A Reviewer 2 Experimental Results

A.1 Additional Dataset and Metrics

Texas. The Texas Hospital Discharge dataset[1] is a large public use data file provided by the Texas Department of State Health Services. We used the preprocessed which consists of 100,000 records uniformly selected from a pre-processed file containing patient data from 2013[2] version from Stadler et al. (2022). We retain 18 attributes and assume a binary classification task by predicting only minor and major mortality risk following the setup of Afonja et al. (2023). Duplicates where also removed. The final size of the dataset was therefore reduced to 75,105 which was split to non-overlapping train/test/validation. Validation size is fixed to 1000 for all dataset.

	# Train	# Test	# N	# C	# Total	Class Ratio
Adult	30932	16858	6	9	15	76:24
Airline	103904	24976	19	5	24	57:43
Texas	60127	13978	7	11	18	81:19

Table 1: **Tabular Dataset statistics.** # N and # C are the numbers of numerical and categorical columns, respectively.

Pairwise Correlation Similarity Accuracy (CorAcc). We evaluate the correlation between data columns using the approach described by Tao et al. (2021) and Afonja et al. (2023). Specifically, we use Cramer's V with bias correction for categorical columns, the Correlation Ratio for numerical-categorical columns, and the Pearson Correlation Coefficient (absolute values) for numerical columns. The ranges for these measures are as follows: Cramer's V and Correlation Ratio are bounded between 0 and 1, while the Pearson Correlation Coefficient spans -1 to 1. Following Tao et al. (2021), correlation values are discretized into four levels: low [0, 0.1), weak [0.1, 0.3), medium [0.3, 0.5), and strong [0.5, 1). The *CorAcc* metric quantifies the similarity between synthetic and original data by measuring the fraction of column pairs where the assigned correlation levels match.

Pairwise Attribute Distribution Similarity (Pair). This metric extends the Normalized Histogram Intersection (HIST) by calculating the histogram intersection for all two-way marginals and averaging the results across all attribute pairs. For numerical columns, we discretize the values into bins of size 20 and 50 before computing the intersections.

		Adult						Airline						Texas					
	Method	F1	AUC	ACC	CorAcc	Pair	HIST	F1	AUC	ACC	CorAcc	Pair	HIST	F1	AUC	ACC	CorAcc	Pair	HIST
	Real data	69.9$_{\pm2}$	91.7$_{\pm1}$	84.0$_{\pm4}$	97.3	97.5$_{\pm0}$	99.1$_{\pm0}$	90.6$_{\pm8}$	96.2$_{\pm5}$	91.8$_{\pm7}$	97.2	98.4$_{\pm0}$	99.4$_{\pm0}$	86.6$_{\pm2}$	98.8$_{\pm0}$	95.0$_{\pm1}$	96.3$_{\pm0}$	98.9$_{\pm0}$	99.5$_{\pm0}$
$\varepsilon = \infty$	CTGAN	59.5$_{\pm6}$	88.5$_{\pm0}$	80.2$_{\pm3}$	74.9$_{\pm4}$	**85.0**$_{\pm1}$	91.2$_{\pm1}$	87.2$_{\pm3}$	94.7$_{\pm2}$	88.9$_{\pm3}$	84.4$_{\pm1}$	**89.3**$_{\pm1}$	**94.4**$_{\pm1}$	80.8$_{\pm4}$	96.8$_{\pm2}$	92.8$_{\pm2}$	**72.8**$_{\pm4}$	**88.3**$_{\pm1}$	93.1$_{\pm0}$
	TVAE	63.2$_{\pm2}$	87.5$_{\pm1}$	77.7$_{\pm4}$	76.6$_{\pm2}$	84.5$_{\pm1}$	**91.5**$_{\pm1}$	85.8$_{\pm5}$	93.0$_{\pm6}$	87.2$_{\pm6}$	78.0$_{\pm3}$	82.8$_{\pm3}$	90.3$_{\pm2}$	76.7$_{\pm7}$	95.3$_{\pm3}$	90.9$_{\pm3}$	61.2$_{\pm4}$	63.4$_{\pm7}$	93.5$_{\pm1}$
	VAE	53.8$_{\pm10}$	86.6$_{\pm1}$	80.5$_{\pm1}$	65.3$_{\pm2}$	60.2$_{\pm2}$	73.3$_{\pm3}$	79.8$_{\pm1}$	91.1$_{\pm1}$	80.0$_{\pm1}$	67.8$_{\pm5}$	60.4$_{\pm1}$	76.8$_{\pm0}$	74.3$_{\pm3}$	95.6$_{\pm1}$	90.6$_{\pm2}$	51.5$_{\pm7}$	57.5$_{\pm1}$	89.3$_{\pm1}$
	GPT-2	**68.9**$_{\pm0}$	**90.7**$_{\pm0}$	**83.7**$_{\pm2}$	**79.9**$_{\pm1}$	83.7$_{\pm0}$	90.7$_{\pm1}$	**89.6**$_{\pm5}$	**95.9**$_{\pm3}$	**91.4**$_{\pm4}$	**86.5**$_{\pm4}$	85.5$_{\pm2}$	90.8$_{\pm1}$	**83.9**$_{\pm2}$	**98.6**$_{\pm0}$	**93.6**$_{\pm1}$	**72.8**$_{\pm1}$	82.8$_{\pm3}$	**94.9**$_{\pm0}$

Table 2: **Non-DP Benchmark.** The Real data baseline represents the optimal achievable performance, determined by evaluating metrics using real training data compared to real test data. **F1**, **AUC** and **ACC** are reported as averages of two ML Models (XGBoost and Logistic Regression). **Pair**, and **Hist** are reported as averages of two bin sizes (Bins 20 and 50). Each model run five times and four synthetic datasets generated per run with standard deviation reported after ±. The best value per column for each ε is shown in **bold** while the second best value is underlined.

A.2 Marginal-based DP Baselines

AIM. Proposed by (McKenna et al., 2022), AIM is a marginal-based model for generating differentially private synthetic data. It is a workload-adaptive algorithm that follows a three-step process: selecting a set

[1] https://www.dshs.texas.gov/thcic/

[2] https://github.com/spring-epfl/synthetic_data_release/blob/master/data/texas.csv

of queries, privately measuring those queries, and generating synthetic data from the noisy measurements. AIM employs innovative techniques to iteratively prioritize the most useful measurements, considering both their relevance to the workload and their importance in approximating the input data.

MST. Proposed by (McKenna et al., 2021), MST was the winning mechanism of the 2018 NIST Differential Privacy Synthetic Data Competition. It is a general approach for differentially private synthetic data generation that follows three main steps: (1) selecting a collection of low-dimensional marginals, (2) measuring these marginals using a noise addition mechanism, and (3) generating synthetic data that accurately preserves the measured marginals.

		Adult						**Airline**						**Texas**					
	Method	**F1**	**AUC**	**ACC**	**CorAcc**	**Pair**	**HIST**	**F1**	**AUC**	**ACC**	**CorAcc**	**Pair**	**HIST**	**F1**	**AUC**	**ACC**	**CorAcc**	**Pair**	**HIST**
$\varepsilon = 1$,	AIM	59.6$_{\pm 6}$	86.8$_{\pm 1}$	80.3$_{\pm 2}$	86.4$_{\pm 1}$	77.1$_{\pm 9}$	88.4$_{\pm 5}$	77.3$_{\pm 5}$	88.9$_{\pm 4}$	78.2$_{\pm 5}$	91.8$_{\pm 1}$	46.7$_{\pm 3}$	68.1$_{\pm 2}$	84.5$_{\pm 1}$	98.3$_{\pm 0}$	94.2$_{\pm 1}$	81.0$_{\pm 2}$	93.0$_{\pm 5}$	98.9$_{\pm 0}$
$\delta = 10^{-5}$	MST	39.6$_{\pm 19}$	76.8$_{\pm 1}$	72.8$_{\pm 2}$	70.0$_{\pm 1}$	74.6$_{\pm 10}$	87.0$_{\pm 5}$	72.2$_{\pm 6}$	83.4$_{\pm 5}$	75.2$_{\pm 4}$	72.7$_{\pm 0}$	46.3$_{\pm 3}$	68.2$_{\pm 2}$	81.7$_{\pm 0}$	94.8$_{\pm 0}$	93.2$_{\pm 0}$	77.0$_{\pm 0}$	97.5$_{\pm 0}$	99.0$_{\pm 0}$
	DP-GAN	33.5$_{\pm 20}$	67.7$_{\pm 9}$	64.2$_{\pm 10}$	39.9$_{\pm 3}$	41.2$_{\pm 4}$	63.7$_{\pm 3}$	40.2$_{\pm 24}$	63.9$_{\pm 13}$	59.8$_{\pm 6}$	37.4$_{\pm 9}$	22.2$_{\pm 13}$	44.7$_{\pm 12}$	13.7$_{\pm 16}$	58.3$_{\pm 12}$	78.3$_{\pm 8}$	36.1$_{\pm 7}$	34.6$_{\pm 5}$	68.3$_{\pm 7}$
	DP-CTGAN	**42.2**$_{\pm 20}$	**78.0**$_{\pm 7}$	**75.7**$_{\pm 3}$	51.3$_{\pm 3}$	59.2$_{\pm 2}$	75.7$_{\pm 2}$	67.1$_{\pm 8}$	76.7$_{\pm 8}$	68.0$_{\pm 6}$	31.7$_{\pm 2}$	62.2$_{\pm 2}$	78.7$_{\pm 2}$	63.9$_{\pm 11}$	91.4$_{\pm 3}$	82.6$_{\pm 19}$	43.9$_{\pm 6}$	66.9$_{\pm 6}$	84.7$_{\pm 3}$
	DP-VAE	0.0$_{\pm 0}$	50.0$_{\pm 0}$	75.6$_{\pm 0}$	48.8$_{\pm 1}$	40.3$_{\pm 1}$	61.8$_{\pm 2}$	26.5$_{\pm 28}$	57.9$_{\pm 13}$	57.3$_{\pm 6}$	46.6$_{\pm 1}$	20.6$_{\pm 0}$	41.8$_{\pm 1}$	0.0$_{\pm 0}$	50.0$_{\pm 0}$	82.5$_{\pm 0}$	62.1$_{\pm 1}$	43.9$_{\pm 1}$	77.9$_{\pm 1}$
	GPT-2																		
	DP-Standard	27.8$_{\pm 15}$	58.5$_{\pm 7}$	65.2$_{\pm 9}$	55.0$_{\pm 1}$	68.4$_{\pm 1}$	85.7$_{\pm 2}$	60.5$_{\pm 7}$	65.3$_{\pm 9}$	62.4$_{\pm 7}$	64.0$_{\pm 2}$	77.0$_{\pm 2}$	90.3$_{\pm 3}$	55.4$_{\pm 10}$	90.2$_{\pm 5}$	77.1$_{\pm 11}$	70.3$_{\pm 1}$	60.6$_{\pm 1}$	92.3$_{\pm 1}$
	DP-2Stage-U	21.2$_{\pm 12}$	48.9$_{\pm 6}$	61.9$_{\pm 13}$	55.0$_{\pm 1}$	**76.1**$_{\pm 1}$	86.7$_{\pm 1}$	**68.5**$_{\pm 9}$	**77.8**$_{\pm 10}$	**72.1**$_{\pm 7}$	65.3$_{\pm 1}$	**80.8**$_{\pm 1}$	90.7$_{\pm 1}$	23.5$_{\pm 14}$	59.8$_{\pm 14}$	67.3$_{\pm 17}$	68.2$_{\pm 0}$	**80.7**$_{\pm 6}$	**93.4**$_{\pm 0}$
	DP-2Stage-O																		
	+adult	-	-	-	-	-	-	55.2$_{\pm 18}$	62.5$_{\pm 19}$	60.0$_{\pm 16}$	**66.8**$_{\pm 1}$	80.1$_{\pm 1}$	**92.5**$_{\pm 1}$	**74.8**$_{\pm 4}$	**96.7**$_{\pm 1}$	**89.1**$_{\pm 3}$	69.9$_{\pm 2}$	60.5$_{\pm 2}$	91.7$_{\pm 1}$
	+airline	30.4$_{\pm 17}$	61.6$_{\pm 8}$	66.7$_{\pm 8}$	55.4$_{\pm 1}$	72.3$_{\pm 1}$	**88.5**$_{\pm 1}$	-	-	-	-	-	-	74.3$_{\pm 5}$	96.4$_{\pm 0}$	88.8$_{\pm 3}$	**70.9**$_{\pm 1}$	62.0$_{\pm 2}$	93.2$_{\pm 0}$
	+texas	31.6$_{\pm 13}$	60.5$_{\pm 7}$	66.4$_{\pm 8}$	**55.6**$_{\pm 1}$	71.3$_{\pm 1}$	86.9$_{\pm 1}$	52.5$_{\pm 13}$	61.0$_{\pm 13}$	58.4$_{\pm 10}$	66.1$_{\pm 2}$	78.7$_{\pm 2}$	90.4$_{\pm 1}$	-	-	-	-	-	-

Table 3: **DP Benchmark.** Utility metrics (F1, AUC, and ACC) are presented as the averages of two ML Models (XGBoost and Logistic Regression). **Pair**, and **Hist** are reported as averages of two bin sizes (Bins 20 and 50). Results are averaged across five model runs and four synthetic datasets per run with standard deviation reported after ±. The best value per column (excluding the marginal-based metrics, AIM and MST) for $\varepsilon = 1$ is shown in **bold** while second best value is underlined.