33] hyperparameters search space.

Parameter	Distribution
Number of layers	Int[2, 8]
Embedding dimension	Int[128, 512]
Number of diffusion timesteps	Int[100, 10000]
Number of training iterations	Int[5000, 30000]
Learning rate of discriminator	LogUniform[1*e*-5, 3*e*-3]
Number of tuning trials	50

Table A6: GReaT [7] hyperparameters search space.

Parameter	Distribution
Temperature	Float[0.6, 0.9]
Number of fine-tuning epochs	Int[100, 300]
Number of training iterations	Int[5000, 30000]
Batch size	Int[8, 32]
Number of tuning trials	20

Table A7: PATE-GAN [31] hyperparameters search space.

Parameter	Distribution
Number of teachers	Int[5, 20]
Number of generator layers	Int[1, 3]
Number of discriminator layers	Int[1, 3]
Generator dimension	Int[50, 200]
Discriminator dimension	Int[50, 200]
Number of iterations	Int[1000, 5000]
Learning rate	LogUniform[1*e*-5, 1*e*-3]
Number of tuning trials	50

Table A8: TableDiffusion [64] hyperparameters search space.

Parameter	Distribution
Number of layers	Int[1, 6]
Number of diffusion timesteps	Int[3, 20]
Number of epochs	Int[5, 20]
Batch size	Int[128, 1024]
Noise prediction	{True, False}
Learning rate	LogUniform[1*e*-4, 1*e*-2]
Number of tuning trials	50

- **Shoppers** is to analyze the intention of online shoppers. (https://archive.ics.uci.edu/dataset/468/online+shoppers+purchasing+intention+dataset)
- **Phishing** is to predict if a webpage is a phishing site. The dataset consists of important features for predicting phishing sites, including information about webpage transactions. (https://archive.ics.uci.edu/dataset/468/online+shoppers+purchasing+intention+dataset)
- **Magic** is a binary classification dataset that simulates the registration of high-energy gamma particles in the atmospheric telescope.
- **Faults** is for fault detection in the steel manufacturing process. (https://archive.ics.uci.edu/dataset/198/steel+plates+faults)
- **Bean** is to predict the type of dray bean based on form, shape, type, and structure. (https://archive.ics.uci.edu/dataset/602/dry+bean+dataset)
- **Obesity** is to estimate the obesity level based on eating habits and physical condition of individuals from Mexico, Peru, and Columbia. (https://archive.ics.uci.edu/dataset/544/estimation+of+obesity+levels+based+on+eating+habits+and+physical+condition)
- **Robot** is a multi-class classification dataset collected as the robot moves around the room, following the wall using ultrasound sensors. (https://archive.ics.uci.edu/dataset/194/wall+following+robot+navigation+data)
- **Abalone** is to predict the age of abalone from physical measurements. (https://archive.ics.uci.edu/dataset/1/abalone)
- **News** is to predict the number of shares in social networks (popularity). (https://archive.ics.uci.edu/dataset/332/online+news+popularity).
- **Insurance** is for prediction on the yearly medical cover cost. The dataset contains a person's medical information. (https://www.kaggle.com/datasets/tejashvi14/medical-insurance-premium-prediction)
- **Wine** is to model wine quality based on physicochemical tests. (https://archive.ics.uci.edu/dataset/186/wine+quality)

C HYPERPARAMETERS SEARCH SPACES OF MACHINE LEARNING MODELS

We tune and use the following machine learning models to evaluate the utility of synthesizers:

- SVM in Table C9.
- Logistic regression (Ridge regression) in Table C10.
- Decision tree in Table C11.
- Random forest in Table C12.
- Multilayer perception (MLP) in Table C13.
- XGBoost [12]in Table C14.
- CatBoost [55] in Table C15.
- Transformers [24] in Table C16.

Table C9: SVM hyperparameters search space.

Parameter	Distribution
Regularization parameter	LogUniform[1e-5, 1e-1]
Kernel	{linear, poly, rbf, sigmoid}
Number of tuning trials	50

Table C10: Logistic regression (Ridge regression) hyperparameters search space.

Parameter	Distribution
Regularization	Float[0, 10]
Maximum number of iterations	Int[100, 1000]
Fit intercept	{True, False}
Number of tuning trials	50

Table C11: Decision tree hyperparameters search space.

Parameter	Distribution
Maximum depth	Int[4, 64]
Minimum samples split	Int[2, 8]
Minimum samples leaf	Int[1, 8]
Number of tuning trials	50

Table C12: Random forest hyperparameters search space.

Parameter	Distribution
Number of estimators	Int[10, 200]
Maximum depth	Int[4, 64]
Minimum samples split	Int[2, 8]
Minimum samples leaf	Int[1, 8]
Number of tuning trials	50

D IMPLEMENTATION OF EXISTING PRIVACY METRICS

We implement two existing private metrics (*i.e.,* DCR [75] and MI attack [61]) for performance comparison. For DCR, we first normalize all attributes of synthetic data and real data, and calculate the distance distribution from each synthetic data point to its nearest real one, using the 5th percentile as the privacy score. For the MI attack, we follow [61] to conduct feature extraction from synthetic data for shadow model training. We use the most widely-used metric: true-positive rate (TPR) at 10% false-positive rates (FPR) [10], *i.e.,* , TPR@10% as the attack score for comparison.

E COMPLETE EXPERIMENTAL RESULTS

E.1 Complete Fidelity Results

Here we include the remaining fidelity results in our evaluation. The overall fidelity evaluation on test data is demonstrated in Table E17.

Table C13: MLP hyperparameters search space.

Parameter	Distribution
Maximum number of iterations	Float[50, 200]
Regularization	LogUniform[1e-5, 1e-1]
Number of tuning trials	50

Table C14: XGBoost hyperparameters search space.

Parameter	Distribution
Step size shrinkage	Float[0.01, 0.2]
Minimum sum of instance weight	Int[1, 10]
Maximum depth	Int[3, 20]
Minimum loss reduction	Float[0, 1]
Number of tuning trials	50

Table C15: CatBoost hyperparameters search space from [24].

Parameter	Distribution
Max depth	UniformInt[3, 10]
Learning rate	LogUniform[1e-5, 1]
Bagging temperature	Uniform[0, 1]
L_2 leaf reg	LogUniform[1, 10]
Leaf estimation iterations	UniformInt[1, 10]
Number of tuning trials	50

Table C16: Transformer hyperparameters space from [24].

Parameter	Distribution
Number of layers	UniformInt[1, 8]
Layer size	Int{64, 128, 256, 512, 1024}
Dropout	{0, Uniform[0, 0.5]}
Learning rate	LogUniform[1e-5, 1e-2]
Weight decay	{0, LogUniform[1e-6, 1e-3]}
Number of tuning trials	50

In addition, the complete fidelity results consist of the Wasserstein distance for two one-way marginals (categorical and numerical distributions) and three two-way marginals (categorical-categorical distribution, numerical-numerical distribution, categorical-numerical distribution).

For fidelity on test data $\mathcal{D}_{\text{train}}$, the detailed Wasserstein distances for one-way marginals are shown in Table E18 and Table E19. The Wasserstein distances for two-way marginals are shown in Table E20, Table E21 and Table E22.

For fidelity on test data $\mathcal{D}_{\text{test}}$, the detailed Wasserstein distances for one-way marginals are shown in Table E23 and Table E24. The Wasserstein distances for two-way marginals are shown in Table E25, Table E26 and Table E27.

Table E17: Fidelity evaluation (*i.e.,* Wasserstein distance) of data synthesis algorithms on test data $\mathcal{D}_{\text{test}}$. The privacy budget ϵ of HP synthesizers is ∞ (the top part), and the budget for DP synthesizers is 1 (the middle part). `HALF` and `HISTOGRAM` are the baselines that serve as the empirical upper bound and lower bound of the fidelity for HP synthesizers. The low score indicates the synthesizer can generate high-quality synthetic data that preserves the marginals of real ones. The best result is in bold.

	Adult	Shoppers	Phishing	Magic	Faults	Bean	Obesity	Robot	Abalone	News	Insurance	Wine
PGM	$0.172_{\pm.004}$	$0.098_{\pm.002}$	$0.026_{\pm.001}$	$0.039_{\pm.002}$	$0.089_{\pm.006}$	$0.044_{\pm.003}$	$0.062_{\pm.003}$	$0.055_{\pm.003}$	$0.062_{\pm.008}$	$0.050_{\pm.004}$	$0.083_{\pm.009}$	$0.075_{\pm.002}$
PrivSyn	$0.025_{\pm.001}$	$0.041_{\pm.003}$	**$0.017_{\pm.002}$**	$0.015_{\pm.002}$	$0.079_{\pm.007}$	$0.037_{\pm.003}$	$0.053_{\pm.005}$	$0.054_{\pm.004}$	**$0.032_{\pm.005}$**	**$0.018_{\pm.001}$**	$0.074_{\pm.006}$	$0.022_{\pm.001}$
TVAE	$0.086_{\pm.002}$	$0.154_{\pm.002}$	$0.028_{\pm.002}$	$0.020_{\pm.003}$	$0.081_{\pm.016}$	$0.050_{\pm.004}$	$0.059_{\pm.003}$	$0.059_{\pm.007}$	$0.046_{\pm.005}$	$0.079_{\pm.001}$	$0.118_{\pm.009}$	$0.045_{\pm.001}$
CTGAN	$0.061_{\pm.003}$	$0.061_{\pm.002}$	$0.069_{\pm.001}$	$0.150_{\pm.004}$	$0.133_{\pm.007}$	$0.139_{\pm.005}$	$0.085_{\pm.004}$	$0.109_{\pm.009}$	$0.066_{\pm.005}$	$0.040_{\pm.001}$	$0.116_{\pm.008}$	$0.034_{\pm.001}$
TabDDPM	**$0.021_{\pm.001}$**	**$0.031_{\pm.001}$**	$0.019_{\pm.001}$	**$0.012_{\pm.002}$**	**$0.058_{\pm.008}$**	**$0.016_{\pm.003}$**	**$0.043_{\pm.003}$**	**$0.028_{\pm.004}$**	$0.034_{\pm.010}$	$0.032_{\pm.001}$	**$0.070_{\pm.009}$**	**$0.017_{\pm.001}$**
GReaT	$0.052_{\pm.002}$	$0.056_{\pm.004}$	$0.072_{\pm.002}$	$0.039_{\pm.003}$	$0.063_{\pm.007}$	$0.021_{\pm.004}$	$0.062_{\pm.008}$	$0.058_{\pm.006}$	$0.037_{\pm.004}$	-	$0.107_{\pm.010}$	$0.024_{\pm.001}$
PGM (ϵ = 1)	$0.179_{\pm.004}$	$0.103_{\pm.001}$	**$0.028_{\pm.001}$**	$0.042_{\pm.004}$	**$0.112_{\pm.005}$**	**$0.048_{\pm.003}$**	**$0.075_{\pm.004}$**	**$0.072_{\pm.007}$**	**$0.080_{\pm.010}$**	$0.051_{\pm.002}$	**$0.093_{\pm.006}$**	**$0.075_{\pm.001}$**
PrivSyn (ϵ = 1)	**$0.049_{\pm.002}$**	**$0.084_{\pm.002}$**	$0.030_{\pm.003}$	**$0.031_{\pm.003}$**	$0.236_{\pm.017}$	$0.128_{\pm.010}$	$0.154_{\pm.013}$	$0.177_{\pm.011}$	$0.111_{\pm.011}$	**$0.044_{\pm.001}$**	$0.152_{\pm.011}$	$0.130_{\pm.005}$
PATE-GAN (ϵ = 1)	$0.139_{\pm.002}$	$0.171_{\pm.002}$	$0.173_{\pm.002}$	$0.155_{\pm.005}$	$0.215_{\pm.004}$	$0.523_{\pm.004}$	$0.089_{\pm.004}$	$0.478_{\pm.007}$	$0.353_{\pm.009}$	$0.061_{\pm.002}$	$0.386_{\pm.011}$	$0.250_{\pm.003}$
TableDiffusion (ϵ = 1)	$0.179_{\pm.002}$	$0.210_{\pm.002}$	$0.121_{\pm.002}$	$0.132_{\pm.005}$	$0.390_{\pm.004}$	$0.149_{\pm.004}$	$0.338_{\pm.005}$	$0.203_{\pm.002}$	$0.226_{\pm.007}$	$0.128_{\pm.001}$	$0.366_{\pm.008}$	$0.098_{\pm.001}$
`HALF`	$0.022_{\pm.002}$	$0.023_{\pm.002}$	$0.016_{\pm.003}$	$0.011_{\pm.003}$	$0.042_{\pm.005}$	$0.015_{\pm.003}$	$0.041_{\pm.006}$	$0.023_{\pm.005}$	$0.028_{\pm.007}$	$0.010_{\pm.002}$	$0.060_{\pm.007}$	$0.014_{\pm.001}$
`HISTOGRAM`	$0.199_{\pm.017}$	$0.101_{\pm.001}$	$0.030_{\pm.001}$	$0.048_{\pm.002}$	$0.113_{\pm.006}$	$0.080_{\pm.003}$	$0.066_{\pm.002}$	$0.065_{\pm.002}$	$0.094_{\pm.009}$	$0.059_{\pm.006}$	$0.081_{\pm.004}$	$0.076_{\pm.001}$

Table E18: Fidelity evaluation of one-way categorical marginals on train data $\mathcal{D}_{\text{train}}$. The best result of each category is in bold.

	Adult	Shoppers	Phishing	Magic	Faults	Bean	Obesity	Robot	Abalone	News	Insurance	Wine
PGM	$0.015_{\pm.002}$	$0.012_{\pm.001}$	**$0.005_{\pm.001}$**	**$0.004_{\pm.003}$**	**$0.013_{\pm.002}$**	**$0.011_{\pm.003}$**	**$0.007_{\pm.001}$**	**$0.008_{\pm.004}$**	**$0.006_{\pm.004}$**	**$0.005_{\pm.000}$**	$0.018_{\pm.007}$	-
PrivSyn	**$0.014_{\pm.002}$**	**$0.011_{\pm.001}$**	$0.006_{\pm.001}$	$0.007_{\pm.005}$	$0.028_{\pm.006}$	$0.016_{\pm.005}$	$0.014_{\pm.002}$	$0.048_{\pm.026}$	$0.018_{\pm.007}$	$0.005_{\pm.001}$	**$0.016_{\pm.002}$**	-
TVAE	$0.089_{\pm.004}$	$0.193_{\pm.001}$	$0.018_{\pm.001}$	$0.012_{\pm.006}$	$0.050_{\pm.010}$	$0.055_{\pm.010}$	$0.040_{\pm.005}$	$0.058_{\pm.000}$	$0.055_{\pm.006}$	$0.084_{\pm.001}$	$0.072_{\pm.009}$	-
CTGAN	$0.067_{\pm.002}$	$0.067_{\pm.002}$	$0.048_{\pm.002}$	$0.265_{\pm.011}$	$0.121_{\pm.006}$	$0.207_{\pm.008}$	$0.061_{\pm.003}$	$0.136_{\pm.006}$	$0.015_{\pm.006}$	$0.036_{\pm.002}$	$0.042_{\pm.003}$	-
TabDDPM	$0.017_{\pm.001}$	$0.014_{\pm.002}$	$0.011_{\pm.001}$	$0.009_{\pm.005}$	$0.016_{\pm.003}$	$0.017_{\pm.003}$	$0.010_{\pm.001}$	$0.009_{\pm.003}$	$0.013_{\pm.006}$	$0.006_{\pm.001}$	$0.018_{\pm.006}$	-
GReaT	$0.056_{\pm.003}$	$0.044_{\pm.002}$	$0.058_{\pm.001}$	$0.005_{\pm.004}$	$0.034_{\pm.008}$	$0.016_{\pm.001}$	$0.038_{\pm.004}$	$0.054_{\pm.004}$	$0.018_{\pm.007}$	-	$0.108_{\pm.006}$	-
PGM (ϵ = 1)	**$0.020_{\pm.003}$**	**$0.019_{\pm.002}$**	**$0.009_{\pm.001}$**	**$0.006_{\pm.003}$**	**$0.039_{\pm.006}$**	**$0.021_{\pm.007}$**	**$0.026_{\pm.002}$**	**$0.030_{\pm.003}$**	**$0.016_{\pm.005}$**	**$0.006_{\pm.001}$**	**$0.032_{\pm.004}$**	-
PrivSyn (ϵ = 1)	$0.028_{\pm.002}$	$0.041_{\pm.006}$	$0.020_{\pm.002}$	$0.041_{\pm.005}$	$0.095_{\pm.009}$	$0.105_{\pm.012}$	$0.113_{\pm.010}$	$0.069_{\pm.035}$	$0.047_{\pm.014}$	$0.029_{\pm.002}$	$0.059_{\pm.005}$	-
PATE-GAN (ϵ = 1)	$0.152_{\pm.001}$	$0.189_{\pm.002}$	$0.136_{\pm.002}$	$0.176_{\pm.009}$	$0.080_{\pm.003}$	$0.630_{\pm.007}$	$0.070_{\pm.003}$	$0.494_{\pm.003}$	$0.189_{\pm.010}$	$0.044_{\pm.002}$	$0.177_{\pm.003}$	-
TableDiffusion (ϵ = 1)	$0.195_{\pm.002}$	$0.243_{\pm.002}$	$0.091_{\pm.002}$	$0.131_{\pm.006}$	$0.360_{\pm.004}$	$0.147_{\pm.008}$	$0.206_{\pm.003}$	$0.140_{\pm.002}$	$0.283_{\pm.008}$	$0.103_{\pm.001}$	$0.310_{\pm.003}$	-
`HALF`	$0.019_{\pm.002}$	$0.018_{\pm.001}$	$0.008_{\pm.002}$	$0.010_{\pm.007}$	$0.014_{\pm.002}$	$0.022_{\pm.006}$	$0.012_{\pm.003}$	$0.008_{\pm.002}$	$0.015_{\pm.006}$	$0.006_{\pm.001}$	$0.021_{\pm.004}$	-
`HISTOGRAM`	$0.017_{\pm.002}$	$0.015_{\pm.001}$	$0.008_{\pm.001}$	$0.010_{\pm.003}$	$0.008_{\pm.002}$	$0.032_{\pm.004}$	$0.008_{\pm.001}$	$0.006_{\pm.004}$	$0.008_{\pm.002}$	$0.006_{\pm.001}$	$0.016_{\pm.001}$	-

Table E19: Fidelity evaluation of one-way numerical marginals on train data $\mathcal{D}_{\text{train}}$. The best result of each category is in bold.

	Adult	Shoppers	Phishing	Magic	Faults	Bean	Obesity	Robot	Abalone	News	Insurance	Wine
PGM	$0.212_{\pm.012}$	$0.097_{\pm.001}$	-	$0.034_{\pm.001}$	$0.037_{\pm.000}$	$0.034_{\pm.001}$	$0.039_{\pm.001}$	$0.040_{\pm.000}$	$0.034_{\pm.000}$	$0.065_{\pm.001}$	$0.029_{\pm.000}$	$0.042_{\pm.000}$
PrivSyn	$0.013_{\pm.001}$	$0.023_{\pm.001}$	-	$0.010_{\pm.000}$	$0.019_{\pm.002}$	$0.026_{\pm.000}$	$0.019_{\pm.002}$	$0.029_{\pm.001}$	$0.010_{\pm.000}$	**$0.013_{\pm.000}$**	$0.023_{\pm.001}$	$0.009_{\pm.000}$
TVAE	$0.024_{\pm.000}$	$0.014_{\pm.001}$	-	$0.014_{\pm.000}$	$0.020_{\pm.001}$	$0.016_{\pm.001}$	$0.026_{\pm.002}$	$0.021_{\pm.001}$	$0.017_{\pm.001}$	$0.018_{\pm.000}$	$0.024_{\pm.004}$	$0.025_{\pm.000}$
CTGAN	**$0.007_{\pm.000}$**	**$0.008_{\pm.000}$**	-	$0.023_{\pm.001}$	$0.039_{\pm.001}$	$0.033_{\pm.001}$	$0.021_{\pm.000}$	$0.034_{\pm.001}$	$0.036_{\pm.001}$	$0.016_{\pm.001}$	$0.062_{\pm.003}$	$0.021_{\pm.001}$
TabDDPM	$0.007_{\pm.001}$	$0.012_{\pm.000}$	-	**$0.006_{\pm.001}$**	**$0.014_{\pm.001}$**	**$0.005_{\pm.001}$**	**$0.009_{\pm.001}$**	**$0.008_{\pm.001}$**	**$0.007_{\pm.001}$**	$0.037_{\pm.000}$	$0.013_{\pm.003}$	**$0.007_{\pm.000}$**
GReaT	$0.012_{\pm.001}$	$0.021_{\pm.002}$	-	$0.033_{\pm.001}$	$0.026_{\pm.002}$	$0.010_{\pm.001}$	$0.032_{\pm.002}$	$0.026_{\pm.001}$	$0.011_{\pm.002}$	-	**$0.010_{\pm.002}$**	$0.013_{\pm.001}$
PGM (ϵ = 1)	$0.216_{\pm.015}$	$0.098_{\pm.001}$	-	$0.035_{\pm.002}$	**$0.040_{\pm.001}$**	**$0.034_{\pm.002}$**	$0.043_{\pm.000}$	**$0.041_{\pm.000}$**	**$0.034_{\pm.001}$**	$0.067_{\pm.002}$	**$0.031_{\pm.001}$**	**$0.043_{\pm.000}$**
PrivSyn (ϵ = 1)	**$0.020_{\pm.001}$**	$0.044_{\pm.002}$	-	**$0.028_{\pm.001}$**	$0.168_{\pm.004}$	$0.072_{\pm.003}$	$0.098_{\pm.003}$	$0.129_{\pm.004}$	$0.093_{\pm.004}$	$0.056_{\pm.001}$	$0.078_{\pm.002}$	$0.103_{\pm.003}$
PATE-GAN (ϵ = 1)	$0.028_{\pm.001}$	**$0.037_{\pm.001}$**	-	$0.063_{\pm.001}$	$0.150_{\pm.000}$	$0.204_{\pm.003}$	**$0.035_{\pm.001}$**	$0.230_{\pm.000}$	$0.236_{\pm.001}$	**$0.040_{\pm.000}$**	$0.324_{\pm.001}$	$0.168_{\pm.000}$
TableDiffusion (ϵ = 1)	$0.046_{\pm.001}$	$0.050_{\pm.000}$	-	$0.067_{\pm.000}$	$0.126_{\pm.000}$	$0.072_{\pm.002}$	$0.244_{\pm.002}$	$0.132_{\pm.000}$	$0.090_{\pm.001}$	$0.067_{\pm.001}$	$0.156_{\pm.001}$	$0.072_{\pm.001}$
`HALF`	$0.004_{\pm.000}$	$0.003_{\pm.001}$	-	$0.004_{\pm.001}$	$0.006_{\pm.001}$	$0.004_{\pm.001}$	$0.008_{\pm.001}$	$0.005_{\pm.000}$	$0.004_{\pm.001}$	$0.004_{\pm.000}$	$0.009_{\pm.002}$	$0.003_{\pm.000}$
`HISTOGRAM`	$0.226_{\pm.016}$	$0.099_{\pm.003}$	-	$0.035_{\pm.002}$	$0.037_{\pm.000}$	$0.035_{\pm.001}$	$0.039_{\pm.000}$	$0.041_{\pm.000}$	$0.034_{\pm.000}$	$0.068_{\pm.002}$	$0.029_{\pm.001}$	$0.042_{\pm.000}$

F DISCUSSION OF PROPOSED PRIVACY METRIC

The definition of the proposed membership inference score (MDS) aligns closely with concepts of memorization in neural networks [19, 71] and the leave-one-out notion of stability in machine learning [8]. However, it diverges in three crucial ways: (i) We measure the worst-case discourse risk as the privacy metric, whereas other studies focus on the difference of individuals or average cases. (ii) Our work specifically addresses privacy concerns in data synthesis, as opposed to other studies that explore discriminative models like classification. (iii) Our approach emphasizes the discrepancy caused

Table E20: Fidelity evaluation of two-way categorical-categorical marginals on train data $\mathcal{D}_{\text{train}}$. The best result of each category is in bold.

	Adult	Shoppers	Phishing	Magic	Faults	Bean	Obesity	Robot	Abalone	News	Insurance	Wine
PGM	0.049±.001	0.044±.002	0.033±.001	-	0.087±.005	-	0.020±.001	-	-	0.025±.001	0.037±.010	-
PrivSyn	0.038±.003	0.033±.001	**0.015±.001**	-	0.151±.016	-	0.038±.003	-	-	0.022±.001	**0.035±.005**	-
TVAE	0.150±.004	0.336±.001	0.030±.001	-	0.087±.014	-	0.076±.006	-	-	0.161±.002	0.135±.010	-
CTGAN	0.127±.001	0.138±.001	0.077±.003	-	0.259±.006	-	0.126±.003	-	-	0.062±.002	0.101±.005	-
TabDDPM	**0.036±.001**	**0.030±.002**	0.019±.001	-	**0.032±.004**	-	**0.019±.001**	-	-	**0.010±.001**	0.037±.007	-
GReaT	0.092±.003	0.074±.002	0.094±.002	-	0.074±.007	-	0.064±.006	-	-	-	0.205±.006	-
PGM (ϵ = 1)	0.081±.002	**0.074±.003**	**0.038±.001**	-	**0.154±.003**	-	**0.066±.001**	-	-	**0.027±.001**	**0.112±.002**	-
PrivSyn (ϵ = 1)	**0.077±.003**	0.118±.007	0.046±.002	-	0.267±.005	-	0.205±.016	-	-	0.066±.003	0.143±.011	-
PATE-GAN (ϵ = 1)	0.278±.002	0.347±.001	0.210±.002	-	0.253±.006	-	0.129±.005	-	-	0.073±.003	0.273±.005	-
TableDiffusion (ϵ = 1)	0.323±.002	0.361±.003	0.155±.002	-	0.620±.002	-	0.348±.004	-	-	0.196±.001	0.466±.003	-
HALF	0.045±.004	0.040±.002	0.014±.002	-	0.027±.002	-	0.023±.004	-	-	0.011±.001	0.041±.006	-
HISTOGRAM	0.116±.002	0.072±.002	0.046±.001	-	0.188±.005	-	0.053±.002	-	-	0.027±.002	0.044±.003	-

Table E21: Fidelity evaluation of two-way categorical-numerical marginals on train data $\mathcal{D}_{\text{train}}$. The best result of each category is in bold.

	Adult	Shoppers	Phishing	Magic	Faults	Bean	Obesity	Robot	Abalone	News	Insurance	Wine
PGM	0.230±.012	0.110±.002	-	0.038±.002	0.058±.002	0.045±.003	0.051±.001	0.049±.004	0.040±.004	0.072±.001	0.048±.007	-
PrivSyn	0.029±.001	0.036±.001	-	0.018±.005	0.065±.004	0.045±.005	0.043±.002	0.100±.022	0.032±.007	**0.021±.001**	0.041±.002	-
TVAE	0.113±.003	0.207±.001	-	0.027±.005	0.074±.010	0.077±.010	0.070±.005	0.082±.001	0.085±.005	0.103±.002	0.103±.008	-
CTGAN	0.077±.002	0.077±.002	-	0.288±.011	0.165±.007	0.246±.009	0.092±.003	0.175±.006	0.059±.006	0.053±.001	0.111±.004	-
TabDDPM	**0.025±.001**	**0.027±.002**	-	**0.015±.006**	**0.033±.003**	**0.023±.003**	**0.021±.001**	**0.021±.003**	**0.022±.007**	0.043±.001	**0.035±.006**	-
GReaT	0.068±.002	0.065±.003	-	0.041±.003	0.064±.008	0.030±.001	0.072±.006	0.084±.005	0.033±.006	-	0.120±.006	-
PGM (ϵ = 1)	0.243±.017	0.119±.002	-	**0.042±.002**	**0.099±.005**	**0.060±.006**	**0.081±.001**	**0.080±.002**	**0.059±.004**	**0.075±.003**	**0.088±.001**	-
PrivSyn (ϵ = 1)	**0.055±.003**	**0.089±.005**	-	0.070±.005	0.269±.010	0.220±.008	0.216±.012	0.212±.037	0.146±.014	0.086±.003	0.161±.005	-
PATE-GAN (ϵ = 1)	0.182±.001	0.228±.001	-	0.240±.009	0.237±.003	0.837±.009	0.113±.003	0.724±.004	0.427±.010	0.085±.002	0.501±.004	-
TableDiffusion (ϵ = 1)	0.242±.003	0.293±.002	-	0.197±.006	0.487±.004	0.226±.008	0.450±.003	0.274±.002	0.373±.009	0.171±.000	0.467±.003	-
HALF	0.025±.002	0.022±.001	-	0.015±.007	0.021±.002	0.028±.006	0.021±.004	0.016±.001	0.020±.005	0.011±.001	0.033±.005	-
HISTOGRAM	0.253±.015	0.117±.003	-	0.060±.003	0.078±.001	0.163±.002	0.064±.001	0.085±.003	0.094±.001	0.076±.002	0.061±.001	-

Table E22: Fidelity evaluation of two-way numerical-numerical marginals on train data $\mathcal{D}_{\text{train}}$. The best result of each category is in bold.

	Adult	Shoppers	Phishing	Magic	Faults	Bean	Obesity	Robot	Abalone	News	Insurance	Wine
PGM	0.425±.023	0.195±.002	-	0.070±.001	0.083±.001	0.072±.001	0.085±.001	0.103±.000	0.069±.001	0.135±.002	0.060±.001	0.090±.001
PrivSyn	0.027±.001	0.048±.003	-	0.024±.000	0.057±.002	0.054±.000	0.057±.002	0.084±.001	0.036±.001	**0.031±.000**	0.052±.001	0.026±.000
TVAE	0.047±.001	0.028±.001	-	0.032±.001	0.045±.002	0.039±.002	0.061±.002	0.050±.001	0.037±.001	0.037±.001	0.055±.007	0.052±.000
CTGAN	0.016±.001	**0.018±.000**	-	0.051±.002	0.084±.001	0.072±.001	0.058±.001	0.080±.001	0.088±.002	0.034±.001	0.134±.005	0.044±.001
TabDDPM	**0.015±.001**	0.026±.001	-	**0.015±.001**	**0.033±.002**	**0.013±.001**	**0.026±.001**	**0.023±.001**	**0.016±.002**	0.076±.001	0.035±.004	**0.015±.001**
GReaT	0.024±.002	0.042±.005	-	0.067±.003	0.055±.004	0.023±.001	0.068±.004	0.055±.002	0.025±.004	-	**0.028±.002**	0.026±.002
PGM (ϵ = 1)	0.432±.030	0.203±.003	-	0.084±.004	**0.098±.001**	**0.079±.003**	**0.099±.001**	**0.109±.000**	**0.099±.001**	0.137±.004	**0.090±.002**	**0.092±.001**
PrivSyn (ϵ = 1)	**0.043±.002**	0.091±.003	-	**0.068±.001**	0.340±.009	0.171±.005	0.205±.005	0.268±.007	0.221±.006	0.115±.002	0.176±.005	0.209±.005
PATE-GAN (ϵ = 1)	0.056±.002	**0.078±.002**	-	0.131±.001	0.302±.001	0.409±.005	0.082±.002	0.461±.001	0.474±.001	**0.082±.001**	0.649±.002	0.335±.000
TableDiffusion (ϵ = 1)	0.092±.001	0.099±.001	-	0.133±.001	0.252±.001	0.147±.003	0.488±.004	0.265±.001	0.181±.002	0.136±.001	0.314±.001	0.145±.001
HALF	0.009±.001	0.007±.002	-	0.013±.001	0.014±.001	0.012±.001	0.021±.002	0.013±.001	0.009±.002	0.011±.001	0.025±.003	0.008±.000
HISTOGRAM	0.453±.031	0.205±.005	-	0.098±.004	0.096±.001	0.120±.002	0.093±.001	0.113±.000	0.142±.000	0.140±.003	0.080±.002	0.094±.000

by the presence or absence of target data in training, in contrast to other works that highlight performance gains from adding samples to the training set.

G MISLEADING CONCLUSION

We find that some statements in data synthesis may be misleading or even incorrect due to limitations of evaluation metrics or methodologies. We highlight some of them as below:

Table E23: Fidelity evaluation of one-way categorical marginals on test data $\mathcal{D}_{\text{test}}$. The best result of each category is in bold.

	Adult	Shoppers	Phishing	Magic	Faults	Bean	Obesity	Robot	Abalone	News	Insurance	Wine
PGM	**0.013±.003**	**0.023±.002**	**0.012±.001**	0.009±.004	0.044±.007	**0.017±.004**	0.031±.004	**0.012±.007**	0.030±.011	**0.005±.001**	0.062±.012	-
PrivSyn	0.015±.001	0.023±.004	0.012±.002	**0.007±.004**	0.040±.004	0.018±.006	0.031±.007	0.016±.005	**0.020±.005**	0.008±.001	**0.057±.009**	-
TVAE	0.090±.003	0.191±.003	0.021±.001	0.009±.005	0.083±.024	0.060±.007	0.036±.003	0.062±.012	0.049±.009	0.084±.001	0.115±.009	-
CTGAN	0.069±.004	0.065±.002	0.054±.001	0.251±.008	0.102±.011	0.205±.008	0.073±.005	0.135±.014	0.028±.009	0.036±.001	0.068±.010	-
TabDDPM	0.017±.001	0.025±.001	0.014±.001	0.009±.003	0.055±.011	**0.017±.004**	**0.029±.004**	0.022±.006	0.033±.022	0.006±.001	0.059±.010	-
GReaT	0.057±.002	0.053±.004	0.055±.001	0.008±.004	**0.031±.007**	0.018±.007	0.038±.008	0.058±.010	0.032±.005	-	0.112±.010	-
PGM (ϵ = 1)	**0.016±.001**	**0.025±.002**	**0.013±.000**	0.011±.007	**0.066±.007**	**0.020±.005**	**0.041±.004**	**0.041±.015**	0.051±.016	0.007±.001	**0.053±.006**	-
PrivSyn (ϵ = 1)	0.032±.003	0.051±.002	0.018±.002	**0.003±.002**	0.138±.025	0.077±.016	0.085±.012	0.086±.016	**0.029±.015**	**0.012±.001**	0.104±.009	-
PATE-GAN (ϵ = 1)	0.152±.003	0.184±.002	0.136±.002	0.183±.009	0.079±.003	0.636±.005	0.073±.004	0.496±.012	0.204±.018	0.042±.001	0.152±.007	-
TableDiffusion (ϵ = 1)	0.194±.002	0.247±.002	0.089±.002	0.133±.008	0.371±.004	0.150±.007	0.194±.004	0.140±.003	0.254±.010	0.102±.000	0.320±.007	-
HALF	0.020±.003	0.023±.003	0.012±.003	0.011±.005	0.037±.007	0.019±.005	0.030±.005	0.018±.009	0.028±.008	0.008±.002	0.048±.011	-
HISTOGRAM	0.017±.002	0.020±.001	0.011±.001	0.011±.003	0.046±.010	0.017±.005	0.025±.001	0.012±.004	0.038±.020	0.009±.001	0.047±.003	-

Table E24: Fidelity evaluation of one-way numerical marginals on test data $\mathcal{D}_{\text{test}}$. The best result of each category is in bold.

	Adult	Shoppers	Phishing	Magic	Faults	Bean	Obesity	Robot	Abalone	News	Insurance	Wine
PGM	0.195±.004	0.095±.001	-	0.033±.000	0.054±.003	0.034±.001	0.043±.001	0.042±.000	0.046±.003	0.052±.004	0.036±.002	0.048±.001
PrivSyn	0.013±.000	0.024±.000	-	0.011±.000	0.032±.005	0.027±.000	0.027±.002	0.031±.003	0.019±.003	**0.011±.000**	0.031±.003	0.012±.001
TVAE	0.025±.000	0.013±.001	-	0.014±.000	0.028±.003	0.017±.001	0.033±.002	0.024±.001	0.016±.001	0.017±.001	0.033±.005	0.029±.001
CTGAN	0.009±.000	**0.009±.001**	-	0.024±.000	0.047±.002	0.034±.001	0.027±.001	0.037±.003	0.046±.001	0.016±.000	0.072±.003	0.022±.001
TabDDPM	**0.008±.000**	0.012±.001	-	**0.006±.001**	**0.021±.002**	**0.006±.001**	**0.021±.001**	**0.013±.002**	**0.017±.004**	0.034±.000	0.024±.006	**0.010±.001**
GReaT	0.013±.000	0.021±.002	-	0.034±.001	0.036±.002	0.009±.000	0.039±.004	0.027±.002	0.019±.003	-	**0.020±.006**	0.015±.001
PGM (ϵ = 1)	0.196±.004	0.094±.000	-	0.033±.001	**0.055±.002**	**0.034±.001**	0.045±.002	**0.043±.001**	**0.046±.002**	0.053±.002	**0.037±.002**	**0.047±.001**
PrivSyn (ϵ = 1)	**0.024±.000**	0.044±.001	-	**0.026±.002**	0.146±.006	0.069±.004	0.106±.005	0.128±.004	0.086±.002	0.040±.001	0.079±.007	0.085±.003
PATE-GAN (ϵ = 1)	0.027±.000	**0.036±.001**	-	0.063±.001	0.166±.003	0.204±.001	**0.033±.003**	0.231±.001	0.250±.001	**0.037±.002**	0.341±.009	0.167±.002
TableDiffusion (ϵ = 1)	0.046±.000	0.048±.000	-	0.066±.001	0.146±.001	0.072±.001	0.243±.002	0.132±.000	0.099±.002	0.060±.001	0.174±.005	0.065±.001
HALF	0.005±.001	0.003±.000	-	0.004±.001	0.015±.002	0.004±.001	0.018±.003	0.011±.001	0.013±.003	0.005±.001	0.018±.003	0.007±.001
HISTOGRAM	0.208±.020	0.094±.001	-	0.033±.000	0.051±.001	0.034±.000	0.043±.001	0.042±.000	0.047±.002	0.061±.007	0.039±.004	0.048±.001

Table E25: Fidelity evaluation of two-way categorical-categorical marginals on test data $\mathcal{D}_{\text{test}}$. The best result of each category is in bold.

	Adult	Shoppers	Phishing	Magic	Faults	Bean	Obesity	Robot	Abalone	News	Insurance	Wine
PGM	0.048±.003	0.063±.002	0.041±.001	-	0.122±.005	-	0.059±.004	-	-	0.025±.001	0.122±.021	-
PrivSyn	**0.038±.001**	0.058±.004	**0.023±.002**	-	0.150±.010	-	0.064±.009	-	-	0.024±.001	**0.107±.008**	-
TVAE	0.153±.003	0.333±.003	0.035±.002	-	0.118±.025	-	0.072±.002	-	-	0.161±.001	0.203±.012	-
CTGAN	0.128±.005	0.137±.002	0.083±.001	-	0.256±.005	-	0.147±.008	-	-	0.063±.002	0.132±.009	-
TabDDPM	0.039±.001	**0.054±.001**	0.025±.002	-	**0.083±.010**	-	**0.055±.005**	-	-	**0.010±.002**	0.113±.010	-
GReaT	0.094±.002	0.091±.005	0.089±.002	-	0.092±.017	-	0.067±.011	-	-	-	0.203±.012	-
PGM (ϵ = 1)	**0.074±.001**	**0.083±.002**	0.043±.001	-	**0.172±.007**	-	**0.085±.006**	-	-	**0.027±.001**	**0.144±.010**	-
PrivSyn (ϵ = 1)	0.082±.004	0.135±.003	**0.041±.003**	-	0.310±.019	-	0.166±.022	-	-	0.031±.002	0.190±.018	-
PATE-GAN (ϵ = 1)	0.278±.003	0.342±.004	0.211±.002	-	0.247±.010	-	0.140±.004	-	-	0.070±.001	0.258±.012	-
TableDiffusion(ϵ = 1)	0.321±.002	0.365±.003	0.153±.002	-	0.620±.007	-	0.332±.006	-	-	0.195±.000	0.494±.008	-
HALF	0.049±.003	0.052±.004	0.021±.003	-	0.060±.008	-	0.056±.009	-	-	0.013±.003	0.099±.008	-
HISTOGRAM	0.116±.002	0.081±.000	0.049±.001	-	0.217±.007	-	0.072±.002	-	-	0.030±.001	0.096±.003	-

- In [33], authors show that the machine learning performance on TabDDPM is even better than that on real data, which implies that synthetic data can be a perfect (even better) substitute for real data. However, this statement may be incorrect due to inadequate model tuning and improper data shuffling practices. Our evaluations show that even simple models, such as linear regression, can achieve better performance on real data than on high-quality synthetic data (*i.e.,* generated by TabDDPM).
- Some studies [31–33, 35] prioritize machine learning efficacy as the primary (if not only) fidelity evaluation metric. This approach is problematic because data synthesis can be *biased* to label attributes, and a high machine learning efficacy score does not necessarily equate to high fidelity in synthetic data.

Table E26: Fidelity evaluation of two-way categorical-numerical marginals on test data $\mathcal{D}_{\textbf{test}}$. The best result of each category is in bold.

	Adult	Shoppers	Phishing	Magic	Faults	Bean	Obesity	Robot	Abalone	News	Insurance	Wine
PGM	0.212±.005	0.119±.002	-	0.043±.004	0.106±.007	0.053±.005	0.080±.004	0.058±.006	0.077±.012	0.060±.005	0.106±.010	-
PrivSyn	0.029±.001	0.049±.003	-	0.019±.004	0.090±.005	0.048±.006	0.069±.006	0.079±.006	0.042±.006	**0.020±.001**	0.097±.008	-
TVAE	0.115±.003	0.205±.003	-	0.025±.005	0.115±.023	0.082±.007	0.075±.004	0.092±.011	0.081±.007	0.101±.001	0.158±.011	-
CTGAN	0.080±.003	0.076±.002	-	0.275±.008	0.157±.012	0.244±.009	0.109±.005	0.178±.015	0.081±.008	0.053±.001	0.150±.009	-
TabDDPM	**0.027±.001**	**0.039±.001**	-	**0.016±.003**	**0.080±.012**	**0.025±.005**	**0.055±.004**	**0.041±.005**	**0.051±.024**	0.040±.001	**0.090±.009**	-
GReaT	0.070±.002	0.074±.004	-	0.045±.004	0.078±.006	0.032±.006	0.081±.008	0.089±.010	0.055±.004	-	0.138±.012	-
PGM (ϵ = 1)	0.219±.005	0.121±.002	-	0.044±.007	**0.139±.006**	**0.059±.005**	**0.098±.004**	**0.093±.013**	**0.102±.016**	0.061±.002	**0.117±.004**	-
PrivSyn (ϵ = 1)	**0.060±.003**	**0.098±.003**	-	**0.033±.002**	0.290±.022	0.201±.015	0.197±.017	0.226±.017	0.124±.017	**0.053±.001**	0.204±.010	-
PATE-GAN (ϵ = 1)	0.182±.002	0.221±.002	-	0.246±.009	0.251±.002	0.842±.006	0.115±.004	0.726±.012	0.455±.018	0.080±.003	0.494±.011	-
TableDiffusion (ϵ = 1)	0.242±.003	0.295±.002	-	0.199±.008	0.519±.005	0.229±.007	0.437±.006	0.275±.004	0.353±.012	0.163±.000	0.495±.008	-
HALF	0.027±.003	0.029±.003	-	0.016±.005	0.056±.006	0.025±.006	0.052±.008	0.035±.008	0.043±.010	0.013±.002	0.077±.011	-
HISTOGRAM	0.236±.019	0.117±.001	-	0.053±.003	0.123±.009	0.151±.005	0.084±.001	0.092±.003	0.122±.012	0.071±.007	0.106±.005	-

Table E27: Fidelity evaluation of two-way numerical-numerical marginals on test data $\mathcal{D}_{\textbf{test}}$. The best result of each category is in bold.

	Adult	Shoppers	Phishing	Magic	Faults	Bean	Obesity	Robot	Abalone	News	Insurance	Wine
PGM	0.392±.008	0.191±.002	-	0.069±.000	0.120±.006	0.073±.001	0.097±.002	0.107±.001	0.094±.006	0.108±.009	0.088±.003	0.101±.002
PrivSyn	0.027±.000	0.050±.001	-	0.026±.001	0.083±.009	0.056±.001	0.073±.002	0.089±.003	0.049±.006	**0.026±.001**	0.079±.004	0.032±.001
TVAE	0.049±.001	0.026±.001	-	0.033±.001	0.064±.005	0.041±.001	0.078±.003	0.058±.002	0.037±.002	0.034±.001	0.082±.007	0.061±.001
CTGAN	0.018±.001	**0.019±.002**	-	0.052±.001	0.103±.004	0.074±.002	0.069±.001	0.087±.004	0.109±.002	0.033±.001	0.157±.008	0.047±.002
TabDDPM	**0.017±.000**	0.025±.001	-	**0.015±.001**	**0.049±.004**	**0.015±.001**	**0.054±.002**	**0.035±.003**	**0.035±.006**	0.070±.001	0.064±.008	**0.023±.001**
GReaT	0.026±.001	0.042±.003	-	0.068±.002	0.079±.003	0.023±.001	0.086±.008	0.059±.004	0.040±.005	-	**0.060±.008**	0.033±.002
PGM (ϵ = 1)	0.393±.009	0.194±.000	-	0.080±.001	**0.130±.004**	**0.079±.001**	0.108±.003	**0.113±.001**	**0.121±.004**	0.109±.003	**0.114±.005**	**0.102±.001**
PrivSyn (ϵ = 1)	**0.049±.001**	0.091±.002	-	**0.062±.003**	0.297±.012	0.167±.007	0.219±.010	0.268±.008	0.205±.007	0.083±.002	0.181±.010	0.174±.006
PATE-GAN (ϵ = 1)	0.055±.001	**0.074±.001**	-	0.130±.001	0.334±.005	0.408±.003	**0.083±.005**	0.461±.002	0.501±.001	**0.077±.004**	0.682±.017	0.334±.003
TableDiffusion (ϵ = 1)	0.093±.001	0.096±.001	-	0.132±.002	0.294±.002	0.147±.002	0.486±.005	0.265±.001	0.200±.005	0.121±.001	0.348±.011	0.132±.001
HALF	0.011±.001	0.008±.000	-	0.013±.001	0.039±.003	0.013±.001	0.050±.004	0.028±.001	0.027±.005	0.012±.001	0.059±.004	0.020±.001
HISTOGRAM	0.417±.041	0.194±.003	-	0.093±.001	0.127±.003	0.118±.001	0.104±.002	0.116±.000	0.168±.003	0.125±.015	0.115±.006	0.104±.002

- Extensive studies [32, 33, 35, 72, 75] use Distance to Closest Records (DCR) to evaluate the privacy of synthetic data and assert their models are safe. However, in this paper we show that DCR fails to serve as an adequate measure of privacy. We also provide evidence that many recently introduced HP methods exhibit significant risks of membership disclosure, which are often ignored by HP synthesis algorithms.

H COMPUTING WASSERSTEIN DISTANCE

When the cost matrix is large and dense, computing this metric can be prohibitively time-consuming. Several options are provided to address this problem. (i) Sinkhorn distance [14] provides a fast approximation to the Wasserstein distance by penalizing the objective with an entropy term. (ii) Sliced-Wasserstein distance [6], which uses Radon transform to linearly project data into one dimension, can be efficiently computed. (iii) Reducing the size of the cost matrix by randomly sampling a small set of points from the probability densities. These computational strategies ensure the Wasserstein distance remains a feasible solution for fidelity evaluation, even for complex and high-dimensional marginals. In practice, when the domain of target probability distribution is large (*i.e.*, $n \geq 5,000$), we find that randomly sampling half of the elements is both effective and efficient.