Mending synthetic data with MAPS: Model Agnostic Post-hoc Synthetic Data Refinement Framework

Yan Li
University of Copenhagen
yanli@di.ku.dk

Jennifer Bartell
University of Copenhagen
bartell@sund.ku.dk

Anders Krogh
University of Copenhagen
akrogh@di.ku.dk

Abstract

Generating high-quality synthetic data with privacy protections remains a challenging ad-hoc process, requiring careful model design and training often tailored to the characteristics of a targeted dataset. We present MAPS, a model-agnostic post-hoc framework that improves synthetic data quality for any pre-trained generative model while ensuring sample-level privacy standards are met. Our two-stage approach first removes synthetic samples that violate privacy by being too close to real data, achieving 0-identifiability guarantees. Second, we employ importance weighting via a binary classifier to resample the remaining synthetic data according to estimated density ratios. We evaluate MAPS across two healthcare datasets (TCGA-metadata, GOSSIS-1-eICU-cardiovascular) and four generative models (TVAE, CTGAN, TabDDPM, DGD), demonstrating significant improvements in fidelity and utility while maintaining privacy. Notably, MAPS achieves substantial improvements in fidelity metrics, with 40 out of 48 statistical tests demonstrating significant improvements in marginal distributional measures and notable enhancements in correlation structure preservation and joint distribution similarity. For example, Joint Jensen-Shannon Distance reduced from ranges of 0.7888-0.8278 to 0.5434-0.5961 on TCGA-metadata and 0.6192-0.7902 to 0.3633-0.4503 on GOSSIS-1-eICU-cardiovascular. Utility improvements are equally impressive, with classification F1 scores improving from ranges of 0.0866-0.2400 to 0.3043-0.3848 on TCGA-metadata and 0.1287-0.2085 to 0.2104-0.2497 on GOSSIS-1-eICU-cardiovascular across different model-dataset combinations. The code of this project is available at `https://github.com/yanlihub/MAPS.git`.

1 Introduction

The proliferation of data-driven applications in privacy-sensitive domains has created an urgent need for synthetic data that preserves statistical fidelity while protecting individual privacy [1, 2]. However, generating high quality synthetic data that satisfies both privacy requirements and downstream task performance remains challenging, requiring practitioners to navigate complex trade-offs between competing objectives - privacy and fidelity [3, 4]. This privacy-fidelity trade-off forces practitioners to choose between models that provide strong privacy guarantees but produce data of limited practical value, and models that generate high fidelity synthetic data but potentially leak sensitive information [5, 6]. This challenge is compounded by the substantial technical expertise required to design, train, and tune generative models effectively [7, 8].

Existing approaches typically fall into two distinct categories, each with significant limitations. Privacy-first methods such as ADS-GAN [5], PATE-GAN [6], and DP-GAN [9] incorporate privacy constraints during training, providing formal privacy guarantees but often producing synthetic data with notably degraded utility. Conversely, fidelity-first approaches like TVAE [10], CTGAN [10],

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: The Second Workshop on GenAI for Health: Potential, Trust, and Policy Compliance.

TabDDPM [11] and DGD [12] focus on generating high fidelity synthetic data but have no additional privacy protections beyond sampling design.

We present MAPS (Model Agnostic Post-hoc Synthetic Data Refinement Framework), a novel two-stage approach to refining synthetic datasets from any generator that addresses both privacy and fidelity concerns through complementary mechanisms. The first stage minimizes re-identification risks by implementing 0-identifiability guarantees, systematically removing synthetic samples that violate privacy by being closer to real samples than those real samples are to their nearest real neighbors. The second stage enhances data fidelity through theoretically grounded importance weighting, where we train a binary classifier to distinguish between real and synthetic data, then use this classifier to estimate density ratios following the likelihood-free importance weighting framework [13]. The framework is currently designed to refine static tabular synthetic data produced by any generation method.

2 Methodology

In this section, We first formalize the problem setup, then detail the privacy filtering mechanism that enforces 0-identifiability constraints, describe the importance weighting approach that improves distributional fidelity through density ratio estimation, and finally present the Sampling-Importance-Resampling procedure for selecting the refined synthetic dataset.

2.1 Problem Formulation

Let $\mathcal{D} = \{x_i\}_{i=1}^{N}$ denote a dataset of N real samples drawn i.i.d. from an unknown distribution $p(x)$, and $\hat{\mathcal{D}} = \{\hat{x}_j\}_{j=1}^{M}$ denote a synthetic dataset of M samples generated by some generative model with distribution $p_\theta(x)$. Our objective is to refine $\hat{\mathcal{D}}$ to produce a subset $\tilde{\mathcal{D}} \subset \hat{\mathcal{D}}$ of size N that (1) provides formal identifiability protections with respect to $\mathcal{D}$, and (2) exhibits improved fidelity to the true data distribution $p(x)$.

2.2 Stage 1: Privacy Filtering

The first stage protects privacy by removing synthetic samples that violate identifiability constraints. We build upon the ϵ-identifiability framework [5] and set it to be 0-identifiability to maximize protection using this privacy standard. Note that we can use any privacy metric that works on single samples in this stage.

For each real sample $x_i \in \mathcal{D}$, we define its distinctness threshold as:

$$r_i = \min_{x_j \in \mathcal{D} \setminus \{x_i\}} \|\mathbf{w} \cdot (x_i - x_j)\| \tag{1}$$

where $\mathbf{w}$ is a feature weight vector that accounts for the relative importance of different features in measuring similarity[1].

For each real sample x_i, we also compute its proximity to the synthetic dataset:

$$\hat{r}_i = \min_{\hat{x}_j \in \hat{\mathcal{D}}} \|\mathbf{w} \cdot (x_i - \hat{x}_j)\| \tag{2}$$

The ϵ-identifiability of synthetic dataset $\hat{\mathcal{D}}$ with respect to real dataset $\mathcal{D}$ requires that:

$$\mathcal{I}(\mathcal{D}, \hat{\mathcal{D}}) = \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}(\hat{r}_i < r_i) < \epsilon \tag{3}$$

where $\mathbb{I}$ is the indicator function.

In MAPS, we enforce the constraint of 0-identifiability by removing all synthetic samples $\hat{x}_j$ such that there exists a real sample x_i for which $\|\mathbf{w} \cdot (x_i - \hat{x}_j)\| < r_i$. This ensures that no synthetic sample is closer to any real sample than that real sample's nearest real neighbor, providing the strongest possible identifiability guarantee using this measure. While recent works have questioned the sufficiency of

[1] In our implementations, we treat all features with the same weight 1.

distance-based privacy metrics as standalone guarantees [14, 15], there remains no consensus on synthetic data privacy evaluation standards, and distance-based measures have demonstrated practical utility in prior works [5, 16, 17]. We adopt this approach for its sample level compatibility with MAPS's post-hoc framework and complement it with comprehensive membership inference attack based privacy evaluation (Section 4.4) to provide a multifaceted assessment of privacy protection.

2.3 Stage 2: Fidelity Enhancement via Importance Weighting

The second stage improves the fidelity of the identifiability-filtered synthetic data through importance weighting and resampling. We train a binary probabilistic classifier $c_\phi(x) : \mathcal{X} \to [0, 1]$ to distinguish between real and synthetic samples, then use this classifier to estimate importance weights.

The classifier outputs an estimate of the probability that a given sample belongs to the real data distribution:

$$c_\phi(x) = P(y = 1|x) \tag{4}$$

where $y = 1$ indicates that sample x comes from the real data distribution $p(x)$.

Using Bayes' theorem, we can express this probability as:

$$c_\phi(x) = \frac{p(x)\pi_1}{p(x)\pi_1 + p_\theta(x)\pi_0} \tag{5}$$

where $\pi_1 = P(y = 1)$ and $\pi_0 = P(y = 0)$ are the prior probabilities of observing samples from the real and synthetic distributions, respectively, in the training set used for the classifier.

The importance weight, representing the density ratio $\frac{p(x)}{p_\theta(x)}$, can be derived as:

$$\hat{w}_\phi(x) = \frac{p(x)}{p_\theta(x)} = \frac{\pi_0}{\pi_1}\frac{c_\phi(x)}{1 - c_\phi(x)} \tag{6}$$

This formulation follows the ideas from the likelihood-free importance weighting framework of [13], enabling us to estimate density ratios without explicit density models.

2.4 Sampling-Importance-Resampling

Using the estimated importance weights, we apply Sampling-Importance-Resampling (SIR) to select a refined synthetic dataset. Given importance weights for the M' identifiability-filtered synthetic samples, we normalize these weights and sample N samples according to the normalized probabilities:

$$p_j = \frac{\hat{w}_\phi(\hat{x}_j)}{\sum_{N=1}^{M'} \hat{w}_\phi(\hat{x}_N)} \tag{7}$$

The final refined synthetic dataset $\tilde{\mathcal{D}}$ contains N samples that satisfy both identifiability constraints and exhibit improved fidelity to the real data distribution. The full algorithm can be found in Appendix A. Figure 1 provides an overview of our two-stage approach to achieve these objectives.

3 Experimental Setup

Generative Models. We investigate MAPS' capabilities using four representative generative models covering the major paradigms in tabular data synthesis:

- **TVAE** [10]: A representative variational autoencoder specifically designed for tabular data generation.
- **CTGAN** [10]: The most widely adopted GAN-based model for tabular synthetic data generation.
- **TabDDPM** [11]: A diffusion-based model representing the latest paradigm in generative modeling for tabular data.
- **DGD** [12]: An encoder-free deep generative decoder with simple architecture that enhances learning of multi-modal data.

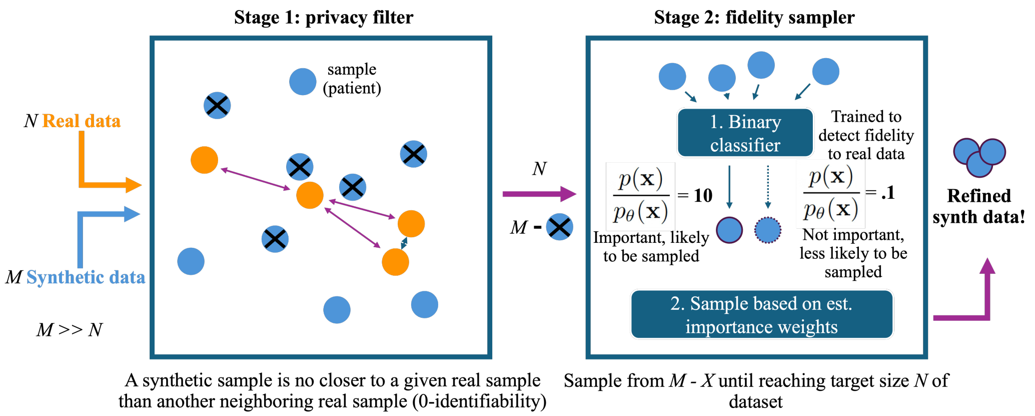
Stage 1: privacy filter
sample (patient)
N Real data
M Synthetic data
M >> N
N
M - ⊗
Stage 2: fidelity sampler
1. Binary classifier
Trained to detect fidelity to real data
$\frac{p(\mathbf{x})}{p_\theta(\mathbf{x})} = 10$
Important, likely to be sampled
$\frac{p(\mathbf{x})}{p_\theta(\mathbf{x})} = .1$
Not important, less likely to be sampled
2. Sample based on est. importance weights
Refined synth data!
A synthetic sample is no closer to a given real sample than another neighboring real sample (0-identifiability)
Sample from M - X until reaching target size N of dataset

Figure 1: Overview of the MAPS two-stage framework. Stage 1 removes synthetic samples violating 0-identifiability (marked with ×, with the number of X). Stage 2 trains a binary classifier to estimate importance weights via density ratios, then applies SIR to produce refined synthetic data that better approximates the real data distribution.

Datasets. We evaluate on two publicly available healthcare datasets:

- **TCGA-metadata**: The largest public cancer dataset covering 33 different cancer types with 11,315 records. We selected 21 variables with minimal missing data for our experiments.
- **GOSSIS-1-eICU-cardiovascular**: A large public cardiovascular ICU dataset with 41,396 records and 68 variables, representing complex clinical data with mixed data types.

The detailed data pre-processing procedure is described in Appendix B.1.

Evaluation Protocol. For each generative model and dataset combination, we generate a synthetic data pool of size $M = 30N$ where N is the number of real samples. This large pool enables meaningful selection during the refinement process. The baseline synthetic dataset consists of N randomly sampled synthetic samples from this pool, while the MAPS-refined dataset also contains N samples but selected through our two-stage refinement process. We assess MAPS effectiveness across three dimensions: (1) Distributional fidelity using statistical tests and similarity measures to evaluate how well synthetic data captures real data distributions, (2) Utility preservation through downstream task performance including clustering and classification tasks using a "train-on-synthetic, test-on-real" evaluation scheme, and (3) Privacy protection via resistance to membership inference attacks to ensure refinement does not compromise privacy standards. A detailed description of the implementations and evaluation methods can be found in Appendix B and Appendix C respectively. Computational requirements and practical considerations for pool size selection are discussed in Appendix B.5.

4 Results and Discussion

In this section, we present a comprehensive evaluation of MAPS across multiple dimensions to demonstrate its effectiveness in improving synthetic data quality while maintaining privacy protections. Our analysis examines distributional fidelity improvements, utility enhancements in downstream tasks, and privacy preservation across two healthcare datasets and four generative models.

4.1 Marginal Distribution Similarity

As an illustrative example of improvements in distributional alignment, we compare selected marginal distributions between the original real data, the raw synthetic data, and the refined synthetic data produced by MAPS. Figure 2 shows examples across both numerical and categorical variables that show the positive impact of refinement with MAPS. For instance, when examining the `initial_weight` variable for TVAE, the raw synthetic data exhibits spurious fluctuations when `initial_weight` is around 100, which are eliminated after refinement. Similarly, for TabDDPM, the problematic bump in the long tail region of the raw synthetic data distribution is corrected in the refined version. Categorical variable improvements manifest as refined synthetic data better reflecting the proportional distributions of real categories, with more accurate frequency representations across all categories.

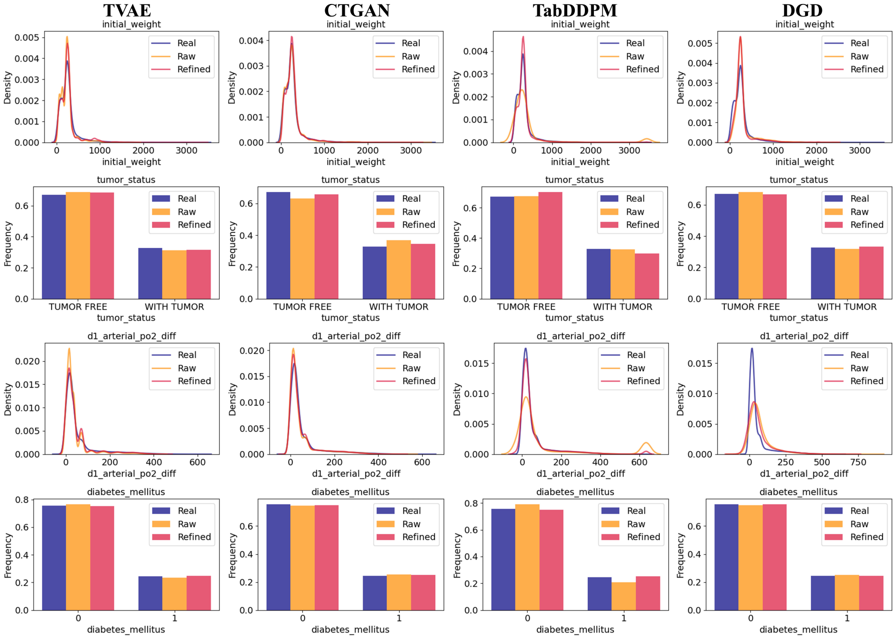

Figure 2: Marginal distribution comparison across models and variables. The 4×4 subplots compare marginal distributions across generative models and variables, including: `initial_weight` and `tumor_status` from TCGA metadata (numerical and categorical, respectively), and `d1_arterial_po2_diff` and `diabetes_mellitus` from GOSSIS-1-eICU-cardiovascular. Real refers to the original data, Raw to raw synthetic data, and Refined to MAPS-refined outputs.

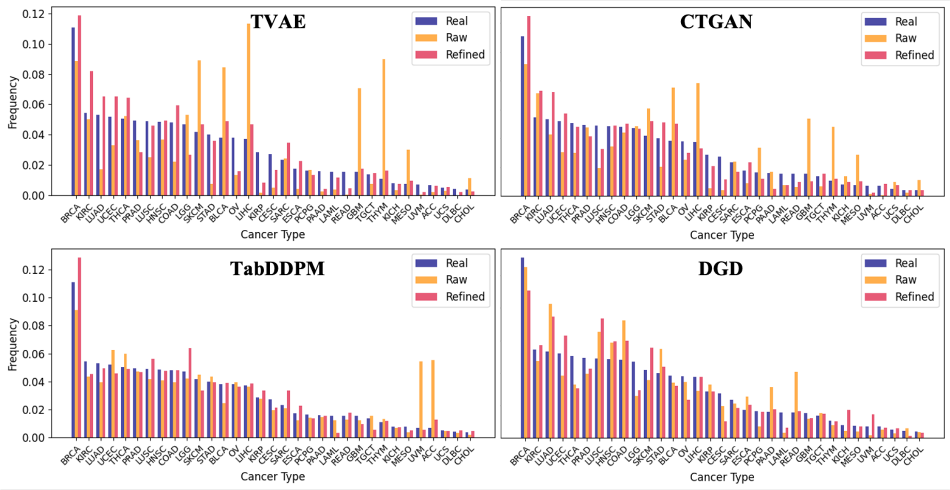

Figure 3: Class distribution fidelity across generative models on TCGA-metadata. The 2×2 subplots show the frequency distribution of 33 cancer types comparing real data, raw synthetic data, and MAPS-refined synthetic data across four generative models. Cancer types are ordered by frequency in the real data.

Preserving class distribution fidelity is critical for downstream utility, particularly when minority classes represent rare but clinically important conditions. Figure 3 demonstrates MAPS's effectiveness in maintaining class distribution across 33 cancer types in the TCGA-metadata dataset. Raw synthetic data from all four models exhibit widespread distortions in the frequency patterns of features, especially for specific cancer types CHOL (TVAE, CTGAN), UVM and ACC (TabDDPM), and READ (DGD). MAPS refinement substantially corrects these distortions, with refined distributions closely tracking true class frequencies across the entire spectrum from the most prevalent to rare cancer types. This preservation is particularly crucial for imbalanced datasets where minority classes constitute only 1-2% of samples, as MAPS is able to maintain their representation rather than diminishing it through refinement. Similar improvements are observed for GOSSIS-1-eICU-cardiovascular across different diagnostic classes (Appendix E).

The quantitative assessment in Table 1 reveals that 40 out of 48 statistical tests demonstrate significant improvements in marginal distributional measures. The Jensen-Shannon Distance shows particularly impressive improvements across all model-dataset combinations: TVAE demonstrates substantial reductions from 0.0807 to 0.0388 on TCGA-metadata and from 0.0652 to 0.0308 on GOSSIS-1-eICU-cardiovascular, while CTGAN achieves similar dramatic decreases from 0.0760 to 0.0373 on TCGA-metadata and from 0.0443 to 0.0269 on GOSSIS-1-eICU-cardiovascular. Total Variation Distance exhibits equally striking improvements, with TVAE reducing from 0.0684 to 0.0312 on TCGA-metadata and CTGAN dropping from 0.0427 to 0.0222 on GOSSIS-1-eICU-cardiovascular.

Table 1: Marginal distribution fidelity assessment. Results show mean ± standard deviation across multiple runs. Arrows indicate improvement direction: ↑ higher is better, ↓ lower is better. Bold values denote statistically significant improvements ($p < 0.05$, paired t-test) between raw and refined results for each model. This criterion is used consistently throughout all result tables.

TCGA-metadata

	TVAE		**CTGAN**		**TabDDPM**		**DGD**	
Metric	**Raw**	**Refined**	**Raw**	**Refined**	**Raw**	**Refined**	**Raw**	**Refined**
Kolmogorov-Smirnov ↑	0.8142±0.0017	**0.8180**±0.0008	0.8107±0.0010	0.8110±0.0015	0.9433±0.0009	**0.9492**±0.0029	0.9082±0.0011	**0.9439**±0.0013
Chi-square Test ↑	0.5422±0.1286	0.6130±0.0406	0.8738±0.0007	0.7518±0.1017	0.7182±0.0914	0.6775±0.0500	0.7222±0.0008	0.7225±0.0006
Jensen-Shannon Dist. ↓	0.0807±0.0008	**0.0388**±0.0012	0.0760±0.0006	**0.0373**±0.0021	0.0479±0.0012	**0.0347**±0.0013	0.0402±0.0005	**0.0362**±0.0008
Total Variation Dist. ↓	0.0684±0.0010	**0.0312**±0.0015	0.0734±0.0005	**0.0328**±0.0029	0.0356±0.0012	**0.0263**±0.0013	0.0391±0.0006	**0.0318**±0.0011
Hellinger Distance ↓	0.0845±0.0009	**0.0397**±0.0012	0.0780±0.0006	**0.0377**±0.0021	0.0499±0.0013	**0.0354**±0.0013	0.0406±0.0005	**0.0366**±0.0008
Inverse KL Divergence ↑	0.9345±0.0026	**0.9808**±0.0007	0.9604±0.0005	**0.9883**±0.0005	0.9821±0.0013	**0.9901**±0.0024	0.9860±0.0005	**0.9910**±0.0004

GOSSIS-1-eICU-cardiovascular

	TVAE		**CTGAN**		**TabDDPM**		**DGD**	
Metric	**Raw**	**Refined**	**Raw**	**Refined**	**Raw**	**Refined**	**Raw**	**Refined**
Kolmogorov-Smirnov ↑	0.8556±0.0003	**0.8738**±0.0004	0.8631±0.0004	**0.8788**±0.0003	0.8746±0.0016	**0.9707**±0.0003	0.8067±0.0008	**0.8352**±0.0006
Chi-square Statistic ↑	0.8061±0.0021	**0.8839**±0.0366	0.9061±0.0302	0.9201±0.0006	0.8096±0.0016	**0.8357**±0.0009	0.8767±0.0283	0.9022±0.0007
Jensen-Shannon Dist. ↓	0.0652±0.0005	**0.0308**±0.0002	0.0443±0.0008	**0.0269**±0.0005	0.0665±0.0007	**0.0418**±0.0005	0.0472±0.0004	**0.0406**±0.0005
Total Variation Dist. ↓	0.0595±0.0003	**0.0244**±0.0004	0.0427±0.0008	**0.0222**±0.0003	0.0596±0.0006	**0.0436**±0.0005	**0.0408**±0.0003	0.0438±0.0005
Hellinger Distance ↓	0.0667±0.0005	**0.0312**±0.0002	0.0455±0.0008	**0.0272**±0.0006	0.0681±0.0007	**0.0422**±0.0005	0.0493±0.0004	**0.0411**±0.0005
Inverse KL Divergence ↑	0.9601±0.0005	**0.9867**±0.0009	0.9705±0.0012	**0.9882**±0.0004	0.9694±0.0004	**0.9811**±0.0003	0.9807±0.0003	**0.9857**±0.0003

Inverse KL Divergence consistently enhances performance, with values moving from the 0.93-0.98 range for raw synthetic data to above 0.98 for all refined models. Notably, CTGAN shows improvements from 0.9604 to 0.9883 on TCGA-metadata and from 0.9705 to 0.9882 on GOSSIS-1-eICU-cardiovascular, while TabDDPM achieves near-perfect scores improving from 0.9821 to 0.9901 on TCGA-metadata. These consistent improvements across different distributional metrics demonstrate MAPS's effectiveness in correcting marginal distributional misalignments across various generative paradigms. Importantly, we observe improvements across all four generative models: TVAE, CTGAN, TabDDPM, and DGD while our results also reveal the inherent variability in output quality among different generation methods. This variability underscores the practical value of having MAPS as a flexible post-hoc add-on that can work with any generative method, regardless of the underlying architecture or training paradigm.

4.2 Joint Distribution and Correlation Structure

Joint distribution and correlation structure improvements via synthetic data refinement with MAPS are equally impressive.

Table 2: Joint distribution similarity and correlation structure preservation. WD = Wasserstein Distance (joint numerical distributions), JSD = Joint Jensen-Shannon Distance (joint categorical distributions), NFN = Normalized Frobenius Norm (correlation structure differences). Lower values indicate better fidelity.

TCGA-metadata

	TVAE		CTGAN		TabDDPM		DGD	
Metric	**Raw**	**Refined**	**Raw**	**Refined**	**Raw**	**Refined**	**Raw**	**Refined**
WD ↓	0.0145±0.0004	**0.0056**±0.0002	0.0130±0.0003	**0.0073**±0.0003	0.1099±0.0048	**0.0089**±0.0005	0.0291±0.0007	**0.0102**±0.0002
JSD ↓	0.7909±0.0041	**0.5562**±0.0020	0.8278±0.0034	**0.5736**±0.0028	0.8239±0.0029	**0.5434**±0.0018	0.7888±0.0045	**0.5961**±0.0040
NFN ↓	0.0770±0.0014	**0.0348**±0.0011	0.0749±0.0006	**0.0376**±0.0020	0.0927±0.0019	**0.0245**±0.0013	0.1330±0.0005	**0.0421**±0.0009

GOSSIS-1-eICU-cardiovascular

	TVAE		CTGAN		TabDDPM		DGD	
Metric	**Raw**	**Refined**	**Raw**	**Refined**	**Raw**	**Refined**	**Raw**	**Refined**
WD ↓	0.0993±0.0024	**0.0921**±0.0003	0.1221±0.0014	**0.1004**±0.0007	1.3132±0.0392	**0.2418**±0.0122	0.3159±0.0061	**0.1540**±0.0022
JSD ↓	0.7547±0.0013	**0.4187**±0.0014	0.7902±0.0022	**0.4503**±0.0010	0.7241±0.0013	**0.3633**±0.0009	0.6192±0.0016	**0.4197**±0.0014
NFN ↓	0.0546±0.0005	**0.0364**±0.0012	0.0523±0.0004	**0.0399**±0.0003	0.1190±0.0011	**0.0342**±0.0004	0.0686±0.0004	**0.0527**±0.0004

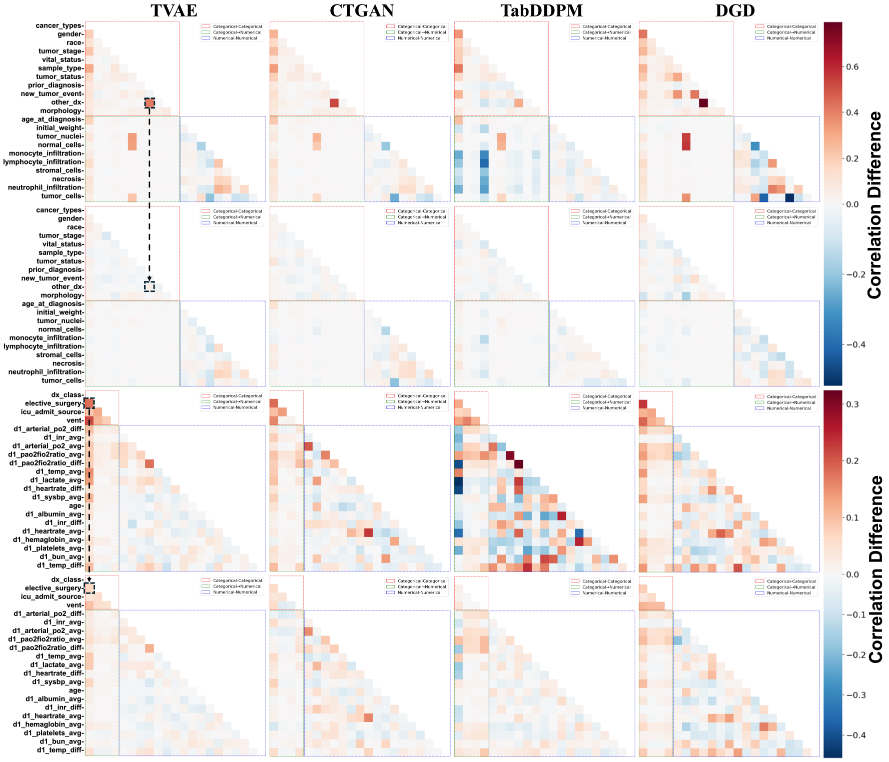

Figure 4: Correlation structure difference analysis. The figure shows 4×4 subplots where columns represent generative models. Rows 1-2 show TCGA-metadata correlation matrix differences (real vs. raw synthetic, real vs. refined synthetic), and rows 3-4 show GOSSIS-1-eICU-cardiovascular differences. Darker colors indicate larger differences from the real data correlation structure.

Table 2 reveals significant improvements across all joint distribution metrics. Joint Jensen-Shannon Distance shows substantial reductions: on TCGA-metadata, improvements range from 24.4% (DGD: 0.7888 to 0.5961) to 34.0% (TabDDPM: 0.8239 to 0.5434), with TVAE achieving 29.7% reduction (0.7909 to 0.5562). GOSSIS-1-eICU-cardiovascular shows even greater gains: TVAE improves 44.5% (0.7547 to 0.4187), CTGAN 43.0% (0.7902 to 0.4503), TabDDPM 49.8% (0.7241 to 0.3633), and DGD 32.2% (0.6192 to 0.4197).

Correlation structure preservation, measured by Normalized Frobenius Norm, shows universal improvements with reductions ranging from 23.2% to 73.6%. Figure 4 demonstrates how MAPS refinement brings synthetic data correlation structures much closer to real data patterns. For

TCGA-metadata, all four models' raw synthetic data fail to adequately capture the correlation between `other_dx` (other diagnosis) and `prior_diagnosis`, while refined synthetic data achieves nearly perfect correlation preservation. For GOSSIS-eICU-cardiovascular, correlations between `elective_surgery` and `dx_class` are weakly captured in raw synthetic data, whereas refined synthetic data demonstrates significantly enhanced correlation fidelity.

4.3 Utility Enhancement Results

The utility improvements observed across clustering and classification tasks directly address one of the most critical concerns for practitioners [18, 19]. Figure 5 demonstrates MAPS's ability to rescue seemingly unusable synthetic data—particularly TabDDPM's transformation from near-zero clustering performance (ARI: 0.0022) to strong clustering agreement (ARI: 0.7384) on TCGA-metadata. Classification task improvements are equally substantial, with F1 scores improving from ranges of 0.0866-0.2400 to 0.3043-0.3848 on TCGA-metadata and from 0.1287-0.2085 to 0.2104-0.2497 on GOSSIS-1-eICU-cardiovascular across different model-dataset combinations. Beyond predictive performance, MAPS also substantially improves feature importance preservation, a critical aspect for model interpretability in healthcare applications with correlations advancing from 0.71-0.84 to 0.92-0.99 for TCGA-metadata and from 0.75-0.78 to 0.82-0.92 for GOSSIS-1-eICU-cardiovascular (detailed analysis in Appendix F).

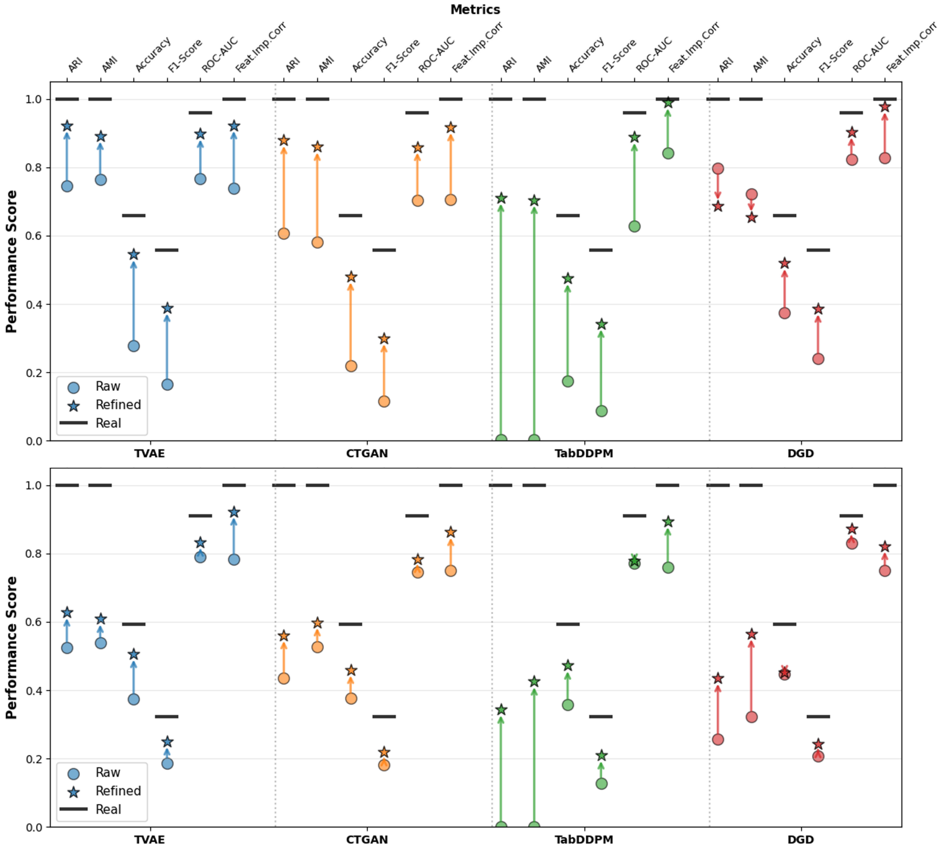

Figure 5: Downstream task performance comparison. The figure shows performance improvements after MAPS refinement across multiple tasks and metrics. Circles represent raw synthetic data performance, stars represent refined synthetic data performance, and the black horizontal line indicates the oracle performance (train on real data, test on real data). The upper and lower panels show results for TCGA-metadata and GOSSIS-1-eICU-cardiovascular datasets, respectively.

4.4 Privacy Protection Analysis

Table 3 demonstrates that MAPS maintains decent privacy protections while improving data quality. Stage 1's 0-identifiability filtering significantly reduces re-identifiability risk, decreasing the identifiability score from an average of 33.62% in raw synthetic datasets to exactly 0% across all models and datasets. However, membership inference attack (MIA) [20] tests reveal nuanced results. While most show maintained or improved privacy protection, TabDDPM and DGD exhibit increased MIA recall on TCGA-metadata (from 0.1056 to 0.2159 and 0.0781 to 0.1679 respectively). This can be attributed to inherent model characteristics: TabDDPM, as a diffusion-based model, tends to memorize training samples more strongly on small heterogeneous tabular data [21], making synthetic samples more vulnerable to MIAs. Since MAPS's importance weighting prioritizes samples closer to the real distribution, it may inadvertently select samples within TabDDPM's memorization zone. DGD's increased MIA recall warrants future investigation. The reasonable precision values (around 0.50) suggest substantial false positives rather than systematic vulnerabilities, indicating the increased recall reflects the fundamental privacy-fidelity tradeoff rather than critical security flaws. DOMIAS attack [22] results show stable performance near 0.5 (random guessing) across all combinations, indicating MAPS does not introduce density-based privacy vulnerabilities.

MAPS's 0-identifiability guarantee provides robust distance based privacy protection, with observed MIA variations reflecting the inherent privacy-fidelity tradeoff. For use cases sensitive to membership inference risk, practitioners must carefully select models or adjust MAPS settings—its modular framework allows tailoring of filtering thresholds and evaluation metrics to specific objectives.

Table 3: Privacy assessment via multiple evaluation metrics. Lower values indicate better privacy protection. IS denotes Identifiability Score (re-identification risk). Standard MIA tests membership inference; DOMIAS targets density-based vulnerabilities.

TCGA-metadata

		TVAE		CTGAN		TabDDPM		DGD	
Metric Type	**Metric**	**Raw**	**Refined**	**Raw**	**Refined**	**Raw**	**Refined**	**Raw**	**Refined**
Distance-based	**IS**	0.4033±0.0036	**0.0000**±0.0000	0.4331±0.0027	**0.0000**±0.0000	0.4465±0.0026	**0.0000**±0.0000	0.3595±0.0040	**0.0000**±0.0000
Standard MIA	**F1**	**0.3352**±0.0005	0.3370±0.0004	0.3368±0.0004	0.3370±0.0002	**0.4084**±0.0012	0.4555±0.0036	**0.3906**±0.0019	0.4391±0.0042
	Precision	0.4758±0.0513	0.4908±0.0246	0.6196±0.0491	**0.5904**±0.0472	0.5046±0.0048	0.4983±0.0072	0.4887±0.0121	0.5048±0.0106
	Recall	**0.0022**±0.0005	0.0042±0.0005	0.0036±0.0004	0.0038±0.0002	**0.1056**±0.0026	0.2159±0.0053	**0.0781**±0.0015	0.1679±0.0046
DOMIAS	**Accuracy**	**0.4945**±0.0011	0.4981±0.0007	0.4973±0.0012	0.4964±0.0011	0.4997±0.0012	**0.4976**±0.0010	**0.4975**±0.0007	0.5004±0.0008
	AUCROC	**0.4934**±0.0022	0.4983±0.0016	0.4978±0.0014	**0.4946**±0.0011	0.4989±0.0009	**0.4975**±0.0004	**0.4946**±0.0023	0.4987±0.0014

GOSSIS-1-eICU-cardiovascular

		TVAE		CTGAN		TabDDPM		DGD	
Metric Type	**Metric**	**Raw**	**Refined**	**Raw**	**Refined**	**Raw**	**Refined**	**Raw**	**Refined**
Distance-based	**IS**	0.4007±0.0018	**0.0000**±0.0000	0.2038±0.0011	**0.0000**±0.0000	0.3429±0.0037	**0.0000**±0.0000	0.0994±0.0013	**0.0000**±0.0000
Standard MIA	**F1**	0.3334±0.0001	0.3335±0.0001	0.3343±0.0002	**0.3340**±0.0002	**0.3519**±0.0005	0.3974±0.0013	0.3369±0.0005	0.3369±0.0005
	Precision	0.2000±0.1458	0.4283±0.0811	0.5797±0.0716	0.5318±0.0570	**0.4902**±0.0102	0.5147±0.0050	0.5560±0.0473	0.5346±0.0273
	Recall	**0.0001**±0.0000	0.0002±0.0001	0.0010±0.0002	**0.0008**±0.0001	**0.0223**±0.0004	0.0849±0.0016	0.0038±0.0004	0.0040±0.0005
DOMIAS	**Accuracy**	0.4990±0.0009	0.4996±0.0005	0.4999±0.0007	0.4999±0.0003	0.4989±0.0003	0.4995±0.0004	0.4991±0.0005	0.4994±0.0006
	AUCROC	0.4984±0.0005	0.4990±0.0003	0.4994±0.0004	0.4992±0.0002	**0.4971**±0.0001	0.4991±0.0002	0.5008±0.0001	**0.5002**±0.0003

5 Conclusion

We present MAPS, a model-agnostic framework for post-hoc synthetic data refinement providing formal 0-identifiability guarantees while improving data quality through importance weighting and resampling. Comprehensive evaluation demonstrates consistent improvements in fidelity, utility, and privacy protection across healthcare datasets and multiple generative models. MAPS's modular design combines sample level and whole set refinement, enabling practitioners to customize components: alternative distance measures in Stage 1 (e.g., DCR [23] or NNDR [23]), different classification architectures in Stage 2, and privacy controls from 0-identifiability to ϵ-identifiability levels, addressing diverse domain needs. However, deploying high-quality synthetic data requires taking greater responsibility regardless of privacy guarantees, and risk assessment should always be conducted [24, 25].

Future work could explore broader applicability to privacy-first generative models and extend the framework to time-series synthetic data. An intriguing direction would be incorporating MAPS's selection criteria directly into the generation process, training generative models to preferentially produce high quality samples that would pass both privacy filtering and importance weighting stages, eliminating the need for large synthetic pools while maintaining refinement benefits.

Acknowledgments

This work was supported by grants NNF20OC0063268 and NNF23OC0083510 from the Novo Nordisk Foundation. Anders Krogh was further supported by grants NNF20OC0062606 and NNF20OC0059939. We thank Martin Bøgsted, Rasmus Rask Kragh Jørgensen, Jakob Bruhn Krøjgaard Skelmose, Jan Reiter Sørensen, and Andreas Bjerregaard Jeppesen for their helpful discussions.

References

[1] Boris van Breugel, Tennison Liu, Dino Oglic, and Mihaela van der Schaar. Synthetic data in biomedicine via generative artificial intelligence. *Nature Reviews Bioengineering*, 2(12):991–1004, 2024.

[2] Allan Tucker, Zhenchen Wang, Ylenia Rotalinti, and Puja Myles. Generating high-fidelity synthetic patient data for assessing machine learning healthcare software. *NPJ digital medicine*, 3(1):147, 2020.

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

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

[5] Jinsung Yoon, Lydia N Drumright, and Mihaela Van Der Schaar. Anonymization through data synthesis using generative adversarial networks (ads-gan). *IEEE journal of biomedical and health informatics*, 24(8):2378–2388, 2020.

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

[7] Brian Belgodere, Pierre Dognin, Adam Ivankay, Igor Melnyk, Youssef Mroueh, Aleksandra Mojsilovic, Jiri Navratil, Apoorva Nitsure, Inkit Padhi, Mattia Rigotti, et al. Auditing and generating synthetic data with controllable trust trade-offs. *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, 2024.

[8] Debalina Padariya, Isabel Wagner, Aboozar Taherkhani, and Eerke Boiten. Privacy-preserving generative models: A comprehensive survey. *arXiv preprint arXiv:2502.03668*, 2025.

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

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

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

[12] Viktoria Schuster and Anders Krogh. The deep generative decoder: Map estimation of representations improves modelling of single-cell rna data. *Bioinformatics*, 39(9):btad497, 2023.

[13] Aditya Grover, Jiaming Song, Ashish Kapoor, Kenneth Tran, Alekh Agarwal, Eric J Horvitz, and Stefano Ermon. Bias correction of learned generative models using likelihood-free importance weighting. *Advances in neural information processing systems*, 32, 2019.

[14] Georgi Ganev and Emiliano De Cristofaro. The inadequacy of similarity-based privacy metrics: Privacy attacks against "truly anonymous" synthetic datasets. In *2025 IEEE Symposium on Security and Privacy (SP)*, pages 4007–4025. IEEE, 2025.

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

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

[17] Nadir Sella, Florent Guinot, Nikita Lagrange, Laurent-Philippe Albou, Jonathan Desponds, and Hervé Isambert. Preserving information while respecting privacy through an information theoretic framework for synthetic health data generation. *npj Digital Medicine*, 8(1):49, 2025.

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

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

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

[21] Yixin Liu, Thalaiyasingam Ajanthan, Hisham Husain, and Vu Nguyen. Self-supervision improves diffusion models for tabular data imputation. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pages 1513–1522, 2024.

[22] Boris Van Breugel, Hao Sun, Zhaozhi Qian, and Mihaela van der Schaar. Membership inference attacks against synthetic data through overfitting detection. *arXiv preprint arXiv:2302.12580*, 2023.

[23] Zilong Zhao, Aditya Kunar, Robert Birke, and Lydia Y Chen. Ctab-gan: Effective table data synthesizing. In *Asian conference on machine learning*, pages 97–112. PMLR, 2021.

[24] Jennifer A Bartell, Sander Boisen Valentin, Anders Krogh, Henning Langberg, and Martin Bøgsted. A primer on synthetic health data. *arXiv preprint arXiv:2401.17653*, 2024.

[25] Jelena Schmidt, Nienke M Schutte, Stefan Buttigieg, David Novillo-Ortiz, Eric Sutherland, Michael Anderson, Bart de Witte, Michael Peolsson, Brigid Unim, Milena Pavlova, et al. Mapping the regulatory landscape for artificial intelligence in health within the european union. *npj Digital Medicine*, 7(1):229, 2024.

[26] Daniel Jarrett, Bogdan Cebere, Tennison Liu, Alicia Curth, and Mihaela van der Schaar. Hyperimpute: Generalized iterative imputation with automatic model selection. 2022.

[27] J Raffa, A Johnson, T Pollard, and B Omar. Gossis-1-eicu, the eicu-crd subset of the global open source severity of illness score (gossis-1) dataset (version 1.0. 0), 2022.

[28] Zhaozhi Qian, Bogdan-Constantin Cebere, and Mihaela van der Schaar. Synthcity: facilitating innovative use cases of synthetic data in different data modalities. *arXiv preprint arXiv:2301.07573*, 2023.

[29] Anton D. Lautrup, Tobias Hyrup, Arthur Zimek, and Peter Schneider-Kamp. Syntheval: a framework for detailed utility and privacy evaluation of tabular synthetic data. *Data Mining and Knowledge Discovery*, 39(1), 2024.

[30] Jean Feydy, Thibault Séjourné, François-Xavier Vialard, Shun-ichi Amari, Alain Trouve, and Gabriel Peyré. Interpolating between optimal transport and mmd using sinkhorn divergences. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 2681–2690, 2019.

Appendix

A MAPS Algorithm

Algorithm 1 MAPS: Model Agnostic Post-hoc Synthetic Data Refinement

```
Require: Real dataset $\mathcal{D} = \{x_i\}_{i=1}^{N}$, Synthetic dataset $\hat{\mathcal{D}} = \{\hat{x}_j\}_{j=1}^{M}$, Target size $N$
Ensure: Refined synthetic dataset $\tilde{\mathcal{D}}$ with $|\tilde{\mathcal{D}}| = N$
// Stage 1: 0-Identifiability Guarantee
Compute feature weights $\mathbf{w}$ based on data characteristics
for each real sample $x_i \in \mathcal{D}$ do
    $r_i \leftarrow \min_{x_j \in \mathcal{D} \setminus \{x_i\}} \|\mathbf{w} \cdot (x_i - x_j)\|$ {Distinctness threshold}
end for
$\mathcal{D}_{filtered} \leftarrow \emptyset$ {Initialize filtered synthetic dataset}
for each synthetic sample $\hat{x}_j \in \hat{\mathcal{D}}$ do
    $is\_safe \leftarrow True$
    for each real sample $x_i \in \mathcal{D}$ do
        if $\|\mathbf{w} \cdot (x_i - \hat{x}_j)\| < r_i$ then
            $is\_safe \leftarrow False$ {Too close to real sample}
            break
        end if
    end for
    if $is\_safe$ then
        $\mathcal{D}_{filtered} \leftarrow \mathcal{D}_{filtered} \cup \{\hat{x}_j\}$
    end if
end for
// Stage 2: Fidelity Enhancement via Importance Weighting
Randomly sample $\mathcal{D}_{filtered\_train} \subset \mathcal{D}_{filtered}$ with $|\mathcal{D}_{filtered\_train}| = N$
$\mathcal{D}_{remaining} \leftarrow \mathcal{D}_{filtered} \setminus \mathcal{D}_{filtered\_train}$ {Samples for SIR}
Create training set: $\mathcal{X}_{train} = \mathcal{D} \cup \mathcal{D}_{filtered\_train}$
Create labels: $y_i = 1$ for $x_i \in \mathcal{D}$, $y_j = 0$ for $\hat{x}_j \in \mathcal{D}_{filtered\_train}$
Train binary classifier $c_\phi$ on $(\mathcal{X}_{train}, \mathbf{y})$
$\pi_1 \leftarrow |\mathcal{D}|/|\mathcal{X}_{train}|$, $\pi_0 \leftarrow |\mathcal{D}_{filtered\_train}|/|\mathcal{X}_{train}|$
for each $\hat{x}_j \in \mathcal{D}_{remaining}$ do
    $p_j \leftarrow c_\phi(\hat{x}_j)$ {Probability of being real}
    $w_j \leftarrow \frac{\pi_0}{\pi_1} \cdot \frac{p_j}{1-p_j}$ {Importance weight}
end for
// Stage 3: Sampling-Importance-Resampling (SIR)
$W_{total} \leftarrow \sum_{\hat{x}_j \in \mathcal{D}_{remaining}} w_j$
for each $\hat{x}_j \in \mathcal{D}_{remaining}$ do
    $p_j^{norm} \leftarrow w_j / W_{total}$ {Normalized sampling probability}
end for
$\tilde{\mathcal{D}} \leftarrow \emptyset$
for $n = 1$ to $N$ do
    Sample index $j$ from $\mathcal{D}_{remaining}$ with probabilities $\{p_j^{norm}\}$
    $\tilde{\mathcal{D}} \leftarrow \tilde{\mathcal{D}} \cup \{\hat{x}_j\}$
end for
return $\tilde{\mathcal{D}}$
```

B Implementation Details

B.1 Data Preprocessing

For TCGA-metadata, missing values are imputed using the ICE imputer from the HyperImpute library [26] before generative model training. For GOSSIS-1-eICU-cardiovascular, we utilize the preprocessed `gossis-1-eicu-only-model-ready.csv.gz` file [27], which contains no missing values.

B.2 Model training

All generative models except DGD are implemented using the Synthcity library [28] with default hyperparameters. TabDDPM is configured with 10,000 training iterations, while DGD is trained with a learning rate of 1e-2, 50 latent dimensions, and 50 GMM components. For TCGA-metadata, DGD uses 2000 epochs; for GOSSIS-1-eICU-cardiovascular, this is reduced to 1000 epochs. All experiments employ a 60:40 train-test split.

B.3 Fidelity Classifier Training

The fidelity classifier employs a 2-layer MLP architecture with hidden layer sizes adapted to dataset characteristics: 240×120 for TCGA-metadata and 480×240 for GOSSIS-1-eICU-cardiovascular. During training, we use balanced sampling (N real samples, N synthetic samples) to ensure $\pi_0/\pi_1 = 1$ in Equation (5). The training data is split 80:20 for training and evaluation. For data preprocessing prior to training, the classifier uses min-max normalization for numerical variables and integer encoding for categorical variables. Training employs the Adam optimizer with learning rate 0.001, batch size 64, and early stopping with patience of 10 epochs to prevent overfitting.

B.4 Importance weights post-processing

The output of the fidelity classifier provides our importance weights, which are subsequently normalized and used for probability-proportional sampling. However, raw importance weights often exhibit significant skewness, where the majority of synthetic samples receive very small weights while a few samples obtain disproportionately large weights. This distribution can lead to problematic oversampling of only a handful of synthetic samples, thereby failing to capture the holistic distributional properties of the real data and potentially increasing sampling variance.

To address this issue, we employ a flattening transformation as suggested by [13]. The flattening process applies a power transformation to the raw importance weights:

$$\hat{w}_{\text{flattened}}(x) = \hat{w}_\phi(x)^\alpha \tag{8}$$

where $\hat{w}_\phi(x)$ represents the original importance weight estimated by the fidelity classifier, and α is a flattening parameter that controls the degree of variance reduction. When $\alpha = 1$, the weights remain unchanged, while smaller values of α increasingly flatten the weight distribution by compressing the range between high and low weights, conversely, larger values of α amplify the differences between weights.

This transformation serves multiple purposes: (1) it tunes the variance of importance weights, leading to more controllable sampling behavior; (2) it prevents extreme weights from dominating the resampling process; and (3) it ensures broader coverage of the synthetic data space while maintaining the relative preference for higher-fidelity samples. The flattening parameter α provides a tunable trade-off between importance weighting effectiveness and sampling diversity.

In our implementation, we employ dataset-specific flattening parameters determined empirically: $\alpha = 1.4$ for all generative models on the TCGA-metadata dataset, and $\alpha = 0.8$ for all models on the GOSSIS-1-eICU-cardiovascular dataset.

B.5 Computational Requirements and Pool Size Considerations

While generating synthetic data is computationally cheap with modern hardware, MAPS requires maintaining a larger synthetic pool for effective refinement. In our experiments, we generate $M = 30N$ synthetic samples using a single NVIDIA A30 GPU, of which approximately 50-60% are filtered in Stage 1 due to 0-identifiability violations, leaving $M' \approx 12N$-$15N$ samples for importance weighting in Stage 2.

The choice of pool size presents a practical tradeoff. Larger pools provide more diverse candidates for the importance weighting stage, enabling better approximation of the true data distribution through selective resampling. However, this comes at the cost of increased memory requirements during both generation and processing stages. For datasets of comparable size to ours (11,315 and 41,396 samples), generating and processing a $30N$ pool is feasible on standard hardware.

For resource constrained settings, practitioners can reduce the pool size (e.g., $M = 5N$ or $10N$) to decrease memory footprint while still benefiting from MAPS's two-stage refinement. Smaller pools will naturally provide less dramatic improvements since the importance weighting stage has fewer high-quality candidates to select from, but the 0-identifiability filtering in Stage 1 remains effective regardless of pool size. The modular nature of MAPS allows users to balance computational constraints against desired fidelity improvements based on their specific application requirements.

C Evaluation details

To comprehensively assess the effectiveness of MAPS across multiple dimensions of synthetic data quality, we employ a multi-faceted evaluation framework that systematically measures improvements in distributional fidelity, downstream task utility, and privacy protection. Our evaluation protocol is designed to capture both marginal and joint distributional properties while ensuring that improvements translate to practical utility in real-world applications. The evaluation framework encompasses three primary dimensions: (1) fidelity metrics that quantify how well synthetic data approximates the statistical properties of real data, (2) utility metrics that assess performance on downstream tasks using a rigorous "train-on-synthetic, test-on-real" protocol, and (3) privacy metrics that ensure refinement does not compromise data privacy protections.

C.1 Fidelity evaluation metrics

Fidelity evaluation forms the cornerstone of our assessment framework, as it directly measures whether MAPS successfully improves the statistical alignment between synthetic and real data. We employ a comprehensive suite of distributional similarity metrics that capture both marginal and joint distributional properties across mixed-type tabular data. The selection of these metrics follows established practices in synthetic data evaluation literature [28, 29, 11], ensuring comprehensive coverage of distributional aspects critical for tabular synthetic data quality assessment. This multi-metric approach ensures robust evaluation across different aspects of distributional fidelity, from univariate marginals to complex multivariate relationships.

C.1.1 Quantifying the marginal distribution similarity

We employ 6 complementary metrics to assess marginal distribution similarity, each targeting specific aspects of distributional alignment:

Kolmogorov-Smirnov Test Statistic: This non-parametric test measures the maximum absolute difference between cumulative distribution functions of real and synthetic data. Applied exclusively to numerical variables, it ranges from 0 to 1, where higher values indicate better distributional alignment (we report 1 - KS statistic for interpretability).

Chi-square Test: This statistical test evaluates the independence hypothesis between real and synthetic categorical distributions through contingency table analysis. Applied to categorical variables, p-values range from 0 to 1, where higher p-values mean weaker evidence against the null hypothesis of distributional equality. When p-values are high, we fail to reject the null hypothesis, suggesting insufficient evidence to conclude that the distributions differ significantly.

Jensen-Shannon Distance: This symmetric divergence measure quantifies the similarity between two probability distributions as the square root of Jensen-Shannon divergence. Applied to categorical variables (or discretized numerical variables), it ranges from 0 to 1, where lower values indicate better similarity.

Total Variation Distance: This metric computes half of the L1 distance between two probability mass functions, representing the overall difference in probability assignments. Applied to categorical variables, it ranges from 0 to 1, where lower values indicate better similarity.

Hellinger Distance: This metric measures the similarity between probability distributions based on the Euclidean distance between their square-rooted probability vectors. Applied to categorical variables, it ranges from 0 to 1, where lower values indicate better similarity.

Inverse KL Divergence: This metric transforms the Kullback-Leibler divergence using the formula $1/(1 + \mathrm{KL}(P||Q))$ to provide a bounded similarity measure. Applied to categorical variables, it ranges from 0 to 1, where higher values indicate better similarity.

These 6 metrics collectively provide a comprehensive assessment of marginal distribution similarity by capturing different aspects of distributional alignment across both numerical and categorical variables, ensuring robust evaluation of synthetic data fidelity at the univariate level.

C.1.2 Quantifying the joint distribution similarity

For joint distribution assessment, we employ two specialized metrics that capture multivariate relationships:

Wasserstein Distance (WD): This optimal transport-based metric measures the minimum cost of transforming one distribution into another in the joint numerical feature space. Applied to all numerical variables simultaneously using the Sinkhorn algorithm for computational efficiency, it ranges from 0 to infinity, where lower values indicate better similarity. The implementation is provided by the GeomLoss package [30] for efficient computation of optimal transport distances.

Joint Jensen-Shannon Distance (JSD): This metric extends Jensen-Shannon divergence to joint categorical distributions by computing divergence over the Cartesian product of all categorical variable combinations. Applied to all categorical variables simultaneously, it ranges from 0 to 1, where lower values indicate better similarity.

These two joint distribution metrics complement the marginal assessments by capturing multivariate dependencies and interaction patterns that are crucial for downstream task performance, providing a complete picture of distributional fidelity across the feature space.

C.2 Mixed-type correlation matrix

We compute correlation matrices using Spearman correlation coefficients for numerical-numerical relationships, Cramér's V for categorical-categorical associations, and correlation ratios for categorical-numerical relationships. The association matrix is not symmetrical due to the asymmetry of correlation ratios.

Normalized Frobenius Norm (NFN): This metric quantifies the difference between correlation structures by computing the Frobenius norm of the difference between real and synthetic correlation matrices, normalized by the square root of the total number of matrix elements. It ranges from 0 to infinity, where lower values indicate better correlation structure preservation.

C.3 Utility Evaluation Protocol

Our utility evaluation employs a rigorous "train-on-synthetic, test-on-real" protocol that directly assesses whether synthetic data can replace real data for downstream applications. This evaluation strategy reflects real-world usage scenarios where synthetic data would be used for model development before deployment on real data.

For clustering evaluation, we use K-means with k=5 clusters (determined by elbow analysis) on numerical variables only. The evaluation protocol splits real data 80:20, with clustering models trained on synthetic data and evaluated on the real test set using Adjusted Rand Index (ARI) and Adjusted Mutual Information (AMI) metrics against true class labels. These metrics range from 0 to 1 (with ARI potentially negative for very poor clusterings), where higher values indicate better clustering agreement with ground truth labels.

Classification tasks employ Random Forest with 100 estimators, using the same train-test split strategy. For TCGA-metadata, the target variable is `cancer_types` while for GOSSIS-1-eICU-cardiovascular, the target variable is `dx_class`. We evaluate performance using accuracy, F1-score (macro-averaged), and ROC-AUC metrics, all ranging from 0 to 1 with higher values indicating better performance.

Feature importance correlation measures Spearman correlation between importance rankings derived from models trained on real data versus synthetic data, providing insight into interpretability preservation. This metric ranges from -1 to 1, where values closer to 1 indicate better preservation of feature relationships crucial for model interpretation. We included this metric because maintaining interpretable feature relationships is critical for applications in sensitive domains like healthcare, where practitioners require consistent and explainable model behavior.

C.4 Privacy Evaluation Protocol

Privacy assessment employs multiple approaches to ensure comprehensive evaluation of data protection guarantees. We evaluate both formal identifiability guarantees through identifiability measures and empirical privacy resistance through attack based assessments.

Identifiability Score (IS): This metric directly measures the proportion of real samples that can be identified through synthetic data, computed as the fraction of real samples whose nearest synthetic neighbor is closer than their nearest real neighbor. Values range from 0 to 1, where 0 indicates perfect 0-identifiability (no real samples pair with a synthetic sample more closely than they pair with a real sample), a value of 1 indicates maximum identifiability risk, where every real sample has a synthetic counterpart that is closer to it than its nearest real neighbor. The IS implementation is provided by the Synthcity library [28].

Membership Inference Attack (MIA) Resistance: We evaluate resistance to standard membership inference attacks using precision, recall, and F1-score metrics, where lower values indicate better privacy protection. These attacks attempt to determine whether specific samples were included in the training data used to generate synthetic samples. The standard MIA implementation is provided by the SynthEval package [29], with evaluation conducted using stratified cross validation with 5 folds to ensure robust privacy assessment across different data splits.

DOMIAS Attack Resistance: We assess resistance to Density-based Membership Inference Attacks through accuracy and AUC-ROC metrics, where values closer to 0.5 (random guessing) indicate better privacy protection. This specialized attack targets density-based vulnerabilities. The DOMIAS attack implementation is provided by the Synthcity package [28].

D PCA plots

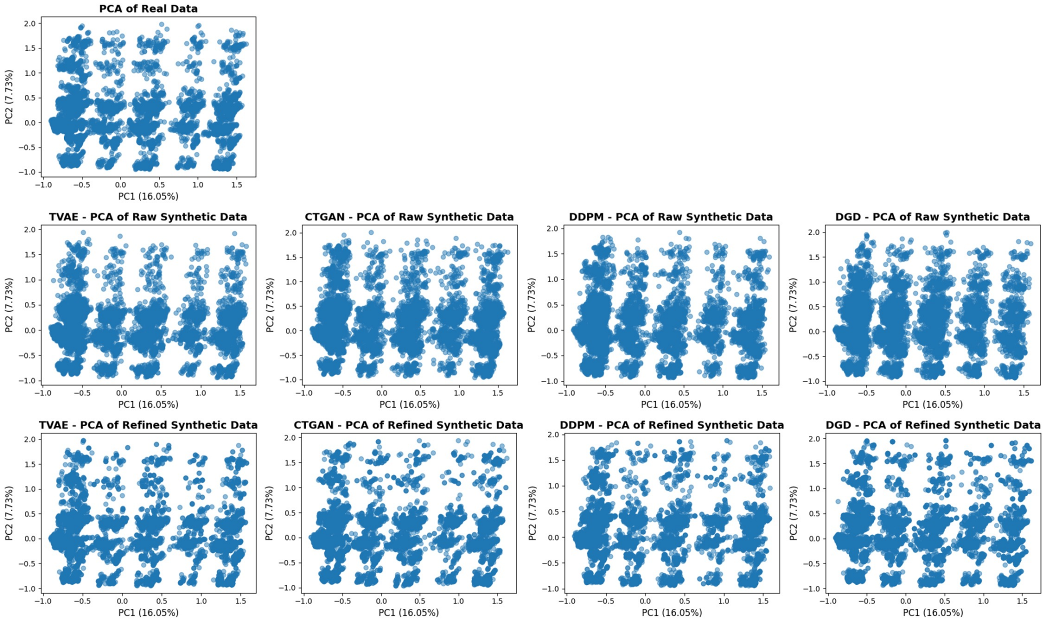

Figure 6: PCA visualization of TCGA-metadata dataset showing the distribution of real data, raw synthetic data, and refined synthetic data in the first two principal components.

The PCA visualizations in Figures 6 and 7 demonstrate how MAPS refinement brings the synthetic data distribution closer to the real data distribution in the principal component space, providing visual confirmation of the quantitative improvements observed in our fidelity metrics.

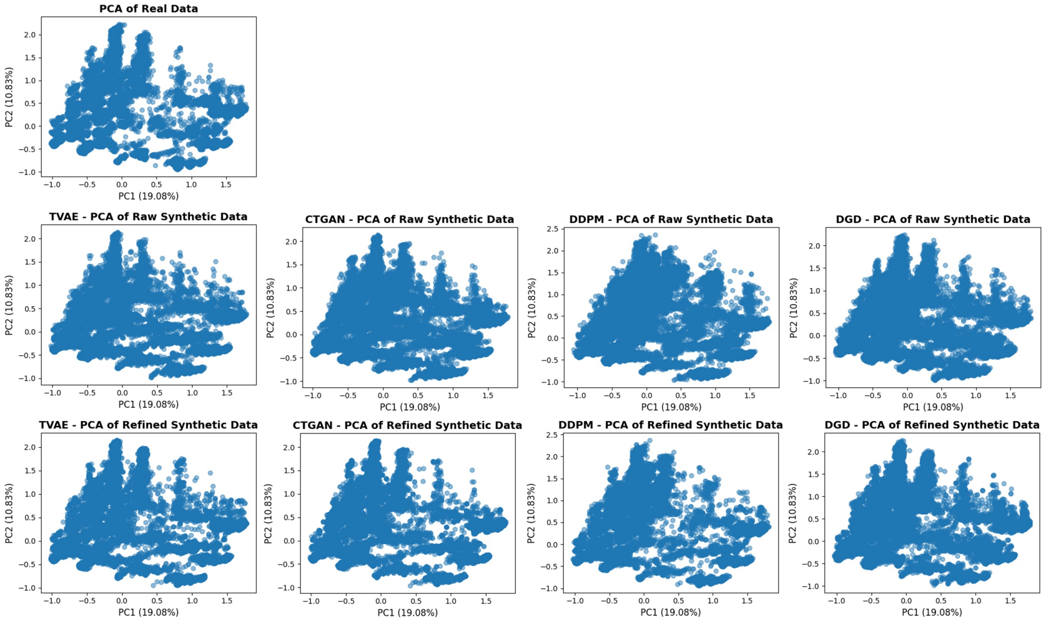

Figure 7: PCA visualization of GOSSIS-1-eICU-cardiovascular dataset showing the distribution of real data, raw synthetic data, and refined synthetic data in the first two principal components.

E Class Distribution Preservation on GOSSIS-1-eICU-cardiovascular

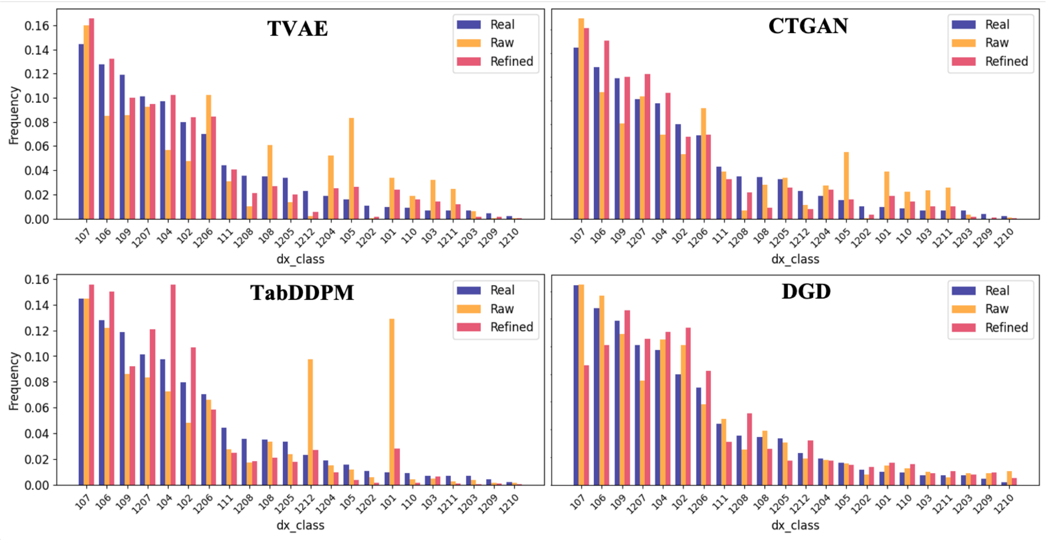

Figure 8: Class distribution fidelity across generative models on GOSSIS-1-eICU-cardiovascular dataset. The 2×2 subplots compare the frequency distribution of diagnostic classes (`dx_class`) between real data, raw synthetic data, and MAPS-refined synthetic data across four generative models. Diagnostic classes are ordered by their frequency in the real data, with class codes representing different cardiovascular conditions.

Figure 8 extends the validation of MAPS's class distribution preservation capabilities to the GOSSIS-1-eICU-cardiovascular dataset, which presents a more complex challenge with its diverse diagnostic classes. The raw synthetic outputs reveal significant distributional deviations across all four baseline models except DGD, with particularly pronounced errors in specific diagnostic classes: class 105, (TVAE, CTGAN) and classes 1211, 101 (TabDDPM). Following MAPS refinement, these distortions are substantially mitigated, as the refined distributions align closely with real class frequencies throughout the entire range: from the most common to the rarest diagnostic categories.

F Feature Importance Preservation Analysis

The feature importance correlation metric measures how well synthetic data preserves the interpretability aspects crucial for healthcare applications. Figure 9 shows that all models demonstrate substantial improvements, with feature importance correlations improving from the 0.71-0.84 range to 0.92-0.99 range for TCGA-metadata, while for GOSSIS-eICU-Cardiovascular dataset, correlations advance from 0.75-0.78 to 0.82-0.92. This indicates that MAPS refined data not only performs better quantitatively but also maintains the feature relationships that clinicians rely on for model interpretation. For instance, examining the feature with the highest importance rank—`tumor_stage` for TCGA-metadata and `d1_arterial_po2_diff` for GOSSIS-eICU-cardiovascular—we observe significant improvements across all models. The arrows demonstrate the progression from raw to refined synthetic data: for TCGA-metadata, the feature importance increases substantially from 0.07 to 0.12 after MAPS refinement, approaching the real data value of 0.14. Similarly, for the GOSSIS-eICU-cardiovascular dataset, the importance value is elevated from 0.03 to 0.04, moving closer to the real data's 0.05.

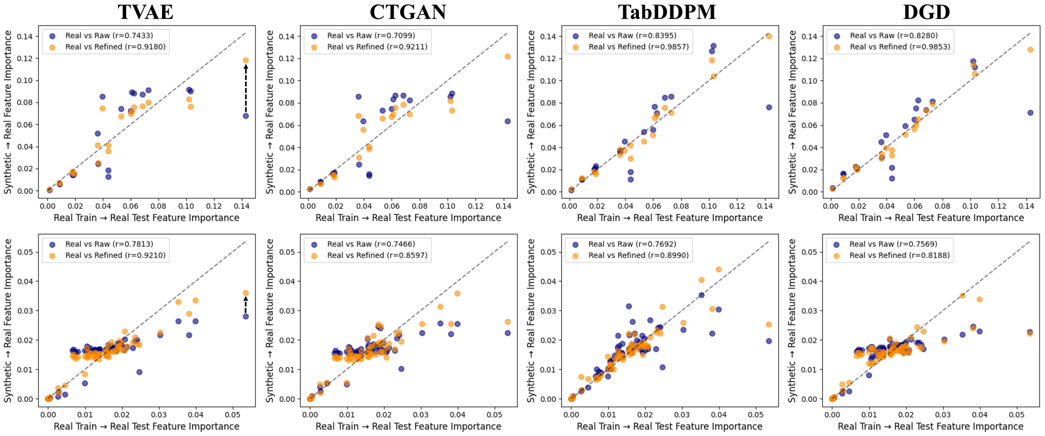

Figure 9: Feature importance correlation analysis. The figure shows the correlation between feature importance rankings derived from models trained on real data versus those trained on synthetic data. Higher correlations indicate better preservation of model interpretability. The first row presents results on TCGA-metadata, and the second row presents results on GOSSIS-1-eICU-cardiovascular.

G Statistical Significance Analysis

Statistical significance testing was performed using paired t-tests across 5 experimental runs. The paired t-test is appropriate for our experimental design as it compares performance between raw and refined synthetic data on the same underlying generative models and datasets.

For each metric, we computed the mean improvement, standard deviation, and p-values across independent experimental runs. Results with $p < 0.05$ are considered statistically significant.

H Detailed Utility Results

Table 4 shows the detailed numerical utility results corresponding to Figure 5.

I Mortality Prediction

To complement the multi-class prediction results presented in the main paper, we evaluate MAPS on mortality prediction, a critical task in clinical research and patient care. While the main results

Table 4: Comprehensive downstream task performance analysis. Results demonstrate MAPS effectiveness across clustering, classification, and feature importance preservation tasks using the "train-on-synthetic, test-on-real" evaluation protocol. The Oracle column represents the upper-bound performance achieved by training and testing on real data, serving as the ground truth benchmark for these tasks.

TCGA-metadata

		TVAE		**CTGAN**		**TabDDPM**		**DGD**		**Oracle**
Task	**Metric**	**Raw**	**Refined**	**Raw**	**Refined**	**Raw**	**Refined**	**Raw**	**Refined**	**Real**
K-Means	**ARI**	0.7458±0.1759	0.8921±0.0082	0.6069±0.0163	**0.8706**±0.0107	0.0022±0.0006	**0.7384**±0.0127	**0.7965**±0.0052	0.6950±0.0181	1.0000
	AMI	0.7654±0.1011	0.8740±0.0057	0.5815±0.0071	**0.8543**±0.0070	0.0020±0.0004	**0.6923**±0.0129	**0.7224**±0.0075	0.6465±0.0222	1.0000
Classification	**Accuracy**	0.2780±0.0160	**0.5398**±0.0157	0.2192±0.0058	**0.4807**±0.0083	0.1742±0.0136	**0.4775**±0.0074	0.3749±0.0040	**0.5173**±0.0167	0.6589±0.0079
	F1 Score	0.1647±0.0101	**0.3848**±0.0234	0.1152±0.0044	**0.3043**±0.0049	0.0866±0.0145	**0.3372**±0.0195	0.2400±0.0093	**0.3707**±0.0189	0.5576±0.0081
	ROC-AUC	0.7664±0.0010	0.8914±0.0077	0.7022±0.0034	**0.8547**±0.0080	0.6272±0.0114	**0.8804**±0.0035	0.8234±0.0083	**0.9082**±0.0037	0.9590±0.0027
	Feat. Imp. Corr.	0.7393±0.0035	**0.9168**±0.0037	0.7067±0.0062	**0.9199**±0.0053	0.8429±0.0030	**0.9854**±0.0012	0.8268±0.0037	**0.9850**±0.0017	1.0000

GOSSIS-1-eICU-cardiovascular

		TVAE		**CTGAN**		**TabDDPM**		**DGD**		**Oracle**
Task	**Metric**	**Raw**	**Refined**	**Raw**	**Refined**	**Raw**	**Refined**	**Raw**	**Refined**	**Real**
K-Means	**ARI**	0.5257±0.0572	0.5810±0.0977	0.4345±0.0015	**0.5476**±0.0192	0.0000±0.0000	**0.3075**±0.0043	0.2569±0.0079	**0.4287**±0.0055	1.0000
	AMI	0.5380±0.0261	0.5923±0.0692	0.5272±0.0031	**0.5848**±0.0156	0.0000±0.0000	**0.3879**±0.0038	0.3236±0.0068	**0.5472**±0.0116	1.0000
Classification	**Accuracy**	0.3745±0.0063	**0.5008**±0.0031	0.3763±0.0042	**0.4545**±0.0035	0.3574±0.0073	**0.4726**±0.0039	0.4476±0.0086	0.4593±0.0107	0.5922±0.0023
	F1 Score	0.1867±0.0043	**0.2497**±0.0016	0.1812±0.0067	**0.2144**±0.0019	0.1287±0.0043	**0.2104**±0.0022	0.2085±0.0080	**0.2466**±0.0041	0.3234±0.0020
	ROC-AUC	0.7895±0.0058	**0.8329**±0.0022	0.7450±0.0027	0.7867±0.0033	0.7722±0.0042	**0.7837**±0.0057	0.8308±0.0029	**0.8711**±0.0051	0.9099±0.0021
	Feat. Imp. Corr.	0.7821±0.0020	**0.9197**±0.0018	0.7492±0.0045	**0.8511**±0.0058	0.7597±0.0087	**0.8863**±0.0074	0.7497±0.0049	**0.8239**±0.0043	1.0000

demonstrate MAPS effectiveness on the more challenging multi-class problem, these mortality prediction results confirm that the framework also provides consistent improvements on this important binary classification task, further validating its broad utility enhancement across different tasks. See below Table 5 for detailed results.

Table 5: Mortality prediction performance on **TCGA-metadata** and **GOSSIS-1-eICU-cardiovascular**. The Oracle column represents the upper-bound performance achieved by training and testing on real data, serving as the ground truth benchmark for these tasks.

TCGA-metadata

		TVAE		**CTGAN**		**TabDDPM**		**DGD**		**Oracle**
Task	**Metric**	**Raw**	**Refined**	**Raw**	**Refined**	**Raw**	**Refined**	**Raw**	**Refined**	**Real**
Mortality Prediction	**Accuracy**	0.8293±0.0061	0.8307±0.0046	0.8245±0.0023	0.8308±0.0062	0.8346±0.0068	0.8370±0.0041	0.8010±0.0039	**0.8184**±0.0032	0.8563±0.0028
	F1 Score	0.7871±0.0099	0.7939±0.0060	0.7933±0.0031	0.7937±0.0088	0.8003±0.0094	0.8035±0.0057	0.7364±0.0095	**0.7772**±0.0058	0.8273±0.0029
	ROC-AUC	0.8847±0.0048	0.8853±0.0043	0.8716±0.0062	**0.8790**±0.0081	0.8875±0.0075	0.8889±0.0044	0.8625±0.0075	**0.8730**±0.0045	0.9103±0.0034
	Feat. Imp. Corr.	0.9662±0.0041	0.9694±0.0031	0.9546±0.0083	**0.9678**±0.0072	0.9822±0.0091	0.9885±0.0105	0.4319±0.0322	**0.9617**±0.0056	1.0000

GOSSIS-1-eICU-cardiovascular

		TVAE		**CTGAN**		**TabDDPM**		**DGD**		**Oracle**
Task	**Metric**	**Raw**	**Refined**	**Raw**	**Refined**	**Raw**	**Refined**	**Raw**	**Refined**	**Real**
Mortality Prediction	**Accuracy**	0.9361±0.0017	**0.9394**±0.0011	0.9350±0.0021	0.9363±0.0019	0.9366±0.0007	**0.9393**±0.0010	0.9285±0.0017	**0.9371**±0.0010	0.9405±0.0010
	F1 Score	0.7191±0.0132	**0.7550**±0.0067	0.7270±0.0144	**0.7422**±0.0082	0.7212±0.0068	**0.7617**±0.0056	0.6207±0.0176	**0.7236**±0.0049	0.7533±0.0073
	ROC-AUC	0.9048±0.0047	**0.9096**±0.0052	0.9016±0.0036	0.9027±0.0055	0.9150±0.0055	0.9167±0.0045	0.8965±0.0063	**0.9112**±0.0053	0.9172±0.0038
	Feat. Imp. Corr.	0.9062±0.0175	**0.9662**±0.0093	0.8749±0.0155	**0.9182**±0.0208	0.5222±0.0865	**0.9662**±0.0145	0.8657±0.0158	**0.9563**±0.0096	1.0000