Journal of Artificial Intelligence Research 79 (2024) 1313-1341 Submitted 08/2023; published 04/2024

Bt-GAN: Generating Fair Synthetic Healthdata via Bias-transforming Generative Adversarial Networks

Resmi Ramachandranpillai R.RAMACHANDRANPILLAI@NORTHEASTERN.EDU
Md Fahim Sikder MD.FAHIM.SIKDER@LIU.SE
David Bergström DAVID.BERGSTROM@LIU.SE
Fredrik Heintz FREDRIK.HEINTZ@LIU.SE
Department of Computer and Information Science (IDA), Linköping University, Sweden

Abstract

Synthetic data generation offers a promising solution to enhance the usefulness of Electronic Healthcare Records (EHR) by generating realistic de-identified data. However, the existing literature primarily focuses on the quality of synthetic health data, neglecting the crucial aspect of fairness in downstream predictions. Consequently, models trained on synthetic EHR have faced criticism for producing biased outcomes in target tasks. These biases can arise from either spurious correlations between features or the failure of models to accurately represent sub-groups. To address these concerns, we present Bias-transforming Generative Adversarial Networks (Bt-GAN), a GAN-based synthetic data generator specifically designed for the healthcare domain. In order to tackle spurious correlations (i), we propose an information-constrained Data Generation Process (DGP) that enables the generator to learn a fair deterministic transformation based on a well-defined notion of algorithmic fairness. To overcome the challenge of capturing exact sub-group representations (ii), we incentivize the generator to preserve sub-group densities through score-based weighted sampling. This approach compels the generator to learn from underrepresented regions of the data manifold. To evaluate the effectiveness of our proposed method, we conduct extensive experiments using the Medical Information Mart for Intensive Care (MIMIC-III) database. Our results demonstrate that Bt-GAN achieves state-of-the-art accuracy while significantly improving fairness and minimizing bias amplification. Furthermore, we perform an in-depth explainability analysis to provide additional evidence supporting the validity of our study. In conclusion, our research introduces a novel and professional approach to addressing the limitations of synthetic data generation in the healthcare domain. By incorporating fairness considerations and leveraging advanced techniques such as GANs, we pave the way for more reliable and unbiased predictions in healthcare applications.

1. Introduction

Clinical Decision Support Systems (Rieke et al., 2020) are important for healthcare organizations to improve care delivery in the era of value-based healthcare, digital innovation, and big data. The adoption of advanced artificial intelligence technology in healthcare systems is gathering interest from many researchers (Ghassemi et al., 2021), (Jiang et al., 2017). One example is called precision medicine – predicting which treatment procedures are likely to succeed on a patient based on numerous traits and the treatment context – is the most prevalent application of classical machine learning in healthcare. These systems can yield benefits in terms of improved accuracy, diagnosis, finding new knowledge about the illness

conditions and their progression, and the ability to provide patients with a more concrete prognosis and treatment plan by the use of historical data in model development.

Data analysis on Electronic Healthcare Records (EHR) (Evans, 2016) is greatly affected by rules such as the Health Insurance Portability and Accountability Act (HIPPA) (Edemekong et al., 2018) in the US and the General Data Protection Regulation (GDPR) (Regulation, 2016) in EU, that preserve patient's privacy. Healthcare organizations are now jointly accountable for any personal data breach since they are responsible for managing all personal data storage and processing in both their organization and that of their suppliers. Over the last three years, Privacy Impact Assessments (PIAs) (Wright, 2012) have become commonplace in the healthcare industry. GDPR has now placed the delivery of PIAs in the public domain, boosting transparency and information risk ownership clarity.

Synthetic data generation (Armanious et al., 2020; Yale et al., 2020) tries to preserve the privacy of patients by producing realistic samples to perform downstream tasks. Generative Adversarial Networks have drawn much attention, however, are often criticized for producing low-quality and less diverse samples(Adiga et al., 2018). The state-of-the-art (SOTA) synthetic healthcare data generation methods (Armanious et al., 2020; Edemekong et al., 2018) have explored the concepts of accuracy, utility, and privacy, but paid less attention to fairness in the Data Generation Process (DGP) and in the subsequent tasks. To facilitate the development of equitable analysis and predictions, the DGP should ensure that the synthetic EHR is fair along with other dimensions (such as utility, privacy, etc.).

There are malignant feature correlations in the medical data, which we call correlation biases. GANs amplify these spurious correlations as studied in (Gupta et al., 2021). Another cause of unfairness can be in the direction of fair resemblance - accurately capturing the diversities in sub-groups. We would like to mention (Bhanot et al., 2021), as it raises some concerns about the trustworthiness of synthetic data in the direction of fair resemblance. We use the term *representation fairness* instead of *fair resemblance* throughout this study as this refers to how different sub-group proportions are represented in the synthetic data as compared to real data. Therefore representation biases are seen when the sub-groups are missing, underrepresented, or overrepresented in the synthetic data.

Figure 1 illustrates a simplified example wherein the degree of representation in the synthetic data is measured using Partial Recall (Kynkäänniemi et al., 2019). The generated distribution shows a substantial recall discrepancy between majority and minority groups, notably expanding as the minority level increases. Consequently, the synthetic data with minority samples exhibit both inadequate quality and coverage problems, potentially resulting in incidental or spurious correlations within the synthetic dataset.

In this paper, we address the problem of synthetic healthcare data fairness; generate fair data from biased data (correlation biases), and promote representation fairness in the target data.

Contributions. The problem definition, design, analysis, and experimental assessment of the proposed framework are the main contributions of this research. Particular contributions include:

- Development of a principled GAN framework for synthetic healthcare data generation with guaranteed fairness in the target in contrast to existing techniques such as HealthGAN and MedGAN, where no fairness is ensured.

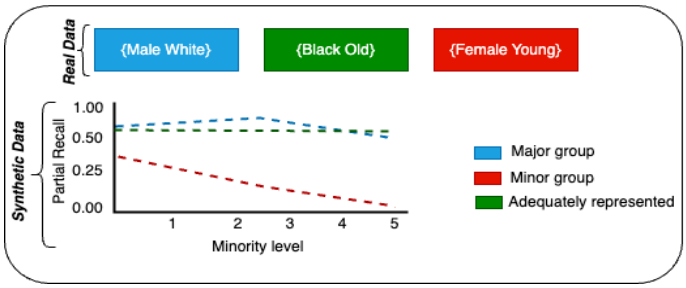


Figure 1: An example of partial recall for groups with underrepresentation (minor), overrepresentation (major), and adequate representation in data generation.

- Problem definition considering how the DGP is affected by various biases from the data (correlation bias) and how the generative model injects additional biases (representation biases) in the training process.
- A comprehensive experimental assessment of the proposed framework on health database and comparison with state-of-the-art methods in terms of data utility and fairness.
- An analysis of the bias amplification (Wang et al., 2019) of the proposed approach and comparison with the state-of-the-art.
- An explainability analysis using SHAP (Nohara et al., 2019) to validate the trustworthiness of the proposed framework.

2. Related Works

We focus on the related literature in terms of i) synthetic data generation in healthcare systems, ii) fair data generation, and (iii) representation issues in GANs in contrast to fairness measures, which we define in Section 4.

GANs in Healthcare Systems. Synthea (Walonoski et al., 2018) simulates patient records from birth to the present day using modules informed by clinicians and real-world statistics. It claims to preserve utility by employing healthcare practitioners and real statistics to construct rules that synthesize the data, which assures privacy. HealthGAN (Yale et al., 2020) generates synthetic health data of multivariate nature. The loss function is based on the Wasserstein distance (Arjovsky et al., 2017) and they used *Synthetic Data Vault (SVD)* (Patki et al., 2016) for categorical encoding. Medical Image translation, MedGAN (Armanious et al., 2020), is an end-to-end framework by merging the adversarial network with non-adversarial losses. The discriminator network is a pre-trained feature extractor, which penalizes the discrepancy. Among the models mentioned above, HealthGAN provides promising results in terms of accuracy and utility (Bhanot et al., 2021).

GANs in Fair Data Generation. The Fairness aware GAN (Fair GAN) (Xu et al., 2018) approach trains a generator to produce fair representations by an additional discriminator that knows the real distribution of the protected features in the training data. DECAF (van Breugel et al., 2021) is designed to generate fair synthetic tabular data with an assumption that the underlying causal structure is known. It uses individual generators

to generate features sequentially based on the causal graph, while de-biasing is done at the inference time.

Improving Under-representation in GANs. GANs frequently experience mode collapse and generate samples with low diversity because of the unstable nature of the min-max game between a generator and a discriminator. Better data coverage has been advocated using objective function-based methods (Kodali et al., 2017), (Mao et al., 2017) and structure-based methods (Radford et al., 2015), (Zhang et al., 2019). Even while they work well to increase data coverage overall, these methods don't give minor modes any extra attention. They frequently fail to recover minor modes when the minority ratio for a given feature is meager. Techniques such as label smoothing have been successfully applied to increase the performance of GANs when the scale of the training data is limited (Zheng et al., 2017). Another category of methods is sampling-based methods (Lee et al., 2021) which give extra attention to minor modes and then promote these modes by score-based sampling. Discriminator Rejection Sampling (DRS) (Azadi et al., 2019) is proposed to apply rejection sampling to filter the synthetic samples based on density ratio estimation. Likewise, GOLD (Mo et al., 2019) reweighted fake samples to improve the representation (or resemblance). Top-k training (Sinha et al., 2020) modifies the generator using only the top-k synthetic samples to promote representation. Similar to DRS, Dia-GAN (Lee et al., 2021) proposed a discrepancy score based on the empirical mean and variance over multiple epochs. We follow a similar approach to tackle representation biases in data generation.

In addition, none of the above methods have successfully handled partially unlabelled data. Medical data often contain missing labels which if not treated properly can adversely affect a patient's health.

In summary, we present Table 1 to compare the methods with different key areas of interest. As far as we know, our Bt-GAN is the first GAN architecture that tackles correlation biases, representation biases, and partially unlabeled data in an end-to-end framework in the underlying DGP.

GAN	**a**	**b**	**c**	**Method**	**Goal**
HealthGAN (Yale et al., 2020)	No	No	No	Wasserstein-distance	Synthetic health data
MedGAN (Armanious et al., 2020)	No	No	No	Progressive-refinement	Synthetic health data
FairGAN (Xu et al., 2018)	Yes	No	No	Adversarial-de-biasing	Fair synthetic data
DECAF (van Breugel et al., 2021)	Yes	No	No	Causal structure	Fair synthetic tabular data
Bt-GAN(ours)	Yes	Yes	Yes	bias-transforming DGP + score-based sampling	Fair and representative synthetic healthdata

Table 1: Comparison of related works with different key areas of interests: (a) Correlation bias, (b) Representation bias, and (c) Partially unlabelled data.

3. Unfairness in Data Generation Process

According to (Wachter et al., 2020), the term *bias-preserving* in fairness literature means that the status quo (or training dataset) is a baseline and the model tries to reproduce the historic performances in the status quo, which only accounts for direct discrimination. But, the *bias-transforming* metrics address indirect discrimination by fixing the structural inequalities in the status quo. In the following, we describe the causes of unfairness in DGP in detail and illustrate how we approach the problem definition with the concepts studied in (Wachter et al., 2020).

3.1 Learning of Protected Attribute Information

Learning protected attribute information is one of the key elements of unfairness in classification or prediction, as described in the literature (Creager et al., 2019), (Locatello et al., 2019), (Song et al., 2019). This situation can be worse in synthetic data generation using GANs as it amplifies the existing biases in the DGP as studied in (Gupta et al., 2021).

Suppose we have an ideal biased dataset $\mathcal{D}_{bias} = \{\tilde{\mathcal{X}}, \tilde{\mathcal{Y}}, \tilde{\mathcal{S}}\}$ where each label $\tilde{\mathcal{Y}}$ is correlated with each sensitive attribute in equal intensity. Then for a well-trained fixed discriminator[1], the generative model, once converged, captures the dependencies from $\mathcal{D}_{bias}$, which means the labels in the synthetic data are also correlated with the sensitive attribute in equal intensity. Based on this and motivated by (Wachter et al., 2020), we define the following in the context of GANs and correlation bias: Let $\mathcal{D}$ be the real dataset containing correlation biases and $\hat{\mathcal{D}}$ be the synthetic data generated by the underlying generator G of a GAN.

Definition 1 - Bias-preserving DGP (Bp-DGP). *A DGP is bias-preserving if and only if the underlying generative model G, once optimized, learns to obtain a transformation from a Multivariate Normal Distribution (MVD) to the real data distribution by preserving the exact correlations from $\mathcal{D}$ and replicate it in $\hat{\mathcal{D}}$* across sensitive groups.

Remark - A Bp-DGP seeks to keep the historic performances from the real data in the output of a target model when trained with synthetic data with equivalent error rates for each group as shown in the real data.

3.2 Representation Bias: Failure of GANs in Capturing the Exact Representations

The synthetic generation based on GANs fails to capture the exact sub-group proportions from the real data as GANs try to match the distributions of real data at the dataset level. We define the following in the context of GANs and representation biases:

Definition 2 - Density-preserving DGP (Dp-DGP). *A DGP is said to be density preserving if and only if the underlying generative model G, once optimized is learned to generate synthetic data $\hat{\mathcal{D}}$ in such a way that the ratio of sub-groups between $\mathcal{D}$ and $\hat{\mathcal{D}}$ should be the same.*

Remark - Due to the min-max objective of GAN optimization, the DGP can inject additional biases in the form of representation biases due to the mode collapse problem as studied in (Gupta et al., 2021).

1. we assume that the data used to train is very large and the model has enough capacity

Correlation biases and representation biases are critical in synthetic data generation based on GANS.

4. Desiderata

4.1 Fairness Definition

Formally, let $\mathcal{D} = \{\mathcal{X}, \mathcal{S}, \mathcal{Y}\}$ be a dataset containing biases, where $X \in \mathcal{X} \subset \mathbb{R}^d$ is a random variable of non-sensitive features, $S \in \mathcal{S}$ be sensitive features and $Y \in \mathcal{Y}$, the labels. Also, let $\mathcal{U}(\mathcal{S}, \mathcal{Y})$ be a definition of algorithmic fairness. We define the following algorithmic fairness measure to eliminate indirect discrimination (Wachter et al., 2020):

Definition 3 - (Statistical Parity (Dwork et al., 2012)). *Suppose we have a function $h : X \to Y', Y' = \{0, 1\}$ for binary classification, and assume S splits X into a majority set $\mathcal{M}$ and a minority set $\mathcal{M}'$ $(X = \mathcal{M} \cup \mathcal{M}')$, then the function h satisfies statistical parity if $P[h(x) = 1 \mid x \in \mathcal{M}] = P[h(x) = 1 \mid x \in \mathcal{M}']$, where x denotes an instance of X and $P[.]$ denotes the probability of an instance.*

We assume the protected attribute is binary for notational convenience as this can be extended to non-binary classification.

4.2 Mutual Information

Definition 4 - (Mutual Information (Veyrat-Charvillon & Standaert, 2009)). *Let W_1 and W_2 be two random variables, the product of marginal distribution be p_{W_1} p_{W_2} and the joint distribution be p_{W_1,W_2}, then the mutual information between W_1 and W_2 can be defined as:*

$$I(W_1; W_2) = H(W_1) - H(W_1 \mid W_2) = \int_{W_1} \int_{W_2} p_{(W_1,W_2)} \log \frac{p_{(W_1,W_2)}}{p_{(W_1)} p_{(W_2)}} \, dW_1 . \, dW_2 \tag{1}$$

Unlike correlation coefficients (Benesty et al., 2009) (which could only estimate the linear dependence), mutual information is used to measure linear as well as non-linear dependencies between two random variables.

Definition 5 - (Relationship between Statistical Parity and Zero Mutual Information (Ghassami et al., 2018)) . *Given a predicted outcome, $Y' = \{0, 1\}$, and a protected attribute S, then,*

$$p\left[Y' \mid S\right] = p\left[Y'\right] \Leftrightarrow p_{Y',S} = p_{Y'} p_S \Leftrightarrow I\left(Y'; S\right) = 0 \tag{2}$$

If the two random variables are independent of each other we get zero mutual information.

4.3 Generative Adversarial Networks(GAN)

The Generative Adversarial Network (GAN) is a prominent member of the generative models family, comprising two essential components: a generator and a discriminator. The generator takes in random noise $z \sim \mathcal{N}(0, 1)$ as input and endeavors to produce realistic data, represented as $x \sim P_{\mathcal{D}}$. On the other hand, the discriminator is responsible

for distinguishing between the generated data and the original data. As time progresses, the generator becomes more adept at deceiving the discriminator, while the discriminator strives to differentiate between real and fake data. This dynamic creates a zero-sum game, where both networks engage in a continuous battle until they reach a state of equilibrium known as the Nash equilibrium. To quantify the performance of the generator (G) and discriminator (D), the following loss function is employed:

$$\min_{\mathrm{G}} \max_{\mathrm{D}} V(\mathrm{G},\mathrm{D}) = E_{x\sim P_{\mathcal{D}}}[log(\mathrm{D}(x))] + E_{z\sim p_z(z)}[log(1-\mathrm{D}(\mathrm{G}(z)))] \tag{3}$$

5. Synthetic Data Fairness

Synthetic data fairness means generating fair data from biased data such that any downstream model trained on fair synthetic data will have fair predictions in real data assuming that the underlying prediction model does not possess any explicit biases.

To achieve synthetic data fairness considering correlation biases (Definition 1) and representation biases (Definition 2) in the DGP, we propose to define Bias-transforming DGP (Bt-DGP) as:

Definition 6 - Bias-transforming DGP (Bt-DGP). *Let G be a generative model and $\mathcal{U}(\mathcal{S},\mathcal{Y})$ be a definition of algorithmic fairness (such as statistical parity in our case). A DGP is bias-transforming if and only if it is Dp-DGP and the underlying generative model G transforms the existing biases in $\mathcal{D}$ in such a way that $\hat{\mathcal{D}}$ is fair, as defined by $\mathcal{U}(\mathcal{S},\mathcal{Y})$ and the utility is maintained with respect to any downstream tasks.*

5.1 Problem Definition

The Synthetic Data Fairness Problem (SDFP) is to generate fair data $\hat{\mathcal{D}} = \{\hat{\mathcal{X}}, \hat{\mathcal{S}}, \hat{\mathcal{Y}}\}$ from $\mathcal{D}$ through Bt-DGP.

Objective: To design a framework that accounts for both representation issues and spurious correlations and to guarantee the fairness of any downstream models trained on the synthetic data.

Approach: For tackling correlation biases we use and exploit Mutual Information (MI) (Hemamou & Coleman, 2022), (Kang et al., 2021), (Ghassami et al., 2018) as described in Section 4.2. For representation biases, we adopt a score-based sampling as detailed in Section 6.

In the next section, we describe how we achieve Bt-DGP by proposing the Bt-GAN framework.

6. Bias-transforming Generative Adversarial Networks

This section presents our Bias-transforming Generative Adversarial Network (Bt-GAN) framework in detail.

The whole process of Bt-GAN can be divided into 3 stages:

1. **Pre-train and Diagnose** - the generator G of a GAN learns to generate high-quality samples from a large real-world dataset (biased and partially labeled). During this

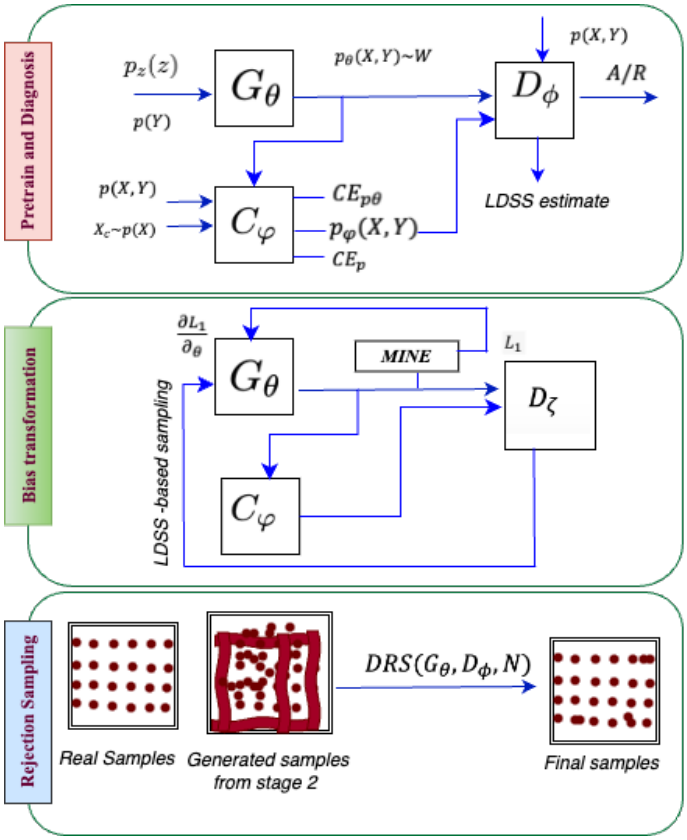


Figure 2: Architecture of Bt-GAN: the utilities of C_φ, G_θ, D_ζ, and D_ϕ are shown. The symbols A and R denote *accept* and *reject* respectively. The discriminator, D_ϕ accepts if it thinks it is from the true data distribution denoted as p.

process, we diagnose representation biases by recording the corresponding sub-group densities from synthetic data.

2. **Bias-transform** - the pre-trained generator G from stage 1 is fine-tuned to generate fair distribution. This involves unlearning the sensitive correlations using a fairness penalty and enforcing representation fairness using score-based weighted sampling based on densities, computed from stage 1. This transformation stage tackles both correlation bias from the data and the representation bias from the GAN, thereby encouraging GAN to learn from the under-represented regions of the data manifold.

3. **Discriminator Rejection Sampling (DRS)** - The score-weighted sampling in step 2 injects new biases towards under-represented regions. We correct it using rejection sampling (Azadi et al., 2019) using the discriminator trained in stage 1.

An overview of Bt-GAN architecture is given in Figure 2 and we describe all the stages in detail as follows:

6.1 Pretrain and Diagnosis

Healthcare data often have missing values, especially partially unlabeled. So, to completely capture the true data distribution from the partially unlabeled data, we employ Triple GAN (Li et al., 2017), a semi-supervised GAN framework involving 3 components: i) a classifier C_φ for the conditional distribution, $p_\varphi(y \mid x)$, ii) a Generator G_θ that characterizes $p_\theta(x \mid y) \approx p(x \mid y)$, and iii) a Discriminator D_ϕ that predicts whether the data pair (x, y) is from $p_\theta(x \mid y)$ (fake) or from $p(x \mid y)$ (real). After a sample x is drawn, C_φ produces $p_\varphi(x, y) = p(x)p_\varphi(y \mid x)$. Then, the joint distribution produced by G_θ becomes, $p_\theta(x, y) = p(y)p_\theta(x \mid y)$. Let x be obtained by the latent variable z, then, $x = G_\theta(y, z), z \sim p_z(z)$, where $p_z(z)$ can be uniform or normal distribution.

The mini-max game for Triple GAN can be formulated as (Li et al., 2017):

$$\begin{aligned} L_{GCD} = \min_{\varphi,\theta} \max_{\phi} V\left(C_\varphi, G_\theta, D_\phi\right) = \\ E_{(x,y)\sim p(x,y)}[\log D(x, y)] + \\ \lambda E_{(x,y)\sim p_\varphi(x,y)}[\log(1 - D(x, y))] + \\ (1 - \lambda) E_{(x,y)\sim p_\theta(x,y)}[\log(1 - D(G(y, z), y))] + L_{CE}, \end{aligned} \tag{4}$$

where, $\lambda \in (0, 1)$ is the balance factor that controls the significance of classification and generation, we set $\lambda = 0.5$ for the entire training process. To make $p(x, y) = p_\varphi(x, y) = p_\theta(x, y)$, a cross entropy loss $L_{CE} = E_{(x,y)\sim p(x,y)}\left[-\log p_\varphi(y \mid x)\right]$ has been added (Li et al., 2017).

During the training of the triple GAN, we record the densities for each selected sub-groups. The densities are estimated by a measure called log disparity of sub-groups (LDS) based on density ratios as follows:

Definition 7 - (Log Disparity of Sub-groups (LDS)).*Let $f(x)$ be a membership function for the binary definition of sub-groups $g_j \in \mathbb{G}, 1 \leq j \leq |\mathbb{G}|$, then $f(x) = 1$, if* x $\in g_j$ *and $f(x) = 0$, if* x $\notin g_j$. *The log disparity of sub-groups, LDS(.) between $p_\mathcal{D}$ and $p_{\hat{\mathcal{D}}}$, is defined as:*

$$\mathrm{LDS}(x_i) = \log\left(\frac{o\left(f(x_i) = 1 \mid x_i \in g_j \in p_{\hat{\mathcal{D}}}\right)}{o\left(f(x_i) = 1 \mid x_i \in g_j \in p_\mathcal{D}\right)}\right), \tag{5}$$

where $o(f(x_i) = 1 \mid x_i \in g_j \in p_{\hat{\mathcal{D}}}) = P(f(x_i) = 1 \mid x_i \in p_{\hat{\mathcal{D}}})/(1 - (P(f(x_i) = 1 \mid x_i \in g_j \in p_{\hat{\mathcal{D}}})))$, and can similarly be calculated for $o\left(f(x_i) = 1 \mid x_i \in g_j \in p_\mathcal{D}\right)$. Log disparity of sub-group can be computed for all the available sub-groups by changing $f(x)$ to different g_j in $\mathbb{G}$.

The discriminator D_ϕ output can be used to estimate the disparity of sub-groups, $LDS(.)$ between $\mathcal{D}$ and $\hat{\mathcal{D}}$.

6.2 Bias Transformation

In the bias transformation step, the pre-trained GAN learns the fair distribution underlying the chosen attributes. The fair distribution means that the generated data should not contain both correlation biases and representation biases.

6.2.1 Tackling Correlation Biases: Unlearning Sensitive Correlations using Information-constrained DGP

GANs learn spurious correlations from data to converge to the true data distribution. Let W be the generated space. Specifically, we define learning of protected attribute information as increasing $I(W;S)$ which is the mutual information between W and S in the DGP. We exploit Definition 5 and formulate a mutual information reduction problem between the generated space W, which is a pair $\{X', Y'\}$, and the vector encoded sensitive features, S (W and S are random variables). So, we propose the following information-constrained minimax objective function:

$$\begin{aligned}\min_{\varphi,\theta}\max_{\phi} V\left(C_\varphi, G_\theta, D_\phi\right) = E_{(x,y)\sim p(x,y)}[\log D(x,y)]+ \\ \lambda E_{(x,y)\sim p_\varphi(x,y)}[\log(1-D(x,y))]+ \\ (1-\lambda)E_{(x,y)\sim p_\theta(x,y)}[\log(1-D(G(y,z),y))]+ \\ L_{CE}; \exists I(W;S)=0,\end{aligned} \tag{6}$$

The condition, $I(W;S)$ depends on the trainable parameter θ. In this setting, S is a constant vector-encoded representation of sensitive features selected by the data manager (or data owner).

For estimating the fairness penalty term $I(W;S) = 0$, we use Mutual Information Neural Estimation (MINE) (Belghazi et al., 2018), and the maximization can be handled by back-propagation. The loss function L_I for MINE is:

$$L_I = E_{\hat{p}(W,S)}\left[T_\eta(W \mid S)\right] - \log E_{\hat{p}_w,\hat{p}_s}\left[e^{T_\eta(W|\hat{S})}\right], \tag{7}$$

where T_η is a statistical neural network. The loss term, L_I can provide an estimate of MI, once the parameter η is maximized and it depends on θ. Also, $\hat{p}$ denotes the empirical estimation of distribution. Therefore:

$$L_{MI} = I[\widehat{W;S}] = \max_{\eta} L_I(\eta,\theta) \tag{8}$$

The fairness penalty term helps the generator in learning a distribution W from Z with the constraint of not containing any malignant information related to S.

So, the final loss, L_F for Bt-GAN would be:

$$L_F = \underbrace{\underbrace{L_{GCD}}_{Semi-supervised-generation} + \underbrace{\alpha L_{MI}}_{MIde-biasing}}_{Fair-generation} \tag{9}$$

The parameter α balances the MI reduction and the quality of generation. We give an ablation study (Figure 3) on how it affects the performance of the model by varying α. The Triple GAN minimax objective is optimized by iteratively modifying the generator, discriminator, and classifier with respective losses.

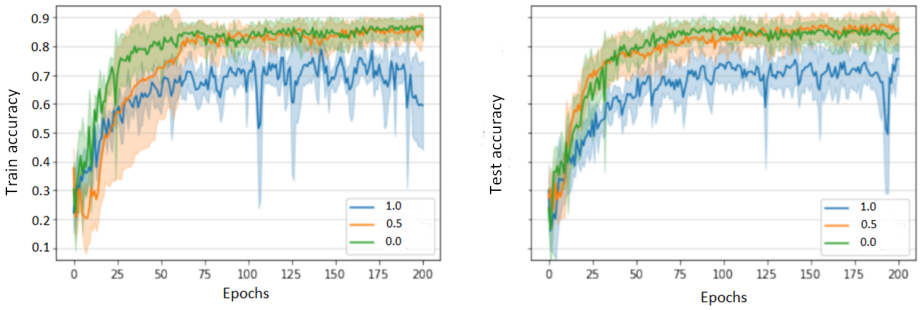


Figure 3: Ablation study showing the effect of different values of α on test and train accuracy. Note that, when $\alpha = 1$, the effect of MI reduction between W and S is large, but the accuracy also drops severely. When $\alpha = 0.5$, the model is performing comparatively better on the mortality prediction task. Also, when $\alpha = 0.0$, the MI reduction part in equation 8 is inactive and thus is the fairness constraint. According to this, we set $\alpha = 0.5$ to balance the quality-fairness trade-off for the entire process.

6.2.2 Improving Data Coverage using Density-preserving Sampling

Following (Lee et al., 2021), we propose a Log Disparity Sub-group Score (LDSS) (based on LDS in equation 4), which measures how close the real and synthetic distributions on sub-groups for training sample x over T iterations, as:

$$LDSS(x_i, T) = \frac{1}{T}\Sigma_{k \in T} LDS(x)_k \tag{10}$$

The aim is to design a sampling probability for an instance i, based on LDSS, and propagate it through SGD to design the batch size, $\mathcal{D}_{\mathrm{B}} = x^k : x^k = x_i, i \sim P_{LDSS}(i)$ for $k = 1, 2, \ldots, B$. ie, each $x_i \in \mathcal{D}$ is sampled with a probability of $P_{LDSS}(i)$.

Definition 8 - ($LDSS$-based Sampling Probability). *For a training dataset $\mathcal{D}$, we denote each instance as x_i for notational convenience, The sampling probability $P_{LDSS}(i)$ of the i^{th} data instance is calculated over a number of steps T by:*

$$P_{LDSS}(i) = \frac{\mathrm{LDSS}\left(x_i, T\right)}{\sum_{k=1}^{|\mathcal{D}|} \mathrm{LDSS}\left(x_k, T\right)} \tag{11}$$

Based on the LDSS, we define the following:

Definition 9 - ($LDSS$-based Density-preserving DGP). *A DGP is LDSS-based Dp-DGP, if and only if the LDSS of sub-groups (for each $g_j \in \mathbb{G}$) between $p_{\mathcal{D}}$ (real data distribution) and $p_{\hat{\mathcal{D}}}$ (synthetic data distribution) is below an acceptable threshold δ.*

We elaborate in Section 8.3 on how the threshold δ has been set to various levels for evaluating the representation biases in synthetic health data.

With reference to Definition 9, we re-iterate Definition 6 as:

Definition 10 - Bias-transforming DGP (Bt-DGP). *Let G be a generative model and $\mathcal{U}(\mathcal{S}, \mathcal{Y})$ be a definition of algorithmic fairness (such as statistical parity in our case). A DGP is said to be bias-transforming if and only if it is LDSS-based Dp-DGP and the underlying generative model G transforms the existing biases in $\mathcal{D}$ in such a way that the $\hat{\mathcal{D}}$ is fair, evaluated by a definition of algorithmic fairness, $\mathcal{U}(\mathcal{S}, \mathcal{Y})$.*

6.3 Discriminator Rejection Sampling

LDSS-based sampling creates biases as the generated distribution $p_{\hat{\mathcal{D}}}$ differs from the real data distribution $p_{\mathcal{D}}$. We employ rejection sampling (Azadi et al., 2019) and use the discriminator D_ϕ from stage 1 (as it knows the real distribution) with an acceptance level $p_{\mathcal{D}}(x)/\operatorname{Lp}_{\hat{\mathcal{D}}}(x)$, for some constant $L > 0$. We use the same architecture for both discriminators (except for the sigmoid activation). Also, to speed up the processing, the parameters of D_ζ are instantiated with D_ϕ.

7. Theorems and Proofs

Theorem 1: *An ideal data generation process*[2] *is Bp-DGP.*

Proof. For a well-trained fixed discriminator of a Triple GAN, D_ϕ, the global convergence of the optimization equation L_{GCD} is achieved when $p_g(x, y) = p(x, y) = p_c(x, y)$, where $p_g(x, y), p(x, y)$, and $p_c(x, y)$ respectively denote generator distribution, real data distribution, and classifier distribution. Since the tree-player game is a hard constraint compared to two-player optimization, it absolutely captures the real dependence from data $\mathcal{D}$ even with missing labels. Note that we have not changed the discriminator. Therefore the synthetic data $\hat{\mathcal{D}}$ contains all the biases in $\mathcal{D}$, and hence it is Bp-DGP.

Theorem 2: *An ideal data generation process is Dp-DGP.*

Proof. Let G be the generator, D be the discriminator, and C be the classifier of a Triple GAN. We assume enough capacity for G, D, and C. Also, we assume that the Generator G does not possess any mode collapse under the ideal DGP and converges when $p_g(x, y) = p(x, y) = p_c(x, y)$ by replicating all the sub-group densities from $\mathcal{D}$. Therefore, an ideal DGP is Dp-DGP.

Theorem 3: *An ideal DGP is both Bp-DGP and Dp-DGP.*

Proof. It can be proved by Theorem 1 and Theorem 2.

Theorem 4: *Given optimal discriminator $D^*_{C,G}(x, y)$, the global minimum of the generator and classifier loss is achieved if and only if $p(x, y) = p_\phi(x, y) = p_\theta(x, y)$.*

We use the following Theorem to prove Theorem 4.

Theorem 4.1. *For any fixed C and G, the optimal D^* of the game defined by the utility function L_{GCD} is:*

$$D^*_{C,G}(x, y) = \frac{p(x, y)}{p(x, y) + p_\gamma(x, y)} \tag{12}$$

where: $p_\gamma(x, y) = (1-\gamma)p_g(x, y)+\gamma p_c(x, y)$ is a mixture distribution for $\gamma \in (0, 1)$. We refer to Lemma 3.1, Lemma 3.2, and Theorem 3.3 of (Li et al., 2017) for the proof for Theorem 4.

Proof. Note that we have made zero changes to the Triple GAN discriminator and therefore we can prove that both C and G will converge to the true data distribution[3] if in each iteration the discriminator has been trained to achieve the optimum for a fixed C and G and C and G subsequently modified to maximize the discriminator loss. Since the discriminator is trained optimum, it penalizes for a very large deviation from real

2. a process in which the underlying generative model exactly converges to the true data distribution
3. we assume all the components have enough capacity

distribution even with the addition of MI loss, thus finding a balance between accuracy and fairness (at $\alpha = 0.5$).

8. Experimental Results and Discussions

Dataset: We use MIMIC-III, a publicly available healthcare database containing de-identified patient admissions between 2001 and 2012 (33798 unique patient stays). We obtained permission to access MIMIC-III for research purposes after completion of an online course (certification number 45456719). More details regarding the cohort construction are given in Appendix.

Benchmarks: The methods we benchmark against are based on SOTA synthetic health data generation schemes and fair data generation. We compare with HealthGAN, as it is the best among the SOTA generative models in healthcare data. Next, we validate the performance against FairGAN on how the fairness measures have been improved in our settings. We could not find any references regarding fair data generation in healthcare, addressing the concerns raised in (Bhanot et al., 2021). We do not include DECAF (van Breugel et al., 2021) as it is based on the underlying causality and finding causal inference knowledge on MIMIC-III is still a work-in-progress (Guzman & Sontag,), and we leave it for future study.

For the ablation study, we use two variants namely Bt-GAN^{-}, which only satisfies $\mathcal{U}(\mathcal{S}, \mathcal{Y})$, and Bt-GAN which accounts for both $\mathcal{U}(\mathcal{S}, \mathcal{Y})$ and LDSS-based Dp-DGP.

Evaluation Metric: We evaluate our proposed model using the following criteria:

- **Data Utility** - We use AUROC, AUPRC, F1 score, and accuracy for evaluating data utility. We train classifiers (Linear Regression (LR) and Random Forest (RF) on synthetic data for downstream tasks and validate the performance. Moreover, we perform sample-level metrics analysis proposed in (Alaa et al., 2022) together with Jenson Shanon Divergence (JSD) (Sinn & Rawat, 2018) and discriminative score.

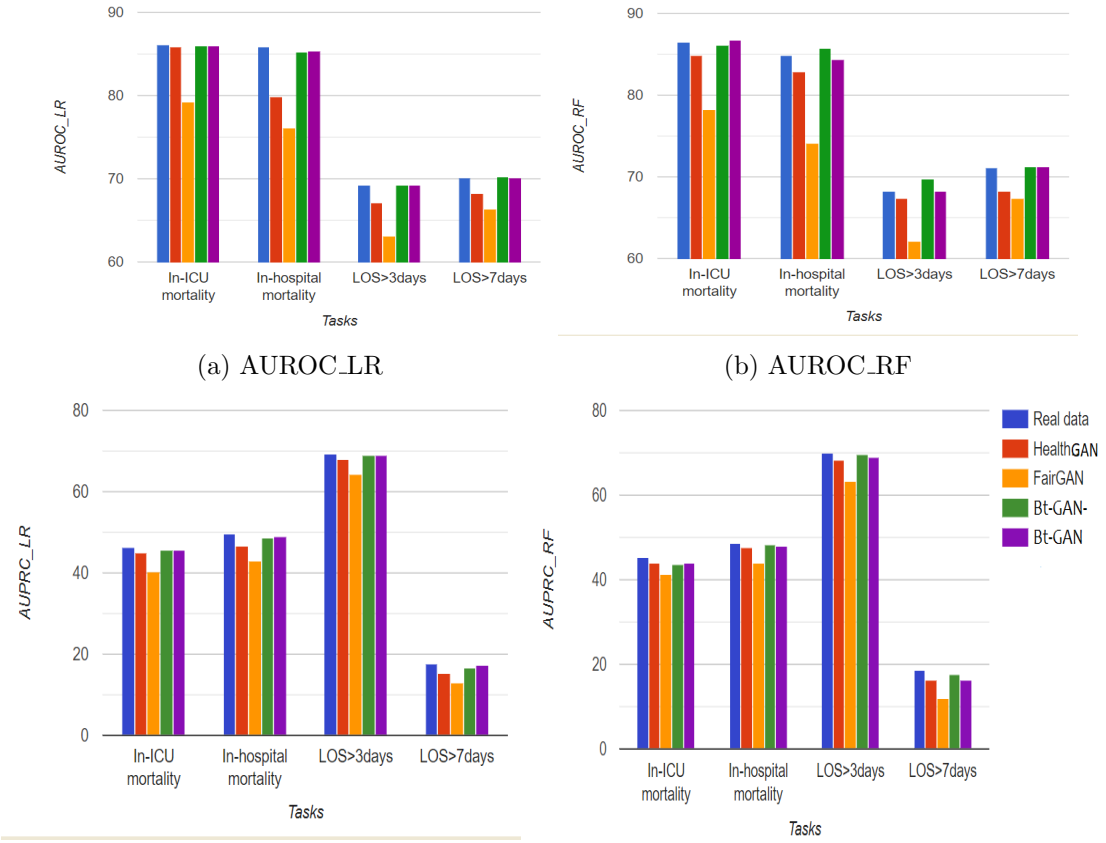
- $\mathcal{U}(S, Y)$ **Fairness** - This metric is used to measure the correlation bias and we use the parity gap and AUROC gap. Moreover, we perform data and model leakage (Wang et al., 2019) to compare the bias amplification in our settings (Details in Appendix).

- **Representation Fairness** - It is compared and evaluated by $LDSS$ calculated in every possible sub-groups. We use pie charts to plot the differences in representations captured by different models.

Experimental Setup: The details of different neural networks are given in Appendix.

8.1 Data Utility Analysis

We focus on four types of binary prediction tasks for analyzing the data utility: (i) In-ICU mortality, (ii) In-hospital mortality, (iii) Length of Stay (LOS) greater than 3 days, and (iv) Length of stay (LOS) greater than 7 days. We use LR and RF models for predictions. Our aim here is to compare the quality and utility of the synthetic data generated by the proposed model, not the classifier accuracy.



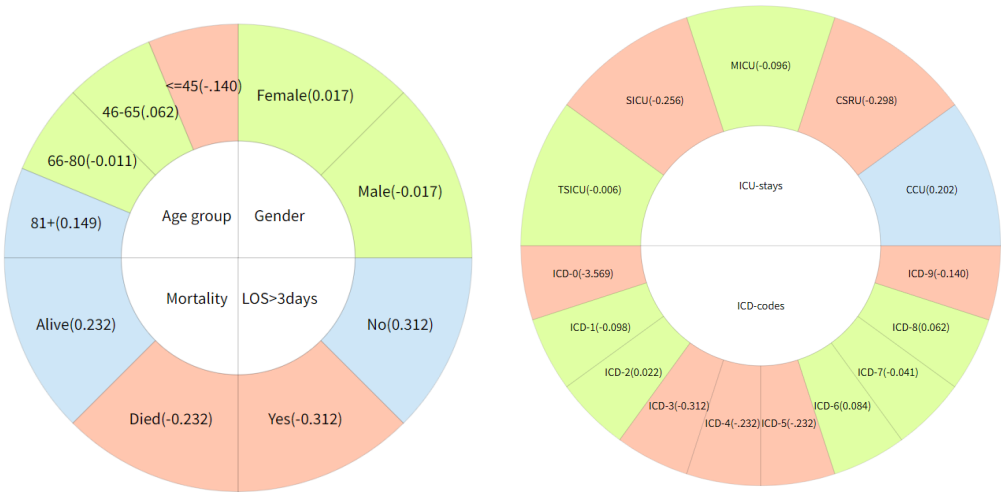
(a) AUROC_LR (b) AUROC_RF

(c) AUPRC_LR (d) AUPRC_RF

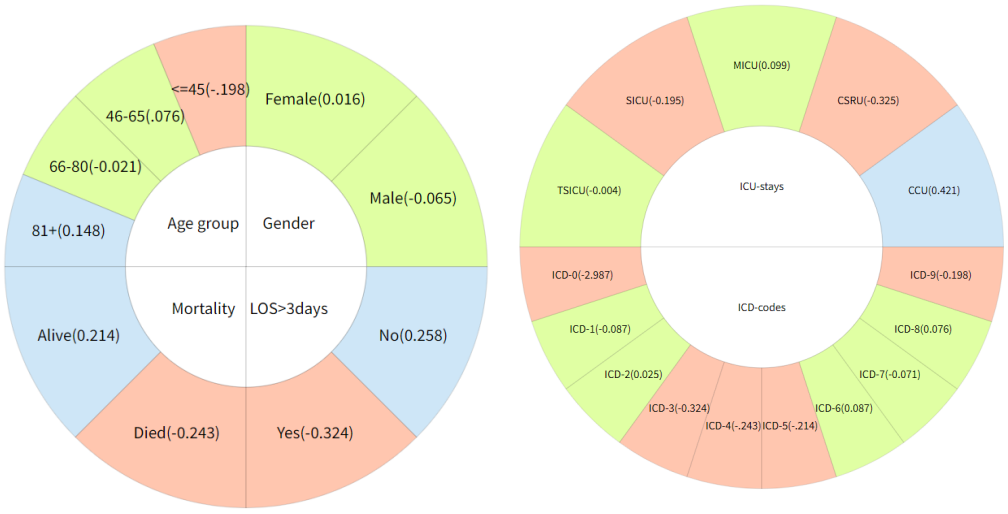
Figure 4: Data utility analysis

	Real Data		**HealthGAN**		**FairGAN**		**Bt-GAN**	
	accuracy.	F1	accuracy.	F1	accuracy.	F1	accuracy.	F1
In-ICU mortality	92.1	37.6	91.3	34.2	89.5	32.4	**91.5**	**34.7**
	91.8	12.1	**91.1**	**12.0**	88.3	11.6	90.9	11.7
LOS > 3*days*	71.2	59.9	**69.4**	**58.3**	67.1	56.3	68.1	57.3
	72.6	59	67.2	**59.1**	66.4	57.9	**68.9**	57.6
In-hospital mortality	90.1	39.6	89.1	37	85.4	32.8	**89.6**	**39.9**
	89.3	17.9	88.3	15.8	86.3	14.3	**90**	**18.1**
LOS > 7*days*	89.9	7.0	87.9	**8.5**	86.1	4.3	**88.4**	6.8
	87.6	1.4	**88.4**	2.1	85.9	0.8	87.3	**2.4**

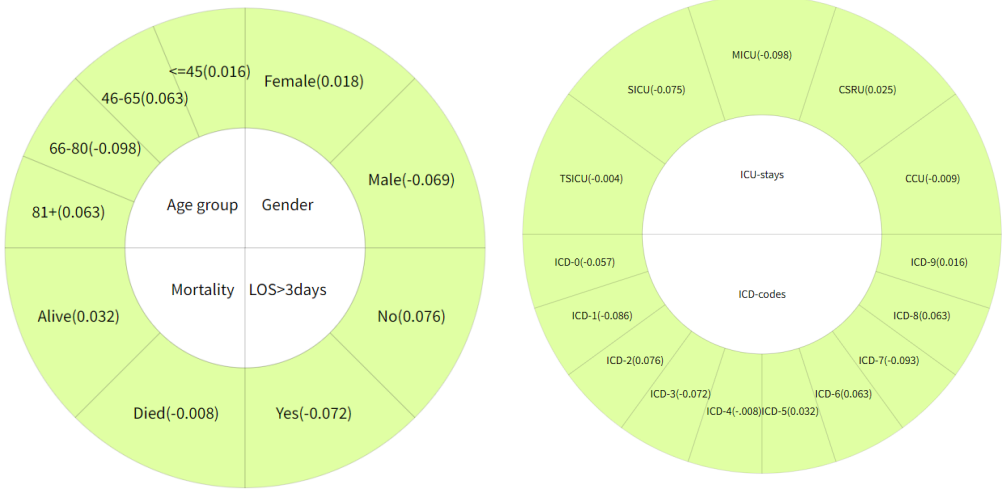
Table 2: Accuracy and F1 on various prediction tasks with real data as reference point. For each task, the first row denotes the predictions by LR, and the second row is the predictions by RF (higher is better for all the values).



(a) HealthGAN



(b) Bt-GAN $^{-}$



(c) Bt-GAN

Figure 5: Representation of sub-groups (The LDSS scores are given in brackets)

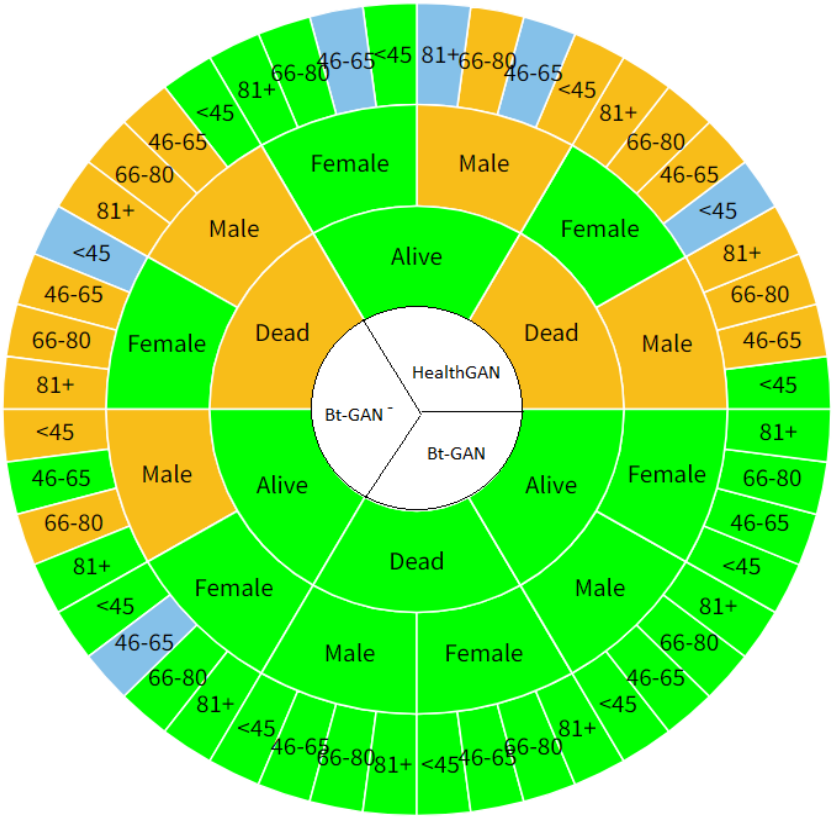
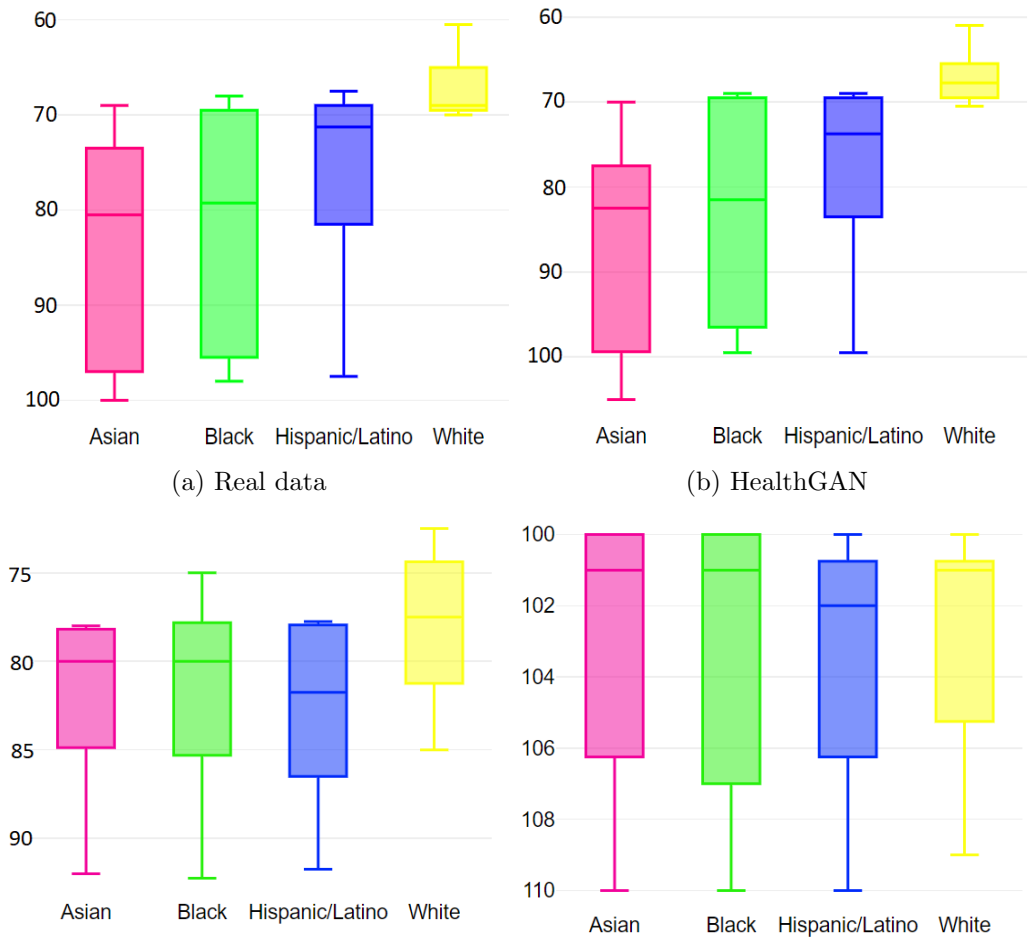


Figure 6: Comparison of log disparity values on the combination of attributes such as mortality, age, and gender. The underrepresented (orange), over-represented (blue), and adequately represented (green) demographic combinations of the proposed models are compared with HealthGAN. The chart areas are respectively divided as models, mortality, gender, and age.

Table 2 reports the accuracy and F1 score of the four downstream tasks. The data generated by HealthGAN and Bt-GAN(ours) maintain the accuracy and F1 score for all the prediction tasks as that of real data, whereas FairGAN fails to keep those. The performance of our model is commendable as it keeps almost the same accuracy and F1 score as that of HealthGAN with added fairness. Additionally, we compare utility in terms of AUROC and AUPRC using LR and RF. Figure 4 shows 4 plots: (a) AUROC_LR, (b) AUROC _RF, (c) AUPRC_LR, and (d) AUPRC_RF. The difference between HealthGAN and Bt-GAN is negligible in all plots, whereas FairGAN achieves the worst performance. Finally, we provide sample-level metrics analysis (Alaa et al., 2022) to verify the quality, fidelity, diversity, and generalization in Table 3. Our model is superior in the discriminative score, β recall, JSD, authenticity, and context FID compared to SOTA.



(a) Real data (b) HealthGAN

(c) FairGAN (d) Bt-GAN

Figure 7: Importance of *ethnicity* among sub-groups by LR using SHAP (feature ranks in the vertical axis)

Model	Discriminative score (↓)	JSD (↓)	α precision (↑)	β recall (↑)	Authenticity (↑)	Context FID (↓)
HealthGAN	0.31±.002	0.032±.01	**0.82±.002**	0.52±.003	0.91±.001	0.89±.021
FairGAN	0.46±.24	0.074±.05	0.56±0.53	0.21±0.12	0.62±.001	3.12±.32
Bt-GAN$^-$	0.35 ±0.01	**0.03±.21**	0.70±.15	0.54±.04	0.88±.13	0.89±.13
Bt-GAN	**0.29 ±0.01**	0.031 ±.001	0.810±.001	**0.61 ±0.032**	**0.92±.010**	**0.81±.12**

Table 3: Quantitative analysis;↑ indicates higher the better, ↓ indicates lower the better, best results are bolded).

8.2 Fairness Analysis

MIMIC-III is composed of a set of sensitive features. As per Equal Credit Opportunity Act [ECOA], gender, age, ethnicity, insurance type, and marital status are considered sensitive

Metrics	Prediction	Real Data	HealthGAN	FairGAN	Bt-GAN
AUROC gap	In-hospital mortality	0.043 ± 0.001	0.082 ± 0.002	0.021 ± 0.001	**0.001 ± 0.001**
	In-ICU mortality	0.03 ± 0.007	0.15 ± 0.035	0.023 ± 0.064	**0.012 ± 0.021**
	LOS>3days	-0.003 ± 0.002	-0.104 ± 0.001	-0.003 ± 0.001	**0.000 ± 0.001**
	LOS>7days	-0.005 ± 0.002	-0.076 ± 0.002	-0.061 ± .001	**-0.013 ± 0.001**
Parity gap	In-hospital mortality	-0.046 ± 0.018	-0.154 ± 0.010	-0.004 ± 0.014	**0.000 ± 0.001**
	In-ICU mortality	-0.031 ± 0.013	-0.331 ± 0.011	-0.005 ± 0.013	**0.000 ± 0.000**
	LOS>3days	0.022 ± 0.012	0.224 ± 0.012	0.022 ± 0.002	**0.000 ± 0.001**
	LOS>7days	-0.004 ± 0.002	-0.004 ± 0.002	**-0.002 ± 0.001**	-0.003 ± 0.001

Table 4: Comparison of fairness gaps between white and black patients with real data as reference (best results are in bold).

information. To compare the fairness of the proposed model with SOTA, we choose *ethnicity* on which we enforce $\mathcal{U}(\mathcal{S}, \mathcal{Y})$ fairness. Though there are no restrictions in choosing the protected features in this study, enforcing equality on age influences the patient's health which may affect mortality or length of stays in the ICU and cause unnecessary medical interventions. So, we recommend the choice of protected features based on the context of applications and suggestions from healthcare experts.

To compare $\mathcal{U}(\mathcal{S}, \mathcal{Y})$ fairness in ethnicity sub-groups, we choose two prediction tasks by LR: (i) In-hospital mortality, and (ii) $LOS > 7days$ on which we analyze the parity and AUROC gap between black and white patients across various synthetic data.

Table 4 compares the parity as well as AUROC gaps in different GANs with real data. The AUROC gap and the parity gap show that predictions on HealthGAN-generated data are biased and amplified compared to real data. In all the cases, a positive value represents a bias towards white patients and a negative value represents a bias towards black patients. The data is fair when the value is close to zero. FairGAN minimizes these biases by adversarial debiasing. The gaps in Bt-GAN are almost close to zero for all the prediction tasks considered. With this being said, the differences in parity, as well as AUROC gaps between HealthGAN-generated data and Bt-GAN, are significant for these predictions.

To further evaluate the data leakage and model leakage (Wang et al., 2019) of our proposed model, we train an attacker (which is an *ethnicity* classifier) on the ground truth labels (of various generated data) and the corresponding downstream predictions (by LR) of the In-ICU mortality task. Note that the leakage of LR on the real dataset is 0.60 ± 001 with an F1 score of 92.3. This shows that the underlying generative model amplifies these leakages as shown in Table 5. Our proposed Bt-GAN models achieve superior performance in controlling these leakages as the bias amplification factor Δ, which is the difference between the model and data leakage is less than zero (details in Appendix). Note that adversarial de-biasing by FairGAN does not account for both leakages, but is mitigated in Bt-GAN through MI-debiasing in the Bt-DGP.

Remark. The adversarial de-biasing-based generation process generates fair data by fooling the discriminator. Our Bt-GAN method uses a module for estimating the mutual information which does not compete with the generator. That means we generate fair health data without fooling the estimator, but by minimizing the information it estimates. The

advantage of our method is that we can train the estimator until convergence in every epoch which is not possible in adversarial de-biasing methods.

Data	Data leakage	Model leakage
HealthGAN	63.17±0.45	65.42±0.58
FairGAN	60.81±0.32	62.97±0.31
Bt-GAN$^-$	50.45±0.10	49.16±0.14
Bt-GAN	**49.90±0.63**	**48.31±0.71**

Table 5: Evaluation of data and model leakage.

8.3 Representation Fairness Analysis

It is essential for health sector research to incorporate demographic information as part of clinical research. One such example could be to test the effect of vaccines on different age groups, gender, or different physical conditions. So, the real data should include appropriate proportions of all the sub-groups to use for clinical research effectively. Thus, similar distributions of sub-groups should be captured when it is being generated.

Our analysis of representation fairness can be performed in two ways: (i) on sub-groups identified by gender, age, in-hospital mortality, and $LOS > 3days$, and (ii) on sub-groups defined by ICD-codes and ICU stay in the emergency care unit. Note that, these sub-groups are selected not based on any clinical research but to analyze how these proportions are being propagated into synthetic data.

Similar to (Bhanot et al., 2021), we set the value of δ (Definition 9) to $\pm\log(0.9)$ (90 percent rule) as a threshold for over/underrepresented groups. Based on this, we divide the range into 4 levels, such as missing $[-\propto$ to $\log(.8)]$ (yellow), underrepresented $(\log(.8)$ to $\log(.9)]$ (orange), adequately represented $(\log(.9)$ to $-\log(.9)]$ (green) and over-represented $(-\log(.9)$ to $-\log(.8)]$ (blue).

Analysis of Sub-groups Identified by Gender, Age, In-hospital Mortality, and $LOS > 3days$. We analyze the individual representation issues and the issues caused by the intersection of these sub-groups in detail. Figure 5 (top row) compares representation issues calculated by $LDSS$ for each of the sub-groups. More representation biases (under/over) are caused by HealthGAN and Bt-GAN$^-$-generated data. Specifically, more over-representation can be seen towards the attributes defined by $age =' 81+'$, $mortality =' Alive'$, and $LOS > 3days =' No'$. The attributes $age ='<= 45'$, $mortality =' Died'$, and $LOS > 3days =' Yes'$ are underrepresented. Suppose, if the majority of the under-represented combinations belong to one particular class and the over-represented in the other, the predictions by a downstream classifier will be severely biased, even though the overall prediction accuracy is acceptable.

Analysis of Sub-groups Identified by ICD Codes and ICU Stay in the Emergency Care. We analyze the representation biases caused by ICD codes and various ICU stays (CCU, CSRU, MICU, SICU, and TSICU) in Figure 5 (bottom row). This analysis is very important in healthcare for setting up emergency care based on appropriate disease characteristics (defined by ICD codes). Among the ICU stays, CSRU and SICU are underrepresented in both the data generated by HealthGAN and Bt-GAN$^-$, whereas CCU

is over-represented. The ICD codes (ICD-0,3,4,5, and 9) are underrepresented in HealthGAN and Bt-GAN$^-$ (more results are in Appendix). Note that, Bt-GAN captures all these sub-groups in exact proportions in contrast to other schemes. Furthermore, we analyze the representation issues caused by the cross-section of sub-groups in Figure 6. Though the representations of gender are correctly captured by all the models, their intersections with other demographics cause certain biases.

8.4 Attribute-based Local Explainability Analysis using SHAP

Motivated by (Cesaro & Gagliardi Cozman, 2019), we analyze the group feature importance for samples in In-hospital mortality prediction by LR and measure how attributed it is across various sub-groups defined by the *ethnicity* group. We then rank the importance among other features within SHAP exploratory analysis as shown in Figure 7. We observe that the attribute *ethnicity* has a great impact (low ranks) on the sub-groups *White* in real data, FairGAN, and HealthGAN resulting in biased explanations (Jain et al., 2020). In Bt-GAN, the effect is more balanced and even reduced, which means the feature *ethnicity* is less important to the model for all the sub-groups. This local analysis is helpful in finding the disparities between the sub-groups in situations where global explanations cannot reveal the inequalities between the sub-groups.

9. Experiment on Fairness Benchmark Datasets

In this section, we evaluate the effectiveness of our model in producing fair synthetic data on fairness benchmark datasets such as Adult Income[4] and COMPAS[5]. Note that the missing labels in both of these datasets are negligible. Therefore, semi-supervised learning has no impact on the generation quality, and thus the terms associated with the classifier in Eq. (4) have no relevance. As a result, Eq. (4) can be replaced by Eq. (3) for any datasets with no (or negligible) missing labels.

9.1 Datasets

UCI Adult Dataset This dataset is based on US census data (1994) and contains 48,842 rows with attributes such as age, sex, occupation, and education level, and the target variable indicates whether an individual has an income that exceeds $50K$ per year or not. In our experiments, we consider the protected attribute to be sex (S = "Sex", Y = "Income").

ProPublica Dataset from COMPAS Risk Assessment System This dataset contains information about defendants from Broward County and contains attributes about defendants such as their ethnicity, language, marital status, sex, etc., and for each individual a score showing the likelihood of recidivism (reoffending). In this experiment, we used a modified version of the dataset. First, attributes such as FirstName, LastName, MiddleName, CASE ID, and DateOfBirth are removed. Studies have shown that this dataset is biased against African Americans. Therefore, ethnicity is chosen to be the protected attribute for this study. Only African American and Caucasian individuals are kept and the

4. https://archive.ics.uci.edu/ml/datasets/adult
5. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

	Datasets	Precision	Recall	AUROC gap	Parity gap
Real Data	Adult	0.902 ± 0.001	0.921 ± 0.002	0.198 ± 0.018	0.121 ± 0.024
	COMPAS	0.903 ± 0.007	0.914 ± 0.007	0.239± 0.002	0.258 ±0.032
DECAF	Adult	0.781 ± 0.018	0.881 ± 0.050	0.011 ± 0.002	0.001 ± 0.001
	COMPAS	0.874 ± 0.010	0.886 ± 0.001	0.021 ± 0.010	0.003 ± 0.011
FairGAN	Adult	0.661 ± 0.140	0.679 ± 0.001	0.089±0.002	0.097 ± 0.018
	COMPAS	0.785 ± 0.010	0.832 ± 0.002	0.045±0.023	0.205 ± 0.055
Bt-GAN (ours)	Adult	**0.898 ± 0.001**	**0.900 ± 0.020**	**0.002± 0.010**	**0.001± 0.020**
	COMPAS	**0.896 ± 0.001**	**0.899 ± 0.003**	**0.001 ± 0.002**	**0.002± 0.001**

Table 6: Data utility and fairness analysis on fairness benchmark datasets with real data as the reference point (higher is better for all the values)

rest are dropped. The target variable in this dataset is a risk decile score provided by the COMPAS system, showing the likelihood of that individual to re-offend, which ranges from 1 to 10. The final modified dataset contains 16,267 records with 16 features. To make the target variable binary, a cut-off value of 5 is considered and individuals with a decile score of less than 5 are considered "Low Chance", while the rest are considered "High Chance". (S = "Ethnicity", Y = "Recidivism Chance").

Competing Methods The methods we compare against are based on fair tabular data generation using GANs. We compare against FairGAN and DECAF as these are state-of-the-art methods in the tabular domain.

9.2 Results

We list the utility and fairness measures in Table 6. The precision and recall of our proposed Bt-GAN are better than FairGAN and DECAF. The fairness evaluated in terms of the AUROC gap and parity gap of Bt-GAN shows that the bias in synthetic data has been reduced to a great extent compared to state-of-the-art models. We achieve a good balance between utility and fairness through a Bt-DGP.

10. Conclusion and Future Works

We proposed Bt-GAN, a semi-supervised Generative Adversarial Network for synthetic healthcare data generation, which accounts for the spurious correlations in the data and promotes representation fairness in the target in an end-to-end framework. This is done by proposing a Bias-transforming generation process, incorporating an information-constrained fairness penalty to tackle correlation biases, a score-based weighted sampling to promote representation fairness, and semi-supervised learning to capture true data distribution under partially unlabeled data. Compared to the SOTA schemes, our method is found to be more reliable to better balance the tradeoff between accuracy and fairness with minimal bias amplification in a more generalized and principled way.

Future Directions: This work is based on GANs and feature-specific constraints such as mutual information. An interesting future direction could be to extend this in the

framework using diffusion models and causality. We leave further experiments to extend our proposed methods from these perspectives as future work.

Ethical and Societal Implications: The notion of fairness is domain and context-dependent, which is more sensitive in healthcare. The proposed methods could be coupled with tasks, where the goal is the progression of trustworthy AI.

Acknowledgments

This work was funded by the Knut and Alice Wallenberg Foundation, the ELLIIT Excellence Center at Linköping-Lund for Information Technology, and TAILOR - an EU project with the aim to provide the scientific foundations for Trustworthy AI in Europe. The computations were enabled by the Berzelius resource provided by the Knut and Alice Wallenberg Foundation at the National Supercomputer Centre.

Appendix A. Quality and Fairness Definitions

Discriminative Score: To calculate the score, we train a classifier(LR) to classify *real* or *fake* with the original data and synthetic data respectively labeled as real and fake and calculated the error.

Jensen-Shannon Divergence (JSD): Here we used JSD to evaluate the quality of the synthetic data as compared to the real data (Menéndez et al., 1997).

α-precision, β-recall, and Authenticity: It is used to evaluate the fidelity, diversity, and generalization of the synthetic data at the sample level as proposed in (Alaa et al., 2022).

Context FID: It is used to measure the difference in statistics between the real and synthetic data with respect to downstream tasks (Jeha et al., 2021).

Data Leakage: It is used to evaluate how much information about the protected attributes can be leaked through task-specific labels. To measure this, we train an attacker f and evaluate it on held-out data. The performance of the attacker, the fraction of instances in $\mathcal{D}$ that leak information about S_i through Y_i, yields an estimate of leakage (Wang et al., 2019):

$$Leakage_{\mathcal{D}} = \frac{1}{|\mathcal{D}|} \sum_{(Y_i, S_i) \in \mathcal{D}} 1\left[f\left(Y_i\right) == S_i\right] \tag{13}$$

Model Leakage: To measure the degree a model, M produces predictions, $\hat{Y}_i = M\left(X_i\right)$, that leak information about the protected variable S_i. We define model leakage as the percentage of examples in $\mathcal{D}$ that leak information about S_i through $\hat{Y}_i$. To measure prediction leakage, we train a different attacker on $\hat{Y}_i$ to extract information about S_i (Wang et al., 2019) :

$$Leakage_{M} = \frac{1}{|\mathcal{D}|} \sum_{\left(\hat{Y}_i, S_i\right) \in \mathcal{D}} 1\left[f\left(\hat{Y}_i\right) == S_i\right)] \tag{14}$$

We measure fairness using two measures; the AUROC gap and the parity gap. The AUROC gap is defined as the difference in AUROC between the selected sub-groups for a particular healthcare task. The parity gap is the difference in statistical parity between two sub-groups for a specified task. In Table 7, we give definitions of the AUROC gap and Parity gap for sub_groups SG_i.

AUROC gap	$\text{AUROC}\,(\text{SG}_1) - \text{AUROC}\,(\text{SG}_2)$
Parity Gap	$\frac{TP_1+FP_1}{N_1} - \frac{TP_2+FP_2}{N_2}$

Table 7: Fairness definitions

Appendix B. Extensions to Other Fairness Notions

Analogous to the connection between statistical parity and zero mutual information, the fairness notions such as equalized opportunity and equalized odds can be designed as a conditional mutual information optimization problem as detailed in (Ghassami et al., 2018).

Appendix C. Architecture Details

Implementation Details: The training epochs use a mini-batch size of 1024. The learning rate is set to .0001 with Adam optimizer. We carried out the experiments using PyTorch in Intel Core i-9, 11th generation, with 128 GB RAM and GPU-2* NVIDIA RTX 2080TI (11 GB). Also, the details architecture used in this study can be found on Table 9.

Dataset Preparation: We extracted records from tables, PATIENTS, ADMISSIONS, ICU STAYS, CHARTEVENTS, LABEVENTS, and OUTPUTEVENTS. Records are then validated using HADM_ID and ICUSTAY_ID, resulting in a total of 33,798 patients with 42276 ICU stays. Among them, we split 28728 patients, 35948 ICU stays for training, and the remaining for testing. We excluded patients with missing HADM_ID and ICUSTAY_ID.

Appendix D. Additional Results

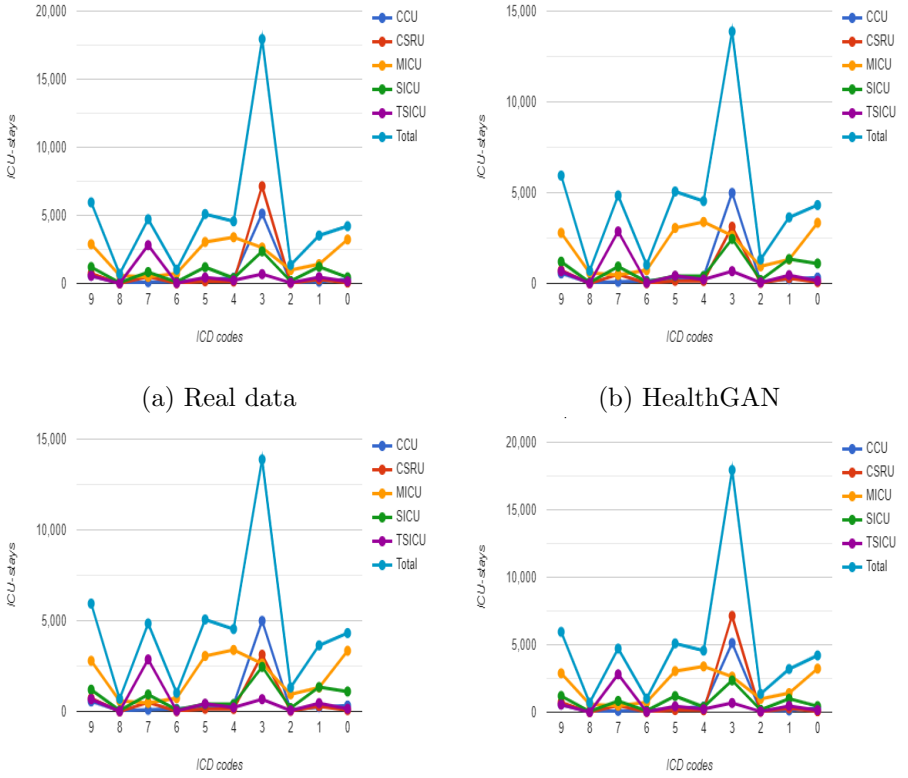
Additional results on the representation analysis are given in Figure 8 and Figure 9.

Appendix E. Cohort Summary

Table 8 details the cohort summary defined by various demographic information.

Appendix F. Impact of Proxy Attributes

We experimented further to study the impact of proxy attributes to check whether these can introduce any biases. We added an extra feature, S_proxy that is strongly correlated with *ethnicity*, particularly for white and black sub-groups. For black patients, $S\_proxy = 1$ for 95 percent of all cases and $S\_proxy = 0$ for the remaining 5 percent. For white patients, the above values are swapped. A correlation plot shows that there is a strong correlation



(a) Real data

(b) HealthGAN

(c) Bt-GAN-

(d) Bt-GAN

Figure 8: Distribution of ICD-codes in ICU stays

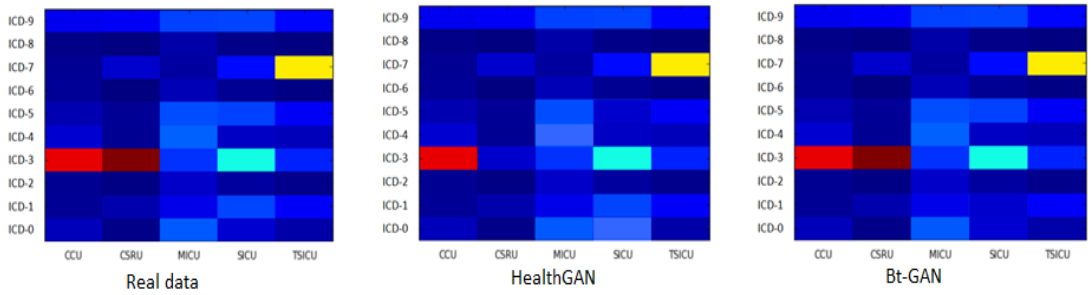


Figure 9: Heatmap representation of ICD codes and ICU stays in real data, HealthGAN generated data and Bt-GAN generated data

between *ethnicity* and *S_proxy* as well as the combination of these to the mortality in the real health data. In our synthetic data, the correlation between 'ethnicity' and *S_proxy* remains the same but the correlation of both of these to the mortality is reduced to a great

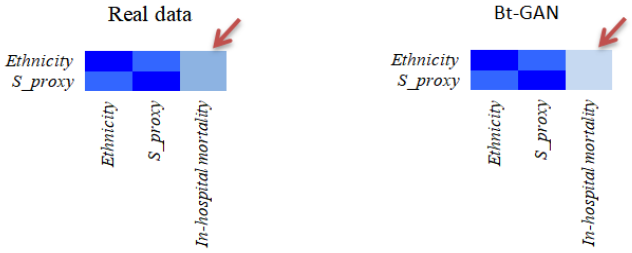


Figure 10: The MI de-biasing accounts for proxy attribute

Attributes	Groups	Percentage
Ethnicity	Black	8%
	White	71%
	Hispanic	3%
	Other	15%
	Asian	2%
Gender	Female	43%
	Male	57%
Insurance	Self-pay	1%
	Government	3%
	Medicaid	8%
	Private	34%
	Medicare	53%
Emergency	TSICU	13%
	CCU	15%
	SICU	16%
	CSRU	20%
	MICU	35%
Total		33798(100%)

Table 8: Cohort summary

extent. So, it is evident that the MI de-biasing accounts for proxy attributes as well (Figure 10).

References

Adiga, S., Attia, M. A., Chang, W.-T., & Tandon, R. (2018). On the tradeoff between mode collapse and sample quality in generative adversarial networks. In *2018 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, pp. 1184–1188. IEEE. doi:10.1109/GlobalSIP.2018.8646478.

Alaa, A., Van Breugel, B., Saveliev, E. S., & van der Schaar, M. (2022). How faithful is your synthetic data? sample-level metrics for evaluating and auditing generative models.

C_φ	G_θ	D_ϕ , D_ζ	T_η
z	No. of features	(W, S)	$-T_n($ W)
(normal distribution)	- Dense layer of 64	- Dense layer of 64	X - Dense layer of 8
- Dense layer of 2	- leaky ReLU	- leaky ReLU	- ReLU
× no.of features -ReLU	- Dense layer of 128	- Dense layer of 64	- Dense layer of 8
BN - Dense layer of 1.5	- leaky ReLU	- leaky ReLU	- ReLU
× no.of features- ReLU	- Dense layer of 256	- Dense layer of 64	- Dense layer of 2
- Dense layer of 1	- leaky ReLU	- leaky ReLU	- ReLU - Y
× no.of features- ReLU	- Dense layer of 1		
- $\text{Gumble}_{0.2}$ softmax(discrete)	- leaky ReLU		

Table 9: Neural Networks for Generator, G_θ, Discriminator, D_ϕ, D_ζ, Classifier, C_φ, and MINE, T_η

In *International Conference on Machine Learning*, pp. 290–306. PMLR.

Arjovsky, M., Chintala, S., & Bottou, L. (2017). Wasserstein generative adversarial networks. In *International conference on machine learning*, pp. 214–223. PMLR.

Armanious, K., Jiang, C., Fischer, M., Küstner, T., Hepp, T., Nikolaou, K., Gatidis, S., & Yang, B. (2020). Medgan: Medical image translation using gans. *Computerized medical imaging and graphics*, *79*, 101684. doi:10.1016/j.compmedimag.2019.101684.

Azadi, S., Olsson, C., Darrell, T., Goodfellow, I., & Odena, A. (2019). Discriminator rejection sampling. In *International Conference on Learning Representations*.

Belghazi, M. I., Baratin, A., Rajeswar, S., Ozair, S., Bengio, Y., Courville, A., & Hjelm, R. D. (2018). Mine: Mutual information neural estimation.. doi:10.48550/arXiv.1801.04062.

Benesty, J., Chen, J., Huang, Y., & Cohen, I. (2009). Pearson correlation coefficient. In *Noise reduction in speech processing*, pp. 1–4. Springer. doi:10.1007/978-3-642-00296-0_5.

Bhanot, K., Qi, M., Erickson, J. S., Guyon, I., & Bennett, K. P. (2021). The problem of fairness in synthetic healthcare data. *Entropy*, *23*(9), 1165. doi:10.3390/e23091165.

Cesaro, J., & Gagliardi Cozman, F. (2019). Measuring unfairness through game-theoretic interpretability. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 253–264. Springer. doi:10.1007/978-3-030-43823-4_22.

Creager, E., Madras, D., Jacobsen, J.-H., Weis, M., Swersky, K., Pitassi, T., & Zemel, R. (2019). Flexibly fair representation learning by disentanglement. In *International conference on machine learning*, pp. 1436–1445. PMLR.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2012). Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pp. 214–226. doi:10.1145/2090236.2090255.

Edemekong, P. F., Annamaraju, P., & Haydel, M. J. (2018). Health insurance portability and accountability act..

Evans, R. S. (2016). Electronic health records: then, now, and in the future. *Yearbook of medical informatics*, *25*(S 01), S48–S61. doi:10.15265/IYS-2016-s006.

Ghassami, A., Khodadadian, S., & Kiyavash, N. (2018). Fairness in supervised learning: An information theoretic approach. In *2018 IEEE International Symposium on Information Theory (ISIT)*, pp. 176–180. IEEE. doi:10.1109/ISIT.2018.8437807.

Ghassemi, M., Oakden-Rayner, L., & Beam, A. L. (2021). The false hope of current approaches to explainable artificial intelligence in health care. *The Lancet Digital Health*, *3*(11), e745–e750. doi:10.1016/S2589-7500(21)00208-9.

Gupta, A., Bhatt, D., & Pandey, A. (2021). Transitioning from real to synthetic data: Quantifying the bias in model.. doi:10.48550/arXiv.2105.04144.

Guzman, U. S., & Sontag, D. An open benchmark for causal inference using the mimic-iii dataset..

Hemamou, L., & Coleman, W. (2022). Delivering fairness in human resources ai: Mutual information to the rescue. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing*, pp. 867–882.

Jain, A., Ravula, M., & Ghosh, J. (2020). Biased models have biased explanations.. doi:10.48550/arXiv.2012.10986.

Jeha, P., Bohlke-Schneider, M., Mercado, P., Kapoor, S., Nirwan, R. S., Flunkert, V., Gasthaus, J., & Januschowski, T. (2021). Psa-gan: Progressive self attention gans for synthetic time series. In *International Conference on Learning Representations*.

Jiang, F., Jiang, Y., Zhi, H., Dong, Y., Li, H., Ma, S., Wang, Y., Dong, Q., Shen, H., & Wang, Y. (2017). Artificial intelligence in healthcare: past, present and future. *Stroke and vascular neurology*, *2*(4). doi:10.1136/svn-2017-000101.

Kang, J., Xie, T., Wu, X., Maciejewski, R., & Tong, H. (2021). Multifair: multi-group fairness in machine learning.. doi:10.48550/arXiv.2105.11069.

Kodali, N., Abernethy, J., Hays, J., & Kira, Z. (2017). On convergence and stability of gans.. doi:10.48550/arXiv.1705.07215.

Kynkäänniemi, T., Karras, T., Laine, S., Lehtinen, J., & Aila, T. (2019). Improved precision and recall metric for assessing generative models. *Advances in Neural Information Processing Systems*, *32*.

Lee, J., Kim, H., Hong, Y., & Chung, H. W. (2021). Self-diagnosing gan: Diagnosing underrepresented samples in generative adversarial networks. *Advances in Neural Information Processing Systems*, *34*.

Li, C., Xu, T., Zhu, J., & Zhang, B. (2017). Triple generative adversarial nets. *Advances in neural information processing systems*, *30*.

Locatello, F., Abbati, G., Rainforth, T., Bauer, S., Schölkopf, B., & Bachem, O. (2019). On the fairness of disentangled representations. *Advances in neural information processing systems*, *32*.

Mao, X., Li, Q., Xie, H., Lau, R. Y., Wang, Z., & Paul Smolley, S. (2017). Least squares generative adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

Menéndez, M., Pardo, J., Pardo, L., & Pardo, M. (1997). The jensen-shannon divergence. *Journal of the Franklin Institute*, *334*(2), 307–318. doi:10.1016/S0016-0032(96)00063-4.

Mo, S., Kim, C., Kim, S., Cho, M., & Shin, J. (2019). Mining gold samples for conditional gans. *Advances in Neural Information Processing Systems*, *32*.

Nohara, Y., Matsumoto, K., Soejima, H., & Nakashima, N. (2019). Explanation of machine learning models using improved shapley additive explanation. In *Proceedings of the 10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*, pp. 546–546. doi:10.1145/3307339.3343255.

Patki, N., Wedge, R., & Veeramachaneni, K. (2016). The synthetic data vault. In *2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, pp. 399–410. IEEE. doi:10.1109/DSAA.2016.49.

Radford, A., Metz, L., & Chintala, S. (2015). Unsupervised representation learning with deep convolutional generative adversarial networks.. doi:10.48550/arXiv.1511.06434.

Regulation, P. (2016). Regulation (eu) 2016/679 of the european parliament and of the council. *Regulation (eu)*, *679*, 2016.

Rieke, N., Hancox, J., Li, W., Milletari, F., Roth, H. R., Albarqouni, S., Bakas, S., Galtier, M. N., Landman, B. A., Maier-Hein, K., et al. (2020). The future of digital health with federated learning. *NPJ digital medicine*, *3*(1), 1–7. doi:10.1038/s41746-020-00323-1.

Sinha, S., Zhao, Z., ALIAS PARTH GOYAL, A. G., Raffel, C. A., & Odena, A. (2020). Top-k training of gans: Improving gan performance by throwing away bad samples. *Advances in Neural Information Processing Systems*, *33*, 14638–14649.

Sinn, M., & Rawat, A. (2018). Non-parametric estimation of jensen-shannon divergence in generative adversarial network training. In *International Conference on Artificial Intelligence and Statistics*, pp. 642–651. PMLR.

Song, J., Kalluri, P., Grover, A., Zhao, S., & Ermon, S. (2019). Learning controllable fair representations. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 2164–2173. PMLR.

van Breugel, B., Kyono, T., Berrevoets, J., & van der Schaar, M. (2021). Decaf: Generating fair synthetic data using causally-aware generative networks. *Advances in Neural Information Processing Systems*, *34*.

Veyrat-Charvillon, N., & Standaert, F.-X. (2009). Mutual information analysis: how, when and why?. In *International Workshop on Cryptographic Hardware and Embedded Systems*, pp. 429–443. Springer. doi:10.1007/978-3-642-04138-9_30.

Wachter, S., Mittelstadt, B., & Russell, C. (2020). Bias preservation in machine learning: the legality of fairness metrics under eu non-discrimination law. *W. Va. L. Rev.*, *123*, 735. doi:10.2139/ssrn.3792772.

Walonoski, J., Kramer, M., Nichols, J., Quina, A., Moesel, C., Hall, D., Duffett, C., Dube, K., Gallagher, T., & McLachlan, S. (2018). Synthea: An approach, method, and software mechanism for generating synthetic patients and the synthetic electronic health care record. *Journal of the American Medical Informatics Association*, *25*(3), 230–238. doi:10.1093/jamia/ocx079.

Wang, T., Zhao, J., Yatskar, M., Chang, K.-W., & Ordonez, V. (2019). Balanced datasets are not enough: Estimating and mitigating gender bias in deep image representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5310–5319.

Wright, D. (2012). The state of the art in privacy impact assessment. *Computer law & security review*, *28*(1), 54–61. doi:10.1016/j.clsr.2011.11.007.

Xu, D., Yuan, S., Zhang, L., & Wu, X. (2018). Fairgan: Fairness-aware generative adversarial networks. In *2018 IEEE International Conference on Big Data (Big Data)*, pp. 570–575. IEEE. doi:10.1109/BigData.2018.8622525.

Yale, A., Dash, S., Dutta, R., Guyon, I., Pavao, A., & Bennett, K. P. (2020). Generation and evaluation of privacy preserving synthetic health data. *Neurocomputing*, *416*, 244–255. doi:10.1016/j.neucom.2019.12.136.

Zhang, H., Goodfellow, I., Metaxas, D., & Odena, A. (2019). Self-attention generative adversarial networks. In *International conference on machine learning*, pp. 7354–7363. PMLR.

Zheng, Z., Zheng, L., & Yang, Y. (2017). Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.