Evaluating the Impact of Medical Image Reconstruction on Downstream AI Fairness and Performance

Matteo Wohlrapp[1,2] MATTEO.WOHLRAPP@CDTM.DE
Niklas Bubeck[2,3] NIKLAS.BUBECK@TUM.DE
Daniel Rueckert[2,3,4] DANIEL.RUECKERT@TUM.DE
William Lotter[1,5] LOTTERB@DS.DFCI.HARVARD.EDU
[1]*Dana-Farber Cancer Institute, Boston MA, USA*
[2]*AI in Medicine, Technical University of Munich, Munich, Germany*
[3]*Munich Center for Machine Learning (MCML), Munich, Germany*
[4]*Department of Computing, Imperial College London, London, UK*
[5]*Harvard Medical School, Boston MA, USA*

Editors: Under Review for MIDL 2026

Abstract

AI-based image reconstruction models are increasingly deployed in clinical workflows to improve image quality from noisy data, such as low-dose X-rays or accelerated MRI scans. However, these models are typically evaluated using pixel-level metrics like PSNR, leaving their impact on downstream diagnostic performance and fairness unclear. We introduce a scalable evaluation framework that applies reconstruction and diagnostic AI models in tandem, which we apply to two tasks (classification, segmentation), three reconstruction approaches (U-Net, GAN, diffusion), and two data types (X-ray, MRI) to assess the potential downstream implications of reconstruction. We find that conventional reconstruction metrics poorly track task performance, where diagnostic accuracy remains largely stable even as reconstruction PSNR declines with increasing image noise. Fairness metrics exhibit greater variability, with reconstruction sometimes amplifying demographic biases, particularly regarding patient sex. However, the overall magnitude of this additional bias is modest compared to the inherent biases already present in diagnostic models. To explore potential bias mitigation, we adapt three strategies from classification literature to the reconstruction setting, but observe limited efficacy. Overall, our findings emphasize the importance of holistic performance and fairness assessments throughout the entire medical imaging workflow, especially as generative reconstruction models are increasingly deployed.

Keywords: Fairness, Image Reconstruction, GANs, Diffusion Models

1. Introduction

AI-based image reconstruction is an increasingly integral component of clinical workflows. These approaches are designed to enhance the quality of noisy medical images such as low-dose X-rays or faster-sampled MRIs, ultimately generating new medical images by imputing patterns learned from the training datasets (Ahishakiye et al., 2021). Notably, there are now over 80 FDA-cleared devices based on this approach (Singh et al., 2025), whose generated images are ultimately interpreted by clinicians.

Traditionally, reconstruction model performance has been evaluated using pixel-level image metrics such as PSNR. However, these metrics provide an incomplete picture, as they

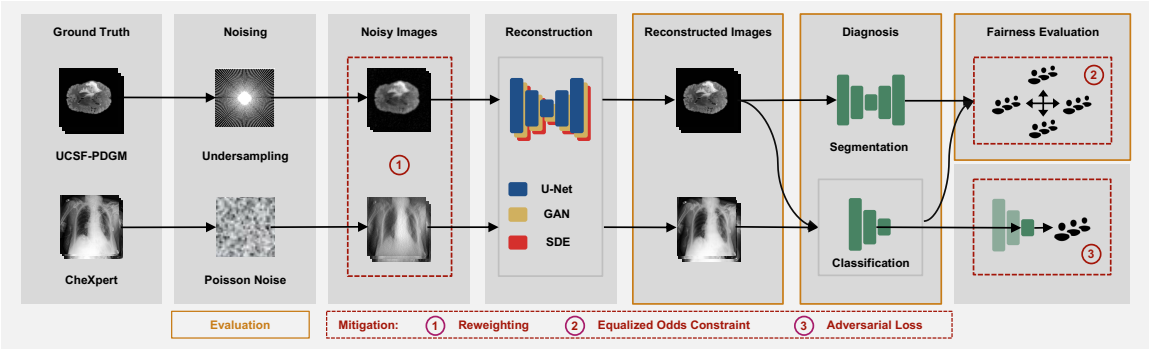


Figure 1: Combined pipeline for downstream bias evaluation and mitigation in medical image reconstruction. MRI and X-ray images undergo realistic simulated degradation and are subsequently reconstructed with three approaches before serving as input to downstream prediction models. Reconstruction quality, downstream performance, and fairness are evaluated. Subsequently, three bias mitigation strategies are applied exclusively during reconstruction fine-tuning.

do not reflect the impact of reconstructed images on subsequent clinical tasks. This gap raises a key unresolved question: *How does AI-based reconstruction influence downstream clinical performance and, in particular, fairness?* The latter is especially important to assess given the risk of generative models in encoding biases (Saumure et al., 2025; Ruggeri and Nozza, 2023; Luccioni et al., 2023; Mehrabi et al., 2021). While some smaller-scale studies have involved clinician review of AI-reconstructed images (Feuerriegel et al., 2023; Lee et al., 2024), this approach is not scalable, especially when investigating nuanced performance differences across subgroups.

In this work, we assess the downstream implications of AI-based reconstruction through an evaluation framework that leverages reconstruction and classification/segmentation AI models applied in tandem. The framework provides a scalable approach to understand how reconstruction errors propagate, while also simulating a realistic clinical scenario as both reconstruction and diagnostic models are increasingly deployed in medical workflows. We apply this framework across three reconstruction approaches (U-Net, GAN, diffusion), two imaging domains (MRI, X-ray), and two tasks (classification and segmentation). We additionally propose and evaluate bias mitigation techniques tailored to reconstruction models. Our findings highlight differences in trends between image metrics and diagnostic accuracy, and the potential of reconstruction models to shift demographic biases.

2. Related Work

Reconstruction Models in Medical Imaging: Medical image reconstruction is a popular AI application due to its promise in increasing image quality while facilitating lower radiation doses and faster scanning times (Ahishakiye et al., 2021). Given pairs of noisy (i.e., undersampled/lower dose) and original images, these models are trained to reconstruct the original from the noisy image. Variations of the U-Net (Ronneberger et al., 2015) are

commonly used as the neural network architecture. In addition to standard losses like mean-squared error (MSE), GAN and diffusion-based approaches are common in the field (Bousse et al., 2024; Heckel et al., 2024).

Fairness Analysis in Medical Imaging: Research on bias in AI-driven healthcare spans various medical domains, with medical imaging receiving considerable attention. In classification tasks, biases are typically revealed by comparing performance across subgroups. Studies cover various imaging modalities, including brain MRI (Stanley et al., 2022; Ioannou et al., 2022), chest X-rays (Seyyed-Kalantari et al., 2021; Glocker et al., 2023; Yang et al., 2024; Lotter, 2024), dermatology images (Chiu et al., 2024; Groh et al., 2021), and retinal images (Burlina et al., 2021). They address sensitive attributes such as sex (Stanley et al., 2022), age (Seyyed-Kalantari et al., 2021), race (Seyyed-Kalantari et al., 2021), and skin tone (Kinyanjui et al., 2020), evaluating disparities using performance metrics such as Area Under the Curve (AUC) (Seyyed-Kalantari et al., 2021), or more dedicated fairness criteria (Yuan et al., 2023). In segmentation, studies have assessed segmentation performance under varying demographic distributions, such as by race and sex representation in training datasets (Ioannou et al., 2022; Lee et al., 2022; Puyol-Antón et al., 2022).

Fairness Analysis of Reconstruction Models: Reconstruction model performance is typically measured using image quality metrics such as Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index Measure (SSIM). Recent studies assessing subgroup biases primarily rely on these metrics, examining how image quality varies across demographic subgroups. For instance, Du et al. (2023b) investigated fairness biases in deep learning-based brain MRI reconstruction, highlighting disparities in image reconstruction quality across different demographic groups using PSNR and SSIM. Similarly, Sheng et al. (2024) explored fairness challenges and potential solutions in ultrasound computed tomography, identifying significant disparities in reconstruction performance linked to subgroup attributes. With limited available literature, bias evaluation in reconstruction models is an emerging area of research for which there is a need to study the implications of image reconstruction on downstream tasks.

Bias Mitigation: In classification, substantial efforts have focused on developing bias mitigation strategies. Data-centric approaches directly modify training datasets, employing methods such as data redistribution (Oguguo et al., 2023), differentiable resampling techniques (Li and Vasconcelos, 2019), harmonization of datasets (Bissoto et al., 2019), and synthetic generation of diverse samples (Wang et al., 2024). Additionally, methods like Just Train Twice (JTT) target misclassified instances to implicitly mitigate subgroup biases without explicit annotations (Liu et al., 2021).

Representation-level strategies aim to learn unbiased feature representations through explicit disentanglement. Techniques include variational autoencoders (Creager et al., 2019), orthogonal disentanglement methods enforcing independence between sensitive attributes and task-specific features (Sarhan et al., 2020; Deng et al., 2023; Chiu et al., 2024; Du et al., 2023a), and group-adaptive architectures employing demographic-specific attention mechanisms (Gong et al., 2020).

Optimization-level methods integrate fairness constraints into model training via adversarial learning, fairness-specific loss functions, or specialized training regimens. Adversarial

methods discourage encoding protected attributes (Zhang et al., 2018; Adeli et al., 2019; Kim et al., 2019; Wang et al., 2020), distributionally robust optimization (Group DRO) targets worst-case subgroup performance (Sagawa et al., 2020), and fairness-specific constraints can be incorporated directly into training (Marcinkevics et al., 2022). Post-processing methods adjust model outputs after training, employing techniques such as calibration and pruning (Wu et al., 2022).

While prior studies have focused mainly on bias mitigation in classification tasks, there remains a critical need to assess analogous strategies for image reconstruction.

3. Methods

Our framework, visualized in Figure 1, encompasses image denoising, downstream task evaluation, fairness assessment, and bias mitigation for medical image reconstruction. The framework uses classification and segmentation models to estimate the effect of reconstruction on downstream task performance and fairness. Additionally, mitigation strategies are applied exclusively at the reconstruction stage to determine their ability to reduce downstream biases without retraining diagnostic models.

3.1. Datasets

We apply our framework to public datasets from two distinct imaging domains:

MRI: UCSF-PDGM includes 501 pre-operative glioma FLAIR volumes from patients with diffuse glioma, along with tumor masks and labels for subtype and grade (Calabrese et al., 2022).

X-Ray: CheXpert comprises 224,316 radiographs from 65,240 patients annotated for 14 thoracic findings (Irvin et al., 2019), of which we use 12 (excluding "Support Devices" and "No Findings" to focus on disease pathologies).

We use a 70/10/20 train/validation/test split stratified by patient for both datasets. For CheXpert, the training set is further divided into non-overlapping sets for reconstruction and classification model training, with percentages of 70/30, respectively. For UCSF-PDGM, the same training data is for both tasks given smaller sample size. Group-wise fairness is assessed for age (dichotomized at the dataset median), sex, and self-reported race (unavailable for UCSF-PDGM). Detailed attribute distributions are reported in Tables 5 and 6 in the Appendix.

3.2. Noising Process

We simulate realistic acquisition degradations as follows:

MRI: k-space data is masked with radial undersampling patterns (Feng, 2022) at acceleration factors 4, 8, and 16, where higher acceleration means greater undersampling.

X-Ray: Standard-dose images are Radon-projected to sinogram space, bow-tie filtered, and corrupted with Poisson noise parameterized by photon count (100,000, 10,000, 3,000), with lower photon count yielding more noise (Gibson et al., 2023).

These ranges approximate realistic acquisition conditions, with examples in the Appendix (Figures 6, 7, 8, and 9).

3.3. Models

We employ three reconstruction models alongside task-specific diagnostic models. Additional information on the compute infrastructure and model hyperparameters can be found in the Appendix.

Reconstruction: To cover deterministic, adversarial, and diffusion regimes, we train from scratch a standard U-Net (Ronneberger et al., 2015) with MSE loss, a Pix2Pix GAN (Isola et al., 2017), and a Stochastic Differential Equations (SDE)-based diffusion model (Luo et al., 2023) for each dataset. We note that the GAN and diffusion models also use a U-Net as the model architecture, but are based on a different training paradigm.

Diagnostic: For classification on UCSF-PDGM, an ImageNet-initialized ResNet50 (He et al., 2015) was trained separately to predict WHO grade and tumor type. The model is trained at the slice-level, and at testing, volume-level predictions are performed individually on each slice and then aggregated using the median. For CheXpert classification, a single ImageNet-initialized DenseNet model (Huang et al., 2017) was trained to jointly predict the 12 findings following Cohen et al. (2021). For segmentation on UCSF-PDGM, we use an ImageNet-initialized U-Net. Segmentation is not evaluated on CheXpert due to the absence of masks. All downstream models are trained on the original, non-degraded images.

3.4. Performance and Fairness Evaluation

Reconstruction quality is measured by PSNR. Downstream performance uses AUROC for classification and Dice for segmentation. For classification fairness, we report the worst case Equalized-Odds (EODD) (Hardt et al., 2016) difference between groups:

$$\begin{aligned} max_{i,j} | & P(\hat{Y} = 1 \mid Y = y, A = a_i) \\ & - P(\hat{Y} = 1 \mid Y = y, A = a_j)|, \quad \forall y \in \{0, 1\}, \\ & \forall \quad \text{attribute A} \in \mathcal{A}, \quad \text{subgroups} \quad a_i. \end{aligned}$$

To compute this metric, model predictions are binarized using a balanced threshold selected to achieve approximately equal sensitivity and specificity in the validation split. Equality of Opportunity (EOP) results are also reported in the Appendix (Figures 16 and 17).

For segmentation fairness, we adapt the Skewed-Error Ratio (SER) (Siddiqui et al., 2024) to Dice:

$$SER_A = \frac{\max_i(1 - \text{Dice}_{a_i})}{\min_j(1 - \text{Dice}_{a_j})}, \quad a_i \in A, \quad \text{A} \in \mathcal{A}$$

Results using an unnormalized Dice difference are also provided in the Appendix (Figures 16 and 17).

Statistical comparisons of subgroup fairness differences were performed using bootstrapped estimates with 1,000 iterations. Bootstrap-derived p-values determined statistical significance using a two-sided $p < 0.05$.

3.5. Bias Mitigation

We adapt three bias mitigation strategies that were originally developed for classification models. Each approach involves fine-tuning only the reconstruction models after the original training described above. Two of the strategies (differentiable equalized-odds and adversarial loss) rely on using the reconstruction and classification models applied in tandem, but the classification network is frozen to exclusively assess the potential for bias mitigation at the reconstruction stage.

Sample Reweighting: A weighted sampler draws each example with inverse joint subgroup frequency during fine-tuning, ensuring that each subgroup (and combination thereof across attributes) is represented with the same frequency. The reconstruction model is fine-tuned using the corresponding original reconstruction loss.

Differentiable Equalized-Odds: For reconstruction output $\hat{x} = f(x)$ and classifier output $\hat{y} = g(\hat{x})$ we minimize: $\mathcal{L}_{\text{EODD}} = \ell_{\text{rec}}(\hat{x}) + \lambda_{\text{fair}} \,\text{EMA}\big(\ell_{\text{BCE}}(\hat{y}) + \text{EODD}^2\big)$, where ℓ_{rec} represents the original reconstruction loss for the model, ℓ_{BCE} represents the binary cross-entropy loss for the frozen classifier, EMA represents an exponential moving average, and EODD represents a differentiable Equalized Odds constraint inspired by Marcinkevics et al. (2022). Specifically, we use the maximum EODD difference of any subgroup as defined above and compute it via soft predictions: $\tilde{y} = \sigma\big((\hat{y}) - \tau)/T\big)$, where the threshold τ and temperature T are set at 0.5 and 0.3, respectively. One loss is computed across all sensitive attributes (i.e., the max EODD over age, sex, and race). In the Appendix, we show that minimizing EODD^2 between subgroups corresponds to minimizing their covariance.

Adversarial Loss: Using the features z_i of the frozen classifier, we append an MLP classifier head h to predict sensitive attributes $(\hat{a}_{x_i}, \hat{b}_{x_i}, \dots) = h(z_i)$. We measure dependence on the sensitive attributes $(a_{x_i}, b_{x_i}, \dots)$ via squared Pearson correlation (Adeli et al., 2019): $\ell_{\text{fair}} = \text{Corr}^2((\hat{a}_{x_i}, \hat{b}_{x_i}, \dots), (a_{x_i}, b_{x_i}, \dots))$. The combined objective is: $\mathcal{L}_{\text{ADV}} = \ell_{\text{rec}} + \lambda_{\text{fair}} \,\text{EMA}\big(\ell_{\text{BCE}} + \ell_{\text{fair}}\big)$.. The weighting factor λ_{fair} is chosen by a one-dimensional log-scale sweep on the CheXpert U-Net baseline measured on the validation split (Appendix Figures 11–14).

4. Results

We first evaluate the impact of reconstruction on downstream task performance before analyzing fairness and the effectiveness of mitigation techniques.

4.1. Impact of Reconstruction on Task Performance

Figure 2 summarizes downstream performance as a function of reconstruction noise. We report segmentation Dice for UCSF-PDGM and the mean AUROC across the 12 CheXpert pathologies. For clarity, the y-axes for PSNR and the task metrics are normalized to the same percentage range. Across all experiments, diagnostic performance remains largely unchanged, even though PSNR decreases substantially with increasing noise. Specifically, the Dice score for UCSF-PDGM segmentation varies by no more than $\sim$3 % across noise conditions, and the mean CheXpert AUROC fluctuates by only 1 %. In contrast, PSNR decreases

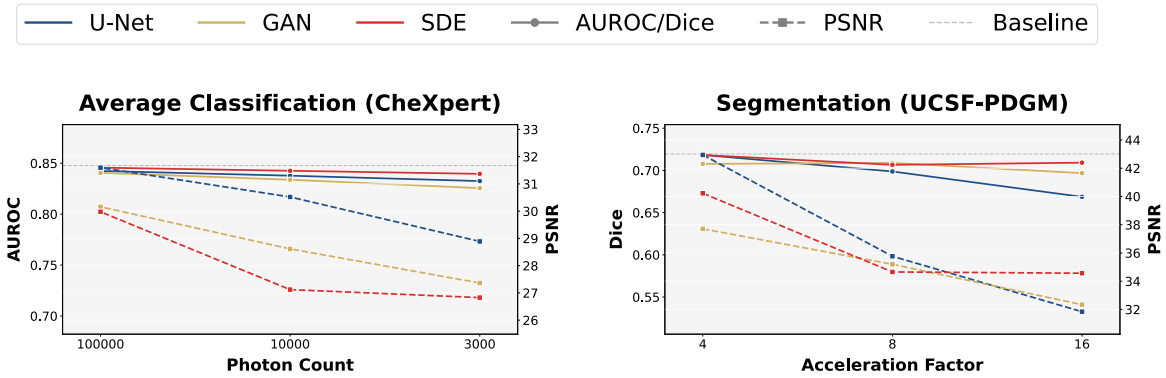


Figure 2: Downstream performance and PSNR at varying noise levels. Axes for PSNR and task performance are scaled to comparable percentage ranges. Although PSNR declines as noise increases, task performance remains stable. Baseline indicates performance on original images.

by over 10 dB (26 %) for UCSF-PDGM and by ∼3 dB (9 %) for CheXpert. Analogous results for UCSF-PDGM classification are presented in the Appendix (Figure 10), where the same pattern—substantial PSNR loss but minimal impact on task performance—holds for all three reconstruction models.

A closer look at CheXpert reveals a mild dependence on baseline task difficulty: pathologies with lower initial AUROC show slightly larger declines. For example, consolidation remains stable with U-Net reconstruction AUROC at 0.91, whereas lung lesion drops from 0.79 to 0.77 as noise increases (see Appendix Table 7 for more details).

4.2. Impact of Reconstruction on Fairness

Although aggregate task performance is largely unaffected, reconstruction could still alter relative performance across demographic subgroups. To test this possibility, we evaluated fairness on the downstream models using acceleration factor 8 for UCSF-PDGM and a photon count of 10,000 for CheXpert, representing the middle noise levels.

	Sex	**Age**	**Race**
Classification	0.05	0.17	0.19
Segmentation	1.13	1.24	–

Table 1: Average baseline fairness of the classifiers (EODD) and the segmentation model (SER) for different sensitive attributes. Sex exhibits the lowest baseline bias.

Fig. 3 displays the distribution of bias shifts when reconstructed images replace the original inputs. To provide a global overview, the histogram represents the bias shifts across all tasks, pathologies, and reconstruction models. As the diagnostic models exhibit bias on the original inputs (Table 1), the bias shifts with reconstruction are plotted on a

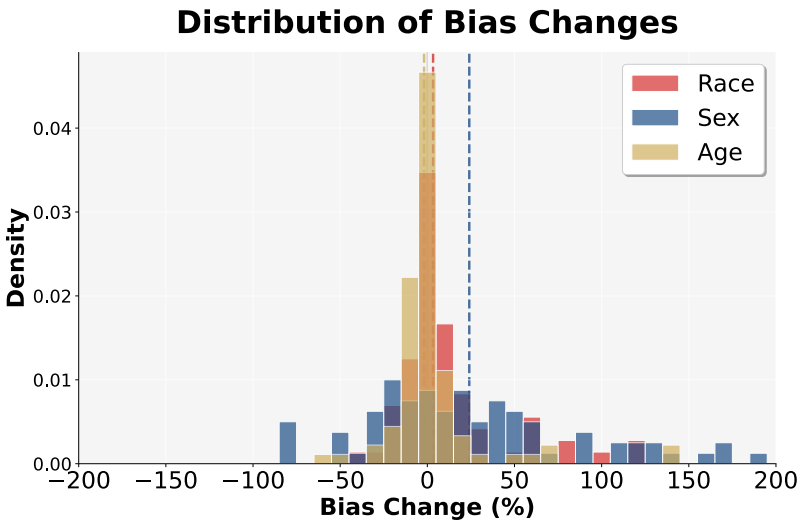


Figure 3: Distribution of bias changes (percent change compared to original images) across all reconstruction models, datasets, and tasks, stratified by sensitive attribute. The vertical lines mark the medians. Most shifts cluster near zero, but sex shows a broader positive tail.

percentage scale compared to the original bias to highlight the relative effects. We find that the mode of these shifts is centered around zero, indicating little bias change in most instances. However, there is a noticeable tail towards positive bias changes, especially for sex, which exhibits a median increase of 24%. This is partly attributable to sex having a lower baseline bias than age and race (Table 1).

The bias changes for each pathology and model are provided in Figures 4 and 5 (represented by the "Reconstruction" value in each plot). Segmentation (Figure 5) shows no significant fairness deviations when using the reconstructed images compared to the original images. UCSF-PDGM classification also exhibits non-significant variations. CheXpert shows more frequent significant shifts. Out of the 36 combinations (12 pathologies x 3 reconstruction models), there were 8 significant changes for sex (all in the positive direction) and 12 significant changes for age (4 in the positive direction). Due to large error bars, there were 0 significant changes for race, but alternative analysis which excluded subgroups with small sample sizes did reveal some significant changes (Appendix A). Overall, the pathology-level findings support the histogram trend with a slight bias increase for sex and a slight decrease for age. The absolute magnitude of the effects were generally modest; however, some are of the order of a 0.05 change in EODD, corresponding to a 5% difference in sensitivity/specificity, which would be meaningful at the population level. Across reconstruction methods, the GAN and SDE-based models revealed smaller bias shifts than the U-Net (Table 2).

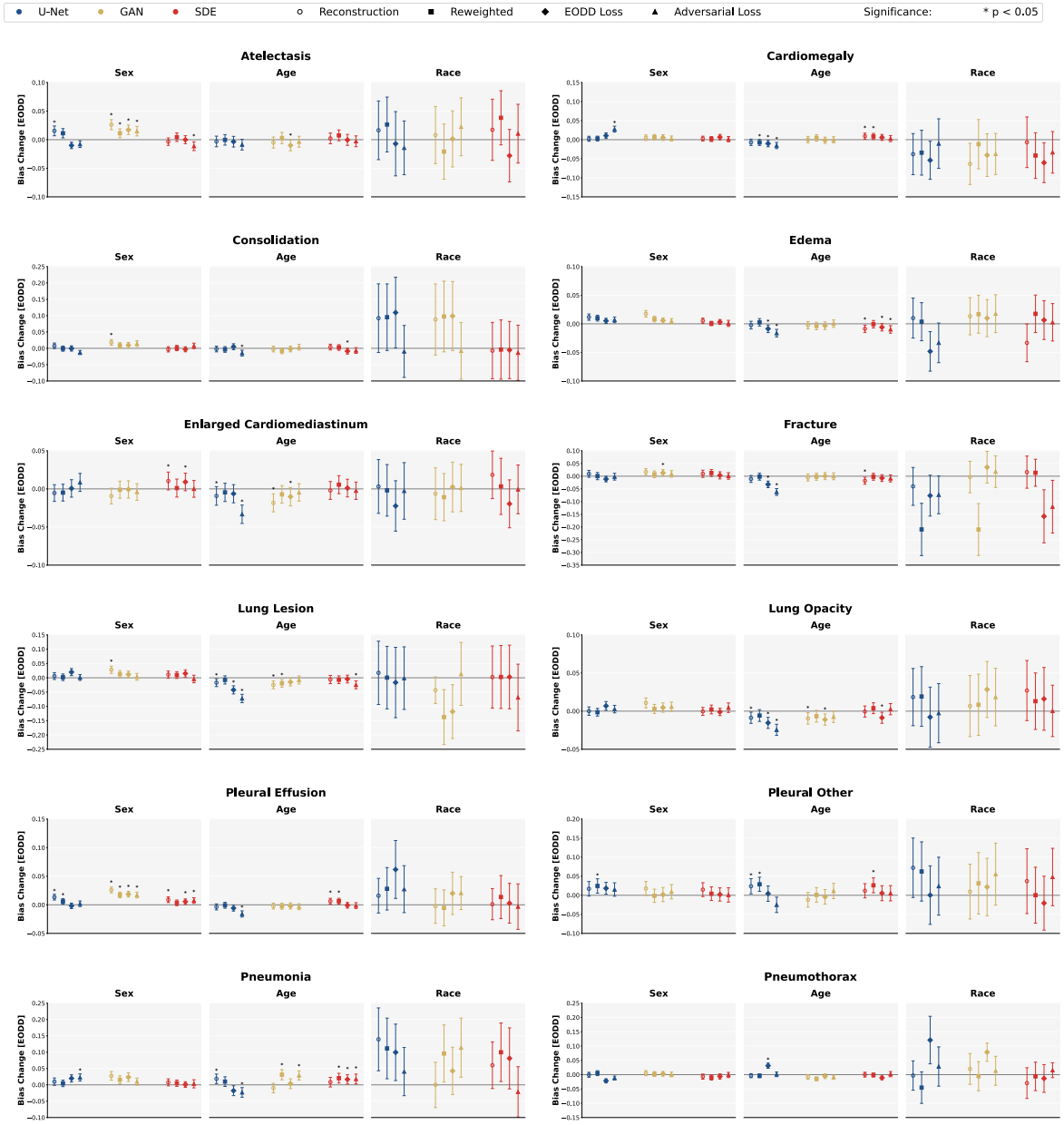


Figure 4: Equalized odds bias change pre- and post-mitigation compared to predictions on original images for CheXpert. Pre-mitigation ("Reconstruction"), bias tends to increase slightly for sex; race exhibits high variance. Post-mitigation—particularly with EODD and adversarial losses—bias declines slightly. Error bars represent standard deviation.

	U-Net	GAN	SDE
Median	2.28	-0.21	1.59
Absolute Median	14.6	11.8	11.5

Table 2: Median of bias change (% change in EODD/SER) by reconstruction approach across all datasets, tasks, and attributes by model. SDE and GAN show a smaller bias shift than U-Net.

4.3. Bias Mitigation

While the impact of reconstruction on fairness was generally modest, applying mitigation strategies at the reconstruction stage could still reduce these effects or even improve the fairness of the underlying diagnostic models. We therefore tested three mitigation techniques, all inspired by classification literature, but applied exclusively during reconstruction model training: sample reweighting, an equalized odds (EODD) constraint, and a subgroup-based adversarial loss.

	Sex	Age	Race
Standard	24.1	-1.88	3.30
Reweighted	10.6	0.03	1.05
EODD	7.56	-2.01	0.52
Adversarial	28.12	-5.39	-1.00

Table 3: Median bias change (% change in EODD/SER) by mitigation strategy across all datasets, tasks, and models. Standard corresponds to the original results without mitigation applied. The EODD constraint shows the greatest reduction for sex, while the Adversarial Loss shows the greatest reduction for age and race.

Table 3 summarizes the bias changes for the mitigated models compared to the standard models. The summary is presented as an aggregation over pathologies and reconstruction model types, with results for each combination presented in Figures 4 and 5. We observe a trend in decreased bias for each mitigation strategy and sensitive attribute, except for adversarial loss and patient sex, where the median bias change increases. EODD showed the largest median fairness improvement for sex, whereas adversarial loss showed the greatest improvement for age and race. Sex-related biases see the most substantial percentage improvements, notably for U-Net and SDE, and less for Pix2Pix (Figure 4). For UCSF-PDGM segmentation, EODD and the adversarial loss reduce bias for most attributes and models, most strongly for U-Net (Figure 5). Classification fairness on UCSF-PDGM exhibits no consistent pattern, with fluctuations in both directions. Overall, while some fairness improvements are observed, the magnitudes are modest compared to the original bias (e.g., a median effect of -8.67% for sex, -0.61% for age, and -2.78% for race across all results) and can depend on the pathology and sensitive attribute.

Fairness gains can incur performance trade-offs, but the trade-offs observed here are modest. Table 4 reports the mean change in PSNR and downstream task performance across

	Reweighted		EODD		Adversarial	
	Chex	**UCSF**	**Chex**	**UCSF**	**Chex**	**UCSF**
PSNR	0.54	-0.75	-0.64	-7.28	-1.22	-12.27
Down.	0.07	-1.97	0.02	-2.94	-0.34	-0.97

Table 4: Mean change (%) in PSNR and downstream performance (AUROC/Dice) per dataset after each mitigation averaged over reconstruction models and tasks. Performance drops are modest, except for PSNR in UCSF-PDGM.

reconstruction models when the mitigation strategies are applied. CheXpert deviations are below 2 % for PSNR and downstream AUROC. Downstream performance in UCSF-PDGM is also only moderately affected by the mitigation strategies, though PSNR shows larger drops with EODD and adversarial mitigation (see Figure 15 in the Appendix). Reweighting incurs the smallest penalties overall.

Additional results using EOP and ΔDice fairness metrics before and after mitigation are provided in the Appendix (Figures 16 and 17) and support the trends described above.

5. Discussion

We developed and applied an analysis framework that integrates reconstruction and prediction models to evaluate the effects of image reconstruction on downstream clinical tasks, quantify fairness implications, and investigate bias mitigation strategies at the reconstruction stage. Our analysis revealed several important insights, as summarized below.

Stability of Downstream Performance: Despite notable reductions in image quality, indicated by decreased PSNR at higher noise levels, downstream segmentation and classification performances remained robust to image reconstruction. This stability suggests that current diagnostic models are largely resilient to reconstruction-induced image degradations, which implies that minor reconstruction noise might not adversely impact clinical diagnostic accuracy. This finding may be surprising given that deep learning classification models are often thought to lack robustness, such as showing changes if the data are heterogeneous or noisy (Chuah et al., 2024). This suggests a nuanced interpretation of robustness, where models may be robust to certain transformations (e.g., reconstruction noise) but not others.

Fairness Implications and Variability: The aggregate effect of reconstruction on fairness was relatively modest, though certain pathologies and sensitive attributes showed significant shifts. These shifts varied in magnitude and direction, with a tendency toward increased bias, especially for patient sex. In most cases, the magnitude represented only a small fraction of the bias already present in the diagnostic models, though some would correspond to a ˜5% difference in sensitivity/specificity between subgroups. Thus, reconstruction can contribute to bias in downstream tasks, but the overall bias appears to be largely driven by the downstream models themselves.

Effectiveness and Dataset-Dependence of Mitigation Techniques: Mitigation strategies, particularly adversarial and Equalized Odds constraints, reduced age and sex biases

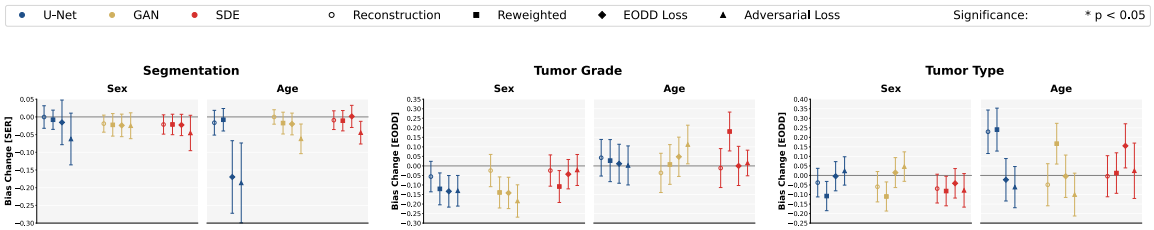


Figure 5: EODD and SER bias change pre- and post-mitigation compared to predictions on original images for UCSF-PDGM tasks. Segmentation shows a trend of a slight decrease in bias when mitigation strategies are applied, particularly for adversarial loss. No consistent trends emerge for the classification tasks. Error bars represent standard deviation.

on CheXpert without measurable performance trade-offs in AUROC or PSNR (Figure 15 in the Appendix). However, similar mitigation strategies yielded inconsistent results on UCSF-PDGM, highlighting that their effectiveness is dataset-specific and dependent on the underlying task complexity and dataset characteristics.

Sensitivity of Model Choice: The SDE and GAN-based reconstruction approaches introduced lower additional bias overall compared to the standard U-Net, which may be counterintuitive given the generative nature of the SDE and GAN models. The U-Net also exhibited larger degradations in downstream performance when fine-tuned with the fairness mitigation strategies(Figure 15 in the Appendix). This sensitivity likely arises from its inherently lower capacity than other methods, limiting simultaneous optimization of image fidelity and fairness constraints.

Summary of Clinical Implications: The robustness of downstream performance to AI-based image reconstruction is encouraging, particularly as these technologies are increasingly integrated into clinical practice. However, some performance drops were observed, especially for more subtle pathologies (e.g., lung lesion), highlighting the importance of rigorous evaluation and real-world monitoring. The potential for fairness shifts also necessitates active monitoring and reporting. This is especially important because model behavior can change as data distributions shift.

Summary of Model Development Implications: Developers of reconstruction models should prioritize downstream task and fairness evaluations alongside traditional pixel-level metrics, recognizing that reconstruction-induced biases, though subtle, can propagate through diagnostic workflows. This is especially the case for patient sex, where anatomical differences can be more prominent and may explain the larger effects observed for this attribute in our results. Bias mitigation strategies applied at the reconstruction stage may help improve fairness, but our results suggest that direct intervention at the classifier stage should be prioritized. Future research should explore multi-stage bias mitigation, integrating reconstruction and classification levels to achieve balanced fairness and performance outcomes.

Limitations: For comprehensiveness, we assessed multiple reconstruction models, downstream tasks, pathologies, and mitigation strategies, but this breadth necessarily creates challenges in data interpretation. As such, we have provided both summary level (e.g., Figure 3) and individual (e.g., Figure 4) results to enhance interpretability. Along with our studied datasets and tasks, it will be important in future work to apply our framework to additional datasets and clinical populations, further probing generalization. Additionally, while the algorithms used to create noisy images in this study simulate realistic acquisition degradations and are common approaches in the field(Feng, 2022; Gibson et al., 2023), they may not fully capture real-world variations.

6. Conclusion

The increasing clinical prevalence of AI-based reconstruction models creates a critical need for quantitative assessments of their potential downstream impact. We performed a scalable evaluation by using reconstruction and diagnostic AI models in tandem across multiple datasets, tasks, pathologies, and model types. We view our results as largely positive for the field – downstream performance was much more robust than image-level metrics to reconstruction noise, and the biases introduced by reconstruction were generally modest. However, some trends of increased bias were observed, especially for patient sex. Altogether, supported by these findings, we argue for the importance of monitoring downstream performance and fairness when using AI-based reconstruction models, and for continued work to mitigate any emerging biases.

References

Ehsan Adeli, Qingyu Zhao, Adolf Pfefferbaum, Edith V. Sullivan, Li Fei-Fei, Juan Carlos Niebles, and Kilian M. Pohl. Representation learning with statistical independence to mitigate bias. *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 2512–2522, 2019. URL https://api.semanticscholar.org/CorpusID:211069024.

Emmanuel Ahishakiye, Martin Bastiaan Van Gijzen, Julius Tumwiine, Ruth Wario, and Johnes Obungoloch. A survey on deep learning in medical image reconstruction. *Intelligent Medicine*, 1(3):118–127, 2021. ISSN 2667-1026. doi: https://doi.org/10.1016/j.imed.2021.03.003. URL https://www.sciencedirect.com/science/article/pii/S2667102621000061.

Alceu Bissoto, Michel Fornaciali, Eduardo Valle, and Sandra Avila. (De) Constructing Bias on Skin Lesion Datasets . In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 2766–2774, Los Alamitos, CA, USA, June 2019. IEEE Computer Society. doi: 10.1109/CVPRW.2019.00335. URL https://doi.ieeecomputersociety.org/10.1109/CVPRW.2019.00335.

Alexandre Bousse, Venkata Sai Sundar Kandarpa, Kuangyu Shi, Kuang Gong, Jae Sung Lee, Chi Liu, and Dimitris Visvikis. A review on low-dose emission tomography post-reconstruction denoising with neural network approaches. *IEEE Transactions on Radiation and Plasma Medical Sciences*, 2024. doi: 10.1109/TRPMS.2023.3349194. URL https://arxiv.org/abs/2401.00232.

Philippe Burlina, Neil Joshi, William Paul, Katia Pacheco, and Neil Bressler. Addressing artificial intelligence bias in retinal diagnostics. *Translational Vision Science Technology*, 10:13, 02 2021. doi: 10.1167/tvst.10.2.13.

Evan Calabrese, Javier E. Villanueva-Meyer, Jeffrey D. Rudie, Andreas M. Rauschecker, Ujjwal Baid, Spyridon Bakas, Soonmee Cha, John T. Mongan, and Christopher P. Hess. The university of california san francisco preoperative diffuse glioma mri dataset. *Radiology: Artificial Intelligence*, 4(6), November 2022. ISSN 2638-6100. doi: 10.1148/ryai.220058. URL http://dx.doi.org/10.1148/ryai.220058.

Ching-Hao Chiu, Yu-Jen Chen, Yawen Wu, Yiyu Shi, and Tsung-Yi Ho. Achieve fairness without demographics for dermatological disease diagnosis. *Medical Image Analysis*, 95: 103188, 2024. ISSN 1361-8415. doi: https://doi.org/10.1016/j.media.2024.103188. URL https://www.sciencedirect.com/science/article/pii/S1361841524001130.

Joshua Chuah, Pingkun Yan, Ge Wang, and Juergen Hahn. Towards the generation of medical imaging classifiers robust to common perturbations. *BioMedInformatics*, 4(2): 889–910, 2024. ISSN 2673-7426. doi: 10.3390/biomedinformatics4020050. URL https://www.mdpi.com/2673-7426/4/2/50.

Joseph Paul Cohen, Joseph D. Viviano, Paul Bertin, Paul Morrison, Parsa Torabian, Matteo Guarrera, Matthew P. Lungren, Akshay Chaudhari, Rupert Brooks, Mohammad Hashir,

and Hadrien Bertrand. Torchxrayvision: A library of chest x-ray datasets and models. In *International Conference on Medical Imaging with Deep Learning*, 2021. URL https://api.semanticscholar.org/CorpusID:240353861.

Elliot Creager, David Madras, Jörn-Henrik Jacobsen, Marissa A. Weis, Kevin Swersky, Toniann Pitassi, and Richard S. Zemel. Flexibly fair representation learning by disentanglement. In *International Conference on Machine Learning*, 2019. URL https://api.semanticscholar.org/CorpusID:174800294.

Wenlong Deng, Yuan Zhong, Qi Dou, and Xiaoxiao Li. On fairness of medical image classification with multiple sensitive attributes via learning orthogonal representations. In *Information Processing in Medical Imaging: 28th International Conference, IPMI 2023, San Carlos de Bariloche, Argentina, June 18–23, 2023, Proceedings*, page 158–169, Berlin, Heidelberg, 2023. Springer-Verlag. ISBN 978-3-031-34047-5. doi: 10.1007/978-3-031-34048-2_13. URL https://doi.org/10.1007/978-3-031-34048-2_13.

Siyi Du, Ben Hers, Nourhan Bayasi, Ghassan Hamarneh, and Rafeef Garbi. Fairdisco: Fairer ai in dermatology via disentanglement contrastive learning. In Leonid Karlinsky, Tomer Michaeli, and Ko Nishino, editors, *Computer Vision – ECCV 2022 Workshops*, pages 185–202, Cham, 2023a. Springer Nature Switzerland. ISBN 978-3-031-25069-9.

Yuning Du, Yuyang Xue, Rohan Dharmakumar, and Sotirios A. Tsaftaris. Unveiling fairness biases in deep learning-based brain mri reconstruction. In *Clinical Image-Based Procedures, Fairness of AI in Medical Imaging, and Ethical and Philosophical Issues in Medical Imaging: 12th International Workshop, CLIP 2023 1st International Workshop, FAIMI 2023 and 2nd International Workshop, EPIMI 2023 Vancouver, BC, Canada, October 8 and October 12, 2023 Proceedings*, page 102–111, Berlin, Heidelberg, 2023b. Springer-Verlag. ISBN 978-3-031-45248-2. doi: 10.1007/978-3-031-45249-9_10. URL https://doi.org/10.1007/978-3-031-45249-9_10.

Li Feng. Golden-angle radial mri: Basics, advances, and applications. *Journal of Magnetic Resonance Imaging*, 56, 04 2022. doi: 10.1002/jmri.28187.

Georg C Feuerriegel, Kilian Weiss, Sophia Kronthaler, Yannik Leonhardt, Jan Neumann, Markus Wurm, Nicolas S Lenhart, Marcus R Makowski, Benedikt J Schwaiger, Klaus Woertler, Dimitrios C Karampinos, and Alexandra S Gersing. Evaluation of a deep learning-based reconstruction method for denoising and image enhancement of shoulder MRI in patients with shoulder pain. *Eur. Radiol.*, 33(7):4875–4884, July 2023.

Nicholas Mark Gibson, Amy Lee, and Martin Bencsik. A practical method to simulate realistic reduced-exposure ct images by the addition of computationally generated noise. *Radiological physics and technology*, 2023. URL https://api.semanticscholar.org/CorpusID:265148810.

Ben Glocker, Charles Jones, Mélanie Bernhardt, and Stefan Winzeck. Algorithmic encoding of protected characteristics in chest x-ray disease detection models. *eBioMedicine*, 89, 2023. URL https://api.semanticscholar.org/CorpusID:256858498.

Sixue Gong, Xiaoming Liu, and Anil K. Jain. Mitigating face recognition bias via group adaptive classifier. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3413–3423, 2020. URL https://api.semanticscholar.org/CorpusID:219687431.

Matthew Groh, Caleb Harris, Luis Soenksen, Felix Lau, Rachel Han, Aerin Kim, Arash Koochek, and Omar Badri. Evaluating deep neural networks trained on clinical images in dermatology with the fitzpatrick 17k dataset. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1820–1828, 2021.

Moritz Hardt, Eric Price, and Nathan Srebro. Equality of opportunity in supervised learning. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, page 3323–3331, Red Hook, NY, USA, 2016. Curran Associates Inc. ISBN 9781510838819.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *arXiv*, abs/1512.03385, 2015. URL http://arxiv.org/abs/1512.03385.

Reinhard Heckel, Mathews Jacob, Akshay Chaudhari, Or Perlman, and Efrat Shimron. Deep learning for accelerated and robust mri reconstruction. *Magnetic Resonance Materials in Physics, Biology and Medicine*, 37(3):335–368, Jul 2024. ISSN 1352-8661. doi: 10.1007/s10334-024-01173-8. URL https://doi.org/10.1007/s10334-024-01173-8.

Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

Stefanos Ioannou, Hana Chockler, Alexander Hammers, and Andrew P. King. A study of demographic bias in cnn-based brain mr segmentation. In *Machine Learning in Clinical Neuroimaging: 5th International Workshop, MLCN 2022, Held in Conjunction with MICCAI 2022, Singapore, September 18, 2022, Proceedings*, page 13–22, Berlin, Heidelberg, 2022. Springer-Verlag. ISBN 978-3-031-17898-6. doi: 10.1007/978-3-031-17899-3_2. URL https://doi.org/10.1007/978-3-031-17899-3_2.

Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, Jayne Seekins, David A. Mong, Safwan S. Halabi, Jesse K. Sandberg, Ricky Jones, David B. Larson, Curtis P. Langlotz, Bhavik N. Patel, Matthew P. Lungren, and Andrew Y. Ng. Chexpert: a large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'19/IAAI'19/EAAI'19. AAAI Press, 2019. ISBN 978-1-57735-809-1. doi: 10.1609/aaai.v33i01.3301590. URL https://doi.org/10.1609/aaai.v33i01.3301590.

Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. *CVPR*, 2017.

Byungju Kim, Hyunwoo Kim, Kyungsu Kim, Sungjin Kim, and Junmo Kim. Learning not to learn: Training deep neural networks with biased data. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014. URL https://api.semanticscholar.org/CorpusID:6628106.

Newton M. Kinyanjui, Timothy Odonga, Celia Cintas, Noel C. F. Codella, Rameswar Panda, Prasanna Sattigeri, and Kush R. Varshney. Fairness of classifiers across skin tones in dermatology. In Anne L. Martel, Purang Abolmaesumi, Danail Stoyanov, Diana Mateus, Maria A. Zuluaga, S. Kevin Zhou, Daniel Racoceanu, and Leo Joskowicz, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020*, pages 320–329, Cham, 2020. Springer International Publishing. ISBN 978-3-030-59725-2.

Dong Ho Lee, Jeong Min Lee, Chang Hee Lee, Saif Afat, and Ahmed Othman. Image quality and diagnostic performance of low-dose liver CT with deep learning reconstruction versus standard-dose CT. *Radiol. Artif. Intell.*, 6(2):e230192, March 2024.

Tiarna Lee, Esther Puyol-Antón, Bram Ruijsink, Miaojing Shi, and Andrew P. King. A systematic study of race and sex bias in cnn-based cardiac mr segmentation. In *Statistical Atlases and Computational Models of the Heart. Regular and CMRxMotion Challenge Papers: 13th International Workshop, STACOM 2022, Held in Conjunction with MICCAI 2022, Singapore, September 18, 2022, Revised Selected Papers*, page 233–244, Berlin, Heidelberg, 2022. Springer-Verlag. ISBN 978-3-031-23442-2. doi: 10.1007/978-3-031-23443-9_22. URL https://doi.org/10.1007/978-3-031-23443-9_22.

Yi Li and Nuno Vasconcelos. Repair: Removing representation bias by dataset resampling. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9564–9573, 2019. doi: 10.1109/CVPR.2019.00980.

Evan Z Liu, Behzad Haghgoo, Annie S Chen, Aditi Raghunathan, Pang Wei Koh, Shiori Sagawa, Percy Liang, and Chelsea Finn. Just train twice: Improving group robustness without training group information. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 6781–6792. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/liu21f.html.

William Lotter. Acquisition parameters influence AI recognition of race in chest x-rays and mitigating these factors reduces underdiagnosis bias. *Nat. Commun.*, 15(1):7465, August 2024.

Alexandra Sasha Luccioni, Christopher Akiki, Margaret Mitchell, and Yacine Jernite. Stable bias: evaluating societal representations in diffusion models. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, 2023. Curran Associates Inc.

Ziwei Luo, Fredrik K. Gustafsson, Zheng Zhao, Jens Sjölund, and Thomas B. Schön. Image restoration with mean-reverting stochastic differential equations. In Andreas Krause,

Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 23045–23066. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/luo23b.html.

Ricards Marcinkevics, Ece Ozkan, and Julia E. Vogt. Debiasing deep chest X-ray classifiers using intra- and post-processing methods. In Zachary Lipton, Rajesh Ranganath, Mark Sendak, Michael Sjoding, and Serena Yeung, editors, *Proceedings of the 7th Machine Learning for Healthcare Conference*, volume 182 of *Proceedings of Machine Learning Research*, pages 504–536. PMLR, 05–06 Aug 2022. URL https://proceedings.mlr.press/v182/marcinkevics22a.html.

Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM Comput. Surv.*, 54(6), July 2021. ISSN 0360-0300. doi: 10.1145/3457607. URL https://doi.org/10.1145/3457607.

Tochi Oguguo, Ghada Zamzmi, Sivaramakrishnan Rajaraman, Feng Yang, Zhiyun Xue, and Sameer Antani. A comparative study of fairness in medical machine learning. In *2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI)*, pages 1–5, 2023. doi: 10.1109/ISBI53787.2023.10230368.

E Puyol-Antón, B Ruijsink, J Mariscal Harana, SK Piechnik, S Neubauer, SE Petersen, R Razavi, P Chowienczyk, and AP King. Fairness in cardiac magnetic resonance imaging: Assessing sex and racial bias in deep learning-based segmentation. *Frontiers in Cardiovascular Medicine*, 9:859310, Apr 2022. doi: 10.3389/fcvm.2022.859310.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In Nassir Navab, Joachim Hornegger, William M. Wells, and Alejandro F. Frangi, editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pages 234–241, Cham, 2015. Springer International Publishing. ISBN 978-3-319-24574-4.

Gabriele Ruggeri and Debora Nozza. A multi-dimensional study on bias in vision-language models. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6445–6455, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.403. URL https://aclanthology.org/2023.findings-acl.403/.

Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. In *International Conference on Learning Representations (ICLR)*, 2020.

Mhd Hasan Sarhan, Nassir Navab, Abouzar Eslami, and Shadi Albarqouni. Fairness by learning orthogonal disentangled representations. In *Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIX*, page 746–761, Berlin, Heidelberg, 2020. Springer-Verlag. ISBN 978-3-030-58525-9. doi: 10.1007/978-3-030-58526-6_44. URL https://doi.org/10.1007/978-3-030-58526-6_44.

R. Saumure, J. De Freitas, and S. Puntoni. Humor as a window into generative ai bias. *Scientific Reports*, 15(1):1326, 2025. doi: 10.1038/s41598-024-83384-6. URL https://doi.org/10.1038/s41598-024-83384-6.

Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 618–626, 2017. doi: 10.1109/ICCV.2017.74.

Laleh Seyyed-Kalantari, Haoran Zhang, Matthew McDermott, Irene Chen, and Marzyeh Ghassemi. Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. *Nature Medicine*, 27, 12 2021. doi: 10.1038/s41591-021-01595-0.

Yi Sheng, Junhuan Yang, Youzuo Lin, Weiwen Jiang, and Lei Yang. Toward fair ultrasound computing tomography: Challenges, solutions and outlook. In *Proceedings of the Great Lakes Symposium on VLSI 2024*, GLSVLSI '24, page 748–753, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400706059. doi: 10.1145/3649476.3660387. URL https://doi.org/10.1145/3649476.3660387.

Ismaeel Siddiqui, Nickolas Littlefield, Luke Carlson, Matthew Gong, Avani Chhabra, Zoe Menezes, George Mastorakos, Sakshi Thakar, Mehrnaz Abedian, Ines Lohse, Kurt Weiss, Johannes Plate, Hamidreza Moradi, Soheyla Amirian, and Ahmad P. Tafti. Fair ai-powered orthopedic image segmentation: addressing bias and promoting equitable healthcare. *Scientific Reports*, 14, 07 2024. doi: 10.1038/s41598-024-66873-6.

Rohan Singh, Monika Bapna, Abdul Rahman Diab, Emily S. Ruiz, and William Lotter. How ai is used in fda-authorized medical devices: a taxonomy across 1,016 authorizations. *npj Digital Medicine*, 8(1):388, Jul 2025. ISSN 2398-6352. doi: 10.1038/s41746-025-01800-1. URL https://doi.org/10.1038/s41746-025-01800-1.

Emma A. M. Stanley, Matthias Wilms, Pauline Mouches, and Nils Daniel Forkert. Fairness-related performance and explainability effects in deep learning models for brain image analysis. *Journal of Medical Imaging*, 9:061102 – 061102, 2022. URL https://api.semanticscholar.org/CorpusID:251876386.

Ryan Wang, Po-Chih Kuo, Li-Ching Chen, Kenneth Patrick Seastedt, Judy Wawira Gichoya, and Leo Anthony Celi. Drop the shortcuts: image augmentation improves fairness and decreases ai detection of race and other demographics from medical images. *eBioMedicine*, 102:105047, 2024. ISSN 2352-3964. doi: https://doi.org/10.1016/j.ebiom.2024.105047. URL https://www.sciencedirect.com/science/article/pii/S2352396424000823.

Zeyu Wang, Klint Qinami, Ioannis Christos Karakozis, Kyle Genova, Prem Nair, Kenji Hata, and Olga Russakovsky. Towards Fairness in Visual Recognition: Effective Strategies for Bias Mitigation . In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8916–8925, Los Alamitos, CA, USA, June 2020. IEEE Computer Society. doi: 10.1109/CVPR42600.2020.00894. URL https://doi.ieeecomputersociety.org/10.1109/CVPR42600.2020.00894.

Yawen Wu, Dewen Zeng, Xiaowei Xu, Yiyu Shi, and Jingtong Hu. Fairprune: Achieving fairness through pruning for dermatological disease diagnosis. In Linwei Wang, Qi Dou, P. Thomas Fletcher, Stefanie Speidel, and Shuo Li, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2022*, pages 743–753, Cham, 2022. Springer Nature Switzerland. ISBN 978-3-031-16431-6.

Yuzhe Yang, Haoran Zhang, Judy W Gichoya, Dina Katabi, and Marzyeh Ghassemi. The limits of fair medical imaging AI in real-world generalization. *Nat. Med.*, 30(10):2838–2848, October 2024.

Chenxi Yuan, Kristin A. Linn, and Rebecca A. Hubbard. Algorithmic fairness of machine learning models for alzheimer disease progression. *JAMA Network Open*, 6(11):e2342203–e2342203, 11 2023. ISSN 2574-3805. doi: 10.1001/jamanetworkopen.2023.42203. URL https://doi.org/10.1001/jamanetworkopen.2023.42203.

Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '18, page 335–340, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450360128. doi: 10.1145/3278721.3278779. URL https://doi.org/10.1145/3278721.3278779.

Appendix A.

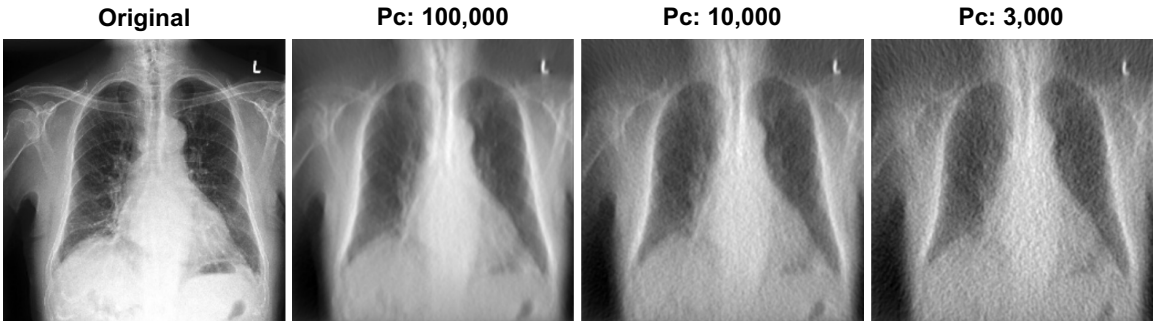


Figure 6: X-Ray images with photon count 100,000, 10,000, 3,000.

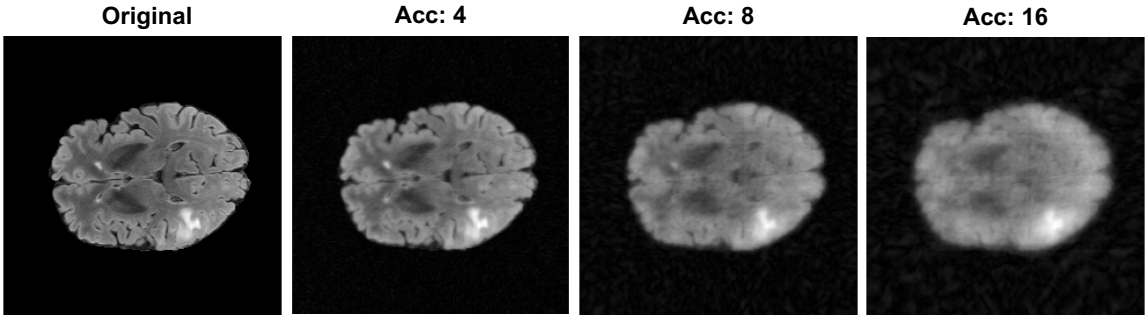


Figure 7: MRI images with acceleration 4,8,16.

Diagnostic Hyperparameters. The segmentation network was optimized with Adam (Kingma and Ba, 2014) using a learning rate of 0.001 and a batch size of 8 without data augmentation for 20 training epochs. The training loss consisted of Dice and L1, equally weighted at 0.5 each. The network used a sigmoid activation and a threshold of 0.5 was used at inference to compute the Dice performance. The model was trained on a per-slice level using all available MRI slices. At inference, Dice performance was computed using slices 60-130, as this range is representative of the regions where the ground truth masks appear and thus is more representative of performance. The Dice scores were computed separately for each slice, then averaged across slices per patient, followed by averaging across patients to compute final performance. For the UCSF-PDGM ResNet classifiers, we trained for 20 epochs with a learning rate of 0.0001 and a batch size of 16 without augmentation. Each task was treated as binary classification (subtype: glioblastoma vs not glioblastoma, grade: (II, III) vs IV) using binary cross entropy loss. All MRI slices were again used for training, followed by using slices 60-130 at inference. Prediction scores were generated separately per each slice, followed by computing a patient-level score as the median across slices to serve as input to patient-level AUROC calculations. The median across slices was

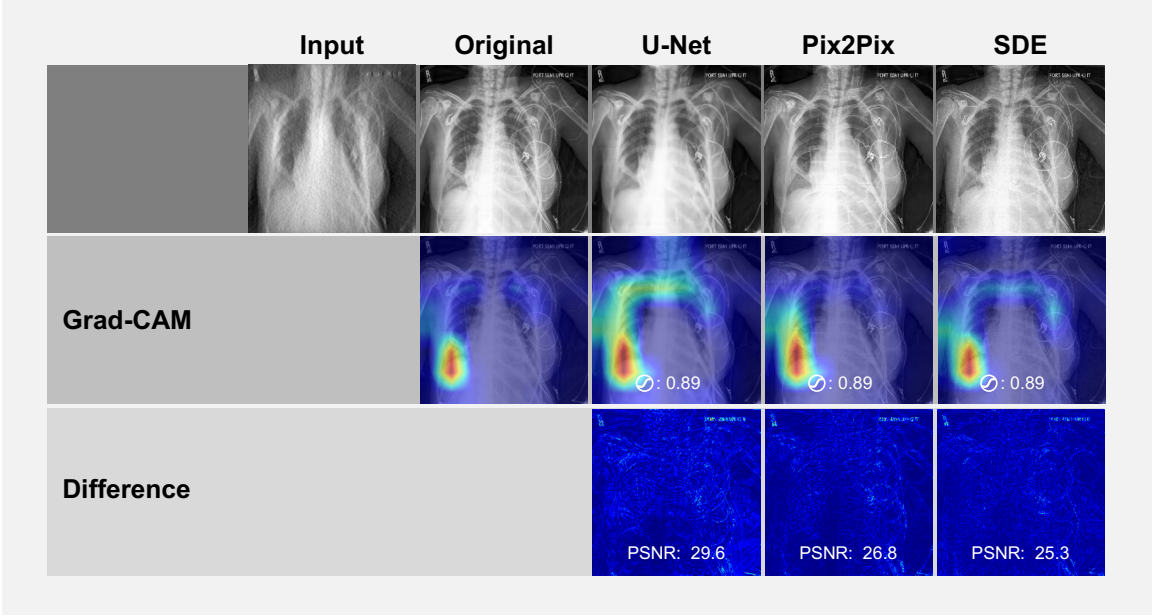


Figure 8: Reconstruction example from photon count 10,000 for the different models. Grad-CAM (Selvaraju et al., 2017) and logit score correspond to the lung lesion prediction of the pre-trained classifier, indicating similar predictions on the reconstructed images.

used to improve robustness to outliers. All UCSF-PDGM diagnostic models were trained using images pre-processed using min-max normalization to the 0-1 range and resized to 256x256. The CheXpert DenseNet classifier was trained using TorchXRayVision (Cohen et al., 2021). The default image preprocessing was used, with an input size of 224x224 pixels and normalization to a range of -1024 to 1024. The model was trained without data augmentation for 50 epochs using the Adam optimizer with a learning rate of 1e-3 and a weight decay of 1e-5.

Reconstruction Hyperparameters. No data augmentation was applied to any of the reconstruction pipelines. A U-Net was trained for 20 epochs on both UCSF-PDGM and CheXpert, using Adam with MSE loss, a learning rate of 0.001, and a batch size of 16. The GAN (Pix2Pix) was trained for 200 epochs on each dataset with Adam, a learning rate 0.0002, and a batch size of 32 to compensate for the smaller data volume. For the SDE model, we employed Adam with a learning rate of 0.0001, a cosine learning-rate schedule, and a batch size of 8; training ran for 40 epochs on CheXpert and 300 epochs on UCSF-PDGM. We note that the number of epochs varied between models because the different approaches take longer to converge (e.g., GANs are inherently less stable than a standard MSE loss), but in each case, the final weights were selected via validation loss monitoring, consistent with standard practice. During mitigation with the EODD-constraint, we employed $\tau = 0.5$ for the threshold, $T = 0.3$ for the temperature, and a momentum value of

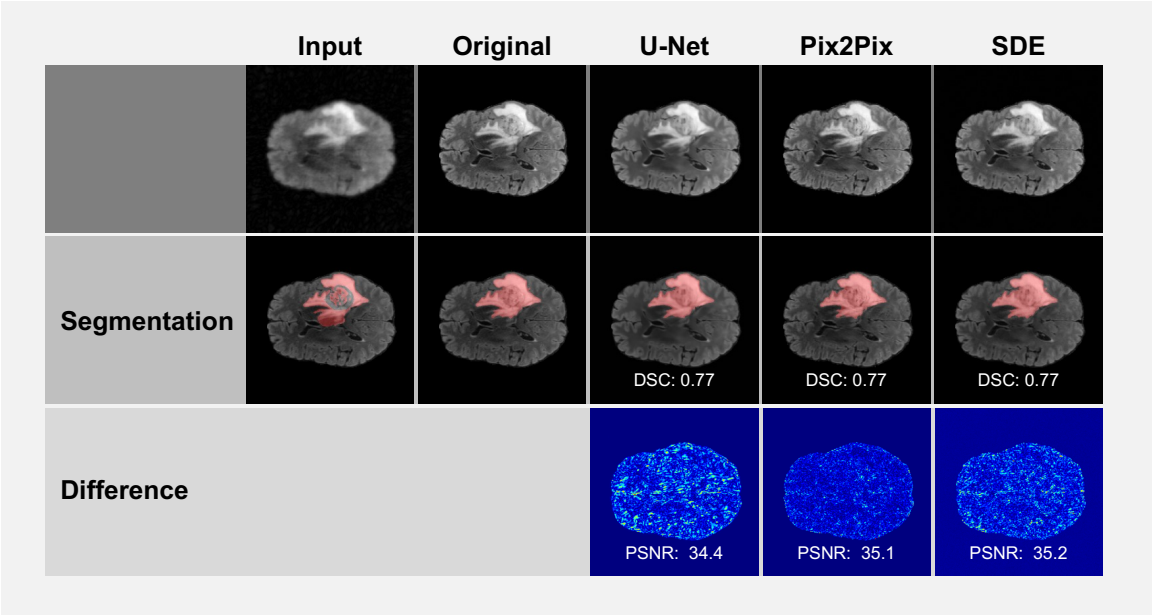


Figure 9: Reconstruction with corresponding segmentation and Dice score of an MRI image with acceleration 8 for the different models.

0.1 for the EMA. The remaining hyperparameters and architectural details were adopted unchanged from the original U-Net (Ronneberger et al., 2015), Pix2Pix (Isola et al., 2017), and SDE (Luo et al., 2023) publications. Image pre-processing consisted of min-max normalization to the 0-1 range and resizing to 256x256 for all reconstruction models.

The models were trained on a single NVIDIA A40 or A100 GPU. The SDE model was computationally most expensive and needed a maximum of 48 hours to train from scratch. For all models, the final weights were chosen based on performance on the validation split during training.

Proof of Proportionality. When the protected attribute A takes more than two categories (e.g., multiple races, genders, or age groups), we compare all pairs a_i, a_j of subgroups. Then, we take the maximum of the pairwise disparities in true positive and false positive rates:

$$\begin{aligned} EODD = \max_{1\le i<j\le k} \Big[& \big|P(\hat{Y}=1 \mid Y=1, A=a_i) \\ & - P(\hat{Y}=1 \mid Y=1, A=a_j)\big| \\ & + \big|P(\hat{Y}=1 \mid Y=0, A=a_i) \\ & - P(\hat{Y}=1 \mid Y=0, A=a_j)\big| \Big] \end{aligned}$$

Each pairwise comparison is handled exactly as in the binary case by treating a_i, a_j as $0, 1$. Therefore, all the steps below—derived under a binary setup—apply pairwise to

any two subgroups. Taking the maximum over these pairwise disparities then yields the multi-group measure.
This proof is based on the derivation by (Marcinkevics et al., 2022), and adjusted for EODD. EODD measures the disparity between subgroups in true positive rate (TPR) and false positive rate (FPR). In the binary case:

$$\begin{aligned} EODD = & P_{X,Y,A}(\hat{Y} = 1|Y = 1, A = 1) \\ & - P_{X,Y|A}(\hat{Y} = 1|Y = 1, A = 0) \\ & + P_{X,Y,A}(\hat{Y} = 1|Y = 0, A = 1) \\ & - P_{X,Y,A}(\hat{Y} = 1|Y = 0, A = 0) \end{aligned}$$

This can be expressed by the following proxy function.

$$EODD = \frac{\sum_{i=1}^{n} f_\theta(x_i) a_i y_i}{\sum_{i=1}^{n} a_i y_i} \tag{1}$$

$$- \frac{\sum_{i=1}^{n} f_\theta(x_i)(1 - a_i) y_i}{\sum_{i=1}^{n} (1 - a_i) y_i} \tag{1}$$

$$+ \frac{\sum_{i=1}^{n} f_\theta(x_i) a_i (1 - y_i)}{\sum_{i=1}^{n} a_i (1 - y_i)} \tag{2}$$

$$- \frac{\sum_{i=1}^{n} f_\theta(x_i)(1 - a_i)(1 - y_i)}{\sum_{i=1}^{n} (1 - a_i)(1 - y_i)} \tag{2}$$

To start, let's define the conditional covariance:

$$\begin{aligned} & \operatorname{cov}(A, X|Y = y) = \\ & \quad \mathbb{E}[(A - \mathbb{E}[A|Y = y])(X - \mathbb{E}[X|Y = y])|Y = y] \\ & \quad = \mathbb{E}[AX|Y = y] - \mathbb{E}[A|Y = y]\mathbb{E}[X|Y = y] \end{aligned} \tag{3}$$

We can use the law of total covariance to prove the validity:

$$\operatorname{cov}(A, X) = \mathbb{E}\Big[\operatorname{cov}(A, X|Y)\Big] \tag{4}$$

$$+ \operatorname{cov}\Big(\mathbb{E}[A|Y], \mathbb{E}[X|Y]\Big) \tag{4}$$

Expanding the first expectation term with (3):

$$\begin{aligned} \mathbb{E}[\operatorname{cov}(A, X|Y)] &= \mathbb{E}\Big[\mathbb{E}[AX|Y] - \mathbb{E}[A|Y]\mathbb{E}[X|Y]\Big] \\ &= \mathbb{E}[AX] - \mathbb{E}[\mathbb{E}[A|Y]\mathbb{E}[X|Y]] \end{aligned} \tag{5}$$

Expanding the second covariance term:

$$\operatorname{cov}(\mathbb{E}[X|Z], \mathbb{E}[Y|Z]) = \mathbb{E}[\mathbb{E}[X|Z]\mathbb{E}[Y|Z]] \tag{6}$$

$$- \mathbb{E}[X]\mathbb{E}[Y] \tag{6}$$

Substituting (5) and (6) into (4):

$$\begin{aligned}\operatorname{cov}(X,Y) &= \mathbb{E}[XY] - \mathbb{E}[\mathbb{E}[X|Z]\mathbb{E}[Y|Z]] \\ &\quad + \mathbb{E}[\mathbb{E}[X|Z]\mathbb{E}[Y|Z]] - \mathbb{E}[X]\mathbb{E}[Y] \\ &= \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] \\ &= \operatorname{cov}(X,Y)\end{aligned}$$

We want to show that $\Delta_{OOD} \propto \widehat{\mathrm{Cov}}(A, f_\theta(X)|Y=1) + \widehat{\mathrm{Cov}}(A, f_\theta(X)|Y=0)$
Let $\sum_i a_i y_i = S_{AY}, \quad \sum_i a_i = S_A, \quad \sum_i y_i = S_Y$.

Expanding EODD:
Expanding (1):

$$\begin{aligned}&\frac{\sum_{i=1}^N f_\theta(x_i) a_i y_i}{\sum_{i=1}^N a_i y_i} - \frac{\sum_{i=1}^N f_\theta(x_i)(1-a_i) y_i}{\sum_{i=1}^N y_i (1-a_i) y_i} \\ &= \frac{1}{S_{AY}} \sum_{i=1}^N f_\theta(x_i) a_i y_i - \frac{1}{S_Y - S_A} \sum_{i=1}^N f_\theta(x_i) \\ &+ \frac{1}{S_Y - S_{AY}} \sum_{i=1}^N f_\theta(x_i) a_i y_i \\ &= \frac{S_Y}{S_{AY}(S_Y - S_{AY})} \sum_{i=1}^N f_\theta(x_i) y_i a_i \\ &- \frac{1}{S_Y - S_{AY}} \sum_{i=1}^N f_\theta(x_i) y_i\end{aligned}$$

Note that:

$$\begin{aligned}&\widehat{\mathrm{Cov}}(A, f_\theta(X)|Y=1) \\ &\quad = \frac{\sum_{i=1}^n f_\theta(x_i) a_i y_i}{\sum_{i=1}^n y_i} \\ &\quad - \frac{\sum_{i=1}^n a_i y_i}{\sum_{i=1}^n y_i} \frac{\sum_{i=1}^n f_\theta(x_i) y_i}{\sum_{i=1}^n y_i} \\ &\quad = \frac{1}{S_Y} \sum_{i=1}^n f_\theta(x_i) a_i y_i \\ &\quad - \frac{S_{AY}}{S_Y^2} \sum_{i=1}^n f_\theta(x_i) y_i.\end{aligned}$$

Showing (5) $\propto \widehat{\mathrm{Cov}}(A, f_\theta(X)|Y=1)$
with factor $\frac{S_Y^2}{S_{AY}(S_Y - S_{AY})}$, independent of f_θ.

Expanding (2):

$$
\begin{aligned}
&\frac{\sum_{i=1}^{n} f_\theta(x_i) a_i (1-y_i)}{\sum_{i=1}^{n} a_i (1-y_i)} \\
&- \frac{\sum_{i=1}^{n} f_\theta(x_i)(1-a_i)(1-y_i)}{\sum_{i=1}^{n} (1-a_i)(1-y_i)} \\
&= \frac{N - S_Y}{(N - S_Y - S_A + S_{AY})(S_A - S_{AY})} \sum_{i=1}^{N} f_\theta(x_i) a_i \\
&\quad - \frac{N - S_Y}{(N - S_Y - S_A + S_{AY})(S_A - S_{AY})} \sum_{i=1}^{N} f_\theta(x_i) a_i y_i \\
&\quad - \frac{1}{N - S_Y - S_A + S_{AY}} \sum_{i=1}^{N} f_\theta(x_i) y_i \\
&\quad - \frac{N}{N - S_Y - S_A + S_{AY}} \sum_{i=1}^{N} f_\theta(x_i)
\end{aligned}
$$

Similarly:

$$
\begin{aligned}
&\widehat{\mathrm{Cov}}(A, f_0(X) | Y = 0) \\
&\quad = \frac{\sum_{i=1}^{N} f_0(x_i) a_i (1-y_i)}{\sum_{i=1}^{N} (1-y_i)} \\
&\quad - \frac{\sum_{i=1}^{N} a_i (1-y_i)}{\sum_{i=1}^{N} (1-y_i)} \cdot \frac{\sum_{i=1}^{N} f_0(x_i)(1-y_i)}{\sum_{i=1}^{N} (1-y_i)} \\
&\quad = \frac{1}{N - S_Y} \sum_{i=1}^{N} f_0(x_i) a_i \\
&\quad - \frac{N}{N - S_Y} \sum_{i=1}^{N} f_0(x_i) a_i y_i \\
&\qquad - \frac{S_A - S_{AY}}{(N - S_Y)^2} \sum_{i=1}^{N} f_0(x_i) \\
&\qquad - \frac{S_A \cdot S_{AY}}{(N - S_Y)^2} \sum_{i=1}^{N} f_0(x_i) y_i
\end{aligned}
$$

Showing (6) $\propto \widehat{\mathrm{Cov}}(A, f_\theta(X) | Y = 0)$ with factor $\frac{(S_A - S_{AY})(N - S_Y - S_A + S_{AY})}{(N - S_Y)^2}$, independent of f_θ.

Therefore, $EODD \propto \widehat{\mathrm{Cov}}(A, f_\theta(X) | Y = 1) + \widehat{\mathrm{Cov}}(A, f_\theta(X) | Y = 0)$.

	AI/AN	Asian	Black	NH/PI	Other	White	
Female, > 62	54	1539	923	314	2518	6456	11804
Female, ≤ 62	39	1739	608	136	1710	9500	13732
Male, > 62	56	1734	1023	240	3553	8984	15590
Male, ≤ 62	27	1924	539	171	1853	11170	15684
	176	6936	3093	861	9634	36110	56810

Table 5: Patient-wise groups used for analysis based on sex, age, and race for the CheXpert dataset. Unequally distributed with very few samples for American Indian or Alaska Native (AI/AN) and Native Hawaiian or Other Pacific Islander (NH/PI).

	Male	Female	
≤ 58	155	92	147
> 58	144	110	254
	299	202	501

Table 6: Patient distribution by sex and age for the UCSF-PDGM dataset. Patients under 58 and females represent minority groups.

Additional Fairness Results

In addition to Equalized Odds and Skewed Error Ratio in the main text, we investigate two additional bias metrics:

Equality of Opportunity (EOP):

$$P(\hat{Y} = 1 \mid Y = 1, A = 0) = P(\hat{Y} = 1 \mid Y = 1, A = 1).$$

We report the worst case Equality of Opportunity (Hardt et al., 2016) difference between groups

$$max_{i,j} |P(\hat{Y} = 1 \mid Y = 1, A = i) - P(\hat{Y} = 1 \mid Y = 1, A = j)|, \quad \forall \quad \mathrm{A} \in \mathcal{A}.$$

EOP is a relaxation of EODD, requiring fairness only concerning the positive class ($Y = 1$).

ΔDice: Given the limited availability of dedicated segmentation fairness metrics, we also compute:

$$\Delta\text{Dice} = \max_{i,j} \left|\text{Dice}_{A_i} - \text{Dice}_{A_j}\right|, A \in \mathcal{A}$$

which represents the maximum difference in Dice across all protected subgroups $\mathcal{A}$.

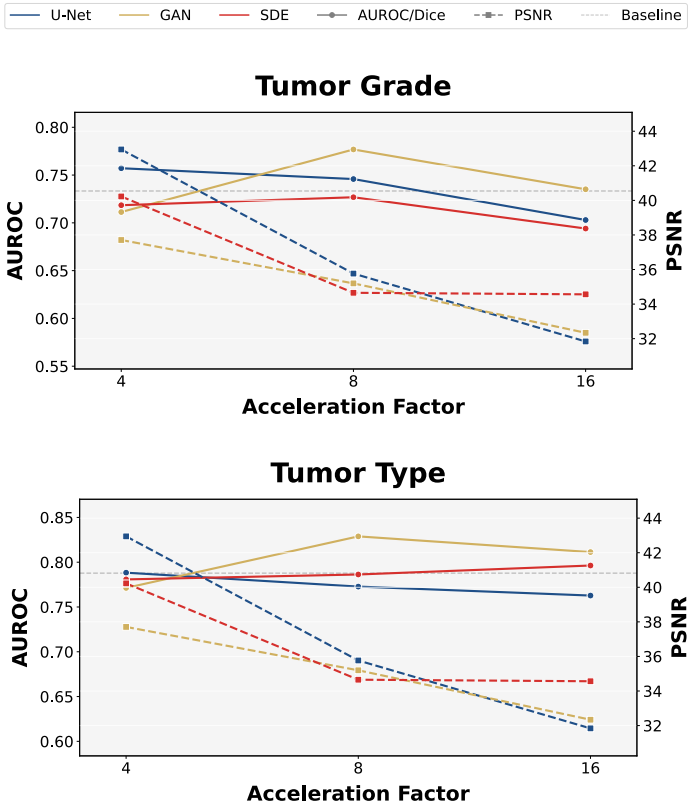
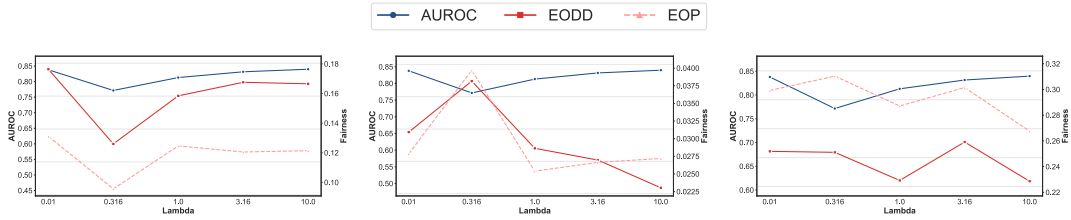


Figure 10: Tumor Type and Tumor Grade and PSNR values for different noise levels on UCSF-PDGM. The image quality and diagnostic performance axes are on a similar percentage scale. Task performance metrics show high stability across models and noise conditions, while PSNR drops with increasing noise.

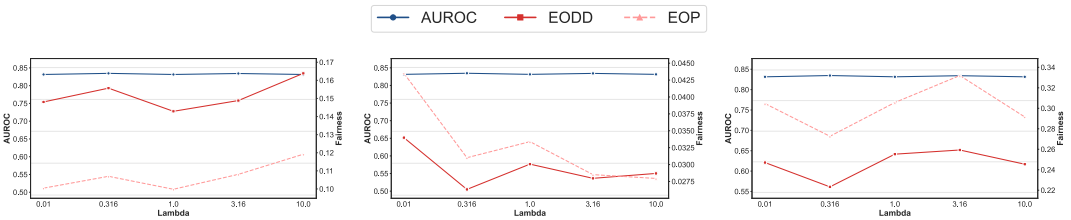
Plots containing the results of these additional evaluations can be found in Figure 16 and 17.

Additionally, Figures 18 and 19 contain results using different race subgroups for CheXpert. Our original evaluations considered each of the original subgroups listed within the dataset (Table 5) when computing the fairness metrics. Given the small counts for the American Indian or Alaska Native and Native Hawaiian or Other Pacific Islander subgroups, leading to large error bars, we also computed these metrics when including these subgroups within the Other subgroup.



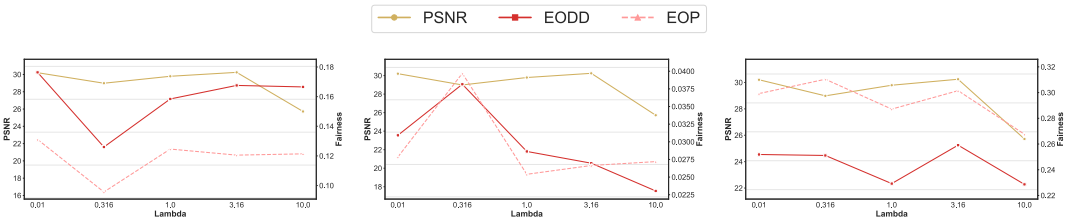
Age Sex Race

Figure 11: Influence of fairness weighting parameter (λ_{fair}) on classifier AUROC performance and fairness metrics for the Equalized Odds (EODD) mitigation constraint, evaluated with U-Net on the CheXpert dataset. There is minor sensitivity of AUROC to lambda; fairness metrics show greater variance but minimal substantial improvement with increased λ.



Age Sex Race

Figure 12: Influence of λ_{fair} on AUROC and fairness metrics for the adversarial fairness loss with U-Net on CheXpert. Similar findings to the EODD loss include minimal AUROC variation and moderate fairness variability without substantial gains.



Age Sex Race

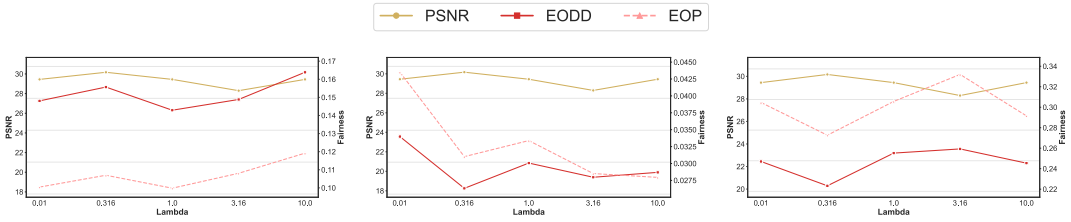
Figure 13: Impact of λ_{fair} on reconstruction quality (PSNR) compared to fairness for the EODD constraint mitigation. PSNR remains stable across lambda variations, while fairness shows slight variation without substantial improvement.

Photon Count	Metrics		Baseline	U-Net	GAN	SDE
100,000	AUROC	Atalectasis	0.87	0.87	0.86	0.87
		Cardiomegaly	0.91	0.91	0.91	0.91
		Consolidation	0.91	0.91	0.91	0.91
		Edema	0.90	0.90	0.90	0.90
		EC	0.79	0.78	0.78	0.79
		Fracture	0.76	0.75	0.75	0.76
		Lung Lesion	0.80	0.79	0.79	0.79
		Lung Opacity	0.88	0.88	0.88	0.88
		Pleural Effusion	0.93	0.92	0.92	0.92
		Pleural Other	0.83	0.82	0.81	0.82
		Pneumonia	0.83	0.83	0.83	0.83
		Pneumothorax	0.77	0.75	0.76	0.77
		Average	0.85	0.84	0.84	0.85
	PSNR			31.60	30.16	29.98
	LPIPS			0.13	0.08	0.08
10,000	AUROC	Atalectasis	0.87	0.87	0.86	0.87
		Cardiomegaly	0.91	0.90	0.90	0.91
		Consolidation	0.91	0.91	0.90	0.91
		Edema	0.90	0.89	0.89	0.90
		EC	0.79	0.78	0.78	0.78
		Fracture	0.76	0.75	0.74	0.75
		Lung Lesion	0.80	0.78	0.78	0.79
		Lung Opacity	0.88	0.88	0.87	0.88
		Pleural Effusion	0.93	0.92	0.91	0.92
		Pleural Other	0.83	0.81	0.80	0.82
		Pneumonia	0.83	0.82	0.82	0.82
		Pneumothorax	0.77	0.75	0.75	0.77
		Average	0.85	0.84	0.83	0.84
	PSNR			30.52	28.62	27.12
	LPIPS			0.19	0.11	0.15
3000	AUROC	Atalectasis	0.87	0.86	0.85	0.86
		Cardiomegaly	0.91	0.90	0.90	0.91
		Consolidation	0.91	0.91	0.90	0.90
		Edema	0.90	0.89	0.89	0.89
		EC	0.79	0.78	0.78	0.78
		Fracture	0.76	0.74	0.73	0.75
		Lung Lesion	0.80	0.77	0.77	0.78
		Lung Opacity	0.88	0.87	0.87	0.87
		Pleural Effusion	0.93	0.91	0.91	0.92
		Pleural Other	0.83	0.80	0.78	0.81
		Pneumonia	0.83	0.82	0.80	0.82
		Pneumothorax	0.77	0.74	0.74	0.77
		Average	0.85	0.83	0.83	0.84
	PSNR			28.89	27.36	26.83
	LPIPS			0.22	0.14	0.15

Table 7: Performance metrics for CheXpert across reconstruction models and photon counts. Includes PSNR, LPIPS, and AUROC scores for multi-label classification tasks across varying noise levels. A subtle trend is observed where pathologies with lower baseline AUROC (e.g., fracture, pneumothorax, lung lesion) experience slightly greater performance degradation under noise. At the same time, more easily detectable conditions (e.g., effusion, cardiomegaly) remain stable. Baseline is the prediction on the ground truth images.

Acceleration	Metrics		Baseline	U-Net	GAN	SDE
4	**AUROC**	**Tumor Type**	0.79	0.79	0.77	0.78
		Tumor Grade	0.73	0.76	0.71	0.72
	Dice		0.72	0.72	0.71	0.72
	PSNR			42.94	37.71	40.23
	LPIPS			0.01	0.02	0.00
8	**AUROC**	**Tumor Type**	0.79	0.77	0.83	0.79
		Tumor Grade	0.73	0.75	0.78	0.73
	Dice		0.72	0.70	0.71	0.71
	PSNR			35.77	35.20	34.65
	LPIPS			0.03	0.02	0.02
16	**AUROC**	**Tumor Type**	0.79	0.76	0.81	0.80
		Tumor Grade	0.73	0.70	0.74	0.69
	Dice		0.72	0.67	0.70	0.71
	PSNR			31.84	32.34	34.56
	LPIPS			0.06	0.04	0.02

Table 8: Performance metrics for UCSF-PDGM across reconstruction models and noise levels. Reports PSNR, LPIPS, Dice, and classification AUROC for tumor type and grade tasks. While PSNR varies with noise and model, downstream segmentation and classification metrics remain relatively stable, indicating robust task performance across conditions.



Age Sex Race

Figure 14: Impact of λ_{fair} on PSNR and fairness for the adversarial fairness loss. Stable PSNR across lambda values with minor fairness variations similar to the EODD loss results.

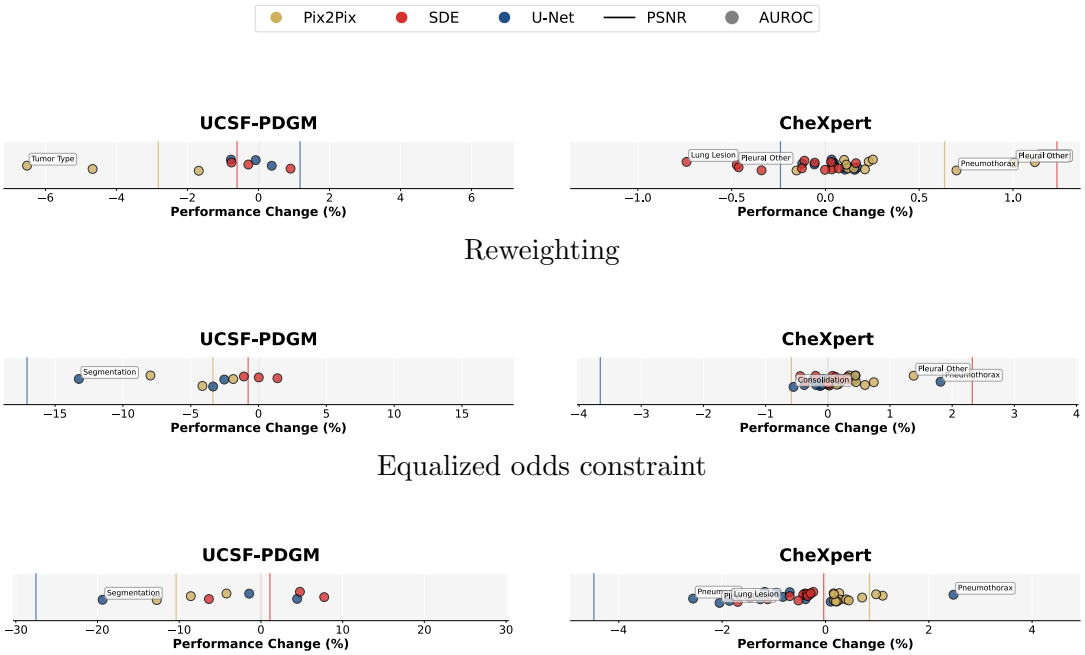


Figure 15: Change in prediction performance after applying bias mitigation techniques. Each row compares two datasets for a given method: (a) Reweighted sampling, (b) Equalized odds constraint, and (c) Adversarial training. UCSF-PDGM experiences more performance degradation. However, all techniques show good stability in task performance, with few outliers in the UCSF-PDGM dataset.

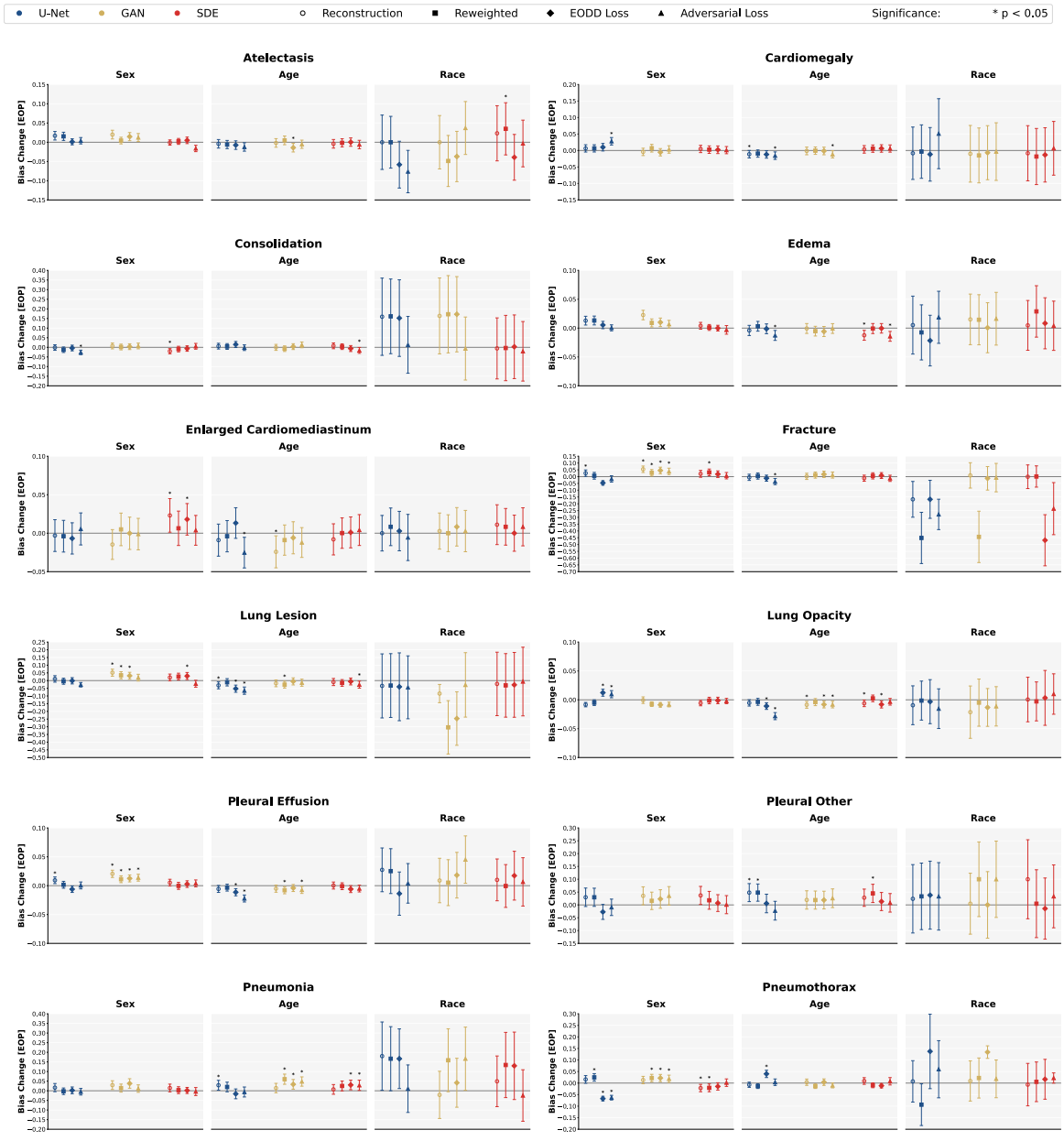


Figure 16: Equality of opportunity (EOP) bias change pre- and post-mitigation compared to predictions on original images for CheXpert classification. Pre-mitigation, bias tends to increase slightly for sex; race exhibits high variance. Post-mitigation—particularly with EODD and adversarial losses—bias declines slightly.

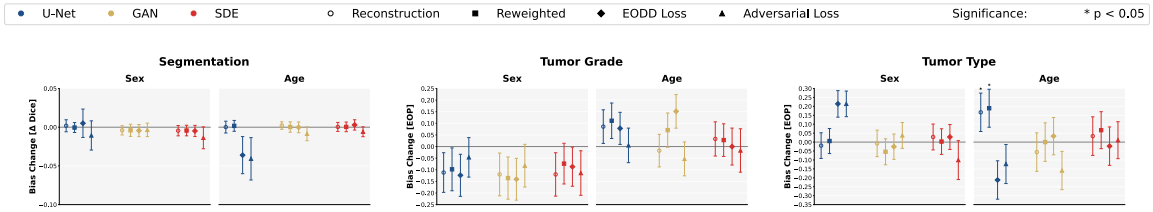


Figure 17: Equality of opportunity (EOP) and Δ Dice bias change compared to predictions on original images pre- and post-mitigation for UCSF-PDGM classification and segmentation.

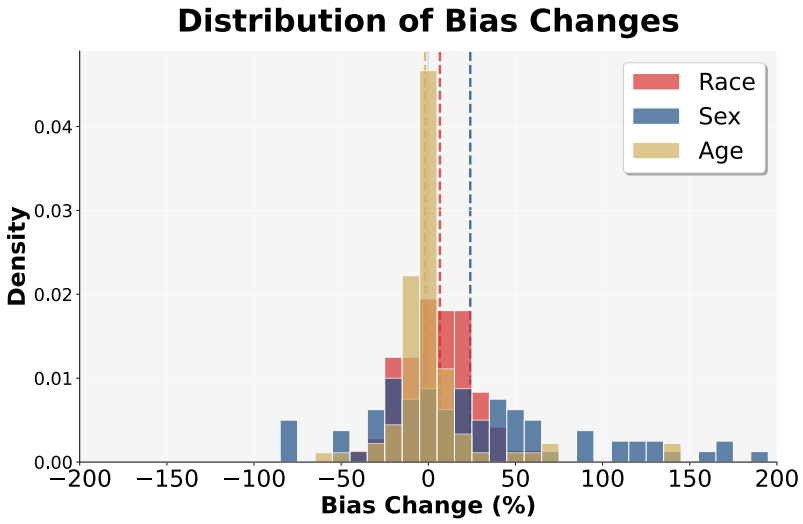


Figure 18: Distribution of bias changes when using alternative race subgroups for CheXpert calculations.

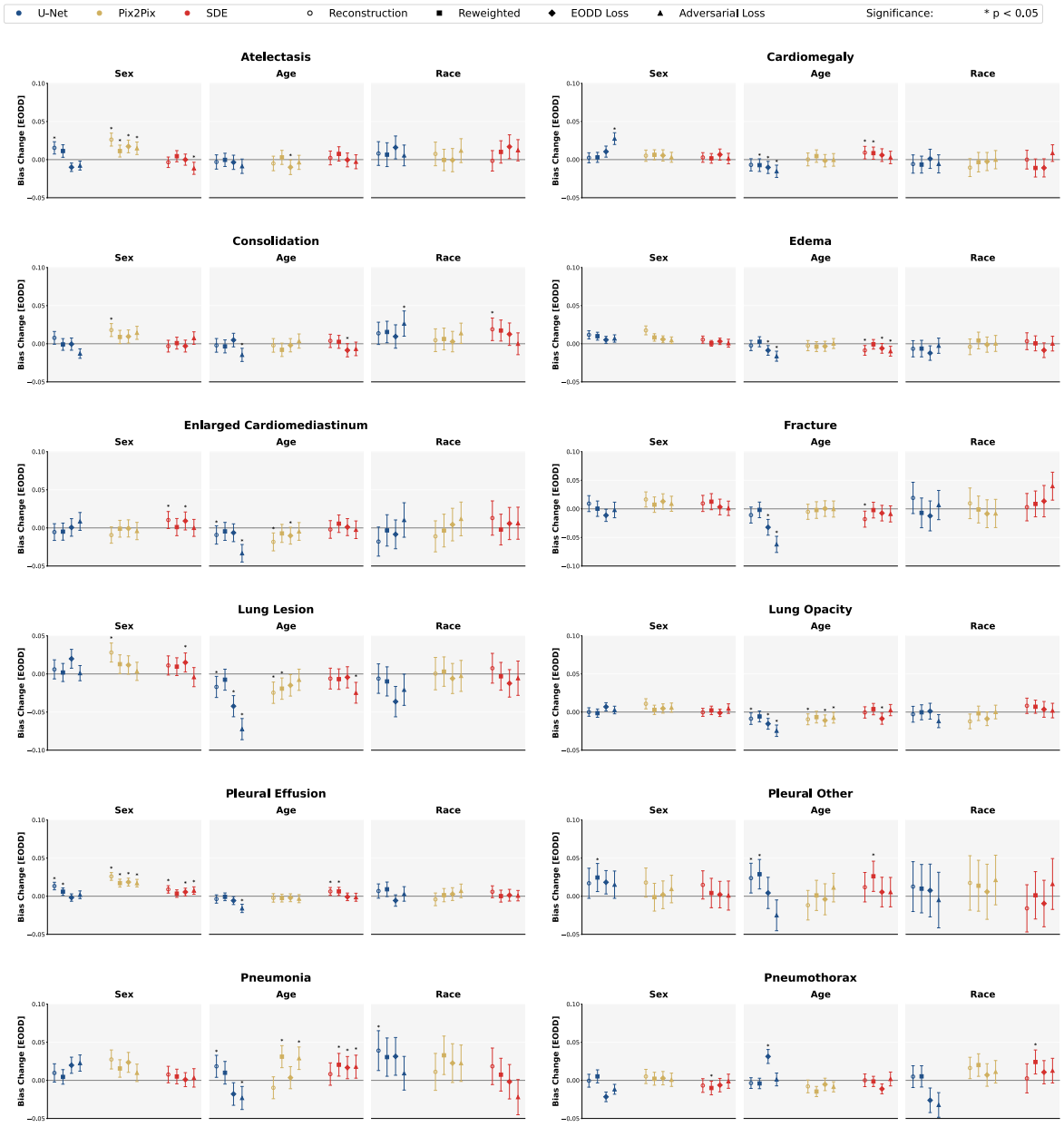


Figure 19: Equalized odds bias change pre- and post-mitigation compared to predictions on original images for CheXpert when using alternative race subgroups.