Quantitative Auditing of AI Fairness with Differentially Private Synthetic Data

ANONYMOUS AUTHOR(S)

Fairness auditing of AI systems can identify and quantify biases. However, traditional auditing using real-world data raises security and privacy concerns. It exposes auditors to security risks as they become custodians of sensitive information and targets for cyberattacks. Privacy risks arise even without direct breaches, as data analyses can inadvertently expose confidential information. To address these, we propose a framework that leverages differentially private synthetic data to audit the fairness of AI systems. By applying privacy-preserving mechanisms, it generates synthetic data that mirrors the statistical properties of the original dataset while ensuring privacy. This method balances the goal of rigorous fairness auditing and the need for strong privacy protections. Through experiments on real datasets like Adult, COMPAS, and Diabetes, we compare fairness metrics of synthetic and real data. By analyzing the alignment and discrepancies between these metrics, we assess the capacity of synthetic data to preserve the fairness properties of real data. Our results demonstrate the framework's ability to enable meaningful fairness evaluations while safeguarding sensitive information, proving its applicability across critical and sensitive domains.

1 Introduction

Fairness in machine learning has become an increasingly important topic as AI systems are widely used across various sectors, including healthcare and criminal justice. These AI systems can significantly influence outcomes that impact individuals and communities.

However, AI systems can exhibit bias, which may lead to increased unfairness. Biased AI systems have the potential to reinforce historical injustices, erode trust in technology, and exacerbate inequality. Therefore, ensuring fairness in AI systems is crucial to prevent the perpetuation of these biases.

For instance, investigative journalists at ProPublica uncovered bias in the COMPAS system, which is used in the United States criminal justice system to predict recidivism. Their analysis revealed that the COMPAS system disproportionately discriminated against certain demographic groups.

These findings have raised significant concerns about the fairness and ethical implications of using such AI systems in critical decision-making processes. Addressing these biases is essential to ensure that AI systems do not perpetuate inequalities.

One effective approach to achieving fairness in AI systems is to conduct audits. Quantitative fairness auditing involves evaluating these systems using various fairness metrics to identify potential biases and inequalities. Some commonly proposed metrics for assessing AI system fairness include demographic parity, equalized odds, and accuracy equality. These metrics provide a quantitative basis for addressing fairness concerns in AI models.

A well-documented example of fairness auditing is the investigation of the COMPAS system. Journalists from ProPublica revealed that COMPAS exhibited bias against certain demographic groups, particularly African Americans. The system was more likely to falsely classify African American defendants as high risk compared to their Caucasian counterparts. This finding not only underscored the necessity of fairness auditing but also highlighted the potential harm of deploying biased AI systems in critical areas.

This case emphasizes the importance of independent audits in ensuring fairness. Evaluations conducted by third parties, such as those performed by ProPublica, are vital because they operate free from the influence of those who develop or deploy the systems. By offering an objective perspective, third-party audits can identify and address biases that internal evaluations might overlook. Without such oversight, AI systems risk perpetuating systemic inequalities, making independent auditing a crucial aspect of accountability in AI development.

While fairness auditing is essential, it presents significant challenges, particularly when it involves sensitive data. To conduct audits, auditors may rely on access to real datasets containing sensitive information. Auditors often require access to real-world datasets containing sensitive attributes, such as race, gender, or income, to evaluate an AI system's fairness. However, this reliance on sensitive data raises concerns about data security and privacy.

This reliance invites data security attacks on auditors, making them targets of unauthorized access. The use of real datasets for fairness auditing creates a point of vulnerability, making auditors potential targets for data breaches and unauthorized access. If sensitive data falls into the wrong hands, it could lead to severe consequences, including identity theft, discrimination, and reputational damage.

Moreover, analyses of real data run the risk of data inference attacks, where confidential information may be deduced from seemingly harmless or aggregated data. Even if the real data is not directly exposed, the results of analyses can sometimes reveal sensitive information through inference attacks. For example, attackers could combine publicly available audit results with external datasets to infer private details about individuals in the original dataset.

Various technologies have been developed to mitigate privacy risks, such as differential privacy. Differential privacy is a mathematical framework that ensures the inclusion or exclusion of any individual in a dataset has a minimal and quantifiable impact on the overall results of data analysis. This is achieved by introducing carefully calibrated random noise into the computations derived from the data, ensuring that individual records remain indistinguishable while preserving aggregate patterns and trends.

Differential privacy is especially useful in situations that involve sensitive data. For example, a differential privacy mechanism could allow researchers to analyze patterns in patient health data without exposing any single patient's medical history. This approach strikes a balance between data utility and privacy, enabling meaningful analysis while safeguarding individuals against privacy breaches or inference attacks.

The key idea is that no matter how much background knowledge an attacker has about the data, they cannot determine with high confidence whether a specific individual is included in the dataset. This is quantified by some parameters which control the amount of noise introduced; smaller parameters provide stronger privacy but can reduce the accuracy of the analysis, while larger parameters allow for more precise results but with weaker privacy guarantees. The challenge is to carefully calibrate the parameters to ensure that privacy is maintained without overly compromising the utility of the data.

One way of auditing while preserving privacy is to use synthetic data. Synthetic data are fake data that is generated from real data. Synthetic data is artificially generated to replicate the statistical properties of real data while protecting privacy. 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. Unlike anonymized data, synthetic data breaks the link to the original data, reducing the risk of re-identification.

Among the generation techniques, differentially private synthetic data offers a way to investigate the data properties under differential privacy guarantees. It stands out for it strong differential privacy guarantees when generating synthetic data. This method combines data synthesis with the principles of differential privacy, ensuring that individuals' sensitive information cannot be inferred from the synthetic data. By introducing controlled randomness during generation, it provides strong privacy guarantees while preserving the overall utility of the data.

Differentially private synthetic data is particularly valuable in sensitive domains like healthcare and criminal justice, enabling researchers to analyze patterns or identify trends without exposing personal details. It also supports fairness auditing by allowing the examination of datasets and models without revealing sensitive attributes. This dual focus on privacy and utility makes it a valuable tool.

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. By using differentially private synthetic data, auditors can conduct fairness analyses without directly exposing sensitive information.

We present an innovative solution to the issue of sensitive data access: auditors generate synthetic data immediately upon receiving the real dataset, which they retain for auditing purposes while discarding the original data.

This method offers a unique solution to reconcile the need for fairness auditing with the imperative to protect privacy. By preserving key statistical properties of the original data, synthetic data allows auditors to assess fairness metrics accurately while mitigating the risks of data breaches and privacy violations.

This technology enables auditing without exposing sensitive information, offering a privacy-preserving way to assess AI systems. However, synthetic data also introduces new challenges. One major concern is whether it accurately reflects the fairness properties of the real data. Whether or not it accurately preserves the fairness properties of the real data is a critical question.

Since fairness metrics depend on the relationships between sensitive attributes and outcomes, any discrepancies in how synthetic data mimics these relationships could lead to inaccurate conclusions in fairness audits. Ensuring that synthetic data preserves these properties is critical for reliable auditing.

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 fairness auditing.

We perform a comparative analysis of fairness metrics computed on real and synthetic data across multiple real-world datasets, including Adult, COMPAS, and Diabetes. By examining the consistency and differences between these metrics, we evaluate how well synthetic data captures the fairness characteristics of real data.

Through our experiments, we evaluate the extent to which synthetic data can approximate fairness metrics computed on original datasets. Our findings contribute to the growing body of research on ethical AI and offer insights for practitioners seeking to ensure fairness in AI systems.

The rest of this paper is organized as follows. Section 2 provides the background of related work. Section 3 provides the preliminary knowledge. Section 4 outlines the motivation behind our proposed framework. Section 5 describes our

methodology for generating differentially private synthetic data and evaluating fairness metrics. Section 6 presents our experimental setup and results. Section 7 discusses the implications of our findings. Section 8 concludes the paper and outlines future work.

2 Related Work

Fairness in machine learning has become a crucial area of research[4]. Various fairness measures have been proposed to quantify[63] the fairness of AI models[8, 14, 15, 15, 24, 29, 47, 65]. 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, 52]. We use a tool produced from [64].

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[56], and Rényi differential privacy[41]. We consider Rényi differential privacy specifically in this work.

The generation of synthetic data[34, 48] under differential privacy constraints allows for data sharing without compromising privacy[55]. There have been several techniques developed for differentially private synthetic data generation[1, 10, 11, 20, 50, 62] in recent years. We use the technique that won the 2018 NIST differential privacy Synthetic Data Challenge competition[39].

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, 54, 61] and ethical concerns[60] 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, 42].

3.1 Fairness Measures

For fairness measures[47, 64], 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	$S \perp \hat{Y} | Y$
Sufficiency	$S \perp Y | \hat{Y}$

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.

Table 2. Fairness measures.

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

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 $|\mathcal{R}_{D_1} \Delta \mathcal{R}_{D_2}| = 1$ 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[33, 58, 59] 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, 41]). *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, 39] 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[36, 37, 45, 57] 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.

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.

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

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.

Model Maker		Auditor		Data Provider
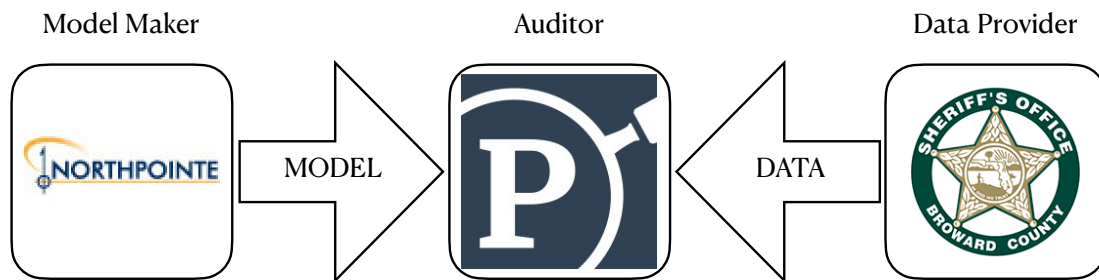
Fig. 1. Roles of each party in the COMPAS example.

In the framework of [64], 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[44] by [35, 38–40] and the fairness checker tool from [64].

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[36, 37]. 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, [39] 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, [39] further manually added certain cliques based on their investigation of the competition dataset. For example, they would add some edges based on some sophisticated heuristics tailored to that particular dataset. In addition, they manually added some particular cliques like (sex,city,income) for the prediction target of income. We do not consider these additional optimizations for the generality of our approach. Meanwhile, their other approach where the auditor does not 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 specific domain knowledge about the dataset may not be always readily obvious; 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 computationally 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 [39], 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.		(b) Perturbed marginal.	
Attributes	Count	Attributes	Count
Female,African-American	1537	Female,African-American	1545.7885
Female,Asian	8	Female,Asian	4.3236
Female,Caucasian	1465	Female,Caucasian	1444.7303
Female,Hispanic	215	Female,Hispanic	208.2250
Female,Native American	15	Female,Native American	26.7479
Female,Other	143	Female,Other	150.8175
Male,African-American	8254	Male,African-American	8296.6274
Male,Asian	63	Male,Asian	92.1107
Male,Caucasian	4621	Male,Caucasian	4643.0108
Male,Hispanic	1236	Male,Hispanic	1274.0831
Male,Native American	42	Male,Native American	28.9542
Male,Other	717	Male,Other	710.3473

5.2 Fairness Checking

After synthesizing the dataset, we used the fairness checker from [64] 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 [47].

We looked at several publicly available datasets, such as Adult[6, 49], COMPAS[32, 46], and Diabetes[27, 43]. These datasets are known to have fairness issues[47].

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 authors from Kaggle. They are finetuned to perform well on the original dataset. For the Adult dataset, we selected three models to experiment: a random forest model is finetuned by hyperparameter search[12]; a logistic regression model is finetuned by performing principal component analysis[2]; and a KNN model is manually finetuned[31]. For the COMPAS dataset, we selected a random forest model finetuned by hyperparameter search[53]. For the Diabetes dataset, we selected a random forest model finetuned by over-sampling and under-sampling techniques[51].

The AI models and both the original dataset and synthetic datasets are then fed to the fairness checker. The synthetic datasets were generated 100 times for each model and the fairness scores averaged across the 100 runs. Sensitive attributes are identified based on manual examination with common sense or by referring to [47]. Then, applicable fairness measures are computed using the checker for both the original and the synthetic.

This research is conducted in Python Jupyter notebooks and is publicly available. The experiments were conducted on a macOS laptop with an Intel Core i9-9880H CPU (2.3GHz) and 32GB of RAM.

6.1 Adult Income Dataset

The Adult[6, 49] 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. We tested three models: a random forest model, a logistic regression model, and a KNN model. The results are shown in Table 5, 6, and 7. The synthetic data generator fit time is 17 minutes and 1 second.

6.2 COMPAS Dataset

The COMPAS[32, 46] 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 8. The synthetic data generator fit time is 33 seconds.

Table 5. Fairness measures experiment results of the Adult dataset on a random forest model. Average difference is 0.0775.

Measure	Original	Synthetic	Difference
Demographic Parity	0.1933	0.0114	0.1818
Overall Accuracy Equality	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

Table 6. Fairness measures experiment results of the Adult dataset on a logistic regression model. Average difference is 0.0829.

Measure	Original	Synthetic	Difference
Demographic Parity	0.0139	0.0047	0.0092
Overall Accuracy Equality	0.0992	0.0149	0.0843
Equalized Odds (False Positive)	0.0017	0.0066	0.0049
Equalized Odds (True Positive)	0.0080	0.0072	0.0008
Conditional Use Accuracy Equality (True Positive)	0.3459	0.0305	0.3154
Conditional Use Accuracy Equality (True Negative)	0.1120	0.0291	0.0829

Table 7. Fairness measures experiment results of the Adult dataset on a KNN model. Average difference is 0.0945.

Measure	Original	Synthetic	Difference
Demographic Parity	0.1721	0.0046	0.1675
Overall Accuracy Equality	0.0473	0.0128	0.0344
Equalized Odds (False Positive)	0.0571	0.0072	0.0499
Equalized Odds (True Positive)	0.1667	0.0110	0.1557
Conditional Use Accuracy Equality (True Positive)	0.1009	0.0288	0.0720
Conditional Use Accuracy Equality (True Negative)	0.1193	0.0319	0.0874

Table 8. Fairness measures experiment results of the COMPAS dataset on a random forest model. Average difference is 0.0727.

Measure	Original	Synthetic	Difference
Demographic Parity	0.1310	0.0809	0.0501
Overall Accuracy Equality	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, 43] 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 9. The synthetic data generator fit time is 1 minute and 24 seconds.

Table 9. Fairness measures experiment results of the Diabetes dataset on a random forest model. Average difference is 0.0078.

Measure	Original	Synthetic	Difference
Demographic Parity	0.0135	0.0015	0.0119
Overall Accuracy Equality	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
Conditional Use Accuracy Equality (True Positive)	0.0133	0.0059	0.0074
Conditional 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 across all models. Different models applied to the same datasets can yield different results, and judging from different metrics, datasets, and models, this indicates that synthetic data is a good approximation of the original data in terms of fairness.

In addition, further experiments showed that, by common sense or 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 random forest experiment, Conditional Use Accuracy Equality (True Positive) has an error of around 0.16. In the COMPAS random forest experiment, both Conditional Use Accuracy Equality measures had an error over 0.1. In the Diabetes random forest 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.

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 due to the potential risks to privacy and the possibility of exposing sensitive information. When evaluating an AI system with personal data, there is a real danger of inadvertently leaking private details about individuals involved in the dataset.

To mitigate this risk, we propose a framework that leverages differentially private synthetic data for fairness audits. Instead of using real data, it generates synthetic data that preserves key patterns of real data but, at the same time, protects individuals' privacy.

The proposed framework, backed by strong privacy guarantees, offers a secure alternative to directly using real data. By applying differential privacy methods, the synthetic data shields individual information from exposure while still enabling thorough fairness evaluations.

Experiments conducted on multiple real-world datasets and across different models show that synthetic data derived from real data retains essential fairness properties. This ensures that any biases present in the original data are captured and reflected in the synthetic counterpart.

While synthetic data is not a flawless substitute for real-world data, it serves as a practical and pragmatic tool for auditing AI fairness in a privacy-preserving manner. Our framework offers an effective means of testing AI systems for fairness while mitigating the privacy concerns surrounding using real data in sensitive applications. Auditors can perform thorough fairness evaluations by generating and using synthetic data while under strong privacy protections.

Future work could focus on applying this framework to larger, more complex datasets, which would help validate the framework's scalability. Furthermore, other privacy-preserving techniques could be explored to complement synthetic data and expand the scope of AI fairness auditing while maintaining privacy. Other means of auditing for certain fairness metrics, such as Conditional Use Accuracy Equality, that are not well suited for synthetic data can be explored. Finally, translating these technical advancements into real-world applications requires policy, legal, and industry collaborations. Research on the legal, ethical, and societal implications of differential privacy and synthetic data will be crucial for broad adoption.

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] Jil Kothari. 2020. Prediction of income with 98% accuracy! https://www.kaggle.com/code/jilkothari/prediction-of-income-with-98-accuracy Accessed: 2024-11-10.

[32] 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.

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

[34] 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).

[35] 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.

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

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

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

[39] 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).

[40] 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.

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

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

[43] Mohammed Mustafa. 2022. Diabetes Prediction Dataset. https://www.kaggle.com/datasets/iammustafatz/diabetes-prediction-dataset Accessed: 2024-11-26.

[44] 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.

[45] 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.

[46] Dan Ofer. 2017. COMPAS Dataset. https://www.kaggle.com/datasets/danofer/compass Accessed: 2024-11-26.

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

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

[49] UC Irvine Machine Learning Repository. 2016. Adult Census Income Dataset. https://www.kaggle.com/datasets/uciml/adult-census-income/data Accessed: 2024-11-26.

[50] 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).

[51] Panjawat Rungtranont. 2023. Diabetes: EDA | Random Forest + HP. https://www.kaggle.com/code/tumpanjawat/diabetes-eda-random-forest-hp Accessed: 2024-11-26.

[52] 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).

[53] Innocent Sky. 2024. Logistics Regression COMPAS Recidivism Racial Bias. https://www.kaggle.com/code/innocentsky/logistics-regression-compas-recidivism-racial-bias Accessed: 2024-11-26.

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

[55] 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).

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

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

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

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

[60] 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.

[61] 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.

[62] 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).

[63] 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.

[64] 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.

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