Quantitative Auditing of AI Fairness with Differentially Private Synthetic Data

ANONYMOUS AUTHOR(S)

Ensuring fairness in AI systems is crucial, but auditing these systems often requires sensitive data, which raises security and privacy concerns. To address this, we introduce a framework that utilizes Differentially Private synthetic data for fairness evaluations. This approach allows auditors to assess AI systems without exposing sensitive information. Through experiments on datasets like Adult, COMPAS, and Diabetes, we compare fairness metrics of synthetic and real data. Our results ... (to be continued)

CCS Concepts: • **General and reference → Metrics**; • **Security and privacy → Usability in security and privacy**; *Privacy-preserving protocols*; • **Information systems** → *Data mining*; *Information systems applications*; • **Computing methodologies** → *Machine learning*; *Artificial intelligence*; • **Mathematics of computing** → *Contingency table analysis*.

Additional Key Words and Phrases: AI, fairness, auditing, differential privacy, synthetic data

ACM Reference Format:

Anonymous Author(s). 2024. Quantitative Auditing of AI Fairness with Differentially Private Synthetic Data. 1, 1 (December 2024), 11 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 Introduction

Fairness in machine learning has become an important topic as AI systems become widely used. Biased AI systems could result in amplified unfairness. Ensuring fairness in AI systems is crucial to prevent these biases from being exacerbated by AI systems. One approach is to audit the fairness of existing AI systems. Quantitative fairness auditing is a process of evaluating the fairness of AI systems with a range of fairness metrics. Many metrics have been proposed to help evaluate the fairness of AI systems.

To conduct audits, auditors may rely on access to real datasets containing sensitive information, raising significant security and privacy concerns. This reliance invites data security attacks on auditors, making them targets of unauthorized access. Analysis on real data runs the risk of data inference attacks, where confidential information may be deduced from seemingly harmless or aggregated data. Various technologies have been developed to mitigate these risks, such as Differential Privacy, where the idea is that whether or not any individual is in a dataset has a limited impact.

We propose a novel auditing framework that leverages Differentially Private synthetic data to evaluate the fairness of AI systems. Our framework lets the auditors assess the fairness of AI systems without exposing sensitive information. Differentially Private synthetic data is a technique that generates synthetic data from real data that preserves some statistical properties of the real data while ensuring Differential Privacy. This technology enables auditing without exposing sensitive information. However, synthetic data also introduces new challenges. Whether or not it accurately preserves the fairness properties of the real data is a critical question.

Our work examines the capabilities of existing Differentially Private synthetic data generation technologies in preserving fairness properties. Through empirical experiments on multiple real datasets, we aim to determine the effectiveness of Differentially Private synthetic data in this respect.

Manuscript submitted to ACM

2 Related Work

Fairness in machine learning has become a crucial area of research[4]. Various fairness measures have been proposed to quantify[62] the fairness of AI models[8, 14, 15, 15, 24, 29, 46, 64]. We consider these fairness measures in this work.

Auditing of AI models is an important step in ensuring that these models are fair[22]. Different tools have been developed to check the fairness of AI models[7, 9, 51]. We use a tool produced from our previous work[63].

Differential Privacy has risen to become the standard of privacy protection in many data analyses[25]. It offers a formal framework for quantifying privacy guarantees when releasing information derived from sensitive data[18, 19]. We use Differential Privacy in this work to protect the privacy of individuals in the data in our auditing framework.

Different flavors of Differential Privacy have been proposed over the years[16], such as Gaussian Differential Privacy[17], Pufferfish Differential Privacy[28], Bayesian Differential Privacy[54], and Rényi Differential Privacy[40]. We consider Rényi Differential Privacy specifically in this work.

The generation of synthetic data[33, 47] under Differential Privacy constraints allows for data sharing without compromising privacy[53]. There have been several techniques developed for Differentially Private synthetic data generation[1, 10, 11, 20, 49, 61] in recent years. We use the technique that won the 2018 NIST Differential Privacy Synthetic Data Challenge competition[38].

There have also been some works that aim to alter the properties of synthetic data, such as introducing bias to them[5, 26]. In this work, we aim to preserve the properties of the original data. Thus, we do not consider these works.

While helpful, there are some technical limitations[13, 23, 52, 60] and ethical concerns[59] with the use of synthetic data. Synthetic data may not always preserve properties of real data. This may cause inaccuracies in downstream tasks such as fairness evaluation. Our work examines these limitations.

3 Preliminaries

Let $\prod_i x_i$ denote the Cartesian product of x_is.

An *attribute* is a symbol A. The set of all attributes is $\mathcal{A}$. An *attribute value* is a symbol a. The set of all attribute values is Ω. An *attribute value space* is a function $\sigma : \mathcal{A} \to 2^{\Omega}$ specifying the set of valid values that attributes can take on. A *row* with respect to σ is a function $r : \mathcal{A} \to \Omega$ where $r(A) \in \sigma(A)$. The set of all rows is $\mathcal{R}$. A *database* is a tuple $D = (\mathcal{A}_D, \Omega_D, \sigma_D, \mathcal{R}_D)$ where $\mathcal{A}_D \subseteq \mathcal{A}$ is the set of attributes in D, $\Omega_D \subseteq \Omega$ is the set of attribute values in D, $\sigma_D : \mathcal{A}_D \to 2^{\Omega_D}$ is the function specifying the valid values that each attribute can take on in D, and $\mathcal{R}_D \subseteq \mathcal{R}$ is the set of rows with respect to σ_D in D with $r_D : \mathcal{A}_D \to \Omega_D$ for all $r_D \in \mathcal{R}_D$ and $r_D(A) \in \sigma_D(A)$ for all $A \in \mathcal{A}_D$. The set of all databases is $\mathcal{D}$.

Let $X \in \mathcal{X}$ be a discrete random variable and $p(x) := \Pr[X = x]$. Its *marginal Shannon entropy* is $H(X) := -\sum_{x\in\mathcal{X}} p(x)\log p(x)$. It quantifies the level of uncertainty of X. Let $Y \in \mathcal{Y}$ be another discrete random variable. Their *joint Shannon entropy* is $H(X,Y) := -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log p(x,y)$.

Let X, Y be discrete random variables. Their *mutual information* is $I(X;Y) := H(X) + H(Y) - H(X,Y)$. It quantifies the level of dependence between X and Y. By definition[56], the joint entropy is greater than or equal to the marginal entropies $H(X,Y) \geq \max(H(X), H(Y))$. Hence, we have an upper bound of mutual information $I(X;Y) \leq \min(H(X), H(Y))$.

Let $\{X_i\}$ be a set of discrete random variables indexed by a graph $G = (V, E)$, where V represents the random variables X_is and E represents dependencies between these random variables. A *Markov random field* is a probability distribution over X_is, such that each random variable X_i, given its neighborhood in G, is conditionally independent of all other variables. Since edges represent dependencies, cliques in a Markov random field represent groups of variables that are all mutually dependent. As a machine learning model, there has been much development in the estimation and inference of Markov random fields[30, 41].

3.1 Fairness Measures

For fairness measures[46, 63], let Y denote the ground truth of an outcome, let $\hat{Y}$ denote the predicated result of an outcome, let S denote the protected attribute, and let ϵ denote some threshold. $Y, \hat{Y}, S$ are binary. For non-binary prediction, such as a score, we use $\hat{V}$.

Fairness measures can be broadly categorized into independence, separation, and sufficiency, which are defined by conditional independence in Table 1. $X \perp Y | Z$ denotes the conditional independence between X and Y conditioning on Z.

Table 1. Fairness categories.

Category	Definition
Independence	$S \perp \hat{Y}$
Separation	$S \perp \hat{Y} \vert Y$
Sufficiency	$S \perp Y \vert \hat{Y}$

These categories can be expanded into forms of probability. For example, the definition of separation is expanded to

$$P[\hat{Y} = 1 | S = 1, Y = 1] = P[\hat{Y} = 1 | S \neq 1, Y = 1]$$
$$P[\hat{Y} = 1 | S = 1, Y = 0] = P[\hat{Y} = 1 | S \neq 1, Y = 0]$$

The definition can be relaxed. Its relaxation, for some parameter ϵ, is

$$|P[\hat{Y} = 1 | S = 1, Y = 1] - P[\hat{Y} = 1 | S \neq 1, Y = 1]| \leq \epsilon$$
$$|P[\hat{Y} = 1 | S = 1, Y = 0] - P[\hat{Y} = 1 | S \neq 1, Y = 0]| \leq \epsilon$$

which is also the definition of a fairness measure called equalized odds.

We consider in this work various fairness measures listed in Table 2.

3.2 Rényi Differential Privacy

A *randomized mechanism* is a randomized algorithm $M : \mathcal{D} \to \mathbb{R}^p$ that takes a database and, after introducing noise, outputs some results.

Let D_1, D_2 be two databases. They are *neighbors*, denoted $D_1 \sim D_2$, if $|\mathcal{R}_{D_1} \Delta \mathcal{R}_{D_2}| \in \{1, 2\}$ where Δ denotes symmetric difference; that is, either one database contains an extra row, or both databases have all but one row in common. In other words, they differ in exactly one row.

Definition 3.1 (Gaussian Mechanism[19]). Let $f : \mathcal{D} \to \mathbb{R}^p$ be a function. The Gaussian Mechanism M adds independent and identically distributed Gaussian noise with mean 0 and standard deviation σ to each component of the p-dimensional vector output of $f(D)$

$$M(D) = f(D) + \mathcal{N}(0_p, \sigma^2 \mathbb{I}_p)$$

Table 2. Fairness measures.

Category	Fairness Measure	Definition
Independence	Disparate Impact	$\frac{P[\hat{Y}=1\|S\neq 1]}{P[\hat{Y}=1\|S=1]} \geq 1-\epsilon$
	Demographic Parity	$\|P[\hat{Y}=1\|S=1]-P[\hat{Y}=1\|S\neq 1]\| \leq \epsilon$
	Conditional Statistical Parity	$\|P[\hat{Y}=1\|S=1,L=l]-P[\hat{Y}=1\|S\neq 1,L=l]\| \leq \epsilon$
	Mean Difference	$\|E[\hat{Y}\|S=1]-E[\hat{Y}\|S\neq 1]\| \leq \epsilon$
Separation	Equalized Odds	$\|P[\hat{Y}=1\|S=1,Y=0]-P[\hat{Y}=1\|S\neq 1,Y=0]\| \leq \epsilon$ $\|P[\hat{Y}=1\|S=1,Y=1]-P[\hat{Y}=1\|S\neq 1,Y=1]\| \leq \epsilon$
	Equal Opportunity	$\|P[\hat{Y}=1\|S=1,Y=1]-P[\hat{Y}=1\|S\neq 1,Y=1]\| \leq \epsilon$
	Predictive Equality	$\|P[\hat{Y}=1\|S=1,Y=0]-P[\hat{Y}=1\|S\neq 1,Y=0]\| \leq \epsilon$
Sufficiency	Conditional Use Accuracy Equality	$\|P[Y=1\|S=1,\hat{Y}=1]-P[Y=1\|S\neq 1,\hat{Y}=1]\| \leq \epsilon$ $\|P[Y=0\|S=1,\hat{Y}=0]-P[Y=0\|S\neq 1,\hat{Y}=0]\| \leq \epsilon$
	Predictive Parity	$\|P[Y=1\|S=1,\hat{Y}=1]-P[Y=1\|S\neq 1,\hat{Y}=1]\| \leq \epsilon$
	Equal Calibration	$\|P[Y=1\|S=1,\hat{V}=v]-P[Y=1\|S\neq 1,\hat{V}=v]\| \leq \epsilon$
N/A	Accuracy Equality	$\|P[Y=\hat{Y}\|S=1]-P[Y=\hat{Y}\|S\neq 1]\| \leq \epsilon$
	Positive Balance	$\|E[\hat{V}\|Y=1,S=1]-E[\hat{V}\|Y=1,S\neq 1]\| \leq \epsilon$
	Negative Balance	$\|E[\hat{V}\|Y=0,S=1]-E[\hat{V}\|Y=0,S\neq 1]\| \leq \epsilon$

where $\mathcal{N}$ is a multivariate normal distribution with mean vector 0_p and covariance matrix $\sigma^2\mathbb{I}_p$ where $\mathbb{I}_p$ is the identity matrix.

Definition 3.2 (Rényi Differential Privacy (RDP)). Let $P_{\mathbf{X}}$ denote the probability distribution induced by the random vector $\mathbf{X}$. A randomized mechanism M satisfies (α,γ)-RDP for $\alpha \geq 1$ and $\gamma \geq 1$ if, for all databases $D_1 \sim D_2$, we have

$$D_\alpha(P_{M(D_1)} \| P_{M(D_2)}) \leq \gamma$$

where $D_\alpha(P_1\|P_2)$ is the Rényi divergence[32, 57, 58] of order α between probability distributions P_1, P_2 over x

$$D_\alpha(P_1\|P_2) := \frac{1}{\alpha-1}\log\int P_1(x)^\alpha P_2(x)^{1-\alpha}\mathrm{d}x$$

Theorem 3.3 (RDP of the Gaussian Mechanism[21, 40]). *The Gaussian Mechanism satisfies $(\alpha, \alpha\frac{\Delta_f^2}{2\sigma^2})$-RDP, where Δ_f denotes the sensitivity[19] of f, which is defined as the maximum L^2-norm difference in the output of f*

$$\Delta_f := max_{D_1\sim D_2}\|f(D_1)-f(D_2)\|_2$$

3.3 Differentially Private Synthetic Data

Let $C \subseteq \mathcal{A}$ be a subset of attributes. Let $\Omega_C = \prod_{i\in C}\Omega_i$. Let x be a row and x_C denote the restriction of x to C. A *marginal*[3, 38] of C on database D is a function $\mu_D : \Omega_C \to \mathbb{N}_0$ such that $\mu_D(t) = \Sigma_{x\in\mathcal{D}}\delta_{t,x_C}$ where δ is the Kronecker function; that is, it is a lookup table of the counts of each possible combination of attribute values. We call marginals of $|C| = n$ attributes n-way marginals.

The task of Differentially Private synthetic data[35, 36, 44, 55] is, given a database D, adding some noise to marginals of D such that it satisfies some Differential Privacy guarantees and outputting another database D', such that the L^1-norm errors between marginals of D and D' is small; that is, their marginals $\mu_D, \mu_{D'}$ are similar.

For example, suppose we have a database with attributes sex and race. The 2-way marginals of the original database and the synthetic database are shown in Table 3. The marginals of the synthetic data is supposed to be similar to that of the original database.

Table 3. Example marginals.

(a) Marginal of original data.

Attributes	Count
Male,White	24
Female,White	33
Male,Black	13
Female,Black	47

(b) Marginal of synthetic data.

Attributes	Count
Male,White	22
Female,White	35
Male,Black	10
Female,Black	46

4 Motivation

Our auditing framework is tripartite. It consists of three parties: the data provider, the model maker, and the third-party auditor.

The data provider is responsible for supplying the raw datasets which should originate from trustworthy sources, such as government agencies like a census bureau.

The model maker develops AI models. These are AI companies or research labs specialized in training and optimizing AI models.

The third-party auditor acts as an evaluator, using our tool to audit the AI models for fairness issues by combining both the datasets and the models. These may be investigative journalists or regulatory bodies.

In the framework of our previous work[63], after obtaining real data from the data provider, the 3rd party auditor holds onto the real data for performing fairness audits, and it supposedly retains it indefinitely for the possibility of any future audits. However, this practice raises both security and privacy concerns.

For security, it creates a point of vulnerability of unauthorized access, and the auditor is now a target of data security attacks. The auditor may not have the necessary resources to defend against these threats. A breach at the auditor's end could result in compromises of individuals' sensitive information.

As for privacy, on the other hand, it introduces risks of information leakage. Releasings of analyses on real data may inadvertently reveal sensitive information by data inference attacks. Attackers can exploit data patterns in outputs or combine outputs with external data to infer sensitive information.

Thus, we introduce a new framework where the auditor generates synthetic data based on real data upon retrieval of the real data, and then holds onto the synthetic data and discards the real data, preventing all further security and privacy violations. The third-party still retains the ability to audit all incoming future models as needed.

5 Methodology

We employed the tools of the winner of the 2018 NIST Differential Privacy Synthetic Data Challenge competition[43] by [34, 37–39] and the fairness checker tool from our previous research[63].

This research is conducted in Python Jupyter notebooks and is publicly available.

5.1 Data Synthesis

The synthesis framework is three-fold, namely, select-measure-generate[35, 36]. We first select the important marginals to preserve, measure them by adding Differential Privacy noise, and then generate synthetic data.

Underneath the hood, the tool employs a Markov random field. The select step corresponds to marking cliques in a Markov random field, and the generate step corresponds to sampling from the fitted Markov random field.

By default, all 1-way marginals are selected to preserve the quantity of each attribute element. We can further preserve correlations by adding n-way marginals. For example, if we want to preserve the relationship between sex and race, we may add the clique (sex,race).

In a perfect world where all correlation information is to be preserved, we may wish to make a complete graph. However, this is intractable as the complexity of the problem would skyrocket. Furthermore, algorithms for Markov random field favor graphs with specific shapes, such as trees.

To circumvent this complexity explosion, instead, [38] devised a technique where the mutual information of all the database attribute pairs is calculated, and then a maximum spanning tree is identified with edge weights being the mutual information to obtain a skeleton tree-shaped Markov random field.

For the competition, [38] further manually added certain cliques based on his investigation of the competition dataset. For example, they manually added the clique (sex,city,income). In addition, they would add some edges based on some sophisticated heuristics tailored to that particular dataset. Meanwhile, his other apporach where the auditor does access the real dataset does not fit our framework.

We hence developed an alternative heuristic for the general-purpose workflow. As mentioned in Section 3, the mutual information of two random variables is bounded by the pair's respective Shannon entropy. Using this property, we add additional edges with weights exceeding a fraction of the minimum of these upper bounds. As a rule of thumb, we have found setting the fraction to be 0.1 to be effective.

For the measure step, we followed the examples provided in the tool's repository. Gaussian noises are added to the selected marginals. Half of the privacy budget is spent on all 1-way marginals and the other half on the selected cliques. These marginals are then fed to the tool to fit the Markov random field. By [38], this procedure satisfies $(\alpha, \frac{\alpha}{2\sigma^2})$-RDP for all $\alpha \geq 1$.

5.2 Fairness Checking

After synthesizing the dataset, we used the fairness checker from [63] to compute the fairness measures of any incoming AI model.

The fairness checker is an open-sourced public domain Python package that computes various fairness measures, such as those mentioned in Table 2.

The checker is designed to be user-friendly and agnostic to the underlying AI model. It is also designed to be easily extensible to accommodate new fairness measures.

The checker simply iterates through the given database D and computes the results based on some given predicates on the rows r_is, and finally outputs the fairness measure values.

Protected groups S, predicted outcomes $\hat{Y}$, and ground truths Y are all formulated as these predicates. These are straightforward logical boolean expressions. Specifically, they are given as Python functions that output boolean values.

For example, if the sensitive attribute is sex and the protected group is female, the protected group predicate would be $S := r(sex) = Female$. This can be easily implemented in Python as a comparison function.

The interpretation of the resulting fairness measure values is dependent on the third-party auditors. The auditors may have different thresholds ϵ for different fairness measures or different AI models.

6 Experiments

To test the viability of our method, we compare the metrics computed from the synthetic dataset against those of the original dataset. We used various datasets with fairness concerns mentioned in [46].

We looked at several publicly available datasets, such as Adult[6, 48], COMPAS[31, 45], and Diabetes[27, 42]. The Adult dataset comes from the 1994 census in the United States and contains about 30000 individuals. The COMPAS dataset comes from an investigative report by ProPublica of the COMPAS criminal recidivism assessment system and contains about 7000 individuals. The Diabetes dataset comes from the hospital readmission data published in the 1994 AI in Medicine journal and contains about 100000 individuals. Since these datasets all have binary outcomes, we did not consider the case of non-binary outcomes in the experiments.

The fairness checker evaluates datasets based on multiple fairness metrics, such as demographic parity and equalized odds. These metrics are computed on some sensitive attributes, predicted outcomes, and ground truths. Examples of sensitive attributes are race and sex. Examples of predicted outcomes and ground truths are loan approval and criminal recidivism.

By comparing these measures between the synthetic and original datasets, we aim to ensure that the synthetic data preserves the fairness properties of the original data. The comparison process is three-fold. It goes as follows.

The dataset is first processed so it can be fed into the synthetic data generator. Some marginals are selected as described in the Section 5, and the synthetic data generator model is fitted to the original data according to the marginals. Then the generator is run multiple times to obtain multiple sets of synthetic data.

Next, several AI models are extracted from various real-life authors from Kaggle. They are finetuned to perform well on the original dataset. For one, a random forest model is finetuned by searching hyperparameters settings[12]. Another random forest model is finetuned with over-sampling methods[50]. Also, a logistic regression model is finetuned by performing principal component analysis[2].

Several AI models and both the original dataset and the rounds of synthetic datasets are fed to the fairness checker. Sensitive attributes are identified based on manual examination with common sense or by referring to [46]. Then, all applicable fairness measures are computed using the checker for both the original and the synthetic.

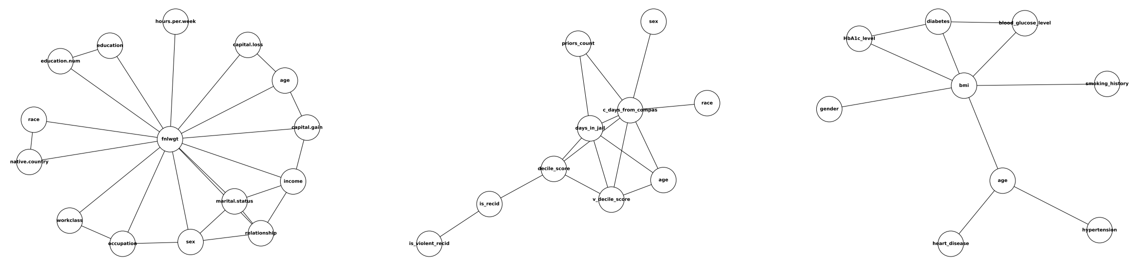
Finally, we analyze the discrepancies between the fairness properties of the original and the synthetic by calculating the difference of their perspective fairness measure values. The average of the differences serves as a summary of the analysis.

6.1 Adult Income Dataset

For the Adult income dataset, the shape of the maximum spanning tree is very shallow, almost resembling a star; it has one internal node and all but one of the leaves have a depth of one. After introducing edges according to our heuristic, we observed an increase in the pairwise edges of the leaves, forming many 3-cliques and two 4-cliques. The resulting graph is shown in Figure 1a.

Marginals based on this graph are then passed to the synthetic data generator for model fitting.

After generating ten rounds of synthetic data and passing them to the checker, their fairness measure values are averaged. Then, we compare them against the values of the original data. The results are shown in Table 4.



(a) Markov random field for marginals of the Adult dataset.

(b) Markov random field for marginals of the COMPAS dataset.

(c) Markov random field for marginals of the Diabetes dataset.

Fig. 1. Markov random field for marginals of the experimented datasets.

Table 4. Fairness measures experiment results of the Adult dataset. With heuristic, average difference is 0.0787, and fit time is 924m15s. Without heuristic, average difference is 0.0775, and fit time is 17m1s.

Measure	Original	Synthetic	Difference	Synthetic(No Heuristic)	Difference(No Heuristic)
Demographic Parity	0.1933	0.1079	0.0853	0.0114	0.1818
Accuracy Eqaulity	0.0250	0.1399	0.1149	0.0169	0.0080
Equalized Odds 1	0.0175	0.0870	0.0695	0.0060	0.0115
Equalized Odds 2	0.0114	0.0622	0.0508	0.0303	0.0189
Accuracy Equality 1	0.1941	0.1466	0.0474	0.0315	0.1626
Accuracy Equality 2	0.1102	0.1450	0.0348	0.0299	0.0803

6.2 COMPAS Dataset

Table 5. Fairness measures experiment results of the COMPAS dataset. With heuristic, average difference is 0.0750, and fit time is 27m33s. Without heuristic, average difference is 0.0727, and fit time is 33s.

Measure	Original	Synthetic	Difference	Synthetic(No Heuristic)	Difference(No Heuristic)
Demographic Parity	0.1310	0.0980	0.0330	0.0809	0.0501
Accuracy Eqaulity	0.0079	0.0129	0.0050	0.0146	0.0067
Equalized Odds 1	0.0249	0.0936	0.0687	0.0802	0.0553
Equalized Odds 2	0.0177	0.1016	0.0838	0.0825	0.0648
Accuracy Equality 1	0.1709	0.0505	0.1203	0.0551	0.1157
Accuracy Equality 2	0.1695	0.0303	0.1392	0.0256	0.1439

Table 6. Fairness measures experiment results of the Diabetes dataset. With heuristic, average difference is 0.0067, and fit time is 1m44s. Without heuristic, average difference is 0.0078, and fit time is 1m24s.

Measure	Original	Synthetic	Difference	Synthetic(No Heuristic)	Difference(No Heuristic)
Demographic Parity	0.0135	0.0038	0.0096	0.0015	0.0119
Accuracy Eqaulity	0.0077	0.0011	0.0065	0.0011	0.0065
Equalized Odds 1	0.0000	0.0006	0.0006	0.0010	0.0010
Equalized Odds 2	0.0086	0.0106	0.0020	0.0094	0.0008
Accuracy Equality 1	0.0133	0.0090	0.0042	0.0059	0.0074
Accuracy Equality 2	0.0216	0.0041	0.0175	0.0019	0.0197

6.3 Diabetes Dataset

7 Evaluation

7.1 Positive Results

7.2 Negative Results

independence

$$P[\hat{Y}|S] - P[\hat{Y}|\overline{S}] = \frac{P[\hat{Y} \cap S]}{P[S]} - \frac{P[\hat{Y} \cap \overline{S}]}{P[\overline{S}]}$$

separation

$$P[\hat{Y}|S, Y] - P[\hat{Y}|\overline{S}, Y] = \frac{P[\hat{Y} \cap S \cap Y]}{P[S \cap Y]} - \frac{P[\hat{Y} \cap \overline{S} \cap Y]}{P[\overline{S} \cap Y]}$$

sufficiency

$$P[Y|S, \hat{Y}] - P[Y|\overline{S}, \hat{Y}] = \frac{P[Y \cap S \cap \hat{Y}]}{P[S \cap \hat{Y}]} - \frac{P[Y \cap \overline{S} \cap \hat{Y}]}{P[\overline{S} \cap \hat{Y}]}$$

accuracy equality

$$P[Y = \hat{Y}|S] - P[Y = \hat{Y}|\overline{S}] = \frac{P[Y = \hat{Y} \cap S]}{P[S]} - \frac{P[Y = \hat{Y} \cap \overline{S}]}{P[\overline{S}]}$$

8 Conclusion

References

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

[2] Prashant Banerjee. [n. d.]. EDA, Logistic Regression, PCA. https://www.kaggle.com/code/prashant111/eda-logistic-regression-pca Accessed: 2024-11-10.

[3] Boaz Barak, Kamalika Chaudhuri, Cynthia Dwork, Satyen Kale, Frank McSherry, and Kunal Talwar. 2007. Privacy, accuracy, and consistency too: a holistic solution to contingency table release. In *Proceedings of the twenty-sixth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*. 273–282.

[4] Solon Barocas, Moritz Hardt, and Arvind Narayanan. 2023. *Fairness and machine learning: Limitations and opportunities*. MIT press.

[5] 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*. 1002–1013.

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

[7] Rachel KE Bellamy, Kuntal Dey, Michael Hind, Samuel C Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilović, et al. 2019. AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. *IBM Journal of*

Research and Development 63, 4/5 (2019), 4–1.

[8] Richard Berk, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth. 2021. Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research* 50, 1 (2021), 3–44.

[9] Sarah Bird, Miro Dudík, Richard Edgar, Brandon Horn, Roman Lutz, Vanessa Milan, Mehrnoosh Sameki, Hanna Wallach, and Kathleen Walker. 2020. Fairlearn: A toolkit for assessing and improving fairness in AI. *Microsoft, Tech. Rep. MSR-TR-2020-32* (2020).

[10] Claire McKay Bowen and Fang Liu. 2020. Comparative study of differentially private data synthesis methods. *Statist. Sci.* 35, 2 (2020), 280–307.

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

[12] Isaac Byrne. [n. d.]. Income Prediction (84.369% Accuracy). https://www.kaggle.com/code/ipbyrne/income-prediction-84-369-accuracy Accessed: 2024-11-10.

[13] Victoria Cheng, Vinith M Suriyakumar, Natalie Dullerud, Shalmali Joshi, and Marzyeh Ghassemi. 2021. Can you fake it until you make it? impacts of differentially private synthetic data on downstream classification fairness. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. 149–160.

[14] Alexandra Chouldechova. 2017. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data* 5, 2 (2017), 153–163.

[15] Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. 2017. Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd acm sigkdd international conference on knowledge discovery and data mining*. 797–806.

[16] Damien Desfontaines and Balázs Pejó. 2019. Sok: differential privacies. *arXiv preprint arXiv:1906.01337* (2019).

[17] Jinshuo Dong, Aaron Roth, and Weijie J Su. 2022. Gaussian differential privacy. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 84, 1 (2022), 3–37.

[18] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. 2006. Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography: Third Theory of Cryptography Conference, TCC 2006, New York, NY, USA, March 4-7, 2006. Proceedings 3*. Springer, 265–284.

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

[20] Ju Fan, Tongyu Liu, Guoliang Li, Junyou Chen, Yuwei Shen, and Xiaoyong Du. 2020. Relational data synthesis using generative adversarial networks: A design space exploration. *arXiv preprint arXiv:2008.12763* (2020).

[21] Vitaly Feldman, Ilya Mironov, Kunal Talwar, and Abhradeep Thakurta. 2018. Privacy amplification by iteration. In *2018 IEEE 59th Annual Symposium on Foundations of Computer Science (FOCS)*. IEEE, 521–532.

[22] Emilio Ferrara. 2023. Fairness and bias in artificial intelligence: A brief survey of sources, impacts, and mitigation strategies. *Sci* 6, 1 (2023), 3.

[23] 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.

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

[25] Honglu Jiang, Jian Pei, Dongxiao Yu, Jiguo Yu, Bei Gong, and Xiuzhen Cheng. 2021. Applications of differential privacy in social network analysis: A survey. *IEEE transactions on knowledge and data engineering* 35, 1 (2021), 108–127.

[26] Lan Jiang, Clara Belitz, and Nigel Bosch. 2024. Synthetic Dataset Generation for Fairer Unfairness Research. In *Proceedings of the 14th Learning Analytics and Knowledge Conference*. 200–209.

[27] Michael Kahn. [n. d.]. Diabetes. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5T59G.

[28] Daniel Kifer and Ashwin Machanavajjhala. 2012. A rigorous and customizable framework for privacy. In *Proceedings of the 31st ACM SIGMOD-SIGACT-SIGAI symposium on Principles of Database Systems*. 77–88.

[29] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. 2016. Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807* (2016).

[30] Daphane Koller. 2009. Probabilistic Graphical Models: Principles and Techniques.

[31] Jeff Larson, Surya Mattu, Lauren Kirchner, and Julia Angwin. 2016. ProPublica Compas Analysis—Data and Analysis for 'Machine Bias'. https://github.com/propublica/compas-analysis.

[32] Yingzhen Li and Richard E Turner. 2016. Rényi divergence variational inference. *Advances in neural information processing systems* 29 (2016).

[33] Yingzhou Lu, Minjie Shen, Huazheng Wang, Xiao Wang, Capucine van Rechem, Tianfan Fu, and Wenqi Wei. 2023. Machine learning for synthetic data generation: a review. *arXiv preprint arXiv:2302.04062* (2023).

[34] Ryan McKenna. 2018. Differential Privacy Synthetic Data Challenge Algorithms. https://github.com/usnistgov/PrivacyEngCollabSpace/tree/master/tools/de-identification/Differential-Privacy-Synthetic-Data-Challenge-Algorithms/rmckenna. Accessed: 2024-12-05.

[35] Ryan McKenna. 2021. Marginal-based Methods for Differentially Private Synthetic Data. https://www.youtube.com/watch?v=UKzh9QgNRxA. Accessed: 2024-11-26.

[36] Ryan McKenna and Terrance Liu. 2022. A Simple Recipe for Private Synthetic Data Generation. https://differentialprivacy.org/synth-data-1/ Accessed: 2024-11-10.

[37] Ryan McKenna, Gerome Miklau, and Daniel Sheldon. 2021. *Private-PGM*. https://github.com/journalprivacyconfidentiality/private-pgm-jpc-778/tree/v2021-10-04-jpc

[38] 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).

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

[40] Ilya Mironov. 2017. Rényi differential privacy. In *2017 IEEE 30th computer security foundations symposium (CSF)*. IEEE, 263–275.

[41] Kevin P Murphy. 2023. *Probabilistic machine learning: Advanced topics*. MIT press.

[42] Mohammed Mustafa. [n. d.]. Diabetes Prediction Dataset. https://www.kaggle.com/datasets/iammustafatz/diabetes-prediction-dataset Accessed: 2024-11-26.

[43] National Institute of Standards and Technology. 2018. 2018 Differential Privacy Synthetic Data Challenge. https://www.nist.gov/ctl/pscr/open-innovation-prize-challenges/past-prize-challenges/2018-differential-privacy-synthetic Accessed: 2024-11-10.

[44] Joseph Near and David Darais. 2021. Differentially Private Synthetic Data. https://www.nist.gov/blogs/cybersecurity-insights/differentially-private-synthetic-data. Accessed: 2024-11-26.

[45] Dan Ofer. [n. d.]. COMPAS Dataset. https://www.kaggle.com/datasets/danofer/compass Accessed: 2024-11-26.

[46] Dana Pessach and Erez Shmueli. 2022. A review on fairness in machine learning. *ACM Computing Surveys (CSUR)* 55, 3 (2022), 1–44.

[47] Trivellore E Raghunathan. 2021. Synthetic data. *Annual review of statistics and its application* 8, 1 (2021), 129–140.

[48] UC Irvine Machine Learning Repository. [n. d.]. Adult Census Income Dataset. https://www.kaggle.com/datasets/uciml/adult-census-income/data Accessed: 2024-11-26.

[49] Lucas Rosenblatt, Xiaoyan Liu, Samira Pouyanfar, Eduardo de Leon, Anuj Desai, and Joshua Allen. 2020. Differentially private synthetic data: Applied evaluations and enhancements. *arXiv preprint arXiv:2011.05537* (2020).

[50] Panjawat Rungtranont. [n. d.]. Diabetes: EDA | Random Forest + HP. https://www.kaggle.com/code/tumpanjawat/diabetes-eda-random-forest-hp Accessed: 2024-11-26.

[51] Pedro Saleiro, Benedict Kuester, Loren Hinkson, Jesse London, Abby Stevens, Ari Anisfeld, Kit T Rodolfa, and Rayid Ghani. 2018. Aequitas: A bias and fairness audit toolkit. *arXiv preprint arXiv:1811.05577* (2018).

[52] Theresa Stadler, Bristena Oprisanu, and Carmela Troncoso. 2022. Synthetic data–anonymisation groundhog day. In *31st USENIX Security Symposium (USENIX Security 22)*. 1451–1468.

[53] 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).

[54] Aleksei Triastcyn and Boi Faltings. 2020. Bayesian differential privacy for machine learning. In *International Conference on Machine Learning*. PMLR, 9583–9592.

[55] Jonathan Ullman. 2022. What is Synthetic Data? https://differentialprivacy.org/synth-data-0/ Accessed: 2024-11-10.

[56] user20160 (https://stats.stackexchange.com/users/116440/user20160). [n. d.]. Joint entropy of multivariate normal distribution less than individual entropy under high correlation. Cross Validated. arXiv:https://stats.stackexchange.com/q/427884 https://stats.stackexchange.com/q/427884 URL:https://stats.stackexchange.com/q/427884 (version: 2019-09-19).

[57] Tim van Erven and Peter Harremoës. 2010. Rényi divergence and majorization. In *2010 IEEE International Symposium on Information Theory*. IEEE, 1335–1339.

[58] Tim Van Erven and Peter Harremos. 2014. Rényi divergence and Kullback-Leibler divergence. *IEEE Transactions on Information Theory* 60, 7 (2014), 3797–3820.

[59] Cedric Deslandes Whitney and Justin Norman. 2024. Real Risks of Fake Data: Synthetic Data, Diversity-Washing and Consent Circumvention. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*. 1733–1744.

[60] Sierra Wyllie, Ilia Shumailov, and Nicolas Papernot. 2024. Fairness feedback loops: training on synthetic data amplifies bias. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*. 2113–2147.

[61] 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).

[62] Min-Hsuan Yeh, Blossom Metevier, Austin Hoag, and Philip Thomas. 2024. Analyzing the Relationship Between Difference and Ratio-Based Fairness Metrics. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*. 518–528.

[63] Chih-Cheng Rex Yuan and Bow-Yaw Wang. 2024. Ensuring Fairness with Transparent Auditing of Quantitative Bias in AI Systems. In *2024 Pacific Neighborhood Consortium Annual Conference and Joint Meetings (PNC)*. IEEE, 25–32.

[64] Indrė Žliobaitė. 2017. Measuring discrimination in algorithmic decision making. *Data Mining and Knowledge Discovery* 31, 4 (2017), 1060–1089.