Generative Counterfactual Augmentation for Bias Mitigation

Jason Uwaeze
Rice University
ju6@rice.edu

Pranav Kulkarni
University of Maryland
pranavk@umd.edu

Vladimir Braverman
Johns Hopkins University
vova@cs.jhu.edu

Michael A. Jacobs
UTHealth Houston
michael.a.jacobs@uth.tmc.edu

Vishwa S. Parekh
UTHealth Houston
vishwa.s.parekh@uth.tmc.edu

Abstract

Deep learning (DL) models trained for chest x-ray (CXR) classification can encode protected demographic attributes and exhibit bias towards underrepresented patient populations. In this work, we propose Generative Counterfactual Augmentation (GCA), a framework for mitigating algorithmic bias through demographic-complete augmentation of training data. We use a StyleGAN3-based synthesis network and SVM-guided latent space traversal to generate structured age and sex counterfactuals for each CXR while preserving disease features. We extensively evaluate GCA for training DL models with the RSNA Pneumonia dataset using controlled underdiagnosis bias injection across age- and sex-groups at varying rates. Our results show up to 23% reduction in FNR disparity, with a mean reduction of 9%, across varying rates of underdiagnosis bias. When evaluated with the external CheXpert and MIMIC-CXR datasets, we observe a consistent FNR reduction and improved model generalizability. We demonstrate that GCA is an effective strategy for mitigating algorithmic bias in DL models for medical imaging, ensuring trustworthiness in clinical settings. Our code is available at https://github.com/Wazhee/GCA

1. Introduction

Algorithmic bias is a significant barrier to the clinical adoption of deep learning (DL) models for disease diagnosis and prognosis [3, 4, 7, 8, 25, 31]. Prior work has shown that DL models trained on chest x-rays (CXRs) can encode protected demographic attributes (such as sex, age, and race) [6, 14, 30] and exhibit disparities in model performance between demographic groups [5, 13, 26]. This has the potential to amplify existing systematic disparities in healthcare and worsen patient outcomes. As a result, training DL models with bias mitigation strategies is critical for ensuring trustworthiness in clinical settings.

Counterfactual generation has recently emerged as a powerful technique for synthetically modifying medical images, enabling transformations such as synthetic aging, causal inference, disease manipulation, and anatomical modifications [2, 20, 21, 23, 29]. The core idea behind counterfactual generation is to address the "what if" questions in medical imaging – what if this patient were female instead of male? What if they were 80 years old instead of 40? As a model-agnostic technique, it enhances generalizability across diverse architectures and training paradigms. Prior work has shown its effectiveness in improving robustness to domain and population shifts [21, 29]. However, its potential as a bias mitigation strategy remains largely unexplored.

In this work, we propose Generative Counterfactual Augmentation (GCA), a framework for mitigating bias to ensure demographic completeness using counterfactual generation. Our method works by augmenting the training dataset with generated structured demographic (age and sex) counterfactuals for each CXR, while preserving disease features. We extensively evaluate GCA through controlled injection of underdiagnosis bias across varying rates, demonstrating its effectiveness in reducing FNR disparities, and show its generalizability to external datasets. Our main contributions are three-fold:

1. A framework for demographic-complete augmentation of training data which generates structured counterfactuals to mitigate algorithmic bias.
2. Comprehensive evaluation of our method's effectiveness for bias mitigation using controlled injection of underdiagnosis bias across age- and sex-groups.
3. Extensive validation of our method's generalizability in mitigating bias across multiple external datasets.

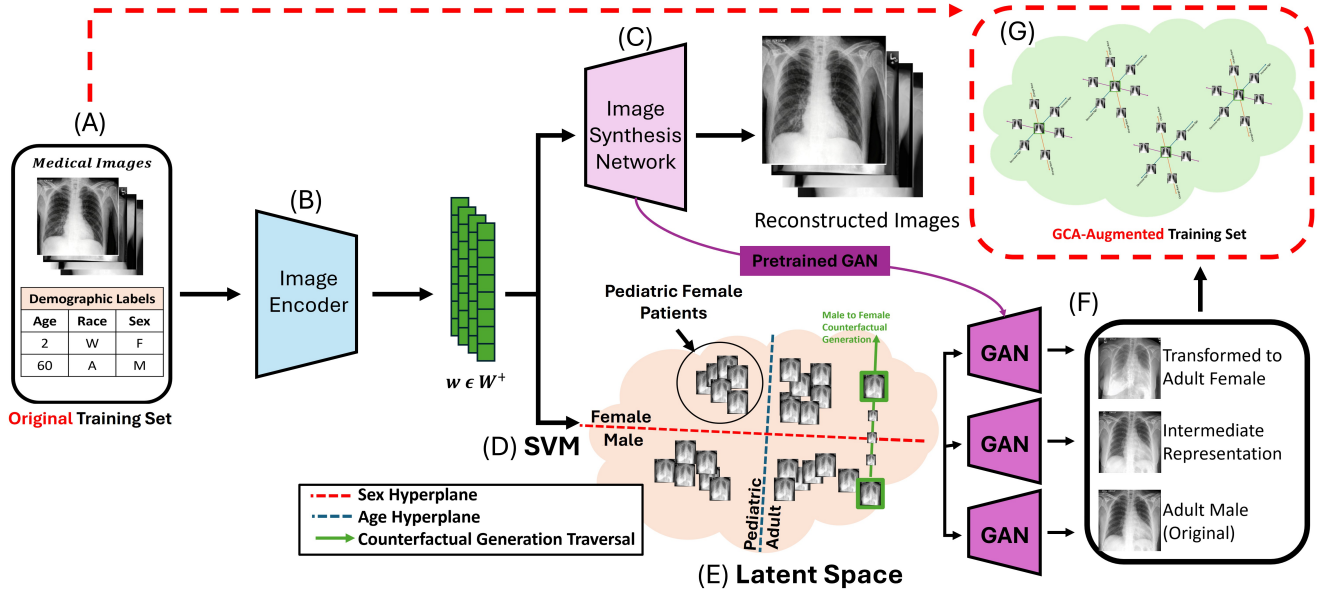


Figure 1. An overview of the GCA framework. **(A)** Medical images with demographic labels. **(B)** Encoder maps images to latent space. **(C)** StyleGAN3-based generator reconstructs images. **(D)** SVMs define demographic hyperplanes. **(E)** Counterfactuals generated via structured traversal. **(F)** Decoder reconstructs demographic variations. **(G)** GCA-augmented dataset ensures demographic completeness.

2. Methods

2.1. Generative Counterfactual Augmentation

Fig. 1 provides a conceptual overview of the GCA framework. GCA systematically generates counterfactuals for each CXR with different demographic attributes (e.g., sex, age) to augment the training dataset (Fig. 1A) into a demographically complete dataset (Fig. 1G). This ensures that DL models trained on the augmented dataset learns from a fully representative population, disentangling protected attributes from disease features across demographic groups. The GCA framework consists of four key components:

2.1.1. Image Generator

We trained an unconditional StyleGAN3 generator, G, to synthesize high-fidelity CXR images. Training was performed using the default parameters from Karras et al. [12] on a single A100 GPU. The generator G comprises two main components. First, a non-linear mapping network, $f_\theta : \mathcal{Z} \to \mathcal{W}$, parameterized by θ, that transforms a randomly sampled normal 512-dimensional vector $\mathbf{z} \in \mathcal{Z}$ into an intermediate 512-dimensional latent vector $\mathbf{w} \in \mathcal{W}$. This transformation constructs a latent space where CXR image features are disentangled, enabling smooth latent space traversals between demographic subgroups. Second, the alias-free synthesis network, G, generates realistic CXR images from the latent representations $\mathbf{w}$. Unlike previous StyleGAN variants, each layer of G is equivariant, making it robust against aliasing and invariant to image morphing transitions (e.g., interpolating between male and female CXR images) [12]. Equivariance in neural networks is defined as $t \circ f = f \circ t$ where f is a nonlinear operation, such as upsampling, and t is a spatial transformation, such as rotation or translation. This property ensures that applying f before t yields the same result as applying t before f.

2.1.2. Image Encoder

While the StyleGAN3 generator, G, can synthesize high-quality CXR images, it lacks the ability to modify real images directly, as it is trained adversarially alongside a discriminator, D, using standard GAN training objectives [11, 12]. This results in uncontrolled image generation, limiting its applicability for counterfactual generation. To overcome this limitation, we introduced an image encoder, E, inspired by Abdal et al. [1], to embed CXR images into the latent space of G. The encoder, E, takes image $i \in I \subseteq \mathbb{R}^{H \times W}$ and pre-trained G as input, then iteratively optimizes a corresponding latent vector $\mathbf{w} \in W^+$ using gradient descent. We optimize E by minimizing a perceptual and mean squared error (MSE) loss function:

$$\mathcal{L}_{\text{percept}}(i_1, i_2) = \sum_{j=1}^{4} \left\| G_j^\phi(i_1) - G_j^\phi(i_2) \right\|_2^2 \quad (1)$$

$$\mathcal{L}_{\text{MSE}}(i_1, i_2) = \frac{\lambda}{N} \left\| i_1 - i_2 \right\|_2^2 \quad (2)$$

$$\mathbf{w}^* = \arg\min_{\mathbf{w}} \left(\mathcal{L}_{\text{percept}}(G(\mathbf{w}), i) + \mathcal{L}_{\text{MSE}}(G(\mathbf{w}), i) \right) \quad (3)$$

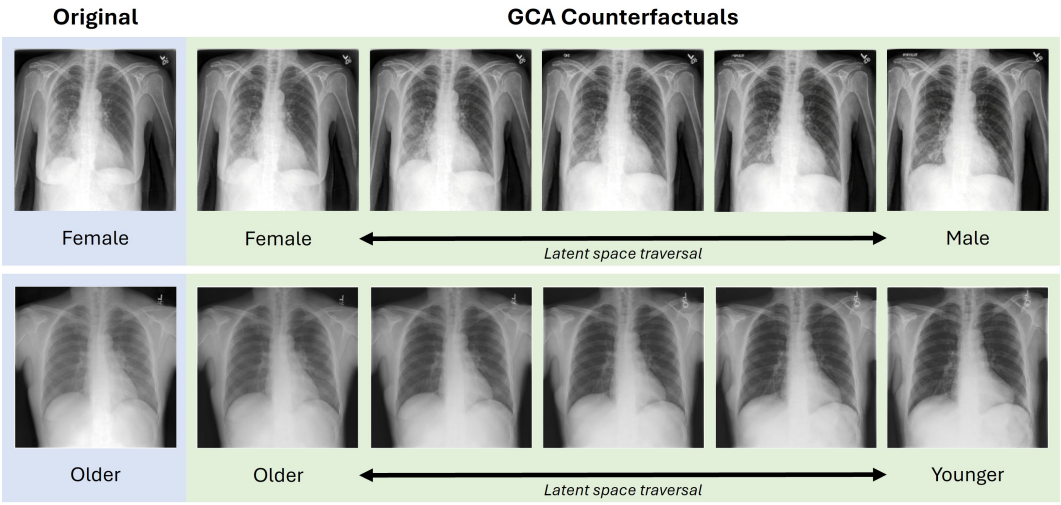


Figure 2. Examples of GCA counterfactuals generated by SVM-guided latent space traversal for sex **(top)** and age **(bottom)** attributes.

where $G_j^{\phi}(\cdot)$ denotes activations of the jth VGG-16 layer pre-trained on ImageNet [22] and $E(i) = \mathbf{w}^* \; \forall i \in I$.

2.1.3. Latent Space Traversal

To generate structured counterfactuals, we trained SVM classifiers in the latent space of G to separate demographic attributes, following Liang et al. [15]. Each SVM learns a hyperplane:

$$\mathbf{w}_A^\top \mathbf{z} + b = 0 \tag{4}$$

where $\mathbf{w}_A$ is the normal vector separating demographic groups. To modify an attribute, we traverse perpendicular to the hyperplane:

$$\mathbf{z}' = \mathbf{z} + \alpha \mathbf{w}_A \tag{5}$$

where α controls the transformation magnitude, ensuring controlled demographic shifts while maintaining anatomical integrity.

The demographic attributes of sex and age differ fundamentally – sex is categorical, while age is continuous. For sex traversal, we trained an SVM to learn a hyperplane, h_{sex}, separating male and female embeddings. Given a latent representation $\mathbf{z}$, we modify sex by moving along the sex axis $\mathbf{w}_{\text{sex}}$:

$$\mathbf{z}' = \mathbf{z} + \alpha \mathbf{w}_{\text{sex}} \tag{6}$$

Unlike sex, age exists along a continuous spectrum, making binary classification infeasible. Instead of training multiple hyperplanes for different age-groups, we trained a single hyperplane, h_{age}, between the youngest (0–20Y) and the oldest (80+Y) groups, enabling smooth interpolation along the age axis w_{age}:

$$\mathbf{z}' = \mathbf{z} + \alpha \mathbf{w}_{\text{age}} \tag{7}$$

Adjusting α allows for synthetic aging or de-aging, ensuring realistic, progressive transformations without abrupt transitions between age groups.

2.1.4. Dataset Augmentation

For each CXR in the training dataset, we generated five intermediate counterfactuals for every demographic attribute (e.g., sex, age) using GCA (Fig. 2). This augments the dataset by $5\times$ to ensure smooth transitions between demographic groups, resulting in a demographically complete training dataset. In the absence of GCA, DL models trained on imbalanced data may associate protected attributes (like demographics) with disease characteristics, resulting in biased predictions [8, 16, 17]. Our method addresses this by ensuring each CXR is represented across multiple demographic attributes, thereby reducing the model's reliance on any single attribute and mitigating bias.

2.2. Experimental Design

2.2.1. Controlled Injection of Underdiagnosis Bias

We implemented structured label perturbations to introduce underdiagnosis bias in a targeted demographic group, allowing us to control for the presence of bias in the training dataset. Given a dataset D_{train}, we define underdiagnosis rate, r, as the rate of underdiagnosis bias characterized by the proportion of positive pneumonia cases mislabeled as "No Findings". Formally, the controlled bias injection transforms the training dataset D_{train} to biased dataset, P_r^T, where r represents the underdiagnosis rate in subgroup T. Then, we independently targeted each sex- and age-group with underdiagnosis rate, $r \in \{0, 0.05, 0.10, 0.25, 0.50, 0.75, 1.00\}$. This produced seven biased datasets per subgroup, denoted as $P \in \{P_0, P_{0.05}, \ldots, P_r\}$, where $P_r \subseteq D_{\text{train}}$ for each r. The

test set remained unchanged to ensure unbiased evaluation.

2.2.2. Model Training and Evaluation

To train GCA, we used the CheXpert [9] and NIH ChestX-ray14 [28] datasets, comprising $n = 224,316$ and $n = 112,120$ CXR images, respectively. To avoid data leakage, we excluded $n = 26,684$ images from the RSNA Pneumonia Detection Challenge dataset, a subset of the NIH dataset. In total, $n = 309,752$ images were used to train the StyleGAN3 generator, G. All images were resized to 256×256 for computational efficiency.

To evaluate the effectiveness of GCA for bias mitigation, we used the RSNA Pneumonia Detection Challenge dataset, which contains $n = 26,684$ frontal CXRs. Following prior work [13, 27], patients labeled with "Consolidation," "Lung Opacity," or "Pneumonia" were categorized as "Pneumonia" in our ground truth labels while all other disease categories were discarded and uncertain labels were classified as "No Findings." We extracted age and sex attributes for all patients, grouping age into five groups (0–20, 20–40, 40–60, 60–80, and 80+ years) [26]. We then applied GCA independently to each demographic attribute to generate two GCA-augmented datasets: **Synth-RSNA-Sex** and **Synth-RSNA-Age**. Additionally, we applied GCA jointly on both attributes to produce a demographic-complete dataset, referred to as **Full-Synth-RSNA**.

We trained pneumonia classifiers using ImageNet pre-trained DenseNet121 models with 5-fold cross validation with 70/10/20 training/validation/testing splits on the RSNA dataset (baseline model) and its GCA-augmented versions: Synth-RSNA-Sex (for sex-group augmentation), Synth-RSNA-Age (for age-group augmentation), and Full-Synth-RSNA (for sex- and age-group augmentation). Each DL model was trained for 100 epochs with batch size of 64, initial learning rate of $5e-5$, and ReduceLROnPlateau scheduler. Model performance was monitored using binary cross-entropy loss. All CXRs were resized to 224×224 and random augmentations are applied during training. Finally, the models are tested on the internal RSNA test set and the external CheXpert ($n = 377,110$) [9] and MIMIC-CXR ($n = 224,316$) [10] datasets.

2.2.3. Metrics and Statistical Analysis

To measure model performance, we use the area under the receiver operating characteristic curve (AUROC) and false negative rate (FNR). Here, FNR refers to the proportion of CXRs with positive pneumonia that were misclassified as negative. We calculate FNR by binarizing the predictions using classification threshold determined by Youden's J statistic:

$$FNR = P(\hat{y} = 0 \mid y = 1) = \frac{FN}{FN + TP} \tag{8}$$

where y denotes the ground truth label and $\hat{y}$ denotes the predicted label. Paired t-tests are used to compare metrics and statistical significance is defined as $p < 0.05$.

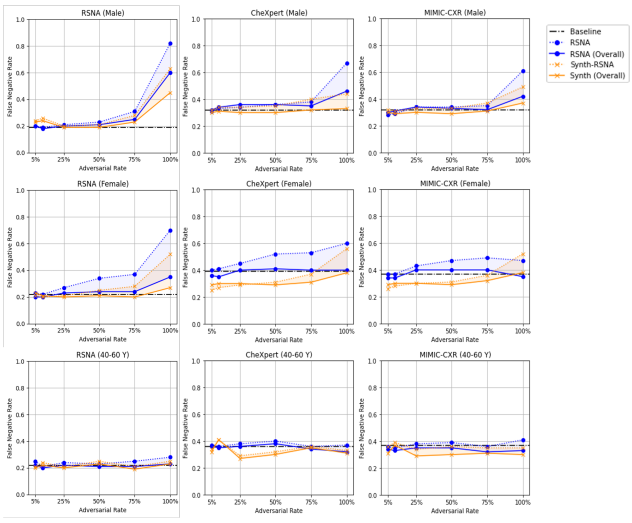


Figure 3. Impact of controlled bias injection on FNR. Models trained on the original (RNSA) and demographically targeted synthetic (Synth-RSNA) datasets are tested on the RSNA, CheXpert, and MIMIC-CXR test sets. The FNR for targeted demographic groups is compared to the overall model's FNR.

To quantify disparities in model performance, and thus, the effectiveness of bias mitigation, we use the vulnerability ν metric [13]. Briefly, ν measures the FNR disparity between a demographic group and the overall model. It is defined as the rate parameter β of logistic regression for the difference in metric of a group and the overall model with increasing rate of bias injected.

$$\mathcal{L}_{\text{MLE}}(\alpha, \beta) = \prod_{i=1}^{n} f(x_i)^{y_i}(1 - f(x_i))^{1-y_i} \tag{9}$$

where $x \triangleq r \in \mathbb{R}^n$ is the rate of underdiagnosis bias, $y \in \mathbb{R}^n$ is the FNR disparity, and $\alpha \in \mathbb{R}$ is the intercept, such that $y \sim f(x; \alpha, \beta)$ denotes the logistic function:

$$y \sim f(x; \alpha, \beta) = \frac{1}{1 + e^{-\alpha - \beta x}} \tag{10}$$

Vulnerability ν can be understood as the magnitude of FNR disparity, where a larger ν corresponds to greater disparity and vice versa. Moreover, a decrease in a group's ν after GCA indicates that the FNR disparity between the group and overall model performance decreased with GCA.

3. Results

3.1. Impact of GCA on Model Performance

We consistently observe that GCA has no impact on AUROC while significantly reducing FNR, as shown in Figs. 3 to 5. Fig. 4 shows that models trained with and without GCA achieved similar AUROC scores (GCA: $0.79 \pm$

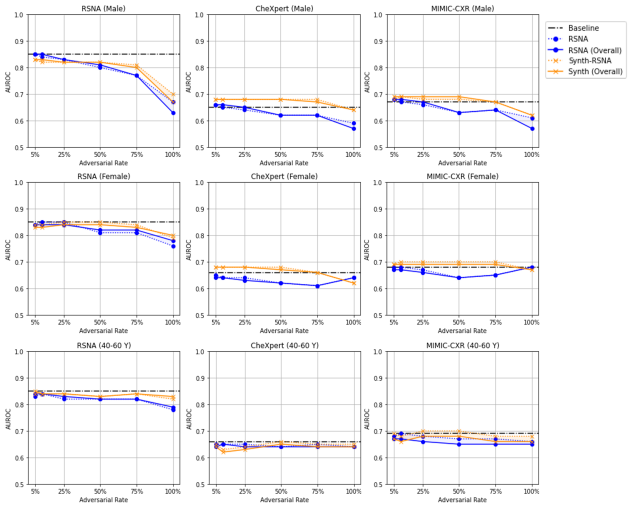


Figure 4. Impact of controlled bias injection on AUROC. Models trained on the original (RNSA) and demographically targeted synthetic (Synth-RSNA) datasets are tested on the RSNA, CheXpert, and MIMIC-CXR test sets. The AUROC for targeted demographic groups is compared to the overall model's AUROC.

0.04, non-GCA: 0.77 ± 0.02), indicating no performance degradation despite training on counterfactuals. Moreover, Fig. 3 shows models trained on the original (RSNA) and demographically targeted synthetic (Synth-RSNA) datasets with 100% underdiagnosis rate achieved a mean FNR of 0.48 ± 0.20 and 0.37 ± 0.15, respectively. This indicates that even when every positive patient in the targeted group was flipped to negative, the GCA-trained model achieved lower FNR than the baseline model.

When evaluating models trained with sex-group augmented Synth-RSNA-Sex dataset (Fig. 3), GCA significantly lowered FNR across all underdiagnosis rates ($r = 0-1$, $p < 0.01$). At high underdiagnosis rates ($r > 0.5$), GCA reduced FNR by 20% ($p < 0.001$), increasing to 32% at $r = 1$ ($p = 0.01$). The FNR reduction was greater in females (29%) than males (16%). The FNR disparity between males and females decreased by 18% ($p < 0.005$) at $r > 0.5$, with a 23% reduction at $r = 1$ ($p = 0.02$) Similarly, when evaluating models trained with age-group augmented Synth-RSNA-Age dataset (Fig. 3), GCA significantly reduced FNR across all underdiagnosis rates ($p < 0.01$), but the overall reduction was smaller (3% for $r > 0.5$, 6% for $r = 1$). The greatest decrease in FNR was observed in vulnerable groups (80+Y and 0-20Y), with FNR reductions of 26% and 24%, respectively ($p < 0.05$). The reduction in FNR disparities between 80+Y and 0-20Y groups was 12% and 13% for $r >= 0.5$ and $r = 1$, respectively.

When evaluating models trained with demographically complete Full-Synth-RSNA dataset (Fig. 5a), GCA significantly reduced FNR across all underdiagnosis rates ($p < 0.01$), but the overall reduction was smaller (3% for $r > 0.5$, 6% for $r = 1$). The greatest defense against bias was observed in vulnerable groups (80+ and 0-20 years), with FNR reductions of 26% and 24%, respectively ($p < 0.05$). Across both demographic groups, FNR increased with r for both GCA and non-GCA models. However, GCA's efficacy strengthened as r increased, with greater FNR reductions at higher underdiagnosis rates, confirming GCA's resilience against high amounts of bias in the training dataset (Figs. 3 and 5a).

3.2. Mitigation of Underdiagnosis Bias

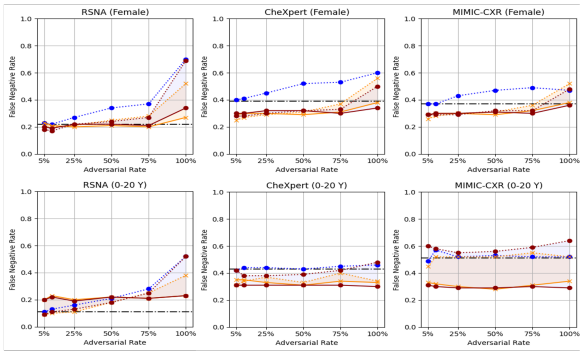
Our results show that models trained with GCA have lower FNR disparities between demographic groups, thus mitigating bias, as shown in Fig. 6. Focusing on the impact of controlled bias injection on targeted groups (diagonals), we observe that the difference $\Delta\nu_{M-F}$ decreases when GCA is applied (from 1.4 for original RSNA to -0.18 and 0.15 for Synth-RSNA-Sex and Full-Synth-RSNA, respectively). Similarly, we consistently observe ν decrease for all age-groups when models are trained with GCA. The 20-40Y group has the greatest decrease in FNR disparity (from $\nu = 3.37$ for original RSNA to $\nu = 1.86$ for both Synth-RSNA-Age and Full-Synth-RSNA), while the 0-20Y group has the least improvement in FNR disparity(from $\nu = 3.64$ for original RSNA to $\nu = 3.63$ and $\nu = 3.40$ for Synth-RSNA-Age and Full-Synth-RSNA, respectively).

3.3. Impact of GCA on Non-Targeted Groups

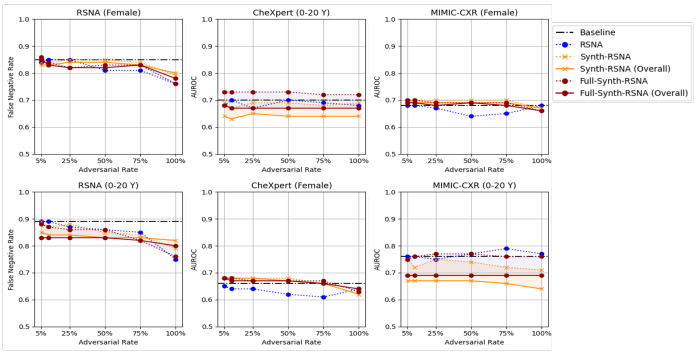
Our findings indicate that GCA does not negatively impact the performance of non-targeted demographic groups, as shown in Fig. 6. We observe similar vulnerability ν values for non-targeted groups, across the original RSNA, Synth-RSNA, and Full-Synth-RSNA datasets. Moreover, non-targeted groups that were adversely affected when a different group was targeted in the original RSNA dataset achieve lower ν when GCA is applied. For example, the 0-20Y is affected when 20-40Y group is targeted in original RSNA. When GCA is applied, ν for 0-20Y decreases considerably, from $\nu = 2.96$ to $\nu = 0.68$ ($\Delta = -2.28$) for both Synth-RSNA-Age and Full-Synth-RSNA.

3.4. Generalizability to External Datasets

We observe that the impact of GCA on bias mitigation remained consistent across the external CheXpert and MIMIC-CXR datasets, as shown in Figs. 3 and 5. Not only did AUROC remain similar (or even slightly improve) in GCA-trained models (Figs. 4 and 5b), but reductions in FNR disparities and vulnerability ν translated to both external datasets, as shown in Figs. 7 and 8. These findings highlight GCA's strong generalizability to external data.



(a) FNR



(b) AUROC

Figure 5. Impact of controlled bias injection for models trained on the original (RSNA), demographically targeted synthetic (Synth-RSNA), and demographically complete synthetic (Full-Synth-RSNA) datasets. Model performance across **(a)** FNR and **(b)** AUROC metrics for targeted demographic groups is compared to the overall model performance.

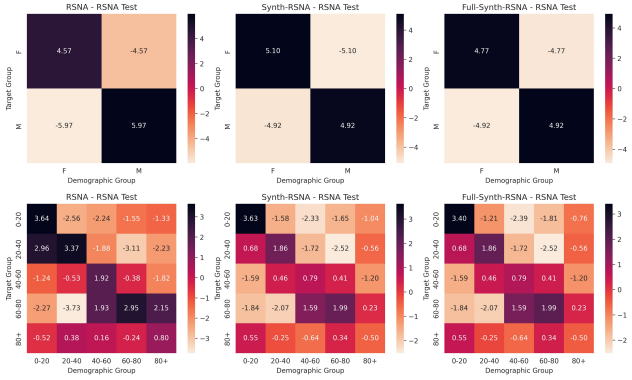


Figure 6. Vulnerability ν of the targeted and non-targeted demographic groups for models trained on the original (RSNA, column 1), demographically targeted synthetic (Synth-RSNA, column 2), and demographically complete synthetic (Full-Synth-RSNA, column 3) datasets and tested on RSNA test set.

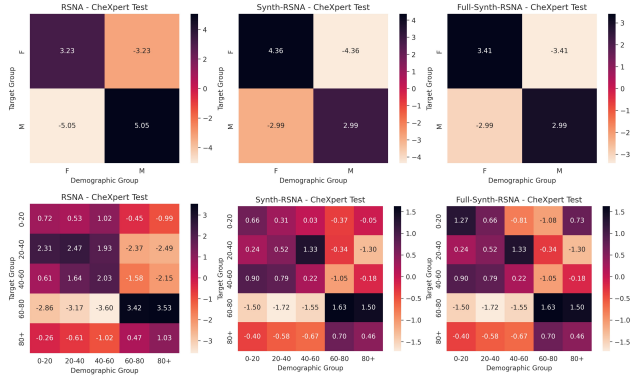


Figure 7. Vulnerability ν of the targeted and non-targeted demographic groups for models trained on the original (RSNA, column 1), demographically targeted synthetic (Synth-RSNA, column 2), and demographically complete synthetic (Full-Synth-RSNA, column 3) datasets and tested on CheXpert test set.

4. Discussion

Our results demonstrate that GCA is an effective strategy for mitigating bias in DL models for medical imaging. By synthesizing structured demographic counterfactuals, GCA augments the training dataset to achieve demographic-completeness, thereby increasing sample size by $5\times$ for Synth-RSNA and $10\times$ for Full-Synth-RSNA, respectively. This augmentation disentangles protected attributes (sex and age) from disease-related characteristics, enabling models trained with GCA to consistently yield lower mean FNRs across age- and sex-groups, even under varying rates of underdiagnosis bias. On average, our method reduced FNR disparities by 9% without introducing artifacts, indicating improved fairness and robustness. Furthermore, we showed that GCA generalizes well to external datasets, consistently achieving higher AUROC and lower FNR on the CheXpert and MIMIC-CXR datasets.

Notably, GCA does not negatively impact the performance of non-targeted demographic groups, despite generating counterfactuals across age and sex attributes. This is because GCA generates label-consistent counterfactuals (e.g., a male CXR augmented to appear female remains labeled as male), forcing the model to learn demographic-invariant features. These counterfactuals do not inject noise into the label space, but instead act as a regularizer, encouraging classifiers to focus on pathology-relevant features. As a result, GCA improves robustness on targeted groups without adversely affecting accuracy on non-targeted groups.

Since GCA requires generating a large volume of counterfactual images – specifically $5 \times N \times M$, where N is the number of augmentations and M is the dataset size, the computational efficiency of the generator G is a crucial consideration. We chose StyleGAN3 [11, 12] as our image syn-

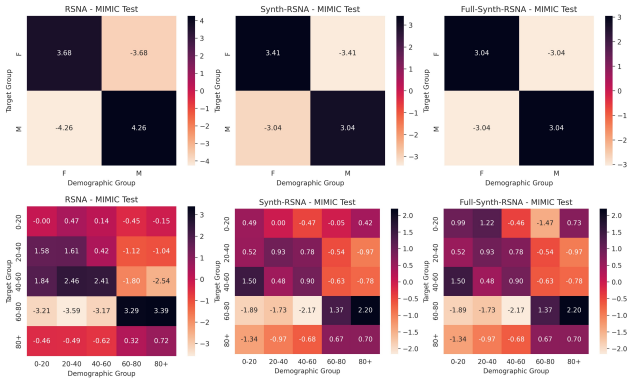


Figure 8. Vulnerability ν of the targeted and non-targeted demographic groups for models trained on the original (RSNA, column 1), demographically targeted synthetic (Synth-RSNA, column 2), and demographically complete synthetic (Full-Synth-RSNA, column 3) datasets and tested on MIMIC-CXR test set.

thesis network due to its favorable balance of image quality, semantic controllability, and low computational overhead. Unlike Hierarchical VAEs (HVAEs), such as Deep Structural Causal Models [18, 20, 21], which often produce blurry or low-resolution outputs, or diffusion models [19, 24], which are computationally expensive and require hundreds of sampling steps per image, StyleGAN3 enables scalable CXR image synthesis with a low one-time training cost with disentangled latent spaces. This allows for controlled demographic traversal for label-consistent counterfactual generation (e.g., changing age or sex while preserving pathology), which is a core aspect of GCA's strategy.

Our work has certain limitations. First, we primarily focus on pneumonia classification using CXRs, and further validation is needed to assess GCA's generalizability across other pathologies and imaging modalities (e.g., CT, MRI). Second, the RSNA dataset only includes age and sex demographic variables, limiting our ability to explore the impact on other protected characteristics like race, scanner, and image acquisition site. For future work, we will aim to extend GCA for broader clinical tasks and incorporate multi-attribute counterfactual generation.

In conclusion, GCA offers a scalable and effective framework for mitigating algorithmic bias in DL models through structured counterfactual generation. Our findings suggest that GCA not only improves model fairness and robustness but also has the potential to be adapted for other imaging modalities and tasks, ensuring trustworthiness in clinical settings.

References

[1] Rameen Abdal, Yipeng Qin, and Peter Wonka. Image2stylegan: How to embed images into the stylegan latent space? In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4432–4441, 2019. 2

[2] Matan Atad, Vitalii Dmytrenko, Yitong Li, Xinyue Zhang, Matthias Keicher, Jan Kirschke, Bene Wiestler, Ashkan Khakzar, and Nassir Navab. Chexplaining in style: Counterfactual explanations for chest x-rays using stylegan. *arXiv preprint arXiv:2207.07553*, 2022. 1

[3] Preetham Bachina, Sean P Garin, Pranav Kulkarni, Adway Kanhere, Jeremias Sulam, Vishwa S Parekh, and Paul H Yi. Coarse race and ethnicity labels mask granular underdiagnosis disparities in deep learning models for chest radiograph diagnosis. *Radiology*, 309(2):e231693, 2023. 1

[4] Elham Beheshtian, Kristin Putman, Samantha M Santomartino, Vishwa S Parekh, and Paul H Yi. Generalizability and bias in a deep learning pediatric bone age prediction model using hand radiographs. *Radiology*, 306(2):e220505, 2022. 1

[5] Mélanie Bernhardt, Charles Jones, and Ben Glocker. Potential sources of dataset bias complicate investigation of underdiagnosis by machine learning algorithms. *Nature Medicine*, 28(6):1157–1158, 2022. 1

[6] Judy Wawira Gichoya, Imon Banerjee, Ananth Reddy Bhimireddy, John L Burns, Leo Anthony Celi, Li-Ching Chen, Ramon Correa, Natalie Dullerud, Marzyeh Ghassemi, Shih-Cheng Huang, et al. Ai recognition of patient race in medical imaging: a modelling study. *The Lancet Digital Health*, 4(6):e406–e414, 2022. 1

[7] Judy Wawira Gichoya, Kaesha Thomas, Leo Anthony Celi, Nabile Safdar, Imon Banerjee, John D Banja, Laleh Seyyed-Kalantari, Hari Trivedi, and Saptarshi Purkayastha. Ai pitfalls and what not to do: mitigating bias in ai. *The British Journal of Radiology*, 96(1150):20230023, 2023. 1

[8] Ben Glocker, Charles Jones, Mélanie Roschewitz, and Stefan Winzeck. Risk of bias in chest radiography deep learning foundation models. *Radiology: Artificial Intelligence*, 5(6): e230060, 2023. 1, 3

[9] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI conference on artificial intelligence*, pages 590–597, 2019. 4

[10] Alistair EW Johnson, Tom J Pollard, Seth J Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Roger G Mark, and Steven Horng. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific data*, 6(1):317, 2019. 4

[11] Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Training generative adversarial networks with limited data. *Advances in neural information processing systems*, 33:12104–12114, 2020. 2, 6

[12] Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Alias-free generative adversarial networks. *Advances in neural information processing systems*, 34:852–863, 2021. 2, 6

[13] Pranav Kulkarni, Andrew Chan, Nithya Navarathna, Skylar Chan, Paul Yi, and Vishwa Sanjay Parekh. Hidden in plain

sight: Undetectable adversarial bias attacks on vulnerable patient populations. In *Medical Imaging with Deep Learning*, pages 793–821. PMLR, 2024. 1, 4

[14] David Li, Cheng Ting Lin, Jeremias Sulam, and Paul H Yi. Deep learning prediction of sex on chest radiographs: a potential contributor to biased algorithms. *Emergency Radiology*, 29(2):365–370, 2022. 1

[15] Hao Liang, Kevin Ni, and Guha Balakrishnan. Visualizing chest x-ray dataset biases using gans. *arXiv preprint arXiv:2305.00147*, 2023. 3

[16] William Lotter. Acquisition parameters influence ai recognition of race in chest x-rays and mitigating these factors reduces underdiagnosis bias. *Nature Communications*, 15(1): 7465, 2024. 3

[17] Pritam Mukherjee, Thomas C Shen, Jianfei Liu, Tejas Mathai, Omid Shafaat, and Ronald M Summers. Confounding factors need to be accounted for in assessing bias by machine learning algorithms. *Nature Medicine*, 28(6):1159–1160, 2022. 3

[18] Nick Pawlowski et al. Deep structural causal models for tractable counterfactual inference. *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. 7

[19] Walter H L Pinaya et al. Medical image synthesis using diffusion models: A comprehensive review. *Medical Image Analysis*, 2022. 7

[20] Fabio De Sousa Ribeiro, Tian Xia, Miguel Monteiro, Nick Pawlowski, and Ben Glocker. High fidelity image counterfactuals with probabilistic causal models. In *International Conference on Machine Learning*, pages 7390–7425. PMLR, 2023. 1, 7

[21] Mélanie Roschewitz, Fabio de Sousa Ribeiro, Tian Xia, Galvin Khara, and Ben Glocker. Counterfactual contrastive learning: robust representations via causal image synthesis. In *MICCAI Workshop on Data Engineering in Medical Imaging*, pages 22–32. Springer, 2024. 1, 7

[22] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115:211–252, 2015. 3

[23] Pedro Sanchez, Antanas Kascenas, Xiao Liu, Alison Q O'Neil, and Sotirios A Tsaftaris. What is healthy? generative counterfactual diffusion for lesion localization. In *MICCAI Workshop on Deep Generative Models*, pages 34–44. Springer, 2022. 1

[24] Patrick Schramowski et al. Diffusion-based generative modeling for counterfactual inference in imaging. *arXiv preprint arXiv:2205.09809*, 2022. 7

[25] Laleh Seyyed-Kalantari, Guanxiong Liu, Matthew McDermott, Irene Y Chen, and Marzyeh Ghassemi. Chexclusion: Fairness gaps in deep chest x-ray classifiers. In *BIOCOMPUTING 2021: proceedings of the Pacific symposium*, pages 232–243. World Scientific, 2020. 1

[26] Laleh Seyyed-Kalantari, Haoran Zhang, Matthew BA McDermott, Irene Y Chen, and Marzyeh Ghassemi. Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. *Nature medicine*, 27(12):2176–2182, 2021. 1, 4

[27] George Shih, Carol C Wu, Safwan S Halabi, Marc D Kohli, Luciano M Prevedello, Tessa S Cook, Arjun Sharma, Judith K Amorosa, Veronica Arteaga, Maya Galperin-Aizenberg, et al. Augmenting the national institutes of health chest radiograph dataset with expert annotations of possible pneumonia. *Radiology: Artificial Intelligence*, 1(1): e180041, 2019. 4

[28] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M Summers. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2097–2106, 2017. 4

[29] Tian Xia, Mélanie Roschewitz, Fabio De Sousa Ribeiro, Charles Jones, and Ben Glocker. Mitigating attribute amplification in counterfactual image generation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 546–556. Springer, 2024. 1

[30] Paul H Yi, Jinchi Wei, Tae Kyung Kim, Jiwon Shin, Haris I Sair, Ferdinand K Hui, Gregory D Hager, and Cheng Ting Lin. Radiology "forensics": determination of age and sex from chest radiographs using deep learning. *Emergency Radiology*, 28:949–954, 2021. 1

[31] Paul H Yi, Preetham Bachina, Beepul Bharti, Sean P Garin, Adway Kanhere, Pranav Kulkarni, David Li, Vishwa S Parekh, Samantha M Santomartino, Linda Moy, et al. Pitfalls and best practices in evaluation of ai algorithmic biases in radiology. *Radiology*, 315(2):e241674, 2025. 1