Where to Intervene? Benchmarking Fairness-Aware Learning on Differentially Private Synthetic Tabular Data [Experiment, Analysis & Benchmark]

Vinícius Gabriel Angelozzi V. de R.
Inria Centre at the University Grenoble Alpes &
Grenoble INP - ENSIMAG UGA
France
vinicius-gabriel.verona@grenoble-inp.org

Héber H. Arcolezi
Inria Centre at the University Grenoble Alpes
France
heber.hwang-arcolezi@inria.fr

ABSTRACT

Machine learning models are increasingly deployed in high-stakes domains, raising concerns about both privacy and fairness. Differential Privacy (DP) has become a gold standard for privacy-preserving data analysis, while fairness-aware mechanisms aim to mitigate discrimination against underrepresented groups. However, these objectives can conflict: DP often amplifies disparities across demographic groups, and little is known about whether established fairness interventions remain effective under DP constraints. In this work, we present, to our knowledge, the first systematic evaluation of fairness interventions on differentially private synthetic tabular data. Our benchmark spans two state-of-the-art marginal-based DP synthesizers (AIM and MST), four datasets commonly used in fairness research, and three classes of fairness interventions (pre-processing, in-processing, and post-processing), evaluated under a wide range of privacy budgets. We compare four pipeline configurations: (*Baseline*) training on original data; (*DP-only*) training on DP synthetic data; (*Fair-only*) applying fairness mechanisms on original data; and (*DP+Fair*) combining fairness mechanisms with DP synthetic data. Our results show that while DP alone can degrade both utility and fairness, interventions can partially recover fairness outcomes. Among them, *post-processing methods emerge as the most effective and stable intervention across different ε values and synthesizers*, often restoring group fairness metrics to levels close to those achieved on non-private data. We release all code, data, and experimental artifacts to ensure full reproducibility and to support future research on the privacy-fairness-utility trade-off.

PVLDB Reference Format:
Vinícius Gabriel Angelozzi V. de R. and Héber H. Arcolezi. Where to Intervene? Benchmarking Fairness-Aware Learning on Differentially Private Synthetic Tabular Data
[Experiment, Analysis & Benchmark]. PVLDB, 19(1): XXX-XXX, 2026.
doi:XX.XX/XXX.XX

PVLDB Artifact Availability:
The source code, data, and/or other artifacts have been made available at https://github.com/vinicius-verona/dp-fair-intervention-benchmark.

Proceedings of the VLDB Endowment, Vol. 19, No. 1 ISSN 2150-8097.
doi:XX.XX/XXX.XX

1 INTRODUCTION

Synthetic tabular data generation is increasingly adopted by the research community [31, 32, 34], regulators [18, 25], and industry [1, 2] as a promising approach to facilitate data sharing and downstream analysis while reducing disclosure risks. These methods aim to approximate the empirical distribution of real datasets and generate artificial records that preserve key statistical properties. However, synthetic data does not guarantee privacy or fairness by default. On the one hand, without formal protections, generative models may memorize or leak sensitive information from the training data, making them vulnerable to membership inference, attribute inference, or reconstruction attacks [26, 29, 46, 52]. At the same time, synthetic data can replicate or even amplify societal biases present in the source data, leading to unfair outcomes in downstream classifiers [10].

Differential privacy (DP) [20, 21] provides rigorous, quantifiable protection by bounding the influence of any single record on model training or statistics. DP-based synthetic data generators [38, 48, 54] address the privacy gap by injecting calibrated noise during training or sampling. Most recent benchmarks have therefore concentrated on the *utility* of DP synthetic data, evaluating predictive performance across generative models, tasks, and privacy levels [28, 44, 45, 47].

Beyond utility, however, the interplay between DP and fairness has received far less attention, and is often more complex [24, 51]. When DP is applied directly to model training (*e.g.*, via DP-SGD [4] or PATE [39]), prior work has shown that noise reduces overall accuracy but disproportionately harms minority or underrepresented groups [9]. This phenomenon, often described as *DP disparate impact*, raises the question of whether similar effects also emerge in models trained on DP synthetic data.

Recent studies confirm that they do. Work on *DP synthetic data* [14, 27, 41] has shown that privacy mechanisms and data characteristics interact in subtle ways, often leading to heterogeneous fairness outcomes across groups. However, these evaluations are largely observational: they measure disparities (*e.g.*, statistical parity or subgroup accuracy) but do not examine whether *fairness-aware learning mechanisms* remain effective when applied to DP synthetic data.

This gap motivates our central research question: **Where should one intervene in the ML pipeline to mitigate unfairness under DP synthetic data?** Should interventions target the data distribution (pre-processing), the training procedure (in-processing), or the model outputs (post-processing)? And critically, do mechanisms

designed for non-private pipelines retain their effectiveness, and at what cost to the utility, when data originate from DP synthesizers?

The question is timely. Several commercial platforms now offer *DP synthetic data* as a product [26, Table 1], including vendors such as Tumult Labs [1] and YData [2]. These platforms are already being used in sensitive domains such as healthcare and finance. As fairness concerns grow, both ethically and legally (*e.g.*, EU AI Act [40]), developers may need to integrate fairness-aware interventions *without redesigning entire DP generation pipelines*.

Contributions. Building on these observations, we design a systematic benchmark that moves beyond *measuring* fairness degradation to *mitigating* it. Our goal is to quantify when and how fairness-aware mechanisms can counteract the disparate impact introduced by DP, while preserving predictive utility. To this end, we contribute both a large-scale empirical evaluation and a reproducible open-source framework. Our main contributions are:

- We present, to our knowledge, the first systematic evaluation of fairness interventions applied to DP synthetic tabular data, moving beyond observational studies toward actionable mitigation. An overview of our benchmark pipeline is shown in Figure 1.
- We evaluate pre-, in-, and post-processing fairness interventions under varying privacy budgets, revealing how the effectiveness of these mechanisms changes when applied to DP synthetic rather than original data.
- Using multiple real-world datasets and state-of-the-art DP synthesizers, we characterize how DP alters fairness-utility dynamics and identify conditions under which interventions succeed or fail to recover fairness.
- We release all code, datasets, and experimental artifacts to support reproducibility and enable future research at the intersection of privacy and fairness (see Section 5.3).

Findings. Analysis of the experiments leads us to conclude that fairness interventions on models trained with DP synthetic data can partially recover fairness metrics that are degraded by DP noise. However, under default parameters, not all mechanisms are stable or consistently effective. Among the methods tested, three stand out: Reweighing (RW) [15], Reject Option Classification (ROC) [33], and Equalized Odds Post-Processing (EqOdds) [30]. RW, a pre-processing method, demonstrates consistency across privacy budgets and datasets, but its fairness corrections systematically incur utility losses. By contrast, the post-processing methods EqOdds and ROC perform more favorably: they not only reduce disparities in group fairness metrics but also often recover part of the utility lost under DP. In-processing mechanisms, however, proved to be the least promising group in this benchmark, offering only limited fairness improvements with little or no utility advantage.

Outline. The rest of this paper is organized as follows: Section 2 reviews the related work, providing context for our contributions. Section 3 introduces the background on generative models, differential privacy, and fairness-aware learning. Section 4 details the benchmark design, the datasets used, the mechanisms compared, and the experiments performed. Section 6 presents the experimental results, where a comprehensive analysis of the results is presented in Section 7. Finally, Section 8 concludes this paper and highlighting future perspectives.

2 RELATED WORK

Privacy and fairness. The interaction between privacy and fairness has been widely studied [3, 8, 9, 16, 22, 28, 35, 36, 49]. A consistent finding is that differentially private learning (*i.e.*, through DP-SGD [4] or PATE [39]) can exacerbate disparities across demographic groups, a phenomenon often referred to as the *disparate impact of DP*. For example, Bagdasaryan et al. [9] show that DP-SGD reduces overall accuracy but disproportionately harms minority groups. Surveys such as [24, 51] highlight both conditions under which DP and fairness may align and evidence the conditions in which they act as conflicting objectives.

DP synthetic data and utility. The release of *DP synthetic data* has motivated a growing line of work benchmarking the *utility* of different generative models. Early studies compare marginal-based synthesizers (*e.g.*, AIM [38], MST [37]) with deep generative approaches (*e.g.*, DP-GANs [48, 50]), showing that marginal-based methods often yield higher predictive utility on tabular datasets [28, 44, 45, 47]. These results position marginal-based models as competitive baselines for structured domains.

DP synthetic data and fairness. More recent work evaluates fairness explicitly. Bullwinkel et al. [14] compare four DP synthesizers across multiple datasets and privacy budgets, finding that most degrade fairness but that MST behaves more favorably than GAN-based alternatives. Similarly, Pereira et al. [41] analyze fairness and utility metrics for both GAN-based and marginal-based synthesizers, showing that the latter tend to preserve subgroup accuracy and often maintain or improve group fairness metrics. Ganev et al. [27] provide a fine-grained analysis of subgroup disparities in DP synthetic data, showing that DP can either amplify or mitigate imbalance depending on the model, and that classifiers trained on such data exhibit reduced performance for minority groups.

Positioning of our work. Prior works on fairness in DP synthetic data [14, 27, 41] have primarily been *observational*: they measures how DP synthetic data affects fairness but do not investigate whether fairness-aware mechanisms can mitigate these effects. To the best of our knowledge, our benchmark is the first to systematically evaluate pre-, in-, and post-processing fairness interventions on models trained with DP synthetic data. In doing so, we move beyond measurement toward actionable strategies for mitigating the disparate impact of differential privacy.

3 PRELIMINARIES

In this section, we briefly review about generative models, differential privacy, and fairness-aware learning.

3.1 Generative Models and Synthetic Data

Generative models aim to approximate the distribution of real data and to produce artificial records that preserve its statistical properties. Let $D = \{x_i\}_{i=1}^{n}$, with $x_i \in \mathcal{X}$, denote the original dataset drawn *i.i.d.* from an unknown distribution p_{data}. A generative model learns a parameterized distribution p_θ such that samples $x' \sim p_\theta$ resemble those from p_{data}. The synthetic dataset $\tilde{D} = \{\tilde{x}_j\}_{j=1}^{m}$ is then released in place of D for downstream analysis, with the goal of enabling utility while mitigating disclosure risk.

While many approaches exist for building generative models, in this work we focus on marginal-based methods [37, 38]. These synthesizers are rooted in Bayesian network formulations that decompose the joint distribution of tabular data into lower-dimensional marginals and conditional dependencies. Such approaches have consistently shown strong performance on structured tabular domains [28, 47], where feature dependencies can be effectively captured by explicit probabilistic modeling. In contrast, deep generative models, while highly successful for image or text synthesis, often face challenges in faithfully modeling heterogeneous tabular data distributions.

3.2 Differential Privacy

Differential privacy is a property of randomized mechanisms that limits how much the output distribution can change when a single individual's record is modified. Intuitively, it enables learning about the population while revealing little about any one person [21]. We adopt the standard ε-DP notion [20].

Definition 1 (ε-Differential Privacy). *Let $\mathcal{X}$ be the data domain and let datasets $D, D' \in \mathcal{X}^n$ be* neighbors *(written $D \sim D'$) if they differ in exactly one individual's record. A randomized mechanism $\mathcal{M} : \mathcal{X}^n \to \mathcal{O}$ satisfies ε-DP if for all measurable $S \subseteq \mathcal{O}$,*

$$\Pr[\mathcal{M}(D) \in S] \leq e^{\varepsilon} \Pr[\mathcal{M}(D') \in S].$$

When $\varepsilon = 0$, outputs for neighboring datasets are identically distributed, implying that the mechanism's output cannot depend on any single record. Larger ε permits greater sensitivity to an individual, weakening privacy. Thus, choosing ε entails a privacy-utility trade-off. Importantly, by the post-processing property of DP [21], any transformation applied after a DP mechanism, such as fairness interventions on DP synthetic data, cannot weaken its privacy guarantee.

3.3 Fairness-Aware Learning

Fairness-aware learning aims to incorporate fairness criteria into predictive models. Let X denote features, $A \in \{0, 1\}$ a protected attribute, $Y \in \{0, 1\}$ the label, and $h : \mathcal{X} \to \{0, 1\}$ (or score $s : \mathcal{X} \to [0, 1]$) the predictor. Models may exhibit biased behavior for diverse reasons: some biases are intrinsic to the data (data-to-model bias), while others emerge from missing representative samples or from limitations and objectives of the learning algorithm [10, 42]. To address these issues, algorithmic interventions are typically organized by *where* they act in the pipeline:

- **Pre-processing.** Methods that transform the training data D into $\hat{D}$ (*e.g.*, reweighting or data transformation) to reduce dependence on the protected attribute A while preserving task utility under the defined fairness notion.
- **In-processing.** Methods that modify the learning algorithm or objective, either by solving a constrained problem

$$\min_h \mathcal{L}(h; D) \quad \text{s.t.} \quad \Delta_{\text{fair}}(h) \leq \tau,$$

  or by adding a fairness penalty $\mathcal{L}(h; D) + \lambda\, \Omega_{\text{fair}}(h)$; examples include regularizers and reduction-based formulations.
- **Post-processing.** Methods that adjust predictions of a trained model (*e.g.*, group-specific thresholds, calibration, or randomized decisions) to satisfy target constraints with minimal utility loss, treating the model as a black box.

4 BENCHMARK DESIGN

In this section, we first describe the overall benchmark structure and experimental configurations, followed by details on datasets and prediction tasks, the DP generative models used, and the fairness mechanisms evaluated.

4.1 Overview and Experimental Configuration

Figure 1 illustrates the overall benchmark pipeline. Starting from the original dataset, we construct four prediction pipelines, corresponding to the configurations described in the following. Each pipeline follows the same three stages, namely, data pre-processing, model training, and inference, but differs in whether DP is applied to the data, and whether fairness interventions are applied before, during, or after training. This setup enables a systematic comparison of privacy, fairness, and utility across intervention points.

(1) **`Baseline`.** A standard machine learning model trained directly on the original (non-private, non-fair) training data D. This serves as the *reference point* to evaluate the impact of privacy and fairness interventions.
(2) **`DP-only`.** The original training data D is replaced with *differentially private synthetic data* $\tilde{D}$. No fairness mechanism is applied. This setting isolates the effect of differential privacy on model performance and fairness metrics.
(3) **`Fair-only`.** A fairness intervention is applied to the original training data, to the learning algorithm, or to the model output, without incorporating any differential privacy. Specifically, we consider three families of interventions (see Section 3.3): *pre-processing*, *in-processing*, and *post-processing*. This setting quantifies the effect of fairness mechanisms in isolation.
(4) **`DP+Fair`.** Privacy and fairness interventions are combined. A fairness mechanism is applied either (i) to the DP synthetic data before training (pre-processing), (ii) during model training (in-processing), or (iii) to the model predictions (post-processing). This setting evaluates whether fairness mechanisms remain effective when operating on DP synthetic data, and whether they can mitigate the fairness degradation introduced by privacy constraints.

4.2 Data and Task

Dataset. To ensure comparability across settings, we restrict our benchmark to *binary classification tasks*. Accordingly, we evaluate on four open datasets widely used in fairness research:

- **Adult** [12] (UCI Census Income) contains $n = 47,621$ individuals and the goal is to predict whether a person's income exceeds \$50K/year based on 10 demographic and occupational attributes. Gender is used as the protected attribute for fairness evaluation.
- **COMPAS** [7] includes $n = 5,050$ defendants and the goal is to predict recidivism risk based on 7 criminal history and demographic attributes. Race is used as the protected attribute for fairness evaluation.

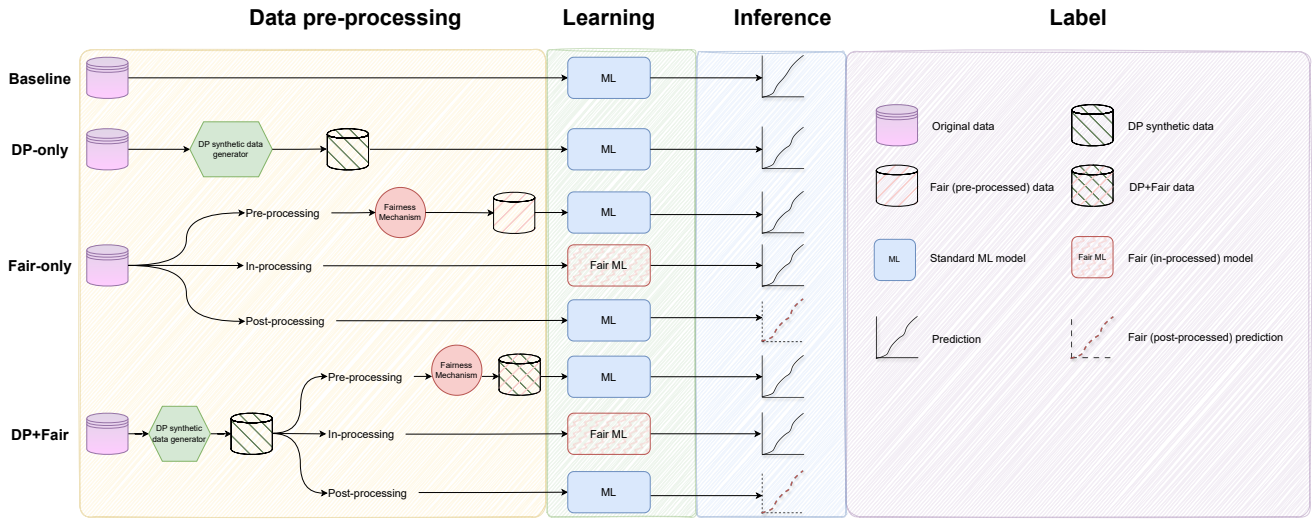

Figure 1: Overview of our benchmark design. We evaluate fairness-aware learning mechanisms applied at three intervention stages, pre-processing, in-processing, and post-processing, across four configurations: (1) `Baseline` (original data, no fairness intervention), (2) `DP-only` (DP synthetic data, no fairness intervention), (3) `Fair-only` (original data with fairness intervention), and (4) `DP+Fair` (fairness interventions on DP synthetic data). This design systematically captures the isolated and combined effects of privacy and fairness interventions.

- **ACSIncome** [19] extends Adult with richer socioeconomic features from the U.S. Census American Community Survey. We select the Utah state subset with $n = 16,337$ individuals. We set the income threshold at the median value (> 38K/year) and use gender as the protected attribute.
- **BiasOnDemand** [11] is a synthetic dataset generator designed to benchmark fairness and bias under controlled conditions. We use it to simulate data distributions with known levels of group imbalance and label bias. In total 6 different bias configurations were tested in 3 different categories: imbalance, historical bias, measurement bias. In total, we generate $n = 30,000$ samples and the goal is to predict the binary value Y conditioned on 2 attributes. The configurations studied are set in a way to: (i) isolate bias on target; (ii) add imbalance to the dataset; (iii) isolate bias to a feature; (iv) add bias to a feature and increase feature correlation and dependence. Furthermore, the values of each configuration parameter were set by experimenting different values until high values of fairness metrics were achieved, indicating strong bias.

Task. All datasets are cast as binary classification tasks. We standardize continuous features, binarize the protected attribute $A \in \{0, 1\}$, and encode the target variable as binary $Y \in \{0, 1\}$. Further details on dataset preprocessing are provided in Appendix A of the full paper [6] and in our open-source code repository.

4.3 DP Generative Models

To generate differentially private synthetic data, we focus on two *marginal-based synthesizers*, AIM [38] and MST [37] described hereafter, both implemented in the SmartNoise library.[1] Our choice is motivated by recent utility-oriented benchmarks [28, 44, 45, 47], which consistently show that marginal-based models outperform deep generative approaches such as DP-GANs [48] on tabular data. Whereas deep generative models excel in unstructured domains like images, they often struggle with the heterogeneous feature types and sparsity typical of structured tabular data. By contrast, marginal-based synthesizers explicitly model low-dimensional marginals and conditional dependencies, leading to higher fidelity and stronger predictive performance in downstream classification tasks. We therefore adopt AIM and MST as representative state-of-the-art DP synthetic data generators.

- **Adaptive Iterative Mechanism (AIM) [38].** AIM is a state-of-the-art differentially private synthetic data generation algorithm. Essentially, it works in a select-measure-generate paradigm. It starts with an iterative greedy selection of sets of queries, performs measurements with noise addition, and at last, generates synthetic data based on these measurements using Private-PGM.
- **Maximum Spanning Tree (MST) [37].** As AIM, MST works in a select-measure-generate paradigm. It starts by creating a complete graph of features and mutual information between features. Later it finds the maximum spanning tree over features using pairwise correlations to select which marginals to measure, and computes differentially private noisy marginals along the tree edges. Finally,

[1] More about the SmartNoise library is available at: https://docs.smartnoise.org/

Private-PGM post-processes these noisy marginals to estimate a joint distribution from which synthetic data is sampled.

4.4 Fairness Mechanisms

To evaluate whether fairness-aware learning mechanisms remain effective on DP synthetic data, we consider representative methods from the three main categories of interventions: *pre-processing*, *in-processing*, and *post-processing*. These methods have been widely studied in the fairness literature and are implemented in open-source libraries such as AIF360 [13], making them suitable benchmarks for our study.

Pre-Processing. These approaches modify the training data before model learning, either by reweighting or transforming features to reduce dependence on the protected attribute.

- **Reweighing (RW) [15].** This method relies on resampling and computing weights for the input samples to decrease discrimination.
- **Disparate Impact Remover (DIR) [23].** This method transforms the original dataset to reduce disparate impact between privileged and unprivileged subgroups. It first detects whether disparate impact exists, then removes the dependence of unprotected features on the protected feature, and finally adjusts the distributions of unprotected features so that both privileged and unprivileged groups have similar distributions.
- **Learning Fair Representations (LFR) [53].** This method creates a probabilistic mapping from a data representation in a given input space to a new representation that reduces the ability to identify protected subgroups while preserving task-relevant information. The objective is to approximate statistical parity across groups while balancing fairness with the predictive utility of the data.

In-Processing. These methods modify the training procedure itself, typically by solving constrained optimization problems that balance accuracy with fairness.

- **Exponentiated Gradient Reduction (EGR) [5].** This method computes an iterative approximation of saddle point of a Lagrangian, by minimizing the classification loss and maximizing the penalty for fairness violation. The idea behind this computation is the same of a zero-sum game with two players.
- **Grid Search Reduction (GSR) [5].** This method shares similar ideas with EGR, but relies on brute-force. Essentially, it builds a grid of Lagrangian multipliers and exhaustively searches for the best solution considering the trade-off between accuracy and fairness.

Post-Processing. These methods operate on the outputs of an already-trained model, adjusting predictions to satisfy fairness constraints.

- **Reject Option Classification (ROC) [33].** This method operates on the model's prediction outputs, adjusting decision boundaries to favor fair outcomes in regions of uncertainty. It aims to reduce discrimination by flipping labels for samples near the decision boundary – particularly when such adjustments enhance fairness for protected groups.
- **Equalized Odds Post-Processing (EqOdds) [13, 30].** This method formulates a linear program to learn probabilities with which the output labels will be changed to satisfy equalized odds constraints while maintaining classification fidelity.
- **Calibrated Equalized Odds Post-processing (CEOP) [43].** This method operates similarly to EqOdds, enforcing parity in error rates – false positive rate and false negative rate remain similar across the protected groups. Additionally, CEOP introduces the concern for performing such tasks while maintaining calibration, *i.e.*, ensuring that the predicted scores remain interpretable as true outcome probabilities.

5 EXPERIMENTAL SETUP

In this section, we provide details on the evaluation metrics, as well as the model training procedure, including data preprocessing, classifier configuration, and protocol for stability and reproducibility.

5.1 Metrics

Privacy levels. We vary the privacy budget across a broad range, $\varepsilon \in \{0.25, 0.5, 0.75, 1.0, 5.0, 10.0, 15.0, 20.0\}$, covering both high-privacy (small ε) and low-privacy (large ε) regimes.

Utility metrics. We report standard utility metrics: accuracy, precision, and recall, computed on the held-out test set.

Fairness metrics. Let $A \in \{0, 1\}$ denote the binary protected attribute, $Y \in \{0, 1\}$ the true label, and $\hat{Y} \in \{0, 1\}$ the predicted label. We evaluate group fairness using three standard metrics:

Definition 2 (Model Accuracy Difference (MAD)). *Difference in overall classification accuracy between groups:*

$$MAD = \Pr[\hat{Y} = Y \mid A = 0] - \Pr[\hat{Y} = Y \mid A = 1].$$

Definition 3 (Equal Opportunity Difference (EOD) [30]). *Difference in true positive rates across groups:*

$$EOD = \Pr[\hat{Y} = 1 \mid Y = 1, A = 0] - \Pr[\hat{Y} = 1 \mid Y = 1, A = 1].$$

Definition 4 (Statistical Parity Difference (SPD)). *Difference in positive prediction rates across groups:*

$$SPD = \Pr[\hat{Y} = 1 \mid A = 0] - \Pr[\hat{Y} = 1 \mid A = 1].$$

All three metrics are defined such that a value of 0 corresponds to perfect parity between groups, while larger deviations indicate increasing disparity.

5.2 Model Training

All datasets are split into training (60%), calibration (20%), and test (20%) subsets using a fixed partition. Privacy interventions (DP synthesizers) and fairness interventions in the form of pre- or in-processing are applied *only to the training set.* The calibration set is left untouched and is used to tune post-processing fairness mechanisms, while the test set remains unseen during training and calibration, serving exclusively for final evaluation (including post-processing applied at inference). This setup simulates realistic deployment and ensures fair comparison across baselines.

Classifier. We use Extreme Gradient Boosting (**XGBoost**) [17] as the base classifier across all experiments. XGBoost is a widely adopted tree-based ensemble method with strong predictive performance and scalability. To maintain consistency across configurations, we use the `binary:logistic` objective and keep all hyperparameters at their default values.

Stability. Since DP mechanisms, train/calibration/test splits, synthetic data generation, and classifier training all involve randomness, we repeat each experiment with 20 independent random seeds. For each dataset (Subsection 4.2), this results in 160 runs, and a total of 1,440 runs per synthesizer. We report the mean and standard deviation of all metrics to ensure robustness and reproducibility.

5.3 Reproducibility and Extensibility

We implement our benchmark on top of the open-source SmartNoise library and the AIF360 fairness toolkit [13]. Our framework integrates DP synthesizers, fairness interventions, and evaluation pipelines in a modular fashion, making it straightforward to add new datasets, generative models, or classifiers. More precisely, our benchmark is organized around two main components: (1) *DP synthetic data generation*, and (2) *execution of experiments*. This design provides a clear separation between data generation and downstream evaluation, making the benchmark both reproducible and easily extensible.

- **Synthetic data generation.** This module integrates existing DP synthesizers (AIM, MST) and can be extended with new generators and different data pre-processers. Users can also add additional datasets to the generation pipeline with minimal configuration.
- **Experimental execution.** This module runs end-to-end experiments, including model training, fairness interventions, and metric evaluation and can be extended by modifying the base classifier.

Together, these two modules ensure that new datasets, synthesizers, classifiers, interventions, or metrics can be integrated with minimal effort. Moreover, all experiments can be reproduced end-to-end using provided scripts, and all figures and tables in this paper can be regenerated from raw logs with a single command. The full codebase, along with documentation and configuration files, is available as open source at: https://github.com/vinicius-verona/dp-fair-intervention-benchmark.

6 RESULTS AND ANALYSIS

In the following subsections, we focus on the **Adult dataset** to illustrate our main findings, considering both AIM and MST synthesizers. This choice allows us to present representative trends while keeping the discussion concise. Results for the remaining datasets are provided in Appendix B of the full paper [6]. Importantly, we highlight that the qualitative findings we discuss in this section are consistent across all datasets.

For each figure in this section, we report fairness and utility metrics across varying privacy budgets (ε). Specifically, each subplot shows one of the six metrics from Section 5.1: three for fairness (MAD, EOD, and SPD) and three for utility (Accuracy, Precision, Recall). These metrics were evaluated under the four experimental settings of Section 4.1 (see Figure 1): `Baseline` (no privacy or fairness intervention), `DP-only` (labeled *DP*), `Fair-only` (labeled *Fair[alg]*), and combined DP with fairness intervention `DP+Fair` (labeled *DP_Fair[alg]*). Solid lines represent the mean, and shaded areas indicate standard deviation across the 20 runs.

6.1 Pre-Processing

We first analyze pre-processing interventions, namely Reweighing (RW) [15], Disparate Impact Remover (DIR) [23], and Learning Fair Representations (LFR) [53]. Figures 2 and 3 report results on the **Adult dataset** under AIM and MST synthesizers.

As shown in these figures, and consistently observed with the additional datasets in Section B of the full paper [6], two broad patterns emerge. First, in `DP+Fair` settings, MST tends to achieve lower disparity values than AIM, with fairness metrics approaching closer to the ideal 0, while AIM remains closer to Baseline values (cf. Figure 3). This apparent advantage, however, is largely explained by a sharp drop in utility. In particular, precision and recall decrease by approximately 30%–40% compared to AIM, indicating that MST's fairness improvements are coupled with severe predictive degradation. Second, across privacy budgets ε, we consistently observe that for AIM, `DP+Fair` curves converge toward the corresponding `Fair-only` curves, effectively correcting both the bias present in the Baseline and the additional bias introduced by `DP-only`. This indicates that the effectiveness of pre-processing mechanisms under DP synthetic data is *synthesizer-dependent*: they are often effective with AIM, but less so with MST.

Reweighing. RW consistently reduced group disparities compared to `DP-only` training, particularly in terms of SPD. For instance, in Figure 3, SPD shifted closer to zero across all privacy budgets, with the strongest gains at lower ε values ($\varepsilon \leq 1$). Like the other pre-processing methods, RW incurred a noticeable utility cost; precision fell to ~ 0.43 compared to a Baseline of ~ 0.70, and accuracy stabilized around 0.75 across all ε values. Nevertheless, among the three pre-processing mechanisms, RW proved to be the most effective overall: it delivered consistent fairness improvements across datasets and synthesizers.

Disparate Impact Remover. DIR produced results that were, in most cases, very close to those of the `DP-only` setting. Across fairness metrics, its corrections were limited: disparities were not substantially reduced compared to `DP-only`, and in some cases both fairness and utility were slightly worsened relative to the `Baseline`. We attribute this behavior partly to unmet preprocessing conditions required by DIR [23], which prevent the algorithm from effectively correcting bias. Preliminary experiments confirmed that, when these conditions are satisfied, DIR can improve fairness and behave similarly to other pre-processing mechanisms. However, since our benchmark enforces a uniform preprocessing pipeline across datasets, we report the default results, *i.e.*, showing that under these conditions, DIR remains largely indistinguishable from `DP-only`.

Learning Fair Representations. LFR exhibited the most aggressive intervention among pre-processing methods. For instance, in both Figures 2 and 3, LFR pushed SPD and EOD close to zero across all ε. However, these fairness gains came at a steep utility cost:

both precision and recall dropped the most among all fairness intervention mechanisms. In some datasets, such as COMPAS, LFR occasionally failed entirely: for certain ε values or random seeds, the transformation collapsed one of the target classes, making model training impossible. Such failures are documented in Appendix B of the full paper [6]. This instability highlights that while LFR can enforce strong fairness, it does so at the cost of fragile and sometimes unreliable utility outcomes.

Summary. Overall, pre-processing mechanisms under `DP+Fair` improve fairness but always incur utility losses. Among them, RW is the most consistent, DIR is largely indistinguishable from `DP-only`, and LFR is powerful but unstable.

6.2 In-Processing

We next analyze in-processing interventions, namely Exponentiated Gradient Reduction (EGR) [5] and Grid Search Reduction (GSR) [5]. Figures 4 and 5 report results on the **Adult dataset** under AIM and MST synthesizers.

As shown in these figures, and consistently observed with the additional datasets in Section B of the full paper [6], two broad patterns emerge. First, compared to pre-processing, in-processing mechanisms exhibit a lower bias correction capability overall. While they improve fairness metrics relative to `DP-only`, the magnitude of correction is modest, and they rarely achieve values close to `Fair-only`. Second, across privacy budgets ε, `DP+Fair` curves sometimes move toward the corresponding `Fair-only` curves, but the convergence is weaker and less consistent than with pre-processing interventions. This indicates that in-processing mechanisms are less effective than pre-processing interventions under DP synthetic data, offering only limited fairness improvements while maintaining utility at levels comparable to `DP-only`.

Exponentiated Gradient Reduction. EGR proved to be the more effective of the two in-processing mechanisms. On Adult–AIM (Figure 4), EGR marginally reduced disparities: MAD decreased from ~ 0.13 under `DP-only` to ~ 0.10, while SPD improved from -0.19 to around -0.14, approximating the values achieved in the `Fair-only` setting. Utility remained stable, with accuracy between 0.81 and 0.84 across ε, and precision and recall showed only slight drops. On Adult–MST (Figure 5), EGR also improved fairness (SPD rising from -0.20 to about -0.05), with benefits more visible than AIM though much more utility drop. Across the three other datasets, EGR consistently corrected the biased baseline models while maintaining utility, confirming its moderated robustness.

Grid Search Reduction. GSR, by contrast, tended to underperform relative to EGR. Specifically, GSR almost completely tracked the `DP-only` curve, offering little evidence of improvement. The exception is for the BiasOnDemand dataset in Section B.3 in the full paper [6], in which GSR showed improvements in the fairness metrics. This tendency to align with `DP-only` outcomes rather than correcting the bias in the baseline was consistent across other datasets as well, confirming that GSR is largely ineffective under DP synthetic data.

Summary. Overall, in-processing mechanisms provide only modest fairness improvements under DP synthetic data, and their effectiveness is weaker than that of pre-processing interventions. EGR is the stronger of the two methods, consistently pushing metrics toward the `Fair-only` curves with relatively limited utility loss, whereas GSR remains largely indistinguishable from `DP-only` across datasets and synthesizers. These findings suggest that among in-processing approaches, **EGR is the only option that offers consistent, though still limited, benefits**.

6.3 Post-Processing

We finally analyze post-processing interventions, namely Reject Option Classification (ROC) [33], Equalized Odds Post-Processing (EqOdds) [13, 30], and Calibrated Equalized Odds (CEOP) [43]. Figures 6 and 7 report results on the **Adult dataset** under AIM and MST synthesizers.

As shown in these figures, and consistently observed with the additional datasets in Section B of the full paper [6], two broad patterns emerge. First, post-processing mechanisms generally achieve stronger bias correction than in-processing and pre-processing, while sometimes recovering utility lost due to DP synthesis. For example, in Figure 7, ROC and EqOdds not only reduced disparities in SPD and EOD but also improved precision and recall compared to the `DP-only` model, approaching Baseline values. This highlights a distinctive property of post-processing: in certain settings, these interventions are able to both mitigate bias and partially restore predictive performance. Second, similar to pre- and in-processing, post-processing outcomes are also sensitive to the choice of algorithm. ROC typically produced stable improvements across fairness metrics, whereas CEOP displayed instability, *i.e.*, sometimes improving fairness on the Adult dataset but failing to deliver consistent gains, or even worsening disparities, on others. This instability was particularly evident in the COMPAS dataset and suggests that CEOP, under default hyperparameters, is less reliable for DP synthetic data.

Reject Option Classification. ROC consistently improved fairness across synthesizers. On Adult–AIM (Figure 6), it reduced MAD from ~ 0.12 under `DP-only` to nearly zero, while simultaneously improving SPD and EOD toward parity. However, due to the privacy-fairness-utility trade-off, ROC slightly decreased utility: accuracy remained above 0.78, with an important loss of precision while improving recall. On Adult–MST (Figure 7), ROC similarly corrected bias while trading-off utility, a trend echoed across Compas and ACSIncome.

Equalized Odds Post-Processing. EqOdds performed comparably to ROC, effectively reducing disparities across MAD, SPD, and EOD. On Adult–AIM (Figure 6), it achieved fairness outcomes close to the `Fair-only` setting, while maintaining accuracy above 0.80. On Adult–MST, EqOdds again reduced disparities but with smaller utility recovery, compared to ROC. Nevertheless, across datasets, EqOdds provided stable and consistent fairness improvements without severely compromising performance.

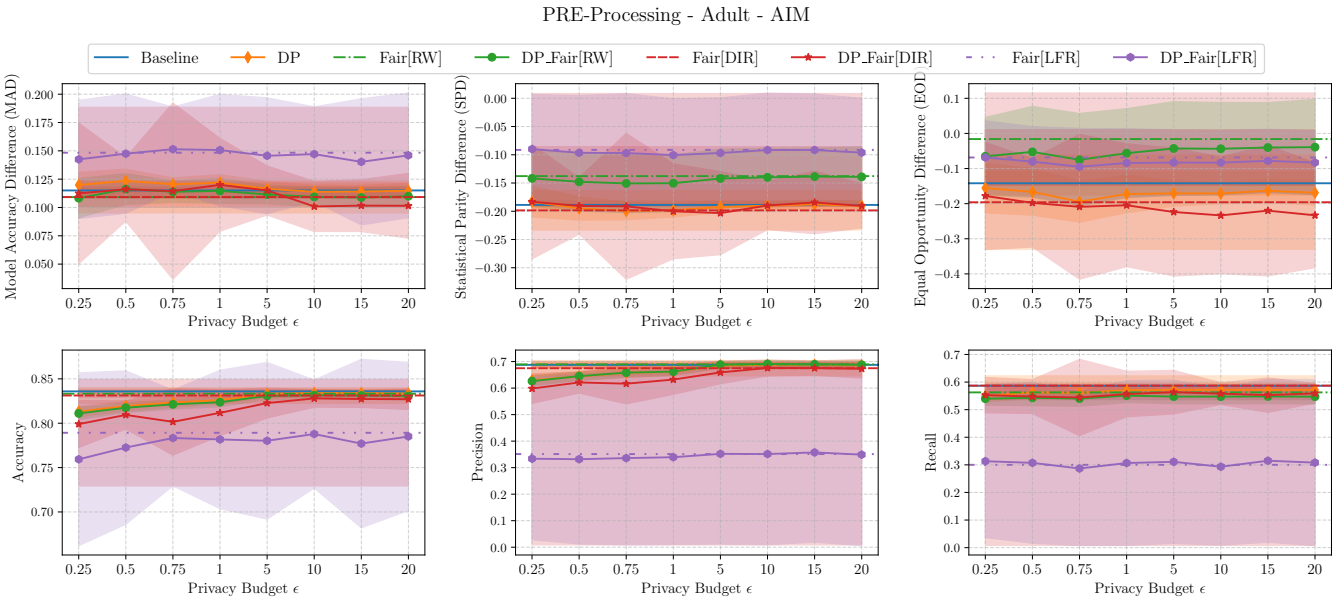

Figure 2: Fairness (MAD, SPD, EOD) and utility (Accuracy, Precision, Recall) metrics across privacy budgets ε on the Adult dataset using the AIM synthesizer with pre-processing fairness mechanisms (RW, DIR, LFR). Solid lines denote means and shaded areas indicate standard deviations across runs.

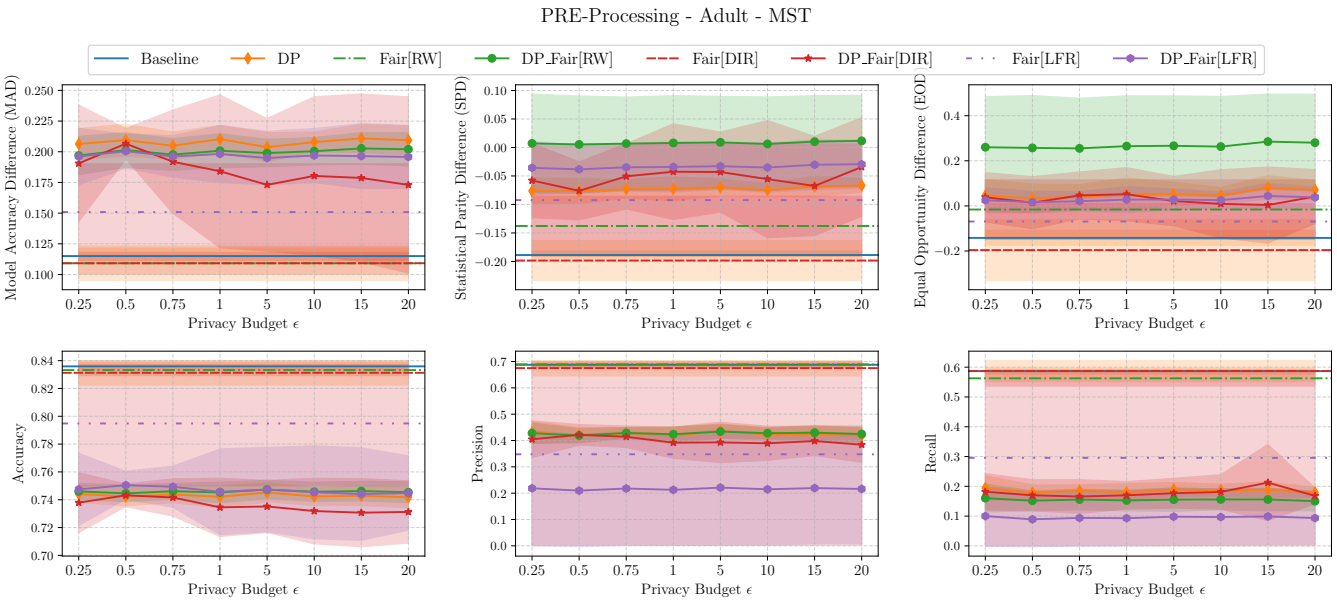

Figure 3: Fairness (MAD, SPD, EOD) and utility (Accuracy, Precision, Recall) metrics across privacy budgets ε on the Adult dataset using the MST synthesizer with pre-processing fairness mechanisms (RW, DIR, LFR). Solid lines denote means and shaded areas indicate standard deviations across runs.

Calibrated Equalized Odds Post-Processing. CEOP, in contrast, was the least stable of the three methods. On Adult–AIM, it occasionally improved fairness, but on Adult–MST (Figure 7) it introduced trade-offs, sometimes improving SPD while worsening EOD. This instability was especially frequent in other datasets (Section B in [6]), where CEOP often degraded both fairness and utility. These findings suggest that CEOP, at least under default parameters, is unreliable in DP synthetic settings and should be applied cautiously.

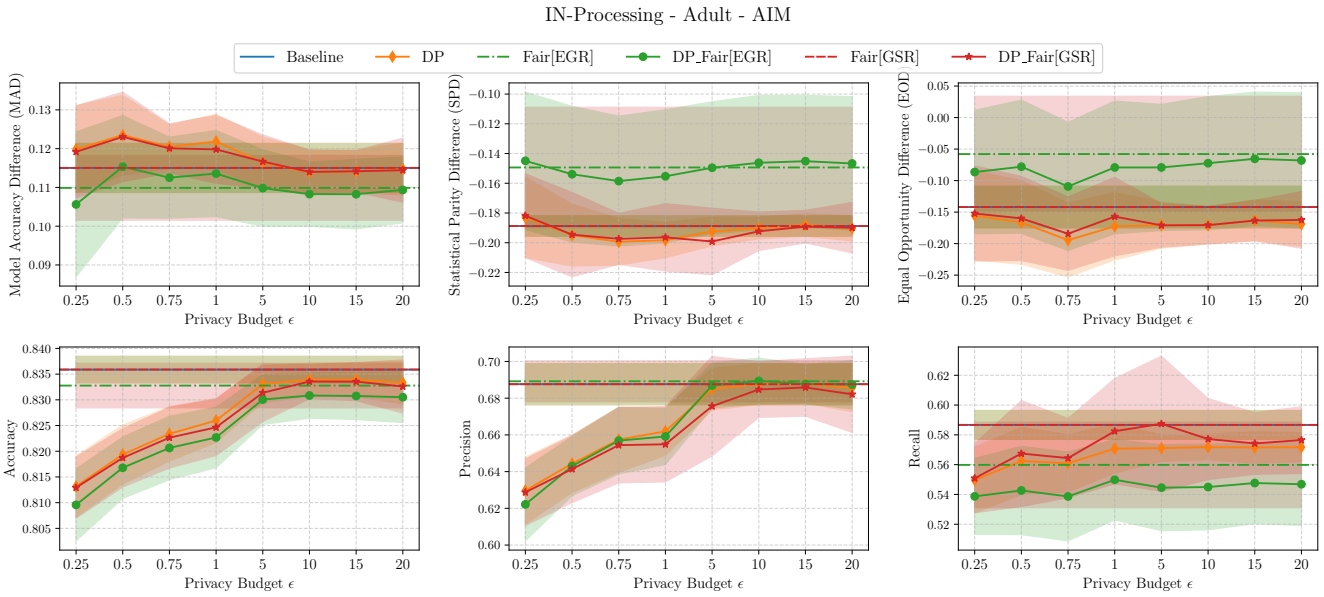

Figure 4: Fairness (MAD, SPD, EOD) and utility (Accuracy, Precision, Recall) metrics across privacy budgets ε on the Adult dataset using the AIM synthesizer with in-processing fairness mechanisms (EGR, GSR). Solid lines denote means and shaded areas indicate standard deviations across runs.

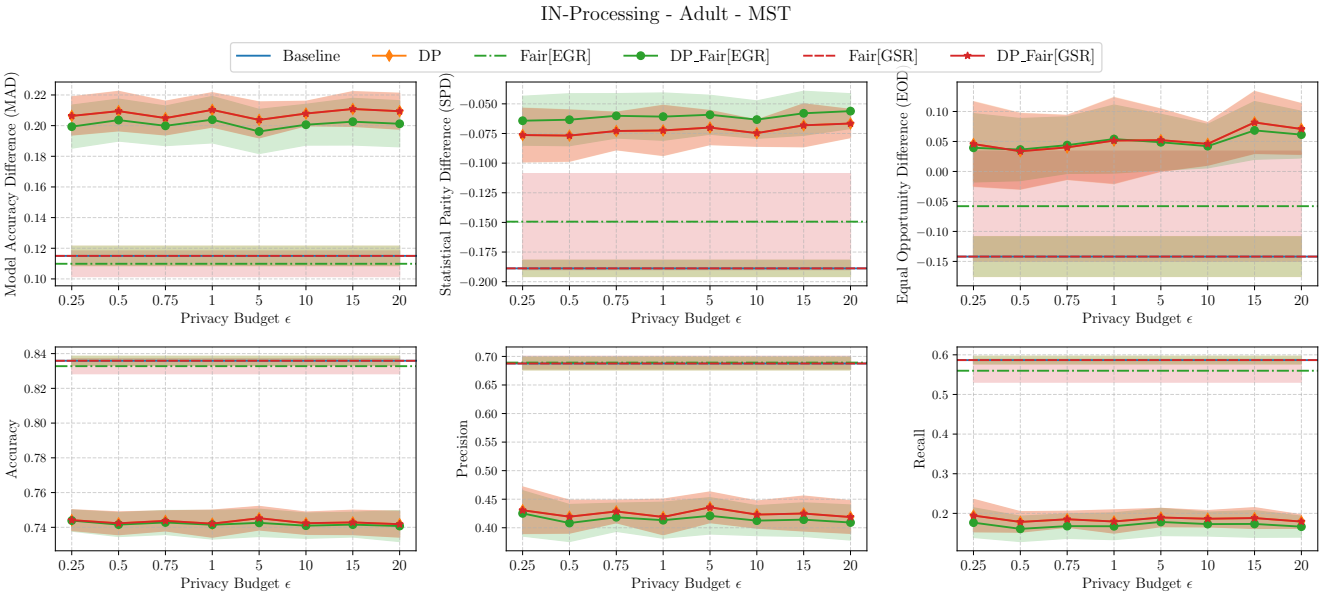

Figure 5: Fairness (MAD, SPD, EOD) and utility (Accuracy, Precision, Recall) metrics across privacy budgets ε on the Adult dataset using the MST synthesizer with in-processing fairness mechanisms (EGR, GSR). Solid lines denote means and shaded areas indicate standard deviations across runs.

Summary. Overall, post-processing mechanisms are the most effective class of fairness interventions under DP synthetic data. ROC and EqOdds consistently reduced disparities while often recovering utility, making them robust across datasets and synthesizers. By contrast, CEOP showed unstable behavior, occasionally worsening outcomes. These results suggest that **post-processing, particularly ROC and EqOdds, is the most reliable strategy for mitigating bias under DP synthetic data**.

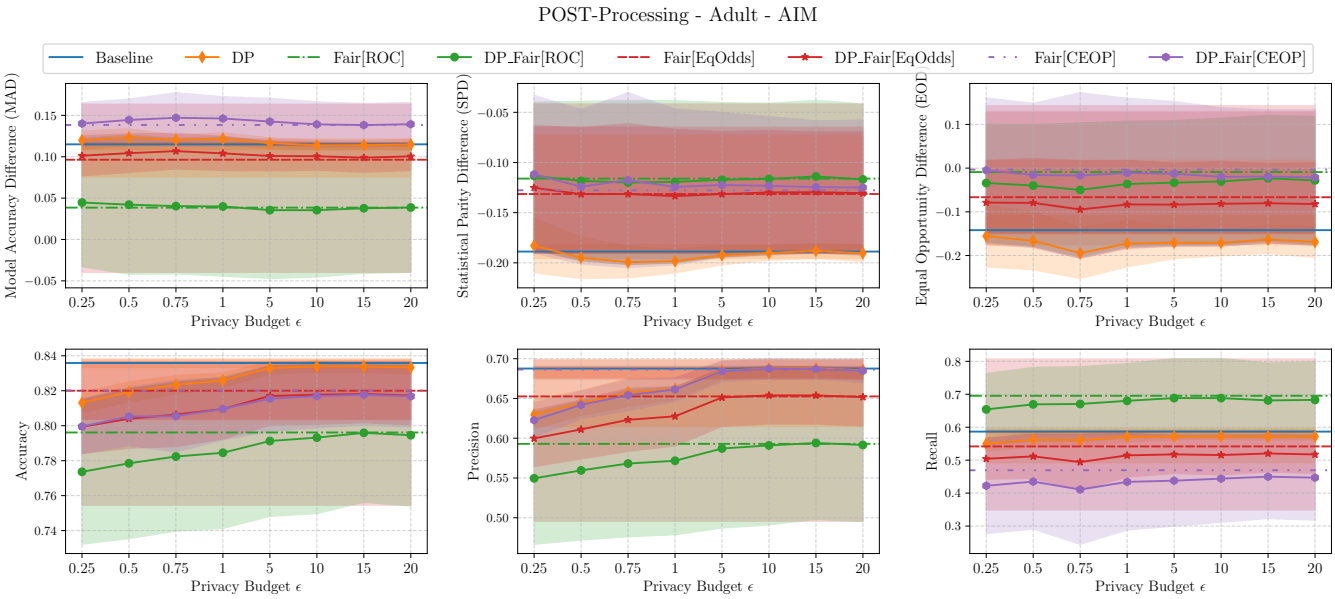

Figure 6: Fairness (MAD, SPD, EOD) and utility (Accuracy, Precision, Recall) metrics across privacy budgets ε on the Adult dataset using the AIM synthesizer with post-processing fairness mechanisms (ROC, EqOdds, CEOP). Solid lines denote means and shaded areas indicate standard deviations across runs.

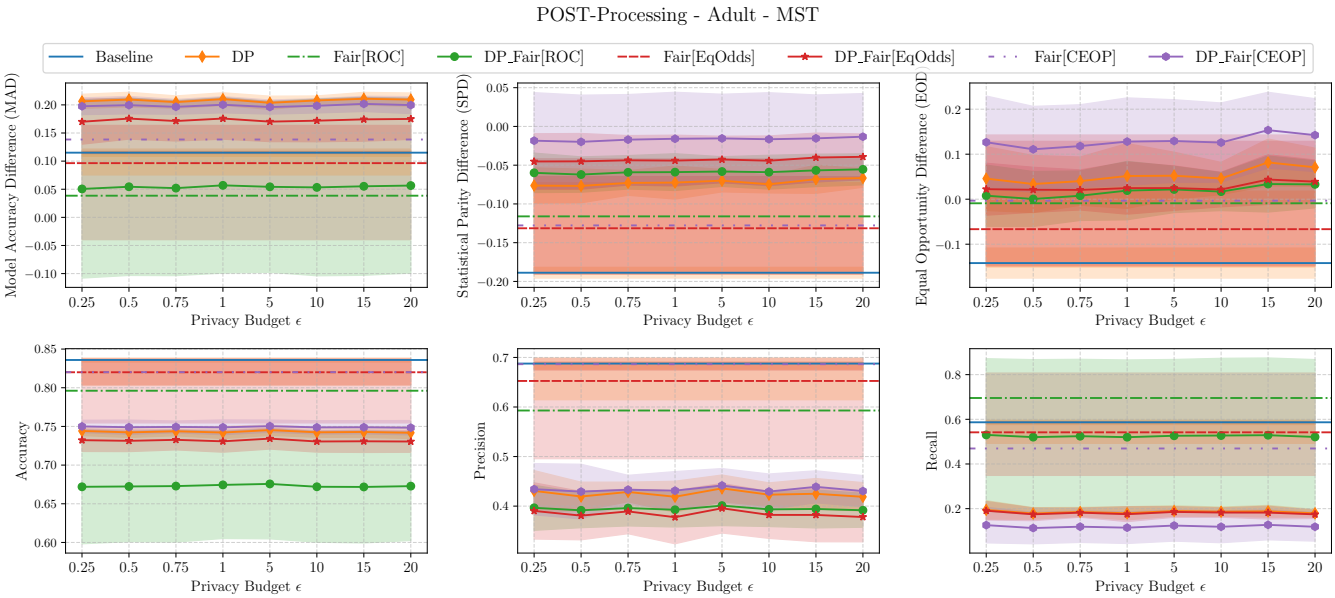

Figure 7: Fairness (MAD, SPD, EOD) and utility (Accuracy, Precision, Recall) metrics across privacy budgets ε on the Adult dataset using the MST synthesizer with post-processing fairness mechanisms (ROC, EqOdds, CEOP). Solid lines denote means and shaded areas indicate standard deviations across runs.

7 DISCUSSION

The results in Section 6 allow us to revisit the central research question: *where should one intervene in the ML pipeline to mitigate unfairness when training on DP synthetic data?* Figures 8 and 9 provide a compact overview at a fixed privacy budget of $\varepsilon = 1$, highlighting the comparative effectiveness of fairness interventions.

Where to intervene. Our benchmark reveals that not all points of intervention are equally effective. Pre-processing methods (RW,

DIR, LFR) can substantially reduce group disparities, but they consistently trade fairness improvements for utility degradation. This is expected, as these methods alter the input distribution itself, often at the expense of signal needed for accurate classification. In contrast, in-processing methods (EGR, GSR) deliver only modest fairness gains. EGR shows slight improvements but fails to reach parity levels close to `Fair-only`, while GSR is often indistinguishable from `DP-only`. By treating fairness as an optimization constraint, in-processing methods appear sensitive to DP-induced noise, which reduces their capacity to enforce constraints reliably. Post-processing stands out: ROC and EqOdds consistently reduce disparities while maintaining accuracy close to `DP-only`, and in some cases even recover recall. ***This makes post-processing the most effective intervention stage overall.***

Synthesizer-dependent effects. The choice of DP synthesizer critically mediates the effectiveness of fairness mechanisms. With AIM, interventions are able to correct much of the bias introduced by DP, and `DP+Fair` outcomes converge toward `Fair-only` results. With MST, disparities are smaller under fairness interventions, but this comes at the cost of sharp utility degradation. As a result, fairness mechanisms have less leverage to improve outcomes. This illustrates an important trade-off: synthesizers that better preserve utility (AIM) create more room for fairness interventions to be effective, while synthesizers that suppress disparities by themselves (MST) do so at the expense of predictive performance. Practitioners should therefore not only choose *where* to intervene, but also recognize that the DP synthesizer determines the headroom for improvements.

Mechanism-specific insights. Among individual mechanisms, several patterns deserve mention. RW generally reduces SPD but at a substantial utility cost. LFR aggressively enforces parity but is unstable: under DP-induced noise it occasionally collapses target classes, making training unreliable. DIR contributes little under uniform preprocessing conditions, suggesting that it may require tailored data pre-processing configurations to be effective. EGR shows moderate robustness across datasets, but its improvements are incremental and fragile under low privacy budgets. GSR largely fails to correct disparities, highlighting the limitations of brute-force reduction approaches under noisy synthetic data. ROC is particularly noteworthy: across datasets and privacy levels, it not only reduces disparities but often increases recall while keeping accuracy stable. This suggests that ROC effectively reallocates decision thresholds in ways that exploit residual predictive signal in DP synthetic data. EqOdds achieves comparable fairness improvements, though with slightly weaker utility recovery, while CEOP is unstable and sometimes counterproductive. Taken together, these observations identify ***ROC as the most reliable mechanism for balancing privacy, fairness, and utility***.

Implications. The broader implication of our findings is that fairness interventions are not uniformly transferable from non-private to DP settings. While post-processing remains reliable, pre- and in-processing methods are more fragile under DP-induced distributional shifts. Moreover, the utility–fairness trade-off interacts strongly with the synthesizer: mechanisms that are effective with AIM may have limited effect under MST. For practitioners, this suggests two guidelines: (i) if utility preservation is a priority, combine AIM with post-processing interventions (especially ROC); (ii) if fairness parity is critical regardless of utility loss, RW or LFR may be considered, but their costs must be acknowledged.

Key takeaway. Overall, our benchmark shows that although DP amplifies fairness–utility trade-offs, carefully chosen interventions can still deliver models that balance privacy guarantees with equitable outcomes.

8 CONCLUSION AND PERSPECTIVES

We introduced the first systematic benchmark of fairness-aware learning on *differentially private (DP) synthetic tabular data*, spanning two state-of-the-art marginal-based synthesizers (AIM [38] and MST [37]), three classes of fairness interventions (pre-processing, in-processing, post-processing), four datasets, and a wide range of privacy budgets. Across configurations, we find that while DP alone generally reduces utility and worsens group disparities, *fairness interventions can partially recover fairness*. Of the three categories, *post-processing methods are the most effective and the most stable across ε values and synthesizers*, often restoring group-fairness metrics close to their non-private counterparts with modest utility cost. These results provide *actionable guidance* on where to intervene in DP-synthetic pipelines.

Limitations. Our study focuses on *tabular, binary classification* with a *binary protected attribute*; while this is the most studied setting in fairness literature [10, 24, 42, 51], extensions to multi-class tasks, regression, and multi-valued or intersecting protected attributes remain open. We only evaluate *marginal-based* DP synthesizers as they outperform *deep generative* DP synthesizers for tabular data [28, 41, 47]. We use a *single base classifier (XGBoost)* with default hyperparameters for comparability; other learners and tuned settings may interact differently with DP and fairness mechanisms. Finally, our fairness evaluation centers on *accuracy disparity and error disparity*; broader notions (*e.g.*, full Equalized Odds, calibration gaps, counterfactual or individual fairness) are not assessed.

Future directions. Future work could (i) extend to *additional synthesizer families*, (ii) explore *multi-task and multi-label* settings with *richer protected attributes*, (iii) evaluate *additional learners* and *hyperparameter regimes*, and (iv) enlarge the metric suite to include *calibration, cost-sensitive utility*, and *individual fairness*. We release code and artifacts to support *reproducibility* and enable follow-up studies on these axes.

ACKNOWLEDGMENTS

This work has been partially supported by the French National Research Agency (ANR), under contract "ANR-24-CE23-6239" JCJC project AI-PULSE.

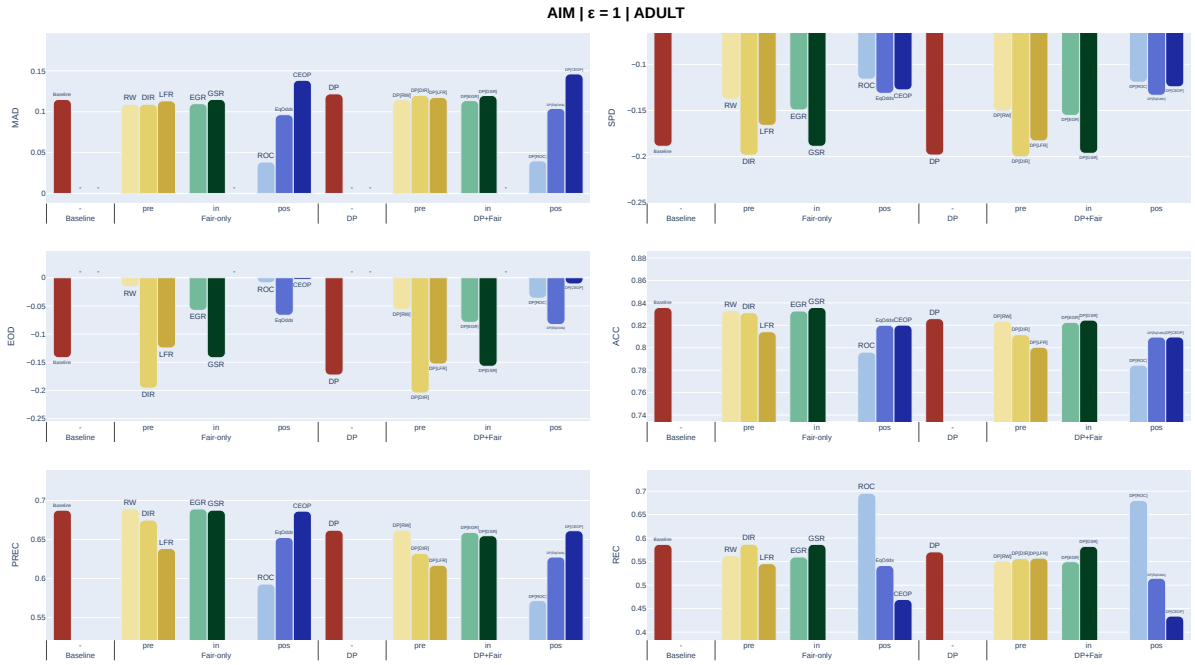

Figure 8: Overview of fairness (MAD, SPD, EOD) and utility (Accuracy, Precision, Recall) metrics on the Adult dataset at privacy budget $\varepsilon = 1$ using the AIM synthesizer across all fairness mechanisms (pre-processing: RW, DIR, LFR; in-processing: EGR, GSR; post-processing: ROC, EqOdds, CEOP). Each bar group compares `Baseline`, `Fair-only`, `DP-only`, and `DP+Fair` configurations.

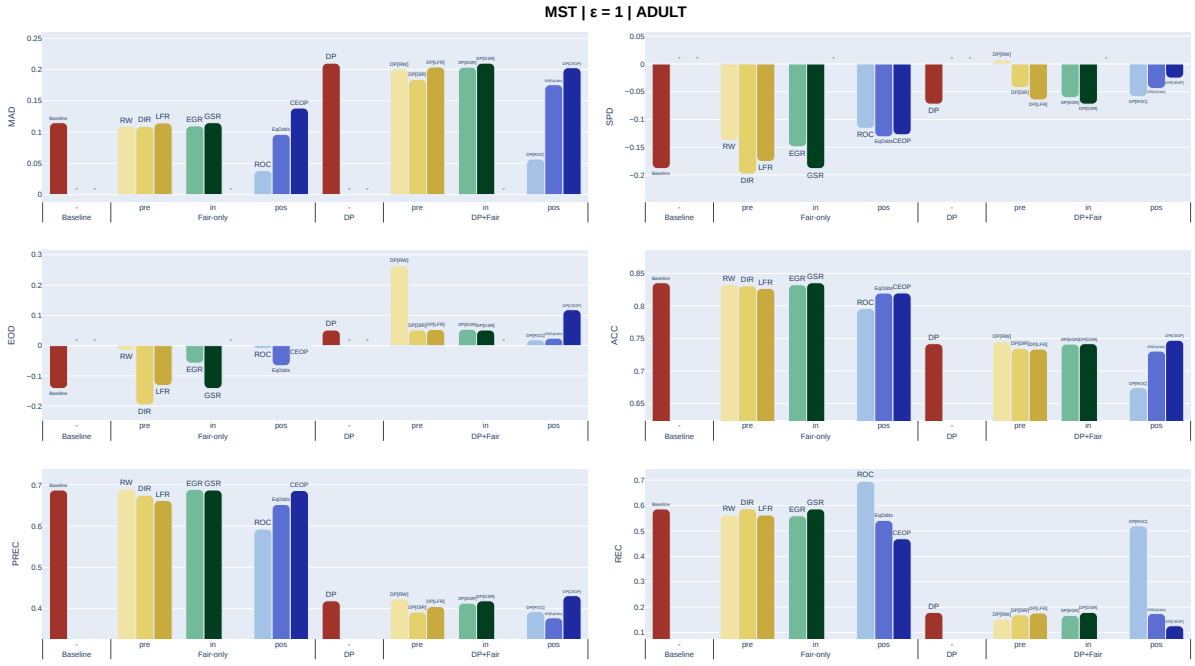

Figure 9: Overview of fairness (MAD, SPD, EOD) and utility (Accuracy, Precision, Recall) metrics on the Adult dataset at privacy budget $\varepsilon = 1$ using the MST synthesizer across all fairness mechanisms (pre-processing: RW, DIR, LFR; in-processing: EGR, GSR; post-processing: ROC, EqOdds, CEOP). Each bar group compares `Baseline`, `Fair-only`, `DP-only`, and `DP+Fair` configurations.

REFERENCES

[1] [n.d.]. Tumult Labs. https://www.tmlt.io/differentially-private-synthetic-data

[2] [n.d.]. YData. https://ydata.ai/products/synthesizer.html

[3] Jan Aalmoes, Vasisht Duddu, and Antoine Boutet. 2025. On the Alignment of Group Fairness with Attribute Privacy. In *International Conference on Web Information Systems Engineering*. Springer, 333–348. https://doi.org/10.1007/978-981-96-0567-5_24

[4] Martin Abadi, Andy Chu, Ian Goodfellow, H. Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. 2016. Deep Learning with Differential Privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security* (Vienna, Austria) *(CCS '16)*. Association for Computing Machinery, New York, NY, USA, 308–318. https://doi.org/10.1145/2976749.2978318

[5] Alekh Agarwal, Alina Beygelzimer, Miroslav Dudik, John Langford, and Hanna Wallach. 2018. A Reductions Approach to Fair Classification. In *Proceedings of the 35th International Conference on Machine Learning (Proceedings of Machine Learning Research)*, Jennifer Dy and Andreas Krause (Eds.), Vol. 80. PMLR, 60–69. https://proceedings.mlr.press/v80/agarwal18a.html

[6] Vinícius Gabriel Angelozzi and Héber H. Arcolezi. 2025. Where to Intervene? Benchmarking Fairness-Aware Learning on Differentially Private Synthetic Tabular Data. Full Version of this paper. Available at: https://github.com/vinicius-verona/dp-fair-intervention-benchmark/blob/dev/paper_full_version.pdf.

[7] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2022. Machine bias. In *Ethics of data and analytics*. Auerbach Publications, 254–264.

[8] Héber H Arcolezi, Mina Alishahi, Adda-Akram Bendoukha, and Nesrine Kaaniche. 2025. Fair Play for Individuals, Foul Play for Groups? Auditing Anonymization's Impact on ML Fairness. *arXiv preprint arXiv:2505.07985* (2025).

[9] Eugene Bagdasaryan, Omid Poursaeed, and Vitaly Shmatikov. 2019. Differential privacy has disparate impact on model accuracy. *Advances in neural information processing systems* 32 (2019).

[10] Solon Barocas, Moritz Hardt, and Arvind Narayanan. 2023. *Fairness and Machine Learning: Limitations and Opportunities*. MIT Press.

[11] Joachim Baumann, Alessandro Castelnovo, Riccardo Crupi, Nicole Inverardi, and Daniele Regoli. 2023. Bias on Demand: A Modelling Framework That Generates Synthetic Data With Bias. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency* (Chicago, IL, USA) *(FAccT '23)*. Association for Computing Machinery, New York, NY, USA, 1002–1013. https://doi.org/10.1145/3593013.3594058

[12] Barry Becker and Ronny Kohavi. 1996. Adult. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5XW20.

[13] Rachel K. E. Bellamy, Kuntal Dey, Michael Hind, Samuel C. Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilovic, Seema Nagar, Karthikeyan Natesan Ramamurthy, John Richards, Diptikalyan Saha, Prasanna Sattigeri, Moninder Singh, Kush R. Varshney, and Yunfeng Zhang. 2018. AI Fairness 360: An Extensible Toolkit for Detecting, Understanding, and Mitigating Unwanted Algorithmic Bias. arXiv:1810.01943 [cs.AI] https://arxiv.org/abs/1810.01943

[14] Blake Bullwinkel, Kristen Grabarz, Lily Ke, Scarlett Gong, Chris Tanner, and Joshua Allen. 2022. Evaluating the fairness impact of differentially private synthetic data. *arXiv preprint arXiv:2205.04321* (2022).

[15] Toon Calders, Faisal Kamiran, and Mykola Pechenizkiy. 2009. Building Classifiers with Independency Constraints. In *2009 IEEE International Conference on Data Mining Workshops*. 13–18. https://doi.org/10.1109/ICDMW.2009.83

[16] Hongyan Chang and Reza Shokri. 2021. On the Privacy Risks of Algorithmic Fairness. In *2021 IEEE European Symposium on Security and Privacy (EuroS&P)*. 292–303. https://doi.org/10.1109/EuroSP51992.2021.00028

[17] 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.

[18] CNIL. [n.d.]. AI system development: CNIL's recommendations to comply with the GDPR. https://www.cnil.fr/en/ai-system-development-cnils-recommendations-comply-gdpr

[19] Frances Ding, Moritz Hardt, John Miller, and Ludwig Schmidt. 2021. Retiring adult: New datasets for fair machine learning. *Advances in Neural Information Processing Systems* 34 (2021).

[20] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. 2006. Calibrating Noise to Sensitivity in Private Data Analysis. In *Theory of Cryptography*, Shai Halevi and Tal Rabin (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 265–284.

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

[22] Tom Farrand, Fatemehsadat Mireshghallah, Sahib Singh, and Andrew Trask. 2020. Neither private nor fair: Impact of data imbalance on utility and fairness in differential privacy. In *Proceedings of the 2020 workshop on privacy-preserving machine learning in practice*. 15–19.

[23] Michael Feldman, Sorelle A. Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. 2015. Certifying and Removing Disparate Impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Sydney, NSW, Australia) *(KDD '15)*. Association for Computing Machinery, New York, NY, USA, 259–268. https://doi.org/10.1145/2783258.2783311

[24] Ferdinando Fioretto, Cuong Tran, Pascal Van Hentenryck, and Keyu Zhu. 2022. Differential Privacy and Fairness in Decisions and Learning Tasks: A Survey. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence*. IJCAI. https://doi.org/10.24963/ijcai.2022/766

[25] FSA. [n.d.]. Feedback Statement on Synthetic Data Call for Input. https://www.fca.org.uk/publications/feedback-statements/fs23-1-feedback-statement-synthetic-data-call-for-input

[26] Georgi Ganev and Emiliano De Cristofaro. 2025. The Inadequacy of Similarity-Based Privacy Metrics: Privacy Attacks Against "Truly Anonymous" Synthetic Datasets. In *2025 IEEE Symposium on Security and Privacy (SP)*. 4007–4025. https://doi.org/10.1109/SP61157.2025.00218

[27] 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*. PMLR, 6944–6959.

[28] Georgi Ganev, Kai Xu, and Emiliano De Cristofaro. 2024. Graphical vs. Deep Generative Models: Measuring the Impact of Differentially Private Mechanisms and Budgets on Utility. In *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security* (Salt Lake City, UT, USA) *(CCS '24)*. Association for Computing Machinery, New York, NY, USA, 1596–1610. https://doi.org/10.1145/3658644.3690215

[29] 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.

[30] Moritz Hardt, Eric Price, and Nati Srebro. 2016. Equality of opportunity in supervised learning. *Advances in neural information processing systems* 29 (2016).

[31] Yuzheng Hu, Fan Wu, Qinbin Li, Yunhui Long, Gonzalo Munilla 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)*. 4696–4713. https://doi.org/10.1109/SP54263.2024.00002

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

[33] Faisal Kamiran, Asim Karim, and Xiangliang Zhang. 2012. Decision theory for discrimination-aware classification. In *2012 IEEE 12th international conference on data mining*. IEEE, 924–929.

[34] Tongyu Liu, Ju Fan, Guoliang Li, Nan Tang, and Xiaoyong Du. 2024. Tabular data synthesis with generative adversarial networks: design space and optimizations. *The VLDB Journal* 33, 2 (2024), 255–280.

[35] Karima Makhlouf, Heber H Arcolezi, Sami Zhioua, Ghassen Ben Brahim, and Catuscia Palamidessi. 2024. On the impact of multi-dimensional local differential privacy on fairness. *Data Mining and Knowledge Discovery* 38, 4 (2024), 2252–2275. https://doi.org/10.1007/s10618-024-01031-0

[36] Karima Makhlouf, Tamara Stefanović, Héber H. Arcolezi, and Catuscia Palamidessi. 2024. A Systematic and Formal Study of the Impact of Local Differential Privacy on Fairness: Preliminary Results. In *2024 IEEE 37th Computer Security Foundations Symposium (CSF)*. 1–16. https://doi.org/10.1109/CSF61375.2024.00039

[37] 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).

[38] Ryan McKenna, Brett Mullins, Daniel Sheldon, and Gerome Miklau. 2022. AIM: an adaptive and iterative mechanism for differentially private synthetic data. *Proc. VLDB Endow.* 15, 11 (July 2022), 2599–2612. https://doi.org/10.14778/3551793.3551817

[39] 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*.

[40] European Parliament and Council. 2024. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act) (Text with EEA relevance). https://eur-lex.europa.eu/eli/reg/2024/1689/oj.

[41] Mayana Pereira, Meghana Kshirsagar, Sumit Mukherjee, Rahul Dodhia, Juan Lavista Ferres, and Rafael de Sousa. 2024. Assessment of differentially private synthetic data for utility and fairness in end-to-end machine learning pipelines for tabular data. *PLOS ONE* 19, 2 (Feb. 2024), e0297271. https://doi.org/10.1371/journal.pone.0297271

[42] Dana Pessach and Erez Shmueli. 2022. A Review on Fairness in Machine Learning. *ACM Comput. Surv.* 55, 3, Article 51 (Feb. 2022), 44 pages. https://doi.org/10.1145/3494672

[43] Geoff Pleiss, Manish Raghavan, Felix Wu, Jon Kleinberg, and Kilian Q Weinberger. 2017. On Fairness and Calibration. In *Advances in Neural Information Processing Systems*, I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.), Vol. 30. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2017/file/b8b9c74ac526fffbeb2d39ab038d1cd7-Paper.pdf

[44] Zhaozhi Qian, Rob Davis, and Mihaela Van Der Schaar. 2023. Synthcity: a benchmark framework for diverse use cases of tabular synthetic data. *Advances in neural information processing systems* 36 (2023), 3173–3188.

[45] Lucas Rosenblatt, Bernease Herman, Anastasia Holovenko, Wonkwon Lee, Joshua Loftus, Elizabeth McKinnie, Taras Rumezhak, Andrii Stadnik, Bill Howe, and Julia Stoyanovich. 2023. Epistemic Parity: Reproducibility as an Evaluation Metric for Differential Privacy. *Proc. VLDB Endow.* 16, 11 (July 2023), 3178–3191. https://doi.org/10.14778/3611479.3611517

[46] Theresa Stadler, Bristena Oprisanu, and Carmela Troncoso. 2022. Synthetic Data – Anonymisation Groundhog Day. In *31st USENIX Security Symposium (USENIX Security 22)*. USENIX Association, Boston, MA, 1451–1468. https://www.usenix.org/conference/usenixsecurity22/presentation/stadler

[47] 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).

[48] Reihaneh Torkzadehmahani, Peter Kairouz, and Benedict Paten. 2019. Dp-cgan: Differentially private synthetic data and label generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*. 0–0.

[49] Archit Uniyal, Rakshit Naidu, Sasikanth Kotti, Sahib Singh, Patrik Joslin Kenfack, Fatemehsadat Mireshghallah, and Andrew Trask. 2021. Dp-sgd vs pate: Which has less disparate impact on model accuracy? *arXiv preprint arXiv:2106.12576* (2021).

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

[51] Kai Yao and Marc Juarez. 2025. SoK: What Makes Private Learning Unfair?. In *2025 IEEE Conference on Secure and Trustworthy Machine Learning*. 841–857. https://doi.org/10.1109/SaTML64287.2025.00052

[52] Zexi Yao, Nataša Krčo, Georgi Ganev, and Yves-Alexandre de Montjoye. 2025. The DCR Delusion: Measuring the Privacy Risk of Synthetic Data. *arXiv preprint arXiv:2505.01524* (2025).

[53] Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. 2013. Learning fair representations. In *International conference on machine learning*. PMLR, 325–333.

[54] Jun Zhang, Graham Cormode, Cecilia M. Procopiuc, Divesh Srivastava, and Xiaokui Xiao. 2017. PrivBayes: Private Data Release via Bayesian Networks. *ACM Trans. Database Syst.* 42, 4, Article 25 (Oct. 2017), 41 pages. https://doi.org/10.1145/3134428