Exploring Bias in Deep Learning-Based MRI Reconstruction

Matteo Wohlrapp[1,2]
Supervised by Niklas Bubeck[1]

[1]AI in Medicine, Technical University of Munich, Munich, Germany
[2]Dana-Farber Cancer Institute, Harvard Medical School, Boston, USA
matteo.wohlrapp@tum.de

Abstract. The growing adoption of artificial intelligence in medical imaging has underscored the importance of fairness, ensuring that applied deep learning models do not introduce new or aggravate existing biases. Recent studies have focused mainly on bias in classification and segmentation tasks, while the influence of reconstruction models remains largely unexplored. In this work, we introduce an evaluation framework to assess the impact of reconstruction on fairness in downstream tasks such as classification and segmentation. By systematically varying noise levels and comparing multiple reconstruction techniques, we reveal that while overall bias changes are modest, certain conditions lead to non-negligible disparities. The effects do not clearly align with pathology or task difficulty and vary by sensitive attribute—showing the most consistent changes for sex and age, while race exhibits high variability. Our research highlights the need to evaluate fairness holistically across the complete imaging pipeline rather than assuming that reconstruction steps are fairness-neutral. The code is provided in our GitHub repository.

Keywords: Bias · Fairness · Reconstruction

1 Introduction

Recent years have marked a substantial increase in artificial intelligence (AI) in healthcare [17]. This surge is evident by the growing number of research publications in the field [27] and the increasing number of AI-based algorithms approved for clinical practice by governmental bodies such as the Food and Drug Administration (FDA) [19]. However, as more models are deployed in clinical settings [8], concerns about discrimination across patient subgroups have emerged [3]. For example, studies have found that deep learning systems diagnosing diabetic retinopathy perform with an accuracy of 73.0% on lighter-skin individuals compared to 60.5% on darker-skin individuals [1]. Similarly, research has shown that female patients, patients under 20 years old, and Black and Hispanic patients are more likely to be underdiagnosed in chest radiographs [28], and that White children are significantly more likely to have their biological sex correctly classified from T-1 weighted brain MRI scans [31]. While bias is not unique to healthcare

applications [26][25][20][22], its consequences in the medical field are dire and can exacerbate existing healthcare disparities [6].

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 metrics across subgroups in various imaging modalities, including brain MRI [31], chest X-rays [11], dermatology images [4], and retinal images [1]. These studies address sensitive attributes such as sex [31], age [28], race [28], and skin tone [16], evaluating disparities using Area Under the Curve (AUC) [28], True Positive Rates (TPR) [11], or dedicated fairness metrics [32]. In segmentation, studies assess segmentation performance under varying demographic distributions, analyzing race and sex representation in training datasets [13][18][23].

In contrast, little attention has been directed toward image reconstruction, which seeks to mitigate the noise introduced by faster MRI sampling or lower-dose CT protocols and enhance image quality prior to downstream analysis or diagnosis. Existing studies have primarily focused on subgroup performance differences using traditional image metrics [7][29]. Accordingly, we aim to complement these image-based evaluations with assessments involving downstream tasks such as classification or segmentation. This is especially relevant because, given the widespread availability of reconstruction algorithms, they could be used with further downstream models for classification or segmentation in a clinical setting.

In this work, we conduct a deeper analysis of bias in medical image reconstruction. Our contributions include (i) an evaluation framework to assess the influence of reconstruction on downstream tasks, (ii) a comprehensive analysis of performance and fairness across various noise levels, datasets, and reconstruction models, and (iii) a quantification of bias introduced during the reconstruction process. These contributions provide valuable insights into the interplay between image reconstruction and subsequent medical image analysis, enabling more equitable AI-driven healthcare solutions.

2 Method

We propose an evaluation framework, visualized in Figure 1, that assesses how reconstruction models affect downstream performance and fairness in medical imaging. Starting from clean MRI and X-ray data, we simulate realistic degradation—undersampling in MRI and dose reduction in X-ray—to produce paired images. These are passed through a set of reconstruction models (U-Net [24], the Generative Adversarial Network (GAN) Pix2Pix [15], and a stochastic differential equation (SDE)-based diffusion model [21]), after which the reconstructed outputs are evaluated on classification and segmentation tasks.

To quantify the impact of reconstruction, we compare image quality between original and reconstructed images, downstream task performance, and fairness metrics measured across sensitive subgroups.

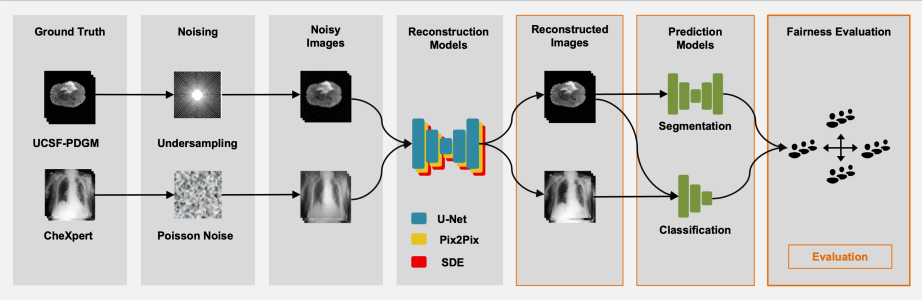

Fig. 1: Overview of the evaluation pipeline. Images are noised, and then noisy MRI and X-ray images are reconstructed and assessed using downstream classification and segmentation tasks. We compare downstream performance and fairness to quantify the impact of the reconstruction step.

2.1 Noising

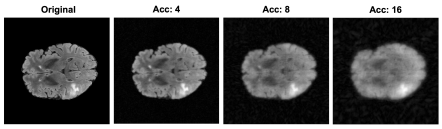

Fig. 2: Undersampled MRI images at various acceleration rates. Simulated degradation shows increasing artifact severity with higher acceleration.

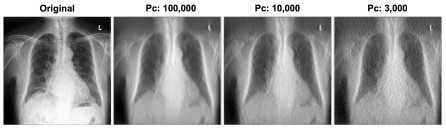

Fig. 3: X-ray images with simulated photon-count-based degradation. Increasing noise is introduced through lower photon counts to mimic low-dose acquisition.

We artificially degrade the images by introducing noise to establish a supervised setting. This process is performed separately for X-ray and MRI data:

MRI For MRI data, we simulate undersampled acquisitions by masking k-space data using a radial sampling pattern similar to the approach described in [9]. By varying the acceleration rate (acc), we obtain images with different degrees of undersampling, as shown in Figure 2.

X-Ray We simulate low-dose X-ray acquisitions from standard-dose images, following the methodology in [10]. Specifically, we use the Radon Transform to project the images into sinogram space and apply a bow-tie filter. We then add Poisson noise, scaled by a specified photon count (pc), to simulate reduced-dose scenarios. Finally, we retrieve the noisy images by applying the Inverse Radon

Transform. Different photon counts yield varying noise levels, as illustrated in Figure 3.

2.2 Fairness Evaluation

Our primary goal is to investigate how reconstruction models might introduce or amplify biases beyond conventional image-quality metrics like Peak Signal-to-Noise Ration (PSNR). We focus on group fairness, which evaluates performance disparities across subgroups defined by sensitive attributes (e.g., age, sex, race)[17]. In cases with more than two subgroups, we report the worst-case fairness outcome corresponding to the most significant observed disparity.

Classification Fairness Metrics We employ two common fairness metrics to evaluate classification performance:

Equalized Odds (EODD) [12].

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

This criterion requires that the predicted outcome $\hat{Y}$ is independent of the sensitive attribute A given the true label Y.

Equality of Opportunity (EOP) [12].

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

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

Segmentation Fairness Metrics We assess segmentation fairness by adapting the Skewed Error Ratio (SER) [30] to use the Dice coefficient instead of the Intersection over Union (IoU) as a relative metric of comparison. Additionally, we quantify the maximum Dice difference across subgroups directly:

Skewed Error Ratio (SER) [30].

$$SER_{\mathcal{A}} = \frac{\max_{A \in \mathcal{A}}(1 - \text{Dice}_A)}{\min_{B \in \mathcal{A}}(1 - \text{Dice}_B)},$$

where A, B denote protected groups and Dice_A is the Dice coefficient for group A. A higher $SER_{A,B}$ indicates a larger discrepancy in segmentation errors among subgroups.

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

$$\Delta\text{Dice} = \max_{A,B \in \mathcal{A}} |\text{Dice}_A - \text{Dice}_B|,$$

which represents the maximum difference in Dice across all protected subgroups $\mathcal{A}$. This measure provides a straightforward assessment of the worst-case fairness gap, and compared to SER does not scale the error based on baseline performance.

3 Experiments and Results

3.1 Datasets

UCSF-PDGM We utilize the University of California San Francisco Preoperative Diffuse Glioma MRI (UCSF-PDGM) dataset [2], a publicly available resource comprising 501 preoperative diffuse glioma cases with acquired with a standardized 3T MRI protocol. The dataset includes advanced diffusion and perfusion imaging, genetic biomarkers (e.g., IDH mutation status, MGMT promoter methylation), and segmentation masks. In our experiments, we focus on FLAIR images. Group-wise fairness is evaluated concerning age (categorically grouped into young and old based on the median age of 58) and sex. A detailed distribution of these sensitive attributes is provided in Table 1 in the supplemental material. We adopt a 70/10/20 split for training, validation, and testing, and we use the same training data for classification and reconstruction due to the limited amount of available data.

CheXpert In addition to the UCSF-PDGM dataset, we perform experiments using the CheXpert dataset [14], a large-scale collection of 224,316 chest radiographs from 65,240 patients, annotated for 14 common thoracic observations (e.g., Atelectasis, Cardiomegaly, Pneumonia). We follow a 70/10/20 split for training, validation, and testing and further partition the training and validation sets in a 70/30 ratio for reconstruction and classification tasks. Group-wise fairness in CheXpert is assessed concerning age (categorically grouped into young and old based on the median age of 62), sex, and race, with the corresponding distributions detailed in Table 2.

3.2 Models

Reconstruction We analyze three denoising models that span the spectrum from classical to generative approaches:

- U-Net [24]: A symmetrical fully convolutional neural network designed for image-to-image tasks.
- Pix2Pix [15]: A conditional GAN for image-to-image translation that leverages a U-Net backbone.
- Image Restoration with Mean-Reverting Stochastic Differential Equations (SDE) [21]: A generative diffusion model employing stochastic differential equations with mean reversion for general-purpose image restoration, built on a U-Net architecture.

Prediction For downstream prediction tasks, we employ separate architectures for classification and segmentation:

- **Classification:** We adopt a ResNet-50 network pre-trained on ImageNet. For the UCSF-PDGM dataset, two classifiers are trained independently to

predict the WHO CNS grade and the final pathological diagnosis (WHO, 2021), while a single joint classifier is utilized for the CheXpert dataset based on [5].
– **Segmentation:** A U-Net model is trained for segmentation on the UCSF-PDGM dataset. Note that segmentation experiments are not conducted on CheXpert due to the absence of segmentation masks.

3.3 Performance Implications of Reconstruction

We analyze the impact of reconstruction on downstream task performance and image quality using qualitative and quantitative evaluations across various noise levels.

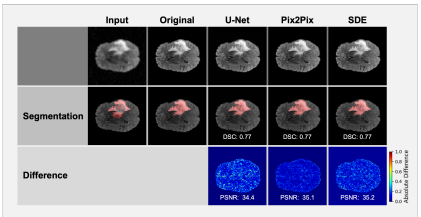

Fig. 4: Qualitative reconstruction results and derived segmentation masks for UCSF-PDGM (acceleration factor 8). High visual similarity across reconstructions with no differences in segmentation performance and slight PSNR difference.

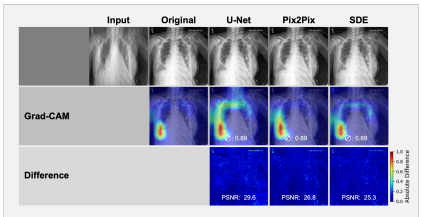

Fig. 5: Qualitative reconstruction results for CheXpert (photon count 10,000). Pix2Pix and SDE produce perceptually realistic images but introduce artifacts or hallucinated structures, while U-Net yields smoother results with fewer anomalies. Grad-CAM maps show largely consistent activation patterns across models.

Qualitative Results For UCSF-PDGM (Figure 4), the reconstructed images from all models appear very similar, and the corresponding segmentation masks yield comparable Dice scores. PSNR values are generally higher for UCSF-PDGM than for CheXpert, with slight variations between the reconstruction models. In contrast, for CheXpert (Figure 5), the reconstructed images differ quite drastically in style. The U-Net produces the smoothest reconstructed images, whereas the Pix2Pix reconstruction appears more realistic but introduces artifacts around finer structures, such as support devices. The SDE approach yields images that are perceptually very close to the ground truth, albeit sometimes generating realistic-looking yet inaccurate elements (e.g., in the depiction

of support devices). Furthermore, Grad-CAM visualizations for pleural effusion indicate relatively similar activation patterns across models, with U-Net exhibiting a slightly larger highlighted area.

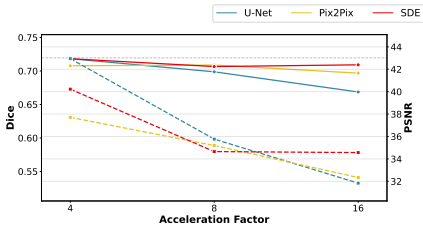

Fig. 7: Segmentation performance and image quality after reconstruction for UCSF-PDGM across noise levels. Dice metrics show high stability across models and noise conditions, while PSNR drops with increasing noise, indicating the robustness of downstream tasks to image degradation.

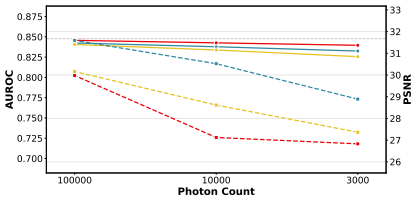

Fig. 8: Averaged classification performance and image quality after reconstruction on CheXpert across noise levels. U-Net maintains stronger image metrics than Pix2Pix and SDE, particularly at higher noise. However, all models preserve general AUROC performance across the noise spectrum, while PSNR drops with increasing noise.

Quantitative Results The complete numerical results for both datasets are summarized in Table 3 (UCSF-PDGM) and Table 4 (CheXpert) in the appendix. In addition, segmentation performance on UCSF-PDGM is detailed in Figure 7, and averaged classification results for CheXpert are shown in Figure 8. For a more granular view, the plots for individual pathologies in CheXpert are provided in Figure 17 in the appendix, and the classification results for UCSF-PDGM are further broken down in Figures 15 and 16.

Across all tests, a similar pattern emerges. The performance metrics for the downstream predictors (both classification and segmentation) remain stable across the full range of noise levels, even as the PSNR values decrease drastically with increased input noise. On average, the models perform better on the UCSF-PDGM dataset than on CheXpert. Notably, while the three reconstruction models perform similarly on UCSF-PDGM, the U-Net consistently outperforms the others on CheXpert. However, regarding image fidelity, the Learned Perceptual Image Patch Similarity (LPIPS) metric [33] reveals a lower score for Pix2Pix than U-Net, with the SDE model achieving the lowest value—indicating that its reconstructions are perceptually closest to the ground truth.

A closer analysis of UCSF-PDGM shows no clear trend in the downstream task performance across the different noise levels. In contrast, CheXpert dis-

plays subtle trends: although the overall drop in performance is relatively low, pathologies with lower baseline performance (e.g., fracture, pneumothorax, lung lesion) tend to experience a more pronounced decline, whereas those with higher baseline performance (e.g., effusion, consolidation, edema) are less affected.

3.4 Fairness Implications of Reconstruction

This section analyzes the fairness for the reconstruction models trained with acceleration eight on UCSF-PDGM and photon count 10,000 on CheXpert.

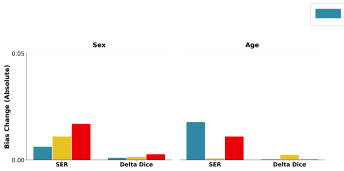

Fig. 10: Bootstrapped absolute segmentation bias change after reconstruction. Absolute changes in fairness metrics are minimal, suggesting segmentation is largely unaffected by reconstruction.

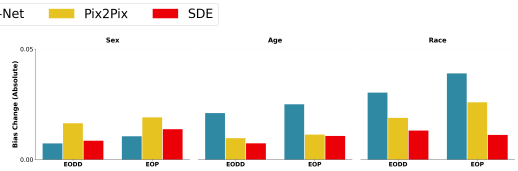

Fig. 11: Bootstrapped average absolute classification bias change across all UCSF-PDGM and CheXpert classification tasks. Overall, there are small bias variations, with the most prominent change by U-Net, followed by Pix2Pix, and SDE.

Figures 10 and 11 display the bootstrapped absolute bias changes for segmentation and classification, respectively. We report the absolute values because no clear pattern emerged, and the average canceled out those individual bias changes. This can bee seen in Figures 12, 13, and 21 in the appendix. In particular, there is almost no observable change for the segmentation fairness metric Δ Dice, while the adapted SER metric shows some variation, albeit inconsistent. For classification, there is more observable change. Although the bias values differ across models for classification, the SDE approach appears to introduce the least additional change, with the U-Net showing the most significant influence—especially regarding Race and Age. Overall, the additional bias introduced by reconstruction remains relatively small.

The importance of the impact of reconstruction on bias also varies with the sensitive attribute. When considering the baseline unfairness—i.e., the unfairness of the classifier on the original images, as detailed in Table 5 in the supplemental material, the relative change is more pronounced for sex, whereas age and race exhibit a higher baseline unfairness. These tables further include the significance of the bias change, measured via a one-sided hypothesis test using bootstrapping.

While the results do not reveal a uniform pattern, they suggest that additional bias is more prominent for sex, with a negative impact on age.

Although the absolute bias values for race are the largest, they are accompanied by considerable standard error. This variability is evident from Figure 13 and from the bold markers in Table 5c in the appendix. Additionally, while no consistent trend is observed across all pathologies, some pathologies (e.g., Consolidation, Pleural Other, or Pneumonia as shown in Figure 13) exhibit relatively large absolute bias values. Notably, these bias changes do not correlate clearly with the baseline performance or fairness, failing to mirror the trends observed in the performance analysis.

Finally, we do not observe any increase in variance or standard error in the predictions on the reconstructed images, as demonstrated in Figures 19 and 20 in the appendix.

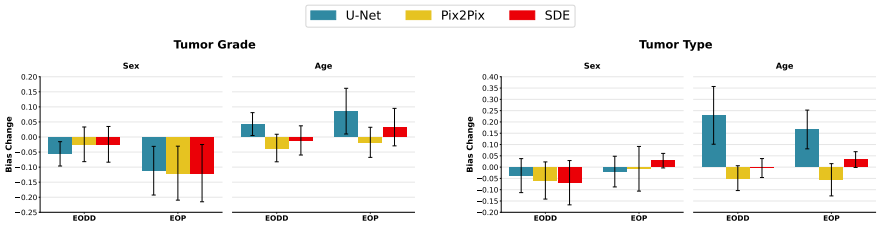

Fig. 12: Bias change in tumor type and tumor grade classification on UCSF-PDGM due to reconstruction. Overall slightly reduced bias for sex with increases for age. Across models and attributes high standard error.

4 Discussion

4.1 Performance Implications of Reconstruction

Qualitative Observations The qualitative analysis reveals that while reconstruction significantly affects low-level image fidelity—as evidenced by large discrepancies in PSNR—the impact on downstream classification is less pronounced. Classifiers are designed to capture more holistic, larger-scale features rather than pixel-level details. In the UCSF-PDGM dataset, as shown in Figure 4, the reconstructed images appear very similar, possibly due to the dataset's smaller size and variability. In contrast, the CheXpert dataset (Figure 5) exhibits greater variability in reconstructed image appearance. Notably, diffusion-based models (SDE) yield exceptionally realistic images; however, their probabilistic nature often leads them to produce hallucinated outputs that reflect the average distribution of images, which can explain the relatively lower PSNR. Meanwhile, the U-Net maintains more discriminative detail, and despite offering realistic textures, the Pix2Pix approach is prone to artifacts and training instabilities such as model collapse.

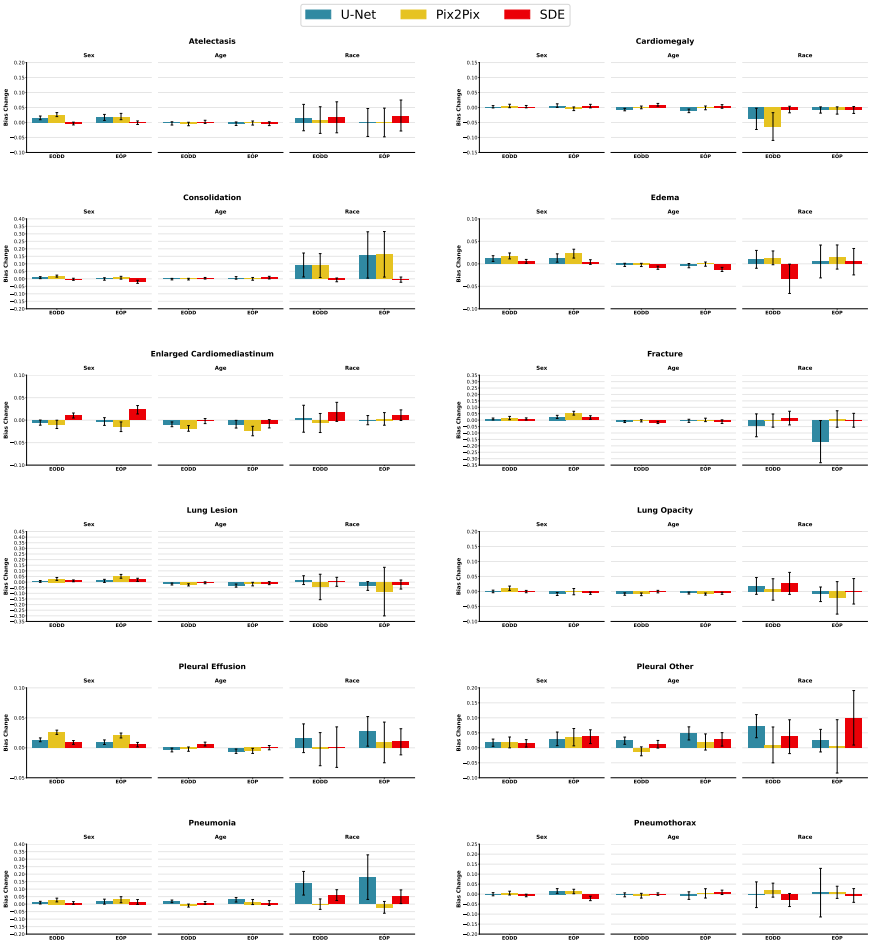

Fig. 13: Classification bias changes after reconstruction on CheXpert, broken down by pathology. Certain pathologies (e.g., Consolidation, Pleural Other, Pneumonia) show large subgroup disparities post-reconstruction. These shifts do not consistently align with baseline performance or fairness.

Quantitative Observations. Quantitatively, our approach is validated by the stability of the downstream performance metrics, which remain robust across a range of noise levels. This stability supports our decision not to retrain classifiers on the reconstructed images, thereby ensuring better comparability. After extensive fine-tuning of all reconstruction models, the U-Net outperforms the Pix2Pix and SDE models on CheXpert, where the variability and sample size are larger (Figure 8). The probabilistic nature of Pix2Pix and SDE appears to introduce increased noise, with the SDE model yielding lower PSNR yet superior perceptual similarity as indicated by lower LPIPS scores (Tables 4 and 3). Furthermore, our analysis suggests that the nature of the pathology may also influence performance trends: tasks involving prominent pathological features (e.g., cardiomegaly, consolidation, edema, enlarged cardio mediastinum) are less affected by the reconstruction process than those involving more subtle features (e.g., fractures, lung lesions, pleural abnormalities, pneumothorax).

4.2 Bias Implications of Reconstruction

Although the thresholds for what constitutes an unfair outcome remain challenging to define, our results suggest that while some additional classification bias is introduced, it is generally minor. The effects are even less pronounced in segmentation, which intuitively seems more sensitive to image changes. This outcome contradicts our expectation that models operating on pixel-level might be more sensitive to reconstruction bias. The importance of the reconstruction bias varies with the sensitive attribute. For example, while the baseline unfairness for age and race is already relatively high (Table 5), it is lower for sex, making the relative change due to reconstruction more relevant. The larger standard errors observed for race may be attributed to smaller subgroup sizes in the CheXpert dataset and most likely also contribute to the lack of significance.

Despite these variations, no clear trends emerge linking bias changes to performance levels or task difficulty, and the changes remain inconsistent across different pathologies (Figures 13, 12). This contrasts the performance results, where we saw at least subtle trends based on the pathological nature.

Interestingly, our results also do not replicate previous findings of pronounced bias in reconstruction models based on traditional image metrics [7][29]. In our experiments, the maximum subgroup PSNR discrepancies remain below 2% (Tables 6, 7 in the appendix), suggesting that the influence of reconstruction on fairness may be less severe than anticipated. Additionally, experiments with alternative tasks such as image-to-image translation and deliberate training dataset skews did not yield significant correlations, underscoring that the differences between reconstruction models (even those with a probabilistic foundation like SDE) are relatively subtle in their impact on fairness.

5 Conclusion

In this report, we have provided a comprehensive analysis of the impact of image reconstruction on downstream task performance and fairness in medical imag-

ing. Our study addresses a critical gap in the literature by rigorously testing how reconstruction models may inadvertently influence bias. Our experiments demonstrate that while the overall bias remains relatively small, there are specific scenarios and sensitive attributes where the effects become more pronounced. Specifically, in contexts such as medical diagnosis, where accurate and equitable treatment of patients is key, even minor biases introduced during the reconstruction process can have significant implications. In those cases, integrating bias mitigation strategies directly into the training of reconstruction methods would be an interesting direction for further research. Our work underscores the need for continued efforts to understand and reduce bias along the complete imaging pipeline to further enhance the equity and reliability of AI-driven healthcare solutions.

References

1. Burlina, P., Joshi, N., Paul, W., Pacheco, K., Bressler, N.: Addressing artificial intelligence bias in retinal diagnostics. Translational Vision Science Technology **10**, 13 (02 2021). https://doi.org/10.1167/tvst.10.2.13
2. Calabrese, E., Villanueva-Meyer, J.E., Rudie, J.D., Rauschecker, A.M., Baid, U., Bakas, S., Cha, S., Mongan, J.T., Hess, C.P.: The university of california san francisco preoperative diffuse glioma mri dataset. Radiology: Artificial Intelligence **4**(6) (Nov 2022). https://doi.org/10.1148/ryai.220058, http://dx.doi.org/10.1148/ryai.220058
3. Chen, R.J., Wang, J.J., Williamson, D.F.K., Chen, T.Y., Lipkova, J., Lu, M.Y., Sahai, S., Mahmood, F.: Algorithmic fairness in artificial intelligence for medicine and healthcare. Nat. Biomed. Eng. **7**(6), 719–742 (Jun 2023)
4. Chiu, C.H., Chen, Y.J., Wu, Y., Shi, Y., Ho, T.Y.: Achieve fairness without demographics for dermatological disease diagnosis. Medical Image Analysis **95**, 103188 (2024). https://doi.org/https://doi.org/10.1016/j.media.2024.103188, https://www.sciencedirect.com/science/article/pii/S1361841524001130
5. Cohen, J.P., Viviano, J.D., Bertin, P., Morrison, P., Torabian, P., Guarrera, M., Lungren, M.P., Chaudhari, A., Brooks, R., Hashir, M., Bertrand, H.: Torchxrayvision: A library of chest x-ray datasets and models. In: International Conference on Medical Imaging with Deep Learning (2021), https://api.semanticscholar.org/CorpusID:240353861
6. Cross, J., Choma, M., Onofrey, J.: Bias in medical ai: Implications for clinical decision-making. PLOS Digital Health **3**(11), e0000651 (Nov 2024). https://doi.org/10.1371/journal.pdig.0000651
7. Du, Y., Xue, Y., Dharmakumar, R., Tsaftaris, S.A.: 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. p. 102–111. Springer-Verlag, Berlin, Heidelberg (2023). https://doi.org/10.1007/978-3-031-45249-9_10, https://doi.org/10.1007/978-3-031-45249-9_10
8. Esteva, A., Chou, K., Yeung, S., Naik, N., Madani, A., Mottaghi, A., Liu, Y., Topol, E., Dean, J., Socher, R.: Deep learning-enabled medical computer vision. npj Digital Medicine **4**, 5 (12 2021). https://doi.org/10.1038/s41746-020-00376-2
9. Feng, L.: Golden-angle radial mri: Basics, advances, and applications. Journal of Magnetic Resonance Imaging **56** (04 2022). https://doi.org/10.1002/jmri.28187
10. Gibson, N.M., Lee, A., Bencsik, M.: A practical method to simulate realistic reduced-exposure ct images by the addition of computationally generated noise. Radiological physics and technology (2023), https://api.semanticscholar.org/CorpusID:265148810
11. Glocker, B., Jones, C., Bernhardt, M., Winzeck, S.: Algorithmic encoding of protected characteristics in chest x-ray disease detection models. eBioMedicine **89** (2023), https://api.semanticscholar.org/CorpusID:256858498
12. Hardt, M., Price, E., Srebro, N.: Equality of opportunity in supervised learning. In: Proceedings of the 30th International Conference on Neural Information Processing Systems. p. 3323–3331. NIPS'16, Curran Associates Inc., Red Hook, NY, USA (2016)

13. Ioannou, S., Chockler, H., Hammers, A., King, A.P.: 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. p. 13–22. Springer-Verlag, Berlin, Heidelberg (2022). https://doi.org/10.1007/978-3-031-17899-3_2, https://doi.org/10.1007/978-3-031-17899-3_2
14. Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., Marklund, H., Haghgoo, B., Ball, R., Shpanskaya, K., Seekins, J., Mong, D.A., Halabi, S.S., Sandberg, J.K., Jones, R., Larson, D.B., Langlotz, C.P., Patel, B.N., Lungren, M.P., Ng, A.Y.: 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). https://doi.org/10.1609/aaai.v33i01.3301590, https://doi.org/10.1609/aaai.v33i01.3301590
15. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. CVPR (2017)
16. Kinyanjui, N.M., Odonga, T., Cintas, C., Codella, N.C.F., Panda, R., Sattigeri, P., Varshney, K.R.: Fairness of classifiers across skin tones in dermatology. In: Martel, A.L., Abolmaesumi, P., Stoyanov, D., Mateus, D., Zuluaga, M.A., Zhou, S.K., Racoceanu, D., Joskowicz, L. (eds.) Medical Image Computing and Computer Assisted Intervention – MICCAI 2020. pp. 320–329. Springer International Publishing, Cham (2020)
17. Lara, M.A.R., Echeveste, R., Ferrante, E.: Addressing fairness in artificial intelligence for medical imaging. Nature Communications **13** (2022), https://api.semanticscholar.org/CorpusID:251371910
18. Lee, T., Puyol-Antón, E., Ruijsink, B., Shi, M., King, A.P.: 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. p. 233–244. Springer-Verlag, Berlin, Heidelberg (2022). https://doi.org/10.1007/978-3-031-23443-9_22, https://doi.org/10.1007/978-3-031-23443-9_22
19. Lin, M.: What's needed to bridge the gap between us fda clearance and real-world use of ai algorithms. Academic radiology **29** (11 2021). https://doi.org/10.1016/j.acra.2021.10.007
20. Luccioni, A.S., Akiki, C., Mitchell, M., Jernite, Y.: Stable bias: evaluating societal representations in diffusion models. In: Proceedings of the 37th International Conference on Neural Information Processing Systems. NIPS '23, Curran Associates Inc., Red Hook, NY, USA (2023)
21. Luo, Z., Gustafsson, F.K., Zhao, Z., Sjölund, J., Schön, T.B.: Image restoration with mean-reverting stochastic differential equations. In: Krause, A., Brunskill, E., Cho, K., Engelhardt, B., Sabato, S., Scarlett, J. (eds.) Proceedings of the 40th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 202, pp. 23045–23066. PMLR (23–29 Jul 2023), https://proceedings.mlr.press/v202/luo23b.html
22. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., Galstyan, A.: A survey on bias and fairness in machine learning. ACM Comput. Surv. **54**(6) (Jul 2021). https://doi.org/10.1145/3457607, https://doi.org/10.1145/3457607

23. Puyol-Antón, E., Ruijsink, B., Mariscal Harana, J., Piechnik, S., Neubauer, S., Petersen, S., Razavi, R., Chowienczyk, P., King, A.: Fairness in cardiac magnetic resonance imaging: Assessing sex and racial bias in deep learning-based segmentation. Frontiers in Cardiovascular Medicine **9**, 859310 (Apr 2022). https://doi.org/10.3389/fcvm.2022.859310
24. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. vol. 9351, pp. 234–241 (10 2015). https://doi.org/10.1007/978-3-319-24574-4_28
25. Ruggeri, G., Nozza, D.: A multi-dimensional study on bias in vision-language models. In: Rogers, A., Boyd-Graber, J., Okazaki, N. (eds.) Findings of the Association for Computational Linguistics: ACL 2023. pp. 6445–6455. Association for Computational Linguistics, Toronto, Canada (Jul 2023). https://doi.org/10.18653/v1/2023.findings-acl.403, https://aclanthology.org/2023.findings-acl.403/
26. Saumure, R., De Freitas, J., Puntoni, S.: Humor as a window into generative ai bias. Scientific Reports **15**(1), 1326 (2025). https://doi.org/10.1038/s41598-024-83384-6, https://doi.org/10.1038/s41598-024-83384-6
27. Senthil, R., Anand, T., Somala, C.S., Saravanan, K.M.: Bibliometric analysis of artificial intelligence in healthcare research: Trends and future directions. Future Healthcare Journal **11**(3), 100182 (2024). https://doi.org/https://doi.org/10.1016/j.fhj.2024.100182, https://www.sciencedirect.com/science/article/pii/S2514664524015728
28. Seyyed-Kalantari, L., Zhang, H., McDermott, M., Chen, I., Ghassemi, M.: Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. Nature Medicine **27** (12 2021). https://doi.org/10.1038/s41591-021-01595-0
29. Sheng, Y., Yang, J., Lin, Y., Jiang, W., Yang, L.: Toward fair ultrasound computing tomography: Challenges, solutions and outlook. In: Proceedings of the Great Lakes Symposium on VLSI 2024. p. 748–753. GLSVLSI '24, Association for Computing Machinery, New York, NY, USA (2024). https://doi.org/10.1145/3649476.3660387, https://doi.org/10.1145/3649476.3660387
30. Siddiqui, I., Littlefield, N., Carlson, L., Gong, M., Chhabra, A., Menezes, Z., Mastorakos, G., Thakar, S., Abedian, M., Lohse, I., Weiss, K., Plate, J., Moradi, H., Amirian, S., P. Tafti, A.: Fair ai-powered orthopedic image segmentation: addressing bias and promoting equitable healthcare. Scientific Reports **14** (07 2024). https://doi.org/10.1038/s41598-024-66873-6
31. Stanley, E.A.M., Wilms, M., Mouches, P., Forkert, N.D.: Fairness-related performance and explainability effects in deep learning models for brain image analysis. Journal of Medical Imaging **9**, 061102 – 061102 (2022), https://api.semanticscholar.org/CorpusID:251876386
32. Yuan, C., Linn, K.A., Hubbard, R.A.: Algorithmic fairness of machine learning models for alzheimer disease progression. JAMA Network Open **6**(11), e2342203–e2342203 (11 2023). https://doi.org/10.1001/jamanetworkopen.2023.42203, https://doi.org/10.1001/jamanetworkopen.2023.42203
33. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: CVPR (2018)

6 Appendix

6.1 Dataset

Table 1: Patient distribution by sex and age for the UCSF-PDGM dataset. Young and female represent minority groups.

	Male	**Female**	
Young	155	92	147
Old	144	110	254
	299	202	501

Table 2: Patient-wise groups 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).

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

6.2 Performance Implications of Reconstruction

Table 3: 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.

Acceleration	Metrics		Baseline	U-Net	Pix2Pix	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 4: 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; EC: Enlarged Cardiomediastinum

Photon Count	Metrics		Baseline	U-Net	Pix2Pix	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

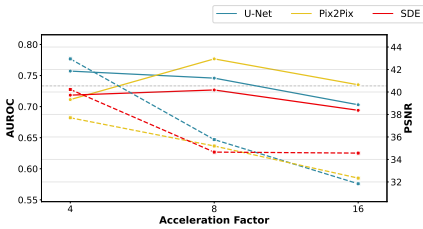

Fig. 15: Tumor grade classification performance and image quality after reconstruction for UCSF-PDGM across noise levels. AUROC shows high stability across models and noise conditions, while PSNR drops with increasing noise, indicating robustness of downstream tasks to image degradation.

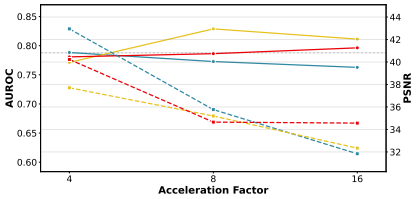

Fig. 16: Tumor type classification performance and image quality for UCSF-PDGM across noise levels. Again, stable AUROC and decreasing PSNR with added noise.

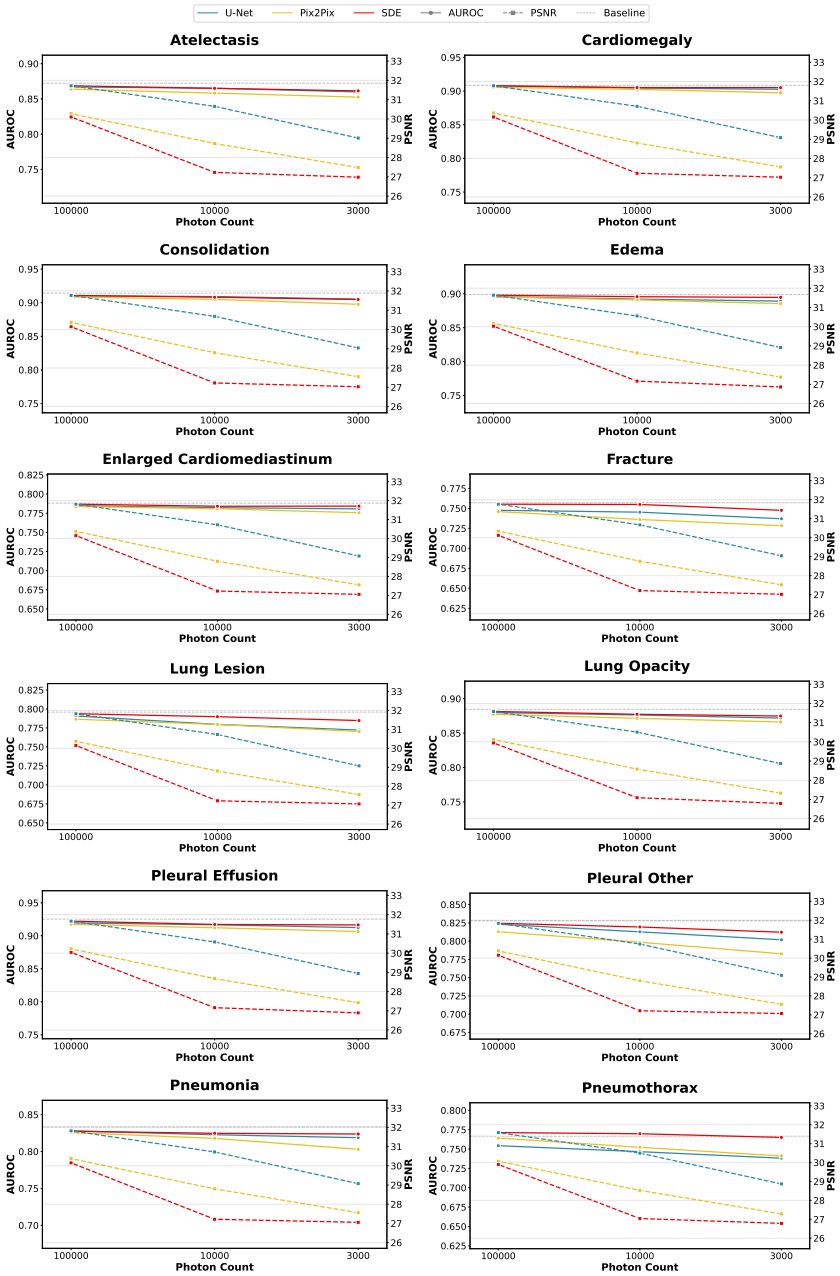

Fig. 17: Classification performance for individual pathologies after reconstruction and image quality for CheXpert across noise levels. Across all pathologies, AUROC remains stable, while PSNR drops with increased noise.

6.3 Fairness Implications of Reconstruction

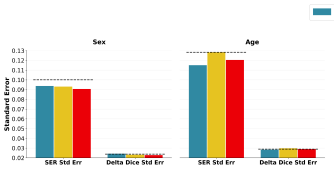

Fig. 19: Standard error of segmentation with baseline standard error for the different reconstruction models. Compared to the baseline, the reconstruction does not introduce additional variance to the segmentation fairness.

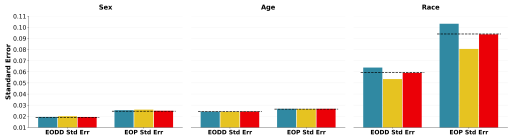

Fig. 20: Average standard error of classification across UCSF-PDGM and CheXpert with baseline standard error for the different reconstruction models. Compared to the baseline, the reconstruction does not introduce additional bias to the classification fairness.

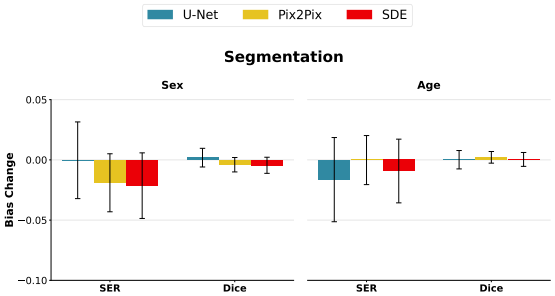

Fig. 21: UCSF-PDGM additional segmentation bias. Very little signal for any of the models and attributes with high standard error.

Table 5: Bias for baseline predictions on ground truth and reconstructed images across sensitive groups. Colored cells denote statistical significance of the difference between baseline and reconstruction (yellow: significantly worse; blue: significantly better); bold entries indicate cases where the standard deviation exceeds the effect size. For sex, there is better baseline fairness, yet slightly significant additional bias post-reconstruction, while age has a smaller baseline fairness but slightly improved bias after reconstruction. High standard deviation across racial groups, with no clear significant bias trends.

$+$, $p < 0.05$ $+$, $0.05 \leq p < 0.1$ $-$, $p < 0.05$ $-$, $0.05 \leq p < 0.1$
Bold indicates standard error larger than absolute effect size

(a) Sex

	Baseline		U-Net		Pix2Pix		SDE	
	EODD	**EOP**	**EODD**	**EOP**	**EODD**	**EOP**	**EODD**	**EOP**
EC	0.030	0.047	**0.027**	**0.044**	**0.023**	0.035	0.041	0.070
Cardiomegaly	0.024	0.040	**0.027**	0.046	**0.026**	**0.035**	**0.026**	**0.044**
Lung Opacity	0.011	0.011	**0.012**	0.005	0.021	**0.010**	**0.011**	0.006
Lung Lesion	0.024	0.033	**0.029**	**0.043**	0.050	0.081	0.034	0.052
Edema	0.007	0.007	**0.013**	**0.014**	0.018	0.023	**0.009**	**0.008**
Consolidation	0.023	0.039	**0.028**	**0.038**	0.038	**0.046**	**0.017**	0.020
Pneumonia	0.017	0.023	**0.022**	**0.033**	0.034	**0.043**	**0.021**	**0.032**
Atelectasis	0.017	0.010	0.030	0.022	0.040	0.024	**0.014**	**0.010**
Pneumothorax	0.043	0.068	**0.046**	0.084	**0.048**	0.081	0.036	0.045
Pleural Effusion	0.015	0.016	0.029	0.025	0.041	0.036	0.024	0.021
Pleural Other	0.040	0.060	0.056	0.089	**0.058**	0.094	0.056	0.096
Fracture	0.046	0.061	0.055	0.086	0.064	0.114	0.056	0.083
Tumor Grade	0.251	0.081	**0.251**	**0.081**	**0.290**	**0.089**	**0.291**	**0.086**
Tumor Type	0.153	0.096	**0.153**	**0.096**	**0.137**	**0.094**	**0.139**	**0.113**

	SER	**Δ Dice**	**SER**	**Δ Dice**	**SER**	**Δ Dice**	**SER**	**Δ Dice**
Segmentation	1.133	0.034	**1.127**	**0.035**	**1.121**	**0.032**	**1.113**	**0.030**

(b) Age

	Baseline		U-Net		Pix2Pix		SDE	
	EODD	**EOP**	**EODD**	**EOP**	**EODD**	**EOP**	**EODD**	**EOP**
EC	0.229	0.197	0.219	0.188	0.210	0.172	**0.227**	**0.189**
Cardiomegaly	0.127	0.081	0.120	0.069	**0.127**	**0.079**	0.136	**0.084**
Lung Opacity	0.157	0.092	0.148	0.086	0.148	0.083	**0.156**	0.085
Lung Lesion	0.262	0.191	0.244	0.159	0.236	0.173	**0.255**	**0.180**
Edema	0.122	0.068	**0.119**	**0.064**	**0.120**	**0.068**	0.113	0.055
Consolidation	0.115	0.070	**0.113**	**0.076**	**0.113**	**0.069**	**0.119**	**0.078**
Pneumonia	0.232	0.190	0.250	0.218	**0.222**	**0.203**	**0.241**	**0.198**
Atelectasis	0.161	0.097	**0.158**	**0.093**	**0.156**	**0.095**	**0.162**	**0.092**
Pneumothorax	0.057	0.019	**0.055**	**0.015**	**0.048**	**0.021**	**0.056**	**0.026**
Pleural Effusion	0.083	0.050	0.079	0.044	**0.080**	0.044	0.089	**0.049**
Pleural Other	0.224	0.166	0.247	0.214	**0.211**	**0.185**	**0.235**	0.193
Fracture	0.325	0.282	0.314	**0.276**	**0.320**	**0.284**	0.307	**0.271**
Tumor Grade	0.211	0.148	**0.211**	**0.148**	**0.181**	**0.090**	**0.198**	**0.113**
Tumor Type	0.419	0.415	0.419	0.415	**0.238**	**0.202**	**0.278**	**0.281**

	SER	**Δ Dice**	**SER**	**Δ Dice**	**SER**	**Δ Dice**	**SER**	**Δ Dice**
Segmentation	1.235	0.058	**1.218**	**0.058**	**1.239**	**0.061**	**1.221**	**0.057**

(c) Race

	Baseline		U-Net		Pix2Pix		SDE	
	EODD	**EOP**	**EODD**	**EOP**	**EODD**	**EOP**	**EODD**	**EOP**
EC	0.284	0.347	**0.297**	**0.348**	**0.282**	**0.349**	**0.304**	**0.360**
Cardiomegaly	0.205	0.182	**0.185**	**0.174**	**0.160**	**0.165**	**0.204**	**0.180**
Lung Opacity	0.148	0.135	**0.164**	**0.126**	**0.155**	**0.119**	**0.170**	**0.141**
Lung Lesion	0.360	0.495	**0.382**	**0.481**	**0.307**	**0.390**	**0.373**	**0.496**
Edema	0.136	0.120	**0.147**	**0.125**	**0.151**	**0.129**	**0.122**	**0.125**
Consolidation	0.200	0.263	**0.278**	**0.403**	**0.263**	**0.387**	**0.199**	**0.262**
Pneumonia	0.226	0.291	0.309	**0.384**	**0.223**	**0.274**	**0.253**	**0.305**
Atelectasis	0.204	0.212	**0.221**	**0.213**	**0.215**	**0.209**	**0.224**	**0.229**
Pneumothorax	0.217	0.259	**0.222**	**0.269**	**0.238**	**0.267**	**0.206**	**0.263**
Pleural Effusion	0.097	0.075	**0.110**	0.103	**0.096**	**0.087**	**0.094**	**0.085**
Pleural Other	0.252	0.309	0.297	**0.314**	**0.254**	**0.305**	**0.265**	**0.355**
Fracture	0.479	0.738	**0.440**	**0.586**	**0.479**	**0.740**	**0.491**	**0.731**

Table 6: Maximum PSNR difference between subgroups in percent and its significance for the UCSF-PDSM dataset. While statistically significant, the difference is marginal.

	U-Net		**Pix2Pix**		**SDE**	
	%	p-**value**	%	p-**value**	%	p-**value**
Age	0.22	0.367	0.45	0.27	0.77	0.002
Sex	1.74	0.003	1.01	0.112	2.21	0.198

Table 7: Maximum PSNR difference between subgroups in percent and its significance for the CheXpert dataset. While statistically significant, the difference is marginal.

	U-Net		**Pix2Pix**		**SDE**	
	%	p-**value**	%	p-**value**	%	p-**value**
Age	0.58	0	0.47	0	0.65	0
Sex	0.18	0	0.70	0	0.67	0
Race	1.57	0	2.78	0	2.63	0