Quantitative Auditing of AI Fairness with Differentially Private Synthetic Data

ANONYMOUS AUTHOR(S)

Auditing AI systems often requires sensitive data, raising security and privacy concerns. To address this, we propose a framework that leverages differentially private synthetic data for fairness evaluations. This approach enables auditors to assess AI systems without exposing sensitive information. Through experiments on real datasets like Adult, COMPAS, and Diabetes, we compare fairness metrics of synthetic and real data. Our results show that synthetic data accurately reflect the fairness properties of real data to a reasonable extent.

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), 13 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.

Auditing can evaluate the fairness of AI systems, sometimes revealing cases of unfairness in them. For example, investigative journalists at ProPublica found that the COMPAS system, which is used to predict recidivism in the criminal justice system in the United States, is biased against certain demographic groups. This highlights the need for fairness auditing conducted by third parties.

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.

To mitigate these risks, the use of synthetic data has been proposed. Synthetic data are fake data that is artificially generated from real data. The principle of this technology is to preserve some statistical properties of the real data while protecting the privacy of the individuals in the real data. Among the generation techniques, differentially private synthetic data offers a way to investigate the data properties under differential privacy guarantees.

Manuscript submitted to ACM

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.

2 Related Work

Fairness in machine learning has become a crucial area of research[4]. Various fairness measures have been proposed to quantify[61] the fairness of AI models[8, 14, 15, 15, 24, 29, 46, 63]. 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[62].

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, 60] 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, 59] and ethical concerns[58] 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 $X \in \mathcal{X}, Y \in \mathcal{Y}$ be discrete random variables. Their *mutual information* $I(X;Y)$ is

$$I(X;Y) = \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} P_{(X,Y)}(x,y) \log \frac{P_{(X,Y)}(x,y)}{P_X(x) P_Y(y)}$$

where $P_{(X,Y)}$ is the joint probability distribution of X, Y and P_X, P_Y are the marginal probability distributions of X, Y respectively. It quantifies the level of dependency between X and 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, 62], 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.

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	$S \perp \hat{Y} \mid Y$
Sufficiency	$S \perp Y \mid \hat{Y}$

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.

Table 2. Fairness measures.

Category	Fairness Measure	Definition
Independence	Demographic Parity	$\lvert P[\hat{Y}=1\lvert S=1]-P[\hat{Y}=1\lvert S\neq 1]\rvert \leq \epsilon$
Separation	Equalized Odds (False Positive)	$\lvert P[\hat{Y}=1\lvert S=1,Y=0]-P[\hat{Y}=1\lvert S\neq 1,Y=0]\rvert \leq \epsilon$
	Equalized Odds (True Positive)	$\lvert P[\hat{Y}=1\lvert S=1,Y=1]-P[\hat{Y}=1\lvert S\neq 1,Y=1]\rvert \leq \epsilon$
Sufficiency	Conditional Use Accuracy Equality (True Positive)	$\lvert P[Y=1\lvert S=1,\hat{Y}=1]-P[Y=1\lvert S\neq 1,\hat{Y}=1]\rvert \leq \epsilon$
	Conditional Use Accuracy Equality (True Negative)	$\lvert P[Y=0\lvert S=1,\hat{Y}=0]-P[Y=0\lvert S\neq 1,\hat{Y}=0]\rvert \leq \epsilon$
N/A	Overall Accuracy Equality	$\lvert P[Y=\hat{Y}\lvert S=1]-P[Y=\hat{Y}\lvert S\neq 1]\rvert \leq \epsilon$

Let D_1, D_2 be two databases. They are *neighbors*, denoted $D_1 \sim D_2$, if $|\mathcal{R}_{D_1} \Delta \mathcal{R}_{D_2}| = 1$ where Δ denotes symmetric difference; that is, one database contains an extra row than the other. 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 $f(D)$

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

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 associated with 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, 56, 57] 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

AI systems are widely used in various aspects in our society. However, these systems may be biased against certain demographic groups. For example, COMPAS, the AI system used in American criminal justice system to predict recidivism, has been found to be biased by ProPublica, a third party media.

To ensure the fairness of AIs, third-party audits such as this is crucial. Auditing can reveal instances of unfairness in AI systems, allowing for corrective actions to be taken. We hence advocate for a framework that centers around third-party audits.

Our auditing framework is tripartite. It consists of three parties: the data provider, the model maker, and the third-party auditor. The COMPAS investigation could fit in with this framework.

The data provider is responsible for supplying the raw datasets which should originate from trustworthy sources, such as government agencies like a census bureau. In the COMPAS example, the data provider is the government agencies that provide criminal justice data, such as the Broward County Sheriff's Office in Florida.

The model maker develops AI models. These are AI companies or research labs specialized in training and optimizing AI models. The model maker in the COMPAS example is Northpointe, Inc., the company that developed the COMPAS system.

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. ProPublica is the third-party auditor in the COMPAS example, using quantitative methods to evaluate the fairness of the COMPAS system using Florida's data.

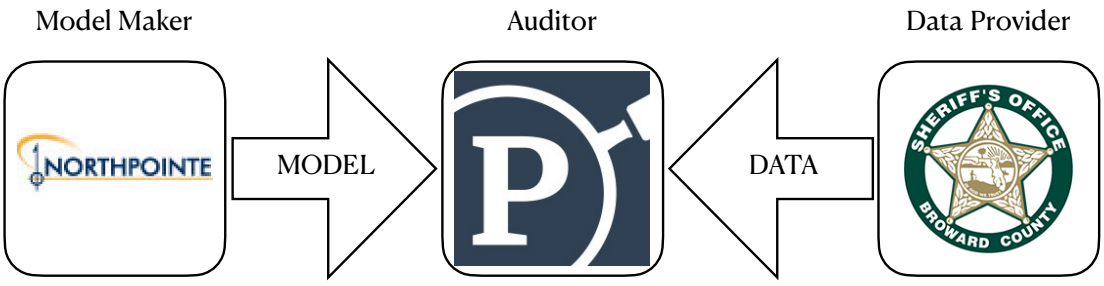

Fig. 1. Roles of each party in the COMPAS example.

In the framework of our previous work[62], 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 to 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. In the COMPAS example, had the data not been public already, ProPublica may be targeted by hackers to access the sensitive criminal justice data.

As for privacy, on the other hand, it introduces risks of information leakage. Releasing analyses of real data may inadvertently reveal sensitive information through data inference attacks. Attackers can exploit data patterns in outputs or combine outputs with external data to infer sensitive information. In the COMPAS example, had the data not been public already, their analyses may inadvertently reveal the identities of individuals in the criminal justice system.

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[62].

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

We shall illustrate the workflow of our methodology with an example of the COMPAS dataset, which contains attributes such as sex and race with the prediction target attribute being recidivism.

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. The marginals are then selected based on the edges of this tree.

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. We do not consider these additional heuristics for

the generality of our approach. Meanwhile, his other apporach where the auditor does access the real dataset does not fit our framework.

While synthetic data generation methods often leverage domain knowledge to improve accuracy, we chose to employ a vanilla approach with only the maximum spanning tree method without incorporating additional heuristics. This decision was made due to two main factors: first, the need for more specific domain knowledge about the dataset makes it difficult to identify meaningful dependencies; second, prediction targets may differ from dataset to dataset, which means that tailored heuristics could not be effectively generalized. By avoiding dataset-specific adjustments, our method remains general and applicable to a wide range of datasets, in addition to being more computatioanlly efficient.

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 $\epsilon = 1$ 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 $\alpha \geq 1$.

In the COMPAS example, we may select the marginal (sex,race). This marginal is then perturbed by Gaussian noise, as shown in Table 4.

Table 4. Example marginals.

(a) Original marginal.

Attributes	Count
Female,African-American	1537
Female,Asian	8
Female,Caucasian	1465
Female,Hispanic	215
Female,Native American	15
Female,Other	143
Male,African-American	8254
Male,Asian	63
Male,Caucasian	4621
Male,Hispanic	1236
Male,Native American	42
Male,Other	717

(b) Perturbed marginal.

Attributes	Count
Female,African-American	1545.7885
Female,Asian	4.3236
Female,Caucasian	1444.7303
Female,Hispanic	208.2250
Female,Native American	26.7479
Female,Other	150.8175
Male,African-American	8296.6274
Male,Asian	92.1107
Male,Caucasian	4643.0108
Male,Hispanic	1274.0831
Male,Native American	28.9542
Male,Other	710.3473

5.2 Fairness Checking

After synthesizing the dataset, we used the fairness checker from [62] 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.

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.

In the COMPAS example, the sensitive attribute is sex and the protected group is female, the protected group predicate would be $S := r(sex) = Female$; prediction is if a person will re-offend $\hat{Y} := \mathcal{M}_{recid}(r)$ where $\mathcal{M}$ is the model's prediction of row r; ground truth is whether a person does re-offend $Y := r(recid)$. These can be easily implemented in Python as a comparison function.

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]. These datasets are known to have fairness issues[46].

The fairness checker evaluates datasets based on multiple fairness metrics, such as demographic parity and equalized odds. 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 according to their mutual information, as described in the Section 5, and the synthetic data generator model is fitted to the original data according to the selected marginals. Then the generator is run 100 times to obtain 100 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].

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

6.1 Adult Income Dataset

The Adult[6, 48] dataset comes from the 1994 census in the United States and contains about 30000 individuals. The dataset has 14 attributes, such as sex, age, and education. The task is to predict whether an individual earns more than $50,000 a year. We set the sensitive attribute as sex and the protected group as female. The results are shown in Table 5.

Table 5. Fairness measures experiment results of the Adult dataset. Average difference is 0.0775, and fit time is 17m1s.

Measure	Original	Synthetic	Difference
Demographic Parity	0.1933	0.0114	0.1818
Overall Accuracy Eqaulity	0.0250	0.0169	0.0080
Equalized Odds (False Positive)	0.0175	0.0060	0.0115
Equalized Odds (True Positive)	0.0114	0.0303	0.0189
Conditional Use Accuracy Equality (True Positive)	0.1941	0.0315	0.1626
Conditional Use Accuracy Equality (True Negative)	0.1102	0.0299	0.0803

6.2 COMPAS Dataset

The COMPAS[31, 45] dataset comes from an investigative report by ProPublica of the COMPAS criminal recidivism assessment system and contains about 7000 individuals. The dataset has 10 applicable attributes, such sex, age, and priors count. The task is to predict whether an individual will re-offend. We set the sensitive attribute as sex and the protected group as female. The results are shown in Table 6.

Table 6. Fairness measures experiment results of the COMPAS dataset. Average difference is 0.0727, and fit time is 33s.

Measure	Original	Synthetic	Difference
Demographic Parity	0.1310	0.0809	0.0501
Overall Accuracy Eqaulity	0.0079	0.0146	0.0067
Equalized Odds (False Positive)	0.0249	0.0802	0.0553
Equalized Odds (True Positive)	0.0177	0.0825	0.0648
Conditional Use Accuracy Equality (True Positive)	0.1709	0.0551	0.1157
Conditional Use Accuracy Equality (True Negative)	0.1695	0.0256	0.1439

6.3 Diabetes Dataset

The Diabetes[27, 42] dataset comes from the hospital readmission data published in the 1994 AI in Medicine journal and contains about 100000 individuals. The dataset has 8 attributes, such as gender, age, and smoking history. The task is to predict whether an individual will be readmitted. We set the sensitive attribute as gender and the protected group as female. The results are shown in Table 7.

Table 7. Fairness measures experiment results of the Diabetes dataset. Average difference is 0.0078, and fit time is 1m24s.

Measure	Original	Synthetic	Difference
Demographic Parity	0.0135	0.0015	0.0119
Overall Accuracy Eqaulity	0.0077	0.0011	0.0065
Equalized Odds (False Positive)	0.0000	0.0010	0.0010
Equalized Odds (True Positive)	0.0086	0.0094	0.0008
Condtional Use Accuracy Equality (True Positive)	0.0133	0.0059	0.0074
Condtional Use Accuracy Equality (True Negative)	0.0216	0.0019	0.0197

7 Evaluation

Before we delve into the results, we must first understand the nature of the fairness measures. The fairness measures are defined assuming all the information is readily available. This, however, might cause trouble for synthetic data, as they can only work with the information at hand.

By close examination of our scenario, we can observe the following: of the three variables, $(\hat{Y}, Y, S)$, two of them, namely (Y, S), are both readily available in the original data. In contrast, $\hat{Y}$ is only available in the output of the given AI model, which cannot be captured by the synthetic data at the time of generation.

Let us inspect the definitions of the fairness measures one by one by expanding them by the definition of conditional probability and marking them according to whether they are in the original data or not; we will box the variables that are not in the original data.

For Demographic Parity, we have

$$P[\hat{Y} = 1|S = 1] - P[\hat{Y} = 1|S \neq 1] = \frac{P[\boxed{\hat{Y}} = 1 \cap S = 1]}{P[S = 1]} - \frac{P[\boxed{\hat{Y}} = 1 \cap S \neq 1]}{P[S \neq 1]}$$

For Equalized Odds (False Positive), we have

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

Equalized Odds (True Positive) follows the same form as Equalized Odds (False Positive).

For Conditional Use Accuracy Equality (True Positive), we have

$$P[Y = 1|S = 1, \hat{Y} = 1] - P[Y = 1|S \neq 1, \hat{Y} = 1] = \frac{P[Y = 1 \cap S = 1 \cap \boxed{\hat{Y}} = 1]}{P[S = 1 \cap \boxed{\hat{Y}} = 1]} - \frac{P[Y = 1 \cap S \neq 1 \cap \boxed{\hat{Y}} = 1]}{P[S \neq 1 \cap \boxed{\hat{Y}} = 1]}$$

Conditional Use Accuracy Equality (True Negative) follows the same form as Conditional Use Accuracy Equality (True Positive).

For Overall Accuracy Equality, we have

$$P[Y = \hat{Y}|S = 1] - P[Y = \hat{Y}|S \neq 1] = \frac{P[Y = \boxed{\hat{Y}} \cap S = 1]}{P[S = 1]} - \frac{P[Y = \boxed{\hat{Y}} \cap S \neq 1]}{P[S \neq 1]}$$

Here we observed that, of the measures considered, only Conditional Use Accuracy Equality has boxed variables that are in both the numerator and the denominator. This is in contrast to the other measures, which only have one boxed variable in the numerator.

We can say less about fairness with the measures with less information, and thus the measures approximated by synthetic data may be less accurate for Conditional Use Accuracy Equality.

7.1 Positive Results

Although the scores may not be fully accurate, and some measures, due to the nature of their definitions, may be harder to approximate, the results showed that synthetic data generally preserves the fairness properties of the original data with a reasonable degree of accuracy.

The average of the differences between the fairness measure values of the original and synthetic datasets is below 0.1 for all datasets. This indicates that synthetic data is a good approximation of the original data in terms of fairness.

In addition, further experiments showed that, in an alternative assumption where the auditor has knowledge of (S, Y) and can add the edge (S, Y) to the Markov random field before fitting, the synthesized data would have been more accurate, even lowering the error of Conditional Use Accuracy Equality in the COMPAS example.

This demonstrated that synthetic data can, to an extent, accurately reflect the fairness properties of the original data. Therefore, fairness analyses performed on synthetic data will yield results mostly consistent with those on the original data.

Auditors hence gained an approach to auditing AI fairness alternative to the traditional method of using real data. When using synthetic data derived from real data, the fairness properties remain meaningful and can still be enough for auditing purposes.

7.2 Negative Results

Looking at the actual experimental results, first and foremost, across all datasets, we observed that Conditional Use Accuracy Equality did indeed show higher errors than other measures.

In the Adult experiment, Conditional Use Accuracy Equality (True Positive) has an error of around 0.16. In the COMPAS experiment, both Conditional Use Accuracy Equality measures had an error over 0.1. In the Diabetes experiment, Conditional Use Accuracy Equality (True Negative) has an error much higher than other measures.

This corroborates our assumption that it is harder to approximate sufficiency measures on synthetic data. As there is less information available for the synthesis process, the measures that require more information will be harder to approximate.

Although further experiments showed that linking (S, Y) before fitting can somewhat alleviate this problem, the auditor may not always have the background knowledge to identify them; moreover, this improvement is not always guaranteed.

This is a fundamental limitation of this approach, and it is important to be aware of this when interpreting the results of fairness analyses. The auditor needs to keep in mind that they are approximating scores, and the scores may not be fully accurate.

8 Conclusion

Auditing AI systems with real data can be dangerous. We propose a new approach that leverages differentially private synthetic data. The proposed framework, with its strong privacy guarantees, provides a safe alternative to the risks of using real data.

Despite the inherent imperfection of synthetic data for some measures, the metrics results remain generally consistent with the original data. Experiments on multiple real-world datasets and real-world models reveal that synthetic data derived from real data maintains fairness properties.

While synthetic data is not perfect, it serves as a practical tool for auditing AI fairness. Our framework provides an effective way to audit AI systems without compromising privacy. Future work includes applications on larger datasets to validate scalability and explorations of other means of privacy preservation in AI auditing.

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. 2019. 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. 2017. 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. 2022. 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. 2017. 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. 2016. 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. 2023. 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] Tim van Erven and Peter Harremoës. 2010. Rényi divergence and majorization. In *2010 IEEE International Symposium on Information Theory*. IEEE, 1335–1339.
[57] Tim Van Erven and Peter Harremos. 2014. Rényi divergence and Kullback-Leibler divergence. *IEEE Transactions on Information Theory* 60, 7 (2014), 3797–3820.
[58] 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.
[59] 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.
[60] 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).
[61] 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.
[62] 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.
[63] Indrė Žliobaitė. 2017. Measuring discrimination in algorithmic decision making. *Data Mining and Knowledge Discovery* 31, 4 (2017), 1060–1089.