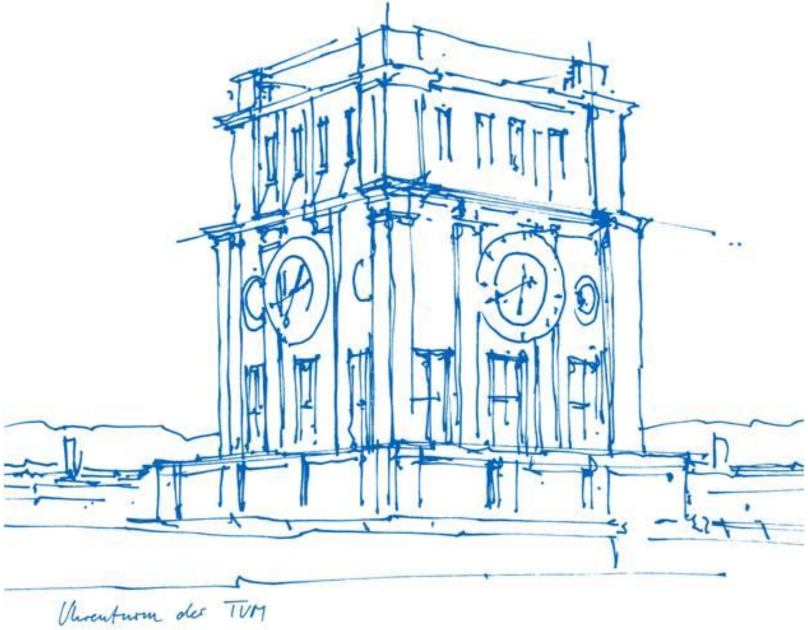
Exploring and Mitigating Bias in Deep-Learning-Based Medical Image Reconstruction

Matteo Wohlrapp

Supervisor: Niklas Bubeck

IDP and Guided Research Final Presentation

Munich, April 16 2025



Agenda

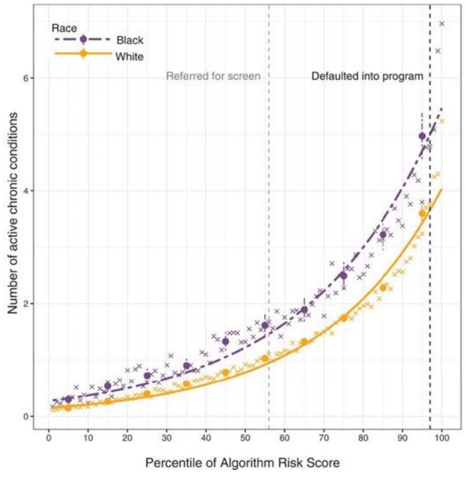
- Fairness Evaluation
  - Method
  - Performance Results
  - Fairness Results
- Bias Mitigation
  - Method
  - Results

Agenda

- Fairness Evaluation
  - Method
  - Performance Results
  - Fairness Results
- Bias Mitigation
  - Method
  - Results

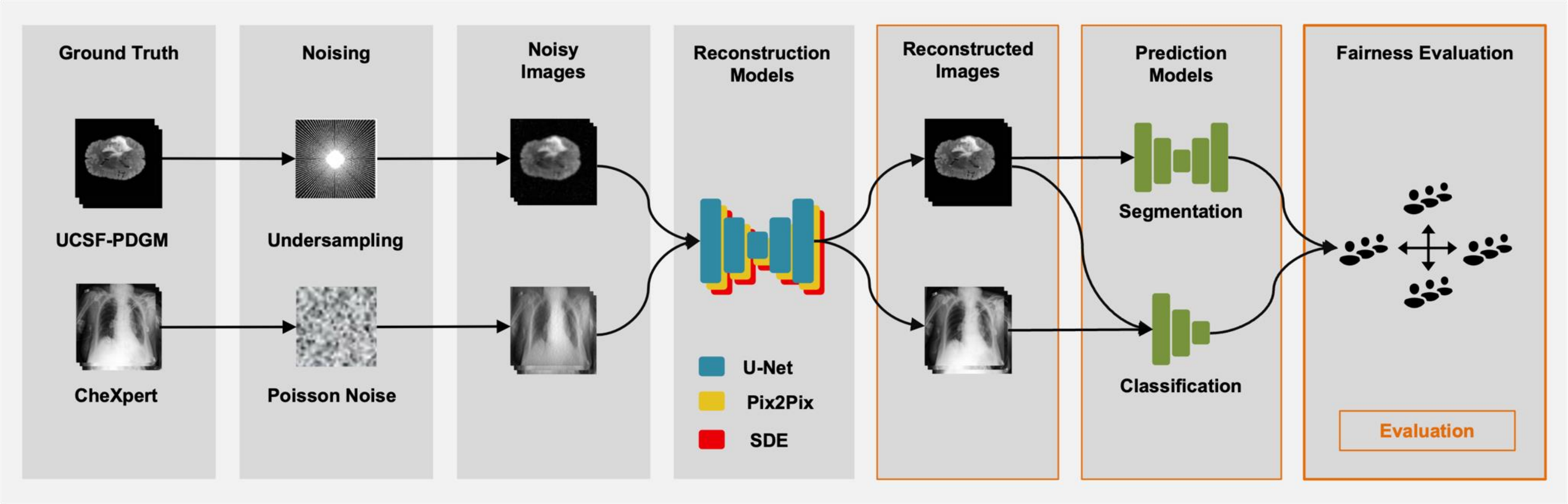
Bias in Reconstruction Models Is Underexplored

- Many pre-trained models are available
- Studies show diagnostic disparities across subgroups
- Extensive research for classification and segmentation
- Limited attention to reconstruction models



Ziad Obermeyer et al.,Dissecting racial bias in an algorithm used to manage the health of populations.Science366,447-453(2019).

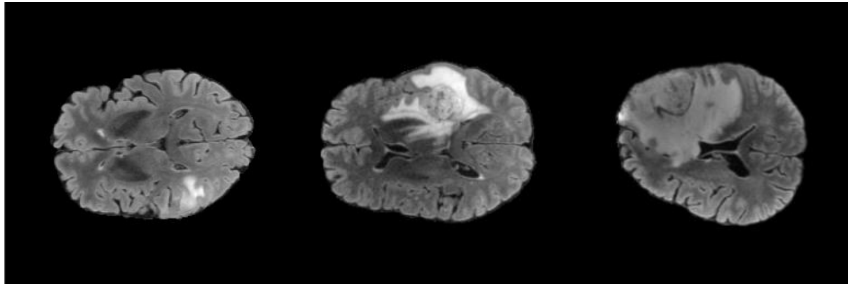
Fairness Influence on Downstream Prediction Models



Datasets From Two Modalities: MRI and X-Ray

UCSF-PDGM

- 501 diffuse glioma cases, FLAIR images
- Several clinical variables, segmentation masks
- Attributes: age (categorical; median 58), sex



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. 2019. *Chexpert: a large chest radiograph dataset with uncertainty labels andxpert comparison. AAAI*

CheXpert

- 224K chest radiographs,
- 14 thoracic observations
- Attributes: age (categorical; median 62), sex, race

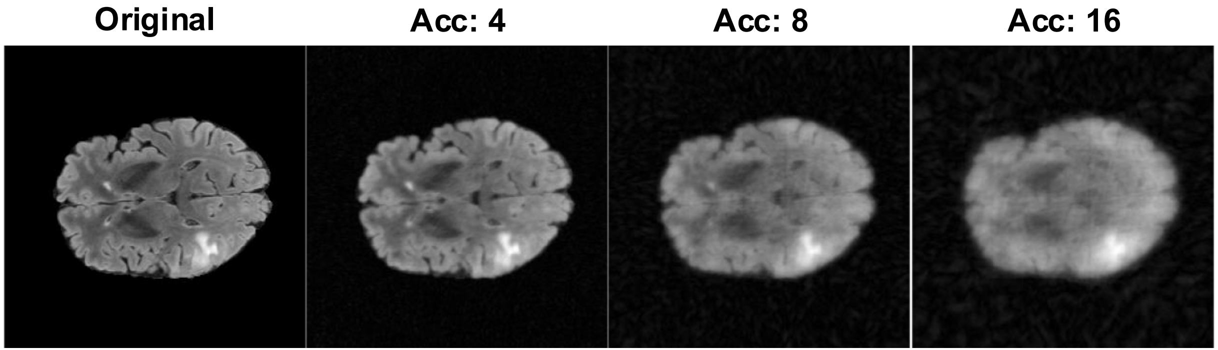
Calabrese, E., Villanueva-Meyer, J.E., Rudie, J.D., Rauschecker, A.M., Baid, U., Bakas, S., Cha, S., Mongan, J.T., Hess, C.P. 2022. *The university of california san francisco preoperative diffuse glioma mri dataset. Radiology: Artificial Intelligence*

Approximating Realistic Noise

MRI

Radial masking of complex frequency space (k-space) to simulate undersampling

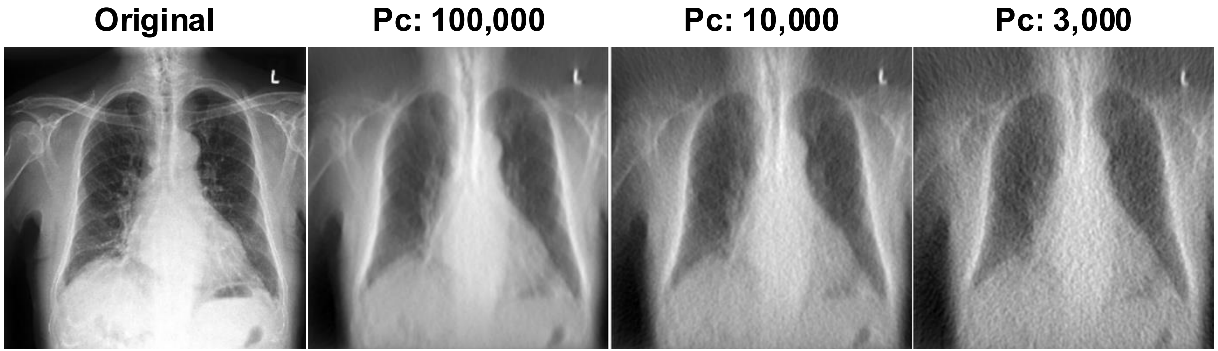
Feng, L.: Golden-angle radial mri: Basics, advances, and applications. Journal of Magnetic Resonance Imaging 56 (04 2022)



X-Ray

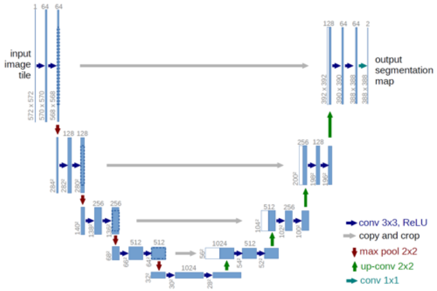
Radon transform followed by Bowtie filter and addition of Poisson noise to simulate electron interference

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).



Classical to Generative Reconstruction Models

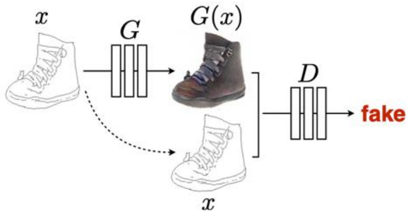
U-Net



Fully convolutional network for image restoration

O. Ronneberger, P. Fischer, T. Brox (2015). U-Net: Convolutional Networks for Biomedical Image Segmentation. MICCAI

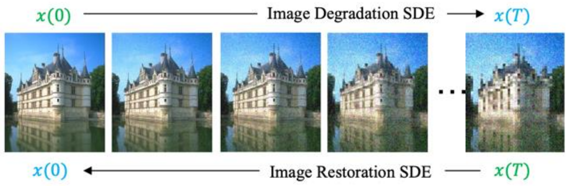
Pix2Pix



Conditional Generative Adversarial Network (GAN) for image-to-image translation

P. Isola, J. -Y. Zhu, T. Zhou, A. A. Efros. (2017). Image-to-Image Translation with Conditional Adversarial Networks. CVPR

SDE



Mean-reverting Stochastic Differential Equations (SDEs)

Z. Luo, F. Gustafsson, Z. Zhao, J. Sjölund, T. Schön. (2023). Image Restoration with Mean-Reverting Stochastic Differential Equations. ICML

Evaluating Performance and Fairness Metrics

Performance Metrics

Reconstruction:
PSNR, LPIPS

Classification:
AUROC

Segmentation:
Dice

Fairness Metrics

Classification

Equalized Odds (EODD) [1] :

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

Equality of Opportunity (EOP) [1] :

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

Segmentation

Skewed Error Ratio (SER) [2] : $SER_{\mathrm{A}} = \dfrac{max_{A \in \mathrm{A}}(1 - Dice_A)}{min_{B \in \mathrm{A}}(1 - Dice_B)}$

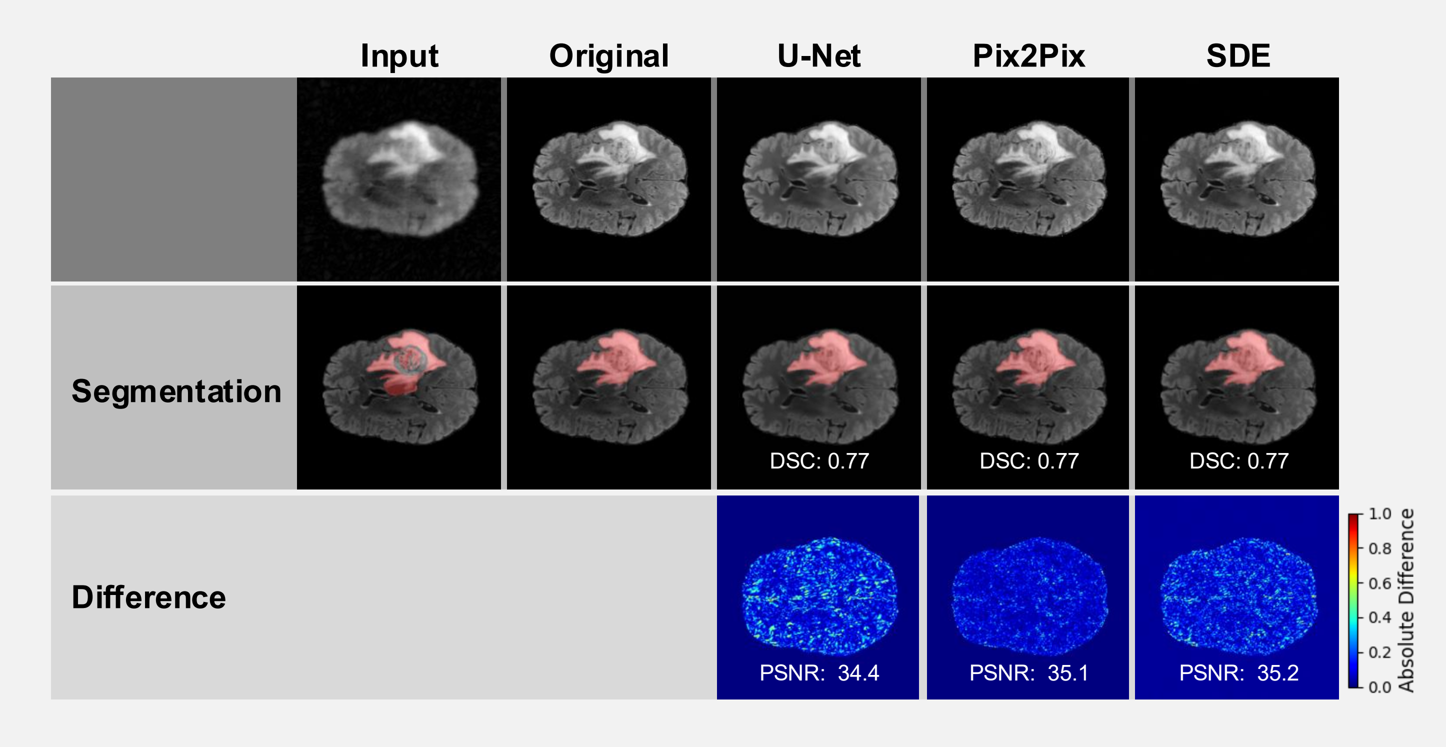
Delta Dice: $\Delta Dice = max_{A,B \in \mathrm{A}} |Dice_A - Dice_B|$

[1]M. Hardt et al. (2016). Equality of opportunity in supervised learning. [2]I. Siddiqui et al. (2024). Fair ai-powered orthopedic image segmentation: addressing bias and promoting equitable healthcare. Scientific Reports

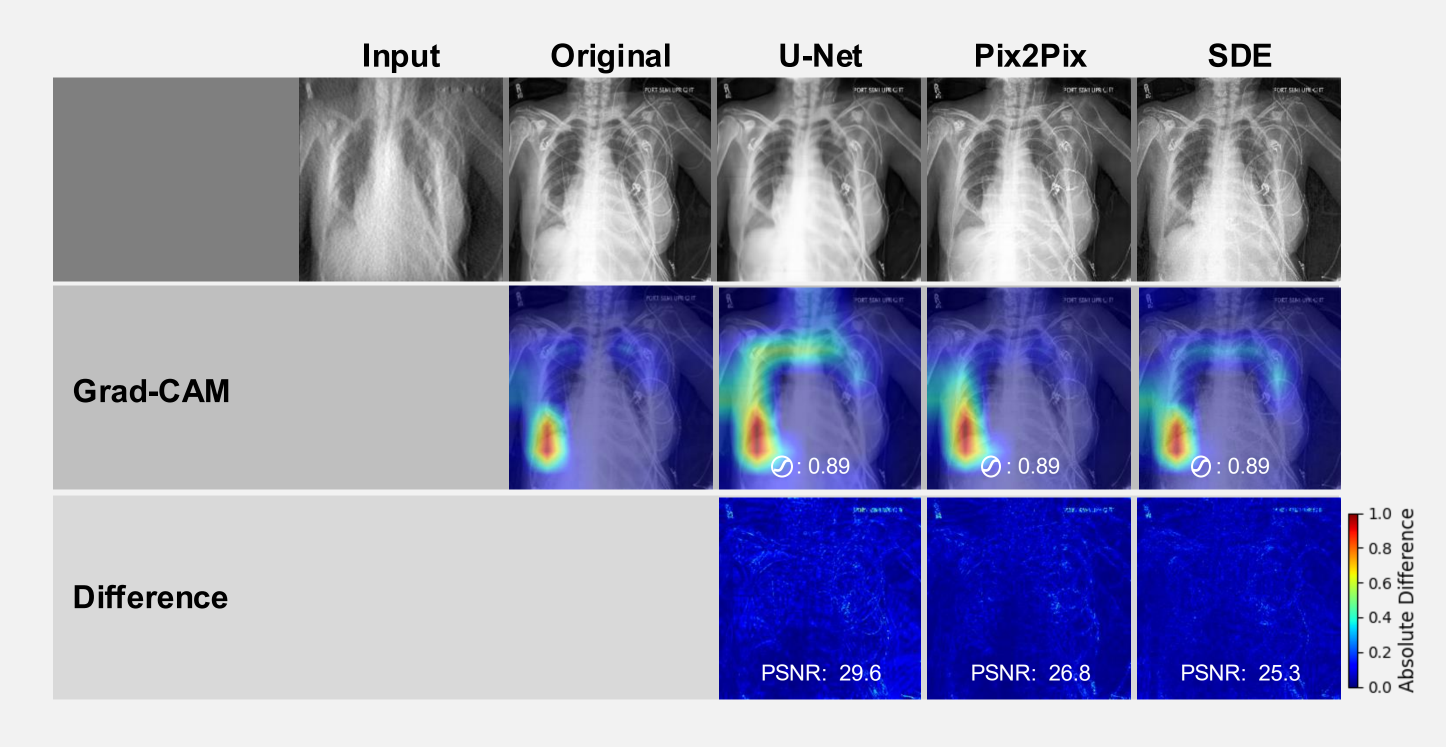
Agenda

- Fairness Evaluation
  - Method
  - Performance Results
  - Fairness Results
- Bias Mitigation
  - Method
  - Results

Similar Appearance Across Models for UCSF-PDGM

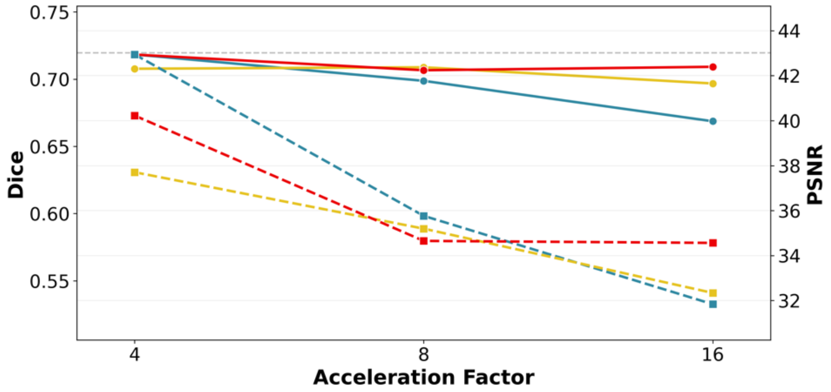


More Variation for CheXpert

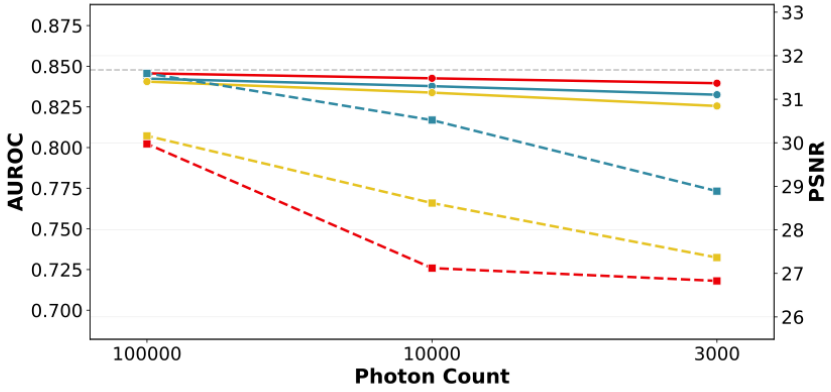


Unlike Downstream Performance, Image Quality Drops

U-Net Pix2Pix SDE AUROC PSNR Baseline

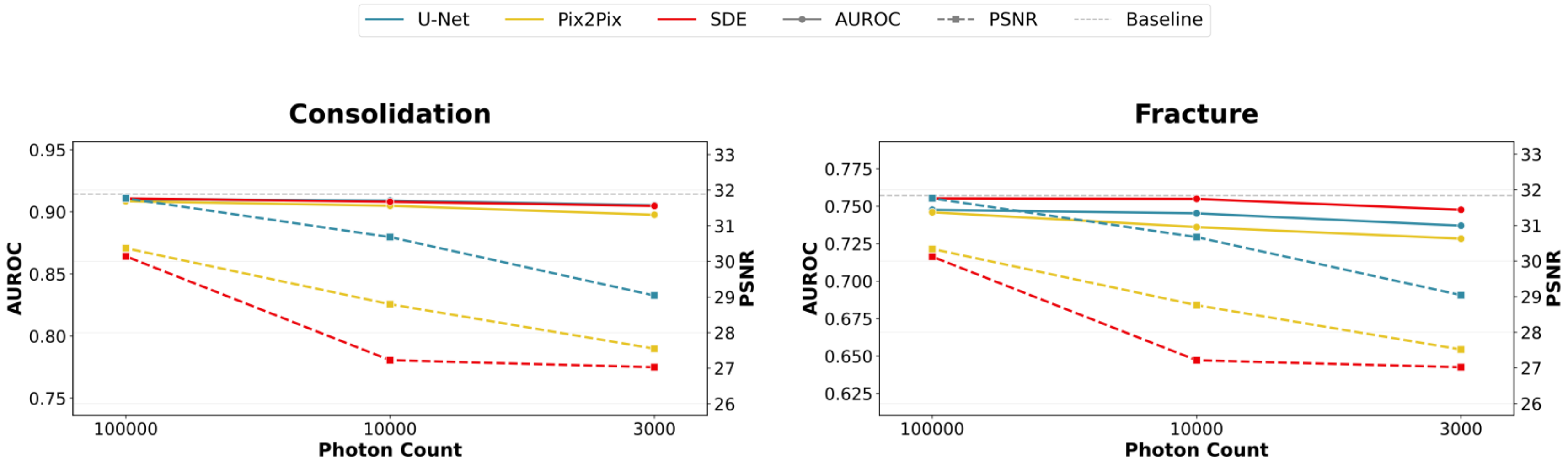


UCSF-PDGM segmentation



Average CheXpert classification

Classifiers of Subtle Pathologies Are More Affected



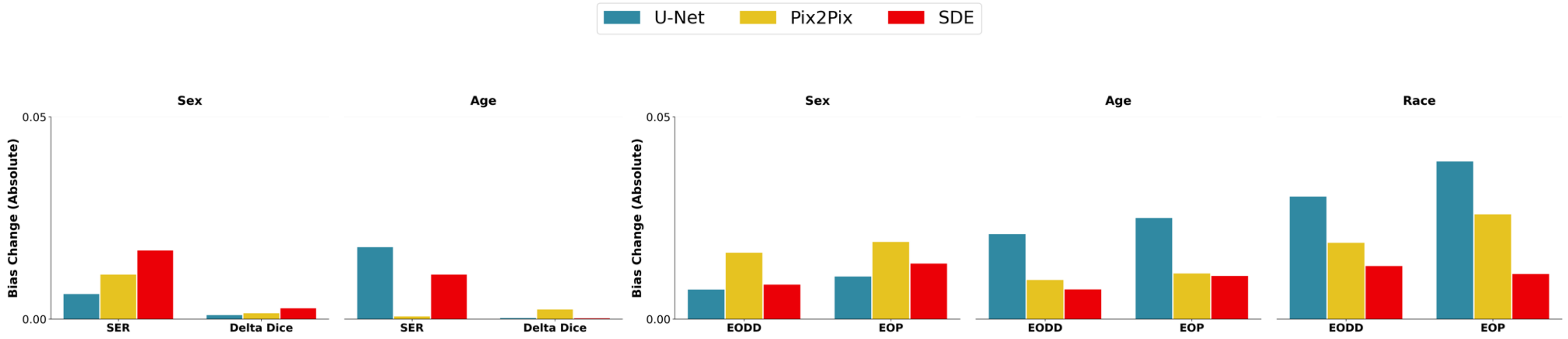
Pathology with higher baseline performance

Pathology with lower baseline performance

Agenda

- Fairness Evaluation
  - Method
  - Performance Results
  - Fairness Results
- Bias Mitigation
  - Method
  - Results

Reconstruction Adds Little Change to Fairness



Absolute bootstrapped bias change for UCSF-PDGM segmentation

Absolute bootstrapped bias change averaged across all classifications

Change Is Still Significant Depending on the Attribute

	Baseline EODD	Baseline EOP	U-Net EODD	U-Net EOP	Pix2Pix EODD	Pix2Pix EOP	SDE EODD	SDE 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

+, $p < 0.05$ +, $0.05 \leq p < 0.1$ −, $p < 0.05$ −, $0.05 \leq p < 0.1$

Bold indicates standard error larger than absolute effect size

Fairness results for attribute sex

	Baseline EODD	Baseline EOP	U-Net EODD	U-Net EOP	Pix2Pix EODD	Pix2Pix EOP	SDE EODD	SDE 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

+, $p < 0.05$ +, $0.05 \leq p < 0.1$ −, $p < 0.05$ −, $0.05 \leq p < 0.1$

Bold indicates standard error larger than absolute effect size

Fairness results for attribute age

Bias Can Be Subject to High Variance

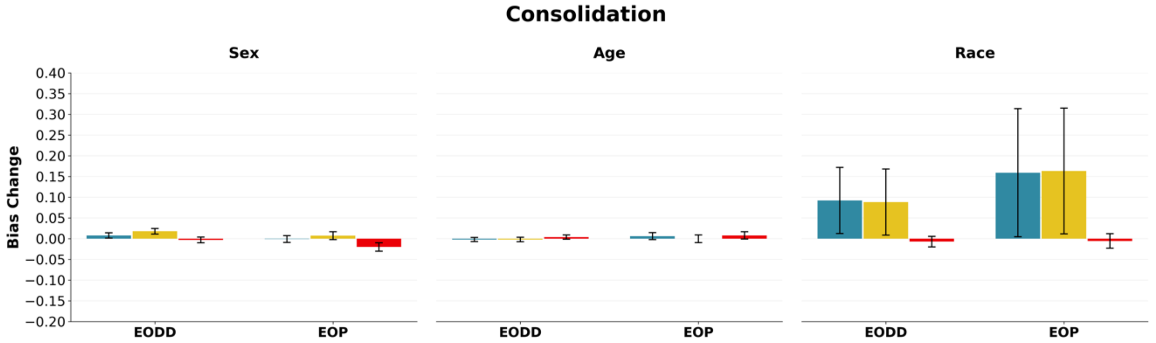
	Baseline		U-Net		Pix2Pix		SDE	
	EODD	EOP	EODD	EOP	EODD	EOP	EODD	EOP
EC	0.284	0.347	**0.297**	**0.348**	**0.282**	**0.349**	**0.304**	**0.360**
Cardiomegaly	0.205	0.182	**0.185**	**0.174**	**0.160**	**0.165**	**0.204**	**0.180**
Lung Opacity	0.148	0.135	**0.164**	**0.126**	**0.155**	**0.119**	**0.170**	**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**

+, $p < 0.05$ +, $0.05 \leq p < 0.1$ −, $p < 0.05$ −, $0.05 \leq p < 0.1$

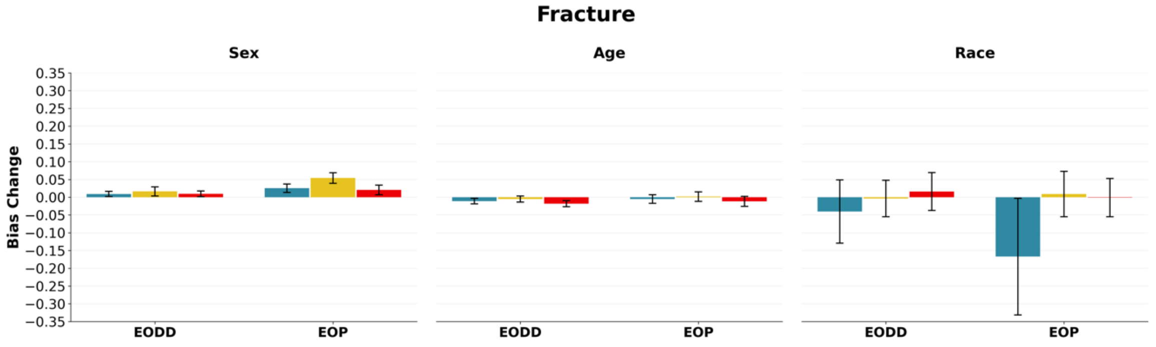
Bold indicates standard error larger than absolute effect size

Fairness results for attribute race

Performance Trend of Pathologies Does Not Continue



Pathology with higher baseline performance



Pathology with lower baseline performance

U-Net Pix2Pix SDE

Low PSNR Difference Contradicts Previous Work[1,2]

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

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

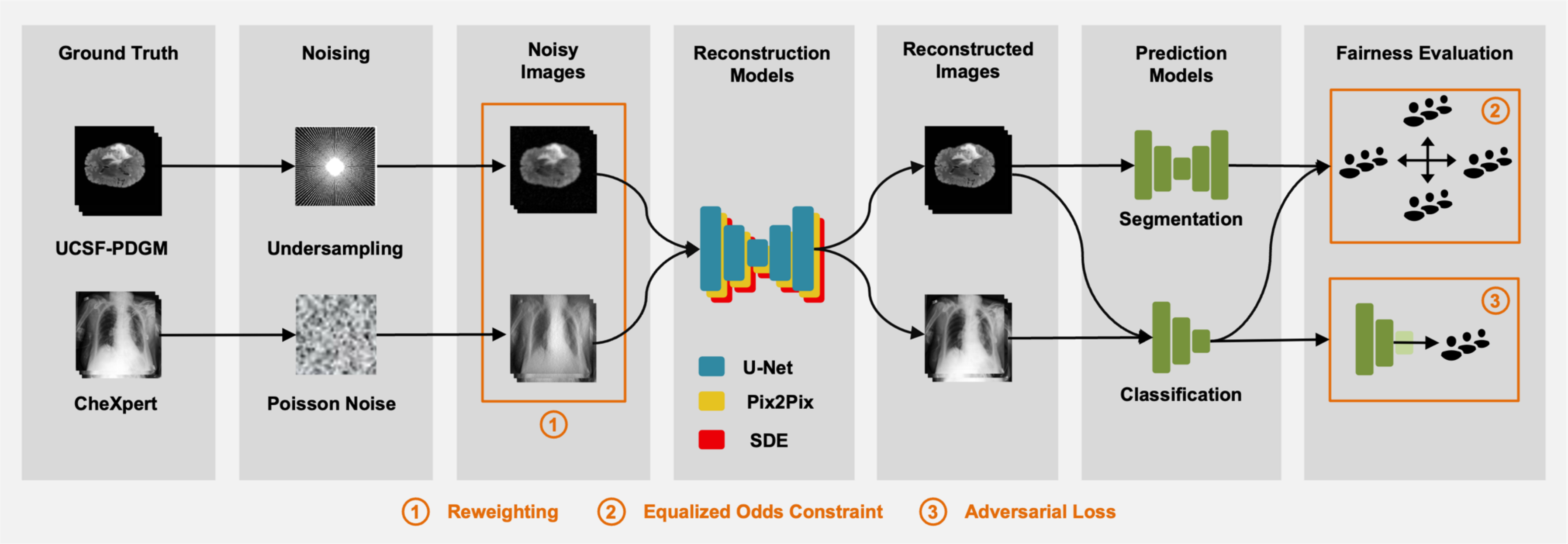
Maximum difference in PSNR values and significance between demographic subgroups for UCSF-PDGM (left) and CheXpert (right)

[1]M. Du. (2023). Unveiling fairness biases in deep learning-based brain mri reconstruction. Clinical Image-Based Procedures, [2]Sheng, Y. (2024). Toward fair ultrasound computing tomography: Challenges, solutions and outlook. Great Lakes Symposium

Agenda

- Fairness Evaluation
  - Method
  - Performance Results
  - Fairness Results
- Bias Mitigation
  - Method
  - Results

Bias Mitigation for Reconstruction Models



Adapting Techniques From Classification Models

Reweighting

$$p_i = \frac{\frac{1}{n_{(g_i^1,\dots,g_i^K)}}}{\sum_{j=1}^{n} \frac{1}{n_{(g_j^1,\dots,g_j^K)}}}$$

where $n_{(g_i^1,\dots,g_i^K)}$ is the number of samples with the exact same sensitive attributes $1,\dots,K$ as sample i

Equalized Odds Constraint

$$EODD = \frac{1}{2}[\mathrm{E}[\hat{y}_i | a_i = 0, y_i = y] - |\mathrm{E}[\hat{y}_i | a_i = 1, y_i = y]]$$

where $y \in \{0, 1\}$, and $\hat{y}_i = \sigma\left(\frac{f_\theta(x_i)-\tau}{T}\right)$, with temperature T, and threshold τ

Marcinkevics, R., Ozkan, E., Vogt, J.E.: Debiasing deep chest X-ray classifiers using intra- and post-processing methods. ML4Health. 2022

Adversarial Loss

$$\mathrm{ADV} = Corr^2(h_\theta(f_\theta(x_i)), a_i)$$

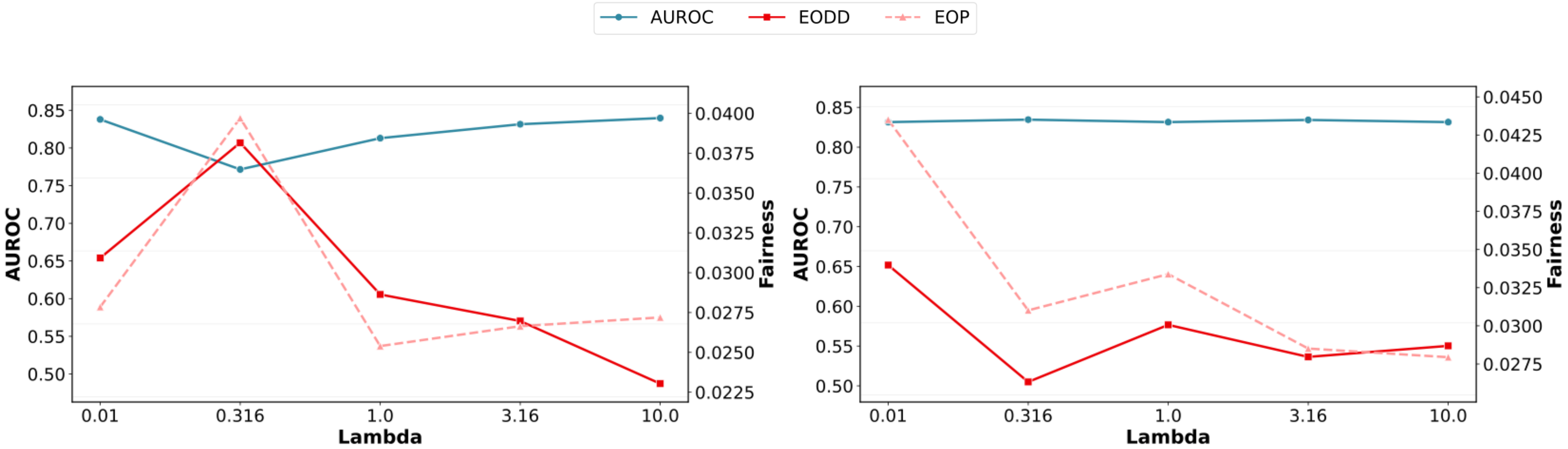
where $Corr^2(u, v) = \left(\frac{\sum_i (u_i - \bar{u})(u_i - \bar{u})}{\sqrt{\sum_i (u_i - \bar{u})^2 \sum_i (v - \bar{v})^2 + \varepsilon}}\right)^2$ is the Pearson Correlation Coefficient, and $\bar{u}$, $\bar{v}$ are the sample means

Adeli E, Zhao Q, Pfefferbaum A, Sullivan EV, Fei-Fei L, Niebles JC, Pohl KM. Representation Learning with Statistical Independence to Mitigate Bias. IEEE CV. 2021

Agenda

- Fairness Evaluation
  - Method
  - Performance Results
  - Fairness Results
- Bias Mitigation
  - Method
  - Results

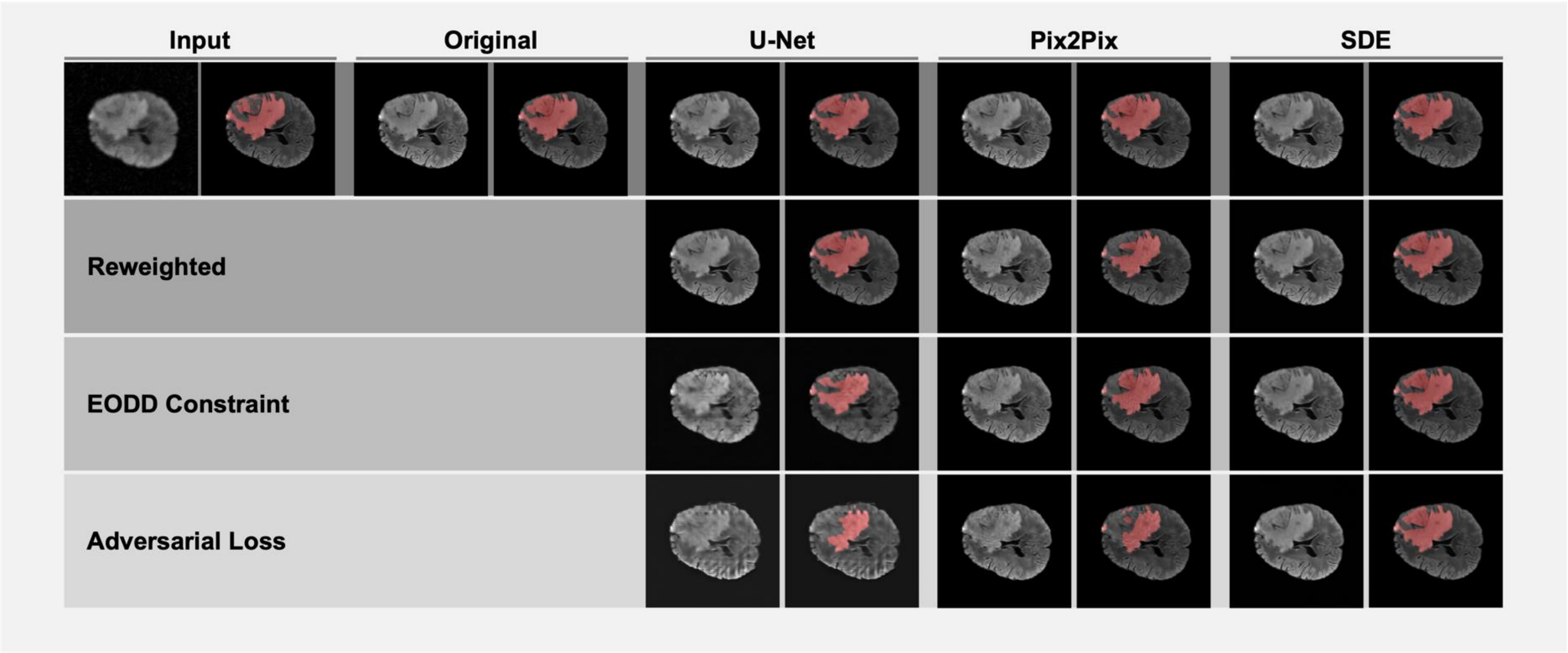
Mitigation Is Little Sensitive to Lambda Values



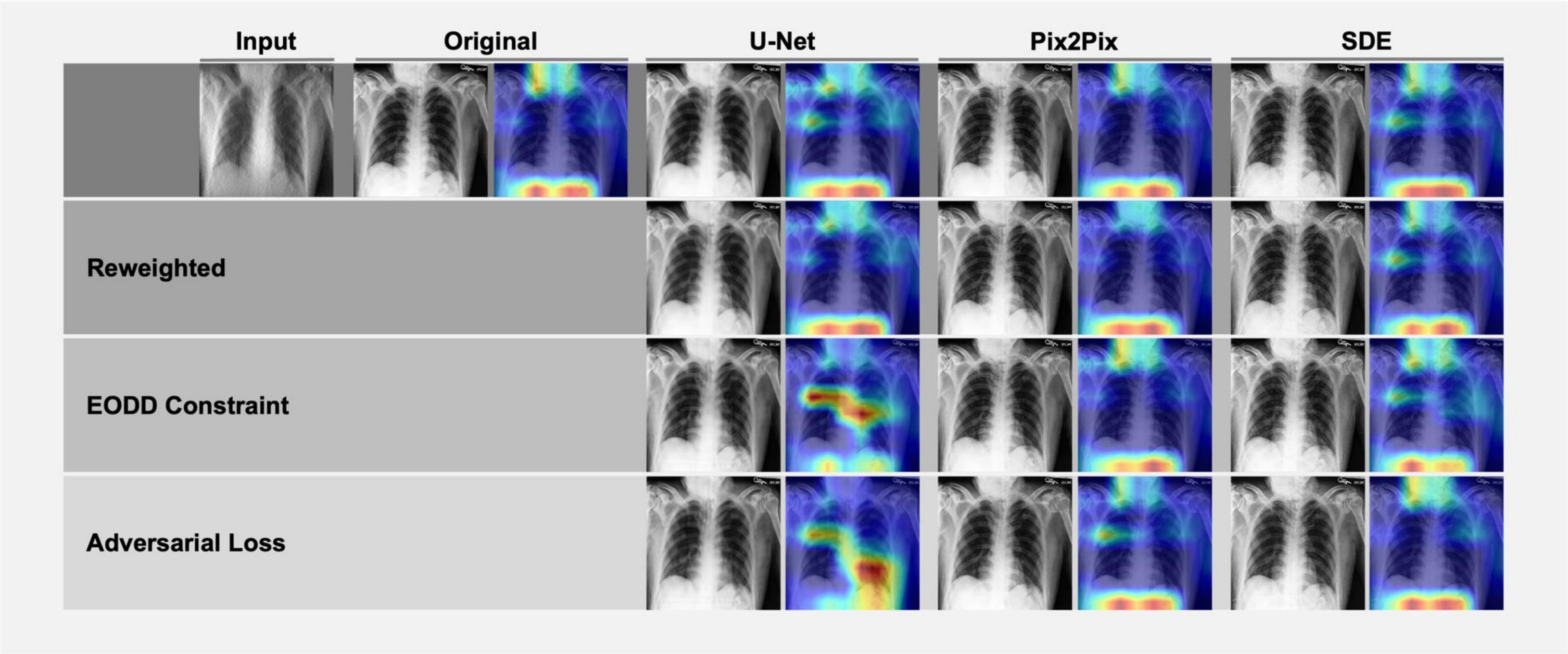
AUROC for different lambdas for the sensitive attribute gender with the EODD fairness constraint

AUROC for different lambdas for the sensitive attribute gender with the adversarial fairness constraint

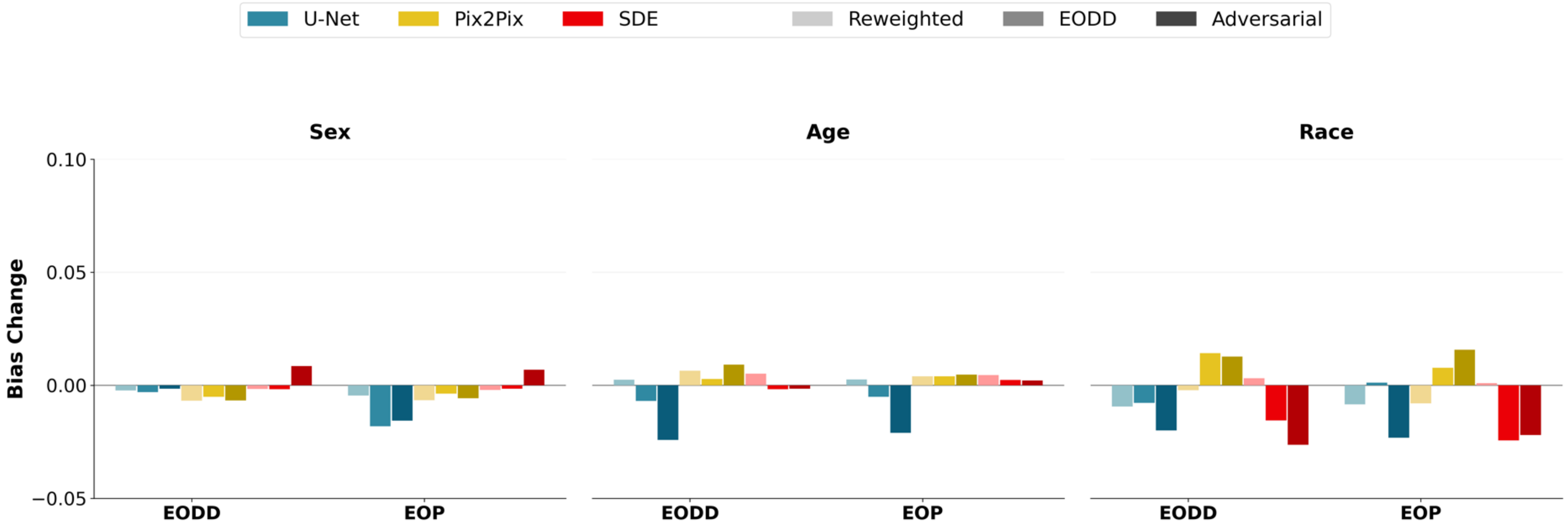
U-Net on UCSF-PDGM Loses Performance



CheXpert Results Are Less Affected

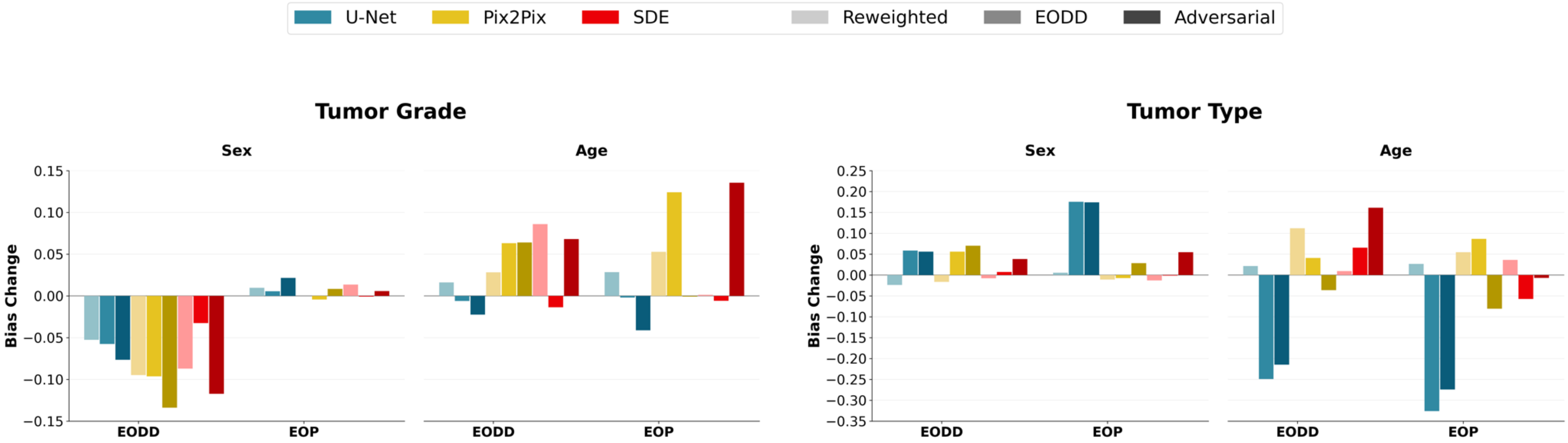


Mitigation Slightly Decreases Bias for CheXpert



Average bias change for all classifiers on the CheXpert dataset

UCSF-PDGM Shows No Clear Trend



Tumor grade bias change for the different mitigation techniques

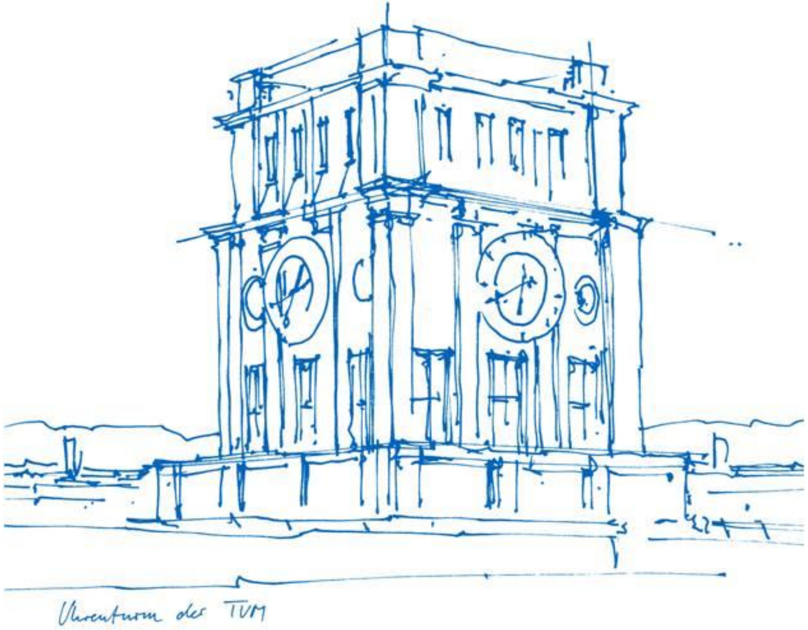
Tumor type bias change for the different mitigation techniques

Downstream Predictors Not Elastic to Reconstruction

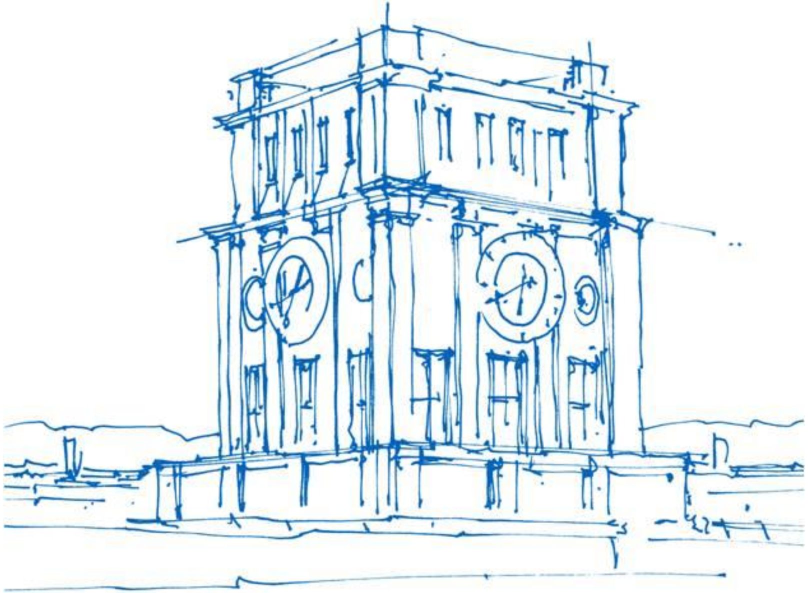
- Downstream prediction models are robust to changes in image quality
- Overall, reconstruction has little but significant effect
- Relative importance depends on the sensitive attribute
- At times, additional bias can be big

- Equalized odds constraint and adversarial loss seem to provide slight mitigation but depend on the dataset
- Overall, it seems like the ’elasticity’ of the reconstruction models is too small, i.e., there is not enough change introduced to make a difference

Thanks/
Questions?

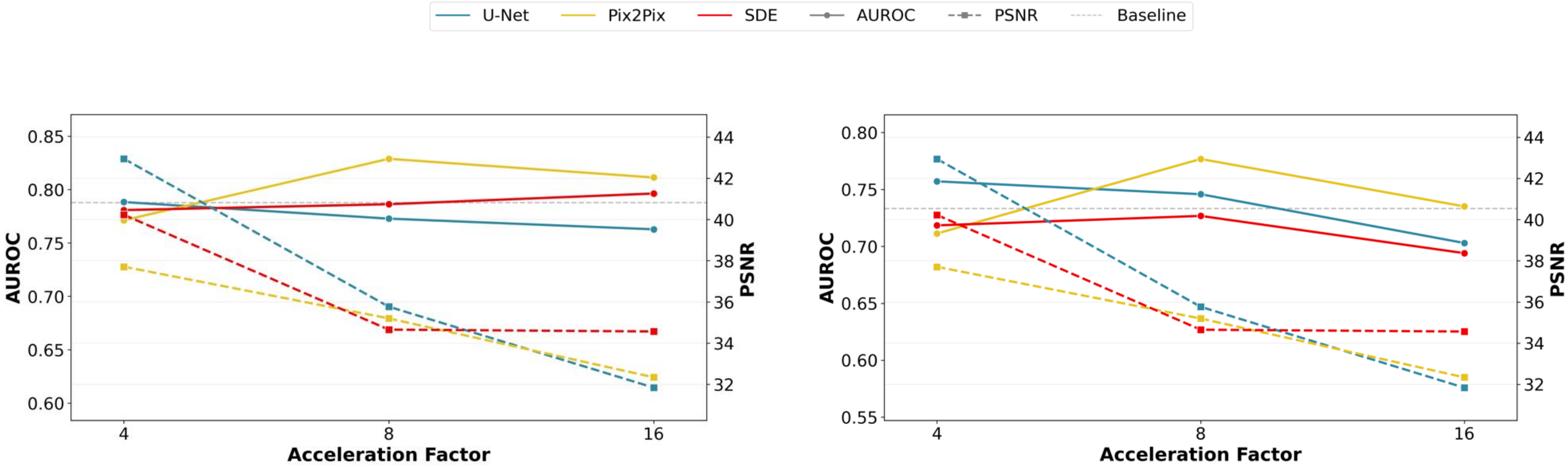


Appendix



Uhrenturm der TUM

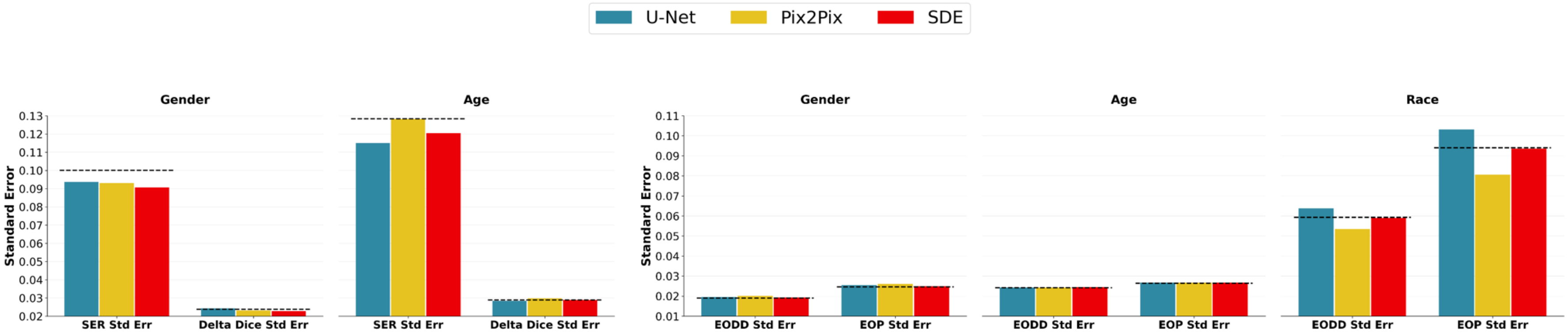
UCSF-PDGM Classification Performance Is Very Similar



Tumor type

Tumor grade

Reconstruction Has No Influence on Variance



UCSF-PDGM segmentation standard error

Classification average standard error

U-Net and Pix2Pix Show Worse Performance for UCSF

Metrics		Baseline	U-Net STD	U-Net RE	U-Net EODD	U-Net ADV
AUROC	Tumor Type	0.788	0.773	0.767	0.753	0.807
	Tumor Grade	0.733	0.746	0.745	0.721	0.735
Dice			0.699	0.701	0.606	0.563
PSNR			35.766	36.185	29.660	25.918
LPIPS			0.030	0.029	0.109	0.103

↓> 0.1 +, 0.05 ≤↓< 0.1 ↑< −0.1 −0.1 <↑≤ −0.05

Metrics		Baseline	Pix2Pix STD	Pix2Pix RE	Pix2Pix EODD	Pix2Pix ADV
AUROC	Tumor Type	0.788	0.829	0.775	0.763	0.723
	Tumor Grade	0.733	0.777	0.740	0.745	0.710
Dice			0.709	0.697	0.696	0.679
PSNR			35.198	34.204	34.012	31.545
LPIPS			0.022	0.028	0.028	0.049

↓> 0.1 +, 0.05 ≤↓< 0.1 ↑< −0.1 −0.1 <↑≤ −0.05

Metrics		Baseline	SDE STD	SDE RE	SDE EODD	SDE ADV
AUROC	Tumor Type	0.788	0.786	0.780	0.778	0.824
	Tumor Grade	0.733	0.727	0.733	0.737	0.783
Dice			0.707	0.705	0.707	0.662
PSNR			34.654	34.443	34.388	35.035
LPIPS			0.016	0.017	0.017	0.014

↓> 0.1 +, 0.05 ≤↓< 0.1 ↑< −0.1 −0.1 <↑≤ −0.05

CheXpert Performance Is Not Affected

Metrics		Baseline	U-Net STD	U-Net RE	U-Net EODD	U-Net ADV
AUROC	Atelectasis	0.872	0.865	0.866	0.864	0.854
	Cardiomegaly	0.909	0.904	0.905	0.902	0.898
	Consolidation	0.914	0.909	0.910	0.904	0.900
	Edema	0.899	0.892	0.892	0.890	0.889
	EC	0.788	0.782	0.782	0.781	0.779
	Fracture	0.757	0.745	0.747	0.749	0.746
	Lung Lesion	0.796	0.780	0.780	0.783	0.765
	Lung Opacity	0.885	0.876	0.877	0.874	0.869
	Pleural Effusion	0.925	0.917	0.917	0.915	0.906
	Pleural Other	0.828	0.813	0.813	0.810	0.796
	Pneumonia	0.833	0.823	0.824	0.822	0.802
	Pneumothorax	0.767	0.747	0.746	0.760	0.765
	Average	0.848	0.838	0.838	0.838	0.831
PSNR			30.521	30.447	29.404	29.153
LPIPS			0.185	0.193	0.178	0.182

↓> 0.1 +, 0.05 ≤↓< 0.1 ↑< −0.1 −0.1 <↑≤ −0.05

Metrics		Baseline	Pix2Pix STD	Pix2Pix RE	Pix2Pix EODD	Pix2Pix ADV
AUROC	Atelectasis	0.872	0.858	0.860	0.862	0.862
	Cardiomegaly	0.909	0.902	0.904	0.904	0.905
	Consolidation	0.914	0.905	0.906	0.905	0.907
	Edema	0.899	0.891	0.893	0.891	0.893
	EC	0.788	0.781	0.782	0.782	0.782
	Fracture	0.757	0.736	0.744	0.742	0.743
	Lung Lesion	0.796	0.780	0.781	0.782	0.783
	Lung Opacity	0.885	0.871	0.873	0.873	0.874
	Pleural Effusion	0.925	0.912	0.913	0.913	0.914
	Pleural Other	0.828	0.798	0.807	0.810	0.807
	Pneumonia	0.833	0.818	0.817	0.822	0.820
	Pneumothorax	0.767	0.752	0.757	0.757	0.758
	Average	0.848	0.834	0.836	0.837	0.837
PSNR			28.615	28.797	28.448	28.859
LPIPS			0.109	0.103	0.109	0.103

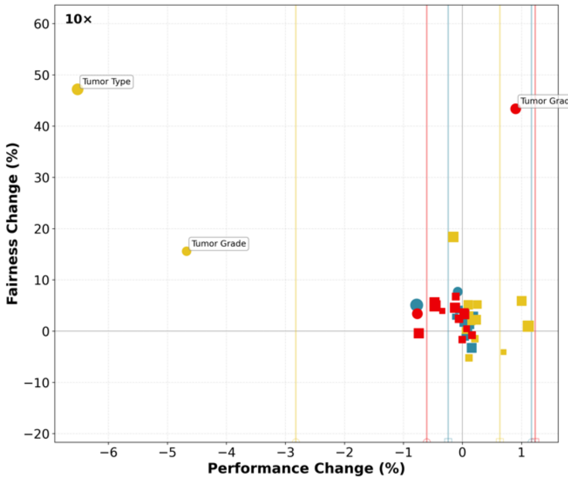
↓> 0.1 +, 0.05 ≤↓< 0.1 ↑< −0.1 −0.1 <↑≤ −0.05

CheXpert Performance Is Not Affected

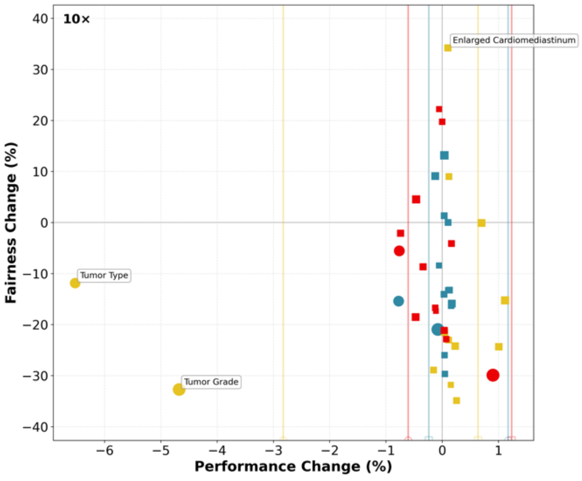
Metrics		Baseline	SDE			
			STD	RE	EODD	ADV
AUROC	Atelectasis	0.872	0.865	0.865	0.867	0.861
	Cardiomegaly	0.909	0.905	0.907	0.905	0.901
	Consolidation	0.914	0.908	0.908	0.910	0.905
	Edema	0.899	0.896	0.895	0.896	0.893
	EC	0.788	0.784	0.784	0.787	0.781
	Fracture	0.757	0.755	0.751	0.752	0.744
	Lung Lesion	0.796	0.790	0.784	0.791	0.788
	Lung Opacity	0.885	0.877	0.877	0.878	0.875
	Pleural Effusion	0.925	0.917	0.918	0.920	0.915
	Pleural Other	0.828	0.819	0.815	0.816	0.810
	Pneumonia	0.833	0.825	0.824	0.825	0.819
	Pneumothorax	0.767	0.770	0.767	0.768	0.757
	Average	0.848	0.843	0.841	0.843	0.837
PSNR			27.121	27.456	27.752	27.112
LPIPS			0.149	0.101	0.110	0.143

↓> 0.1   +, 0.05 ≤↓< 0.1   ↑< −0.1   −0.1 <↑≤ −0.05

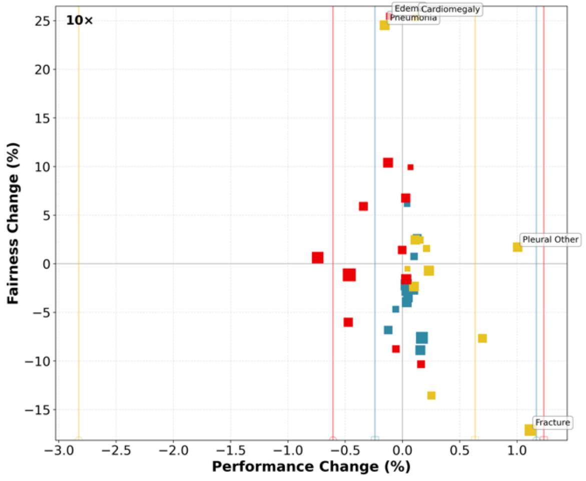
Age Shows Less Variance Than Sex and Race



Attribute age, with EODD fairness

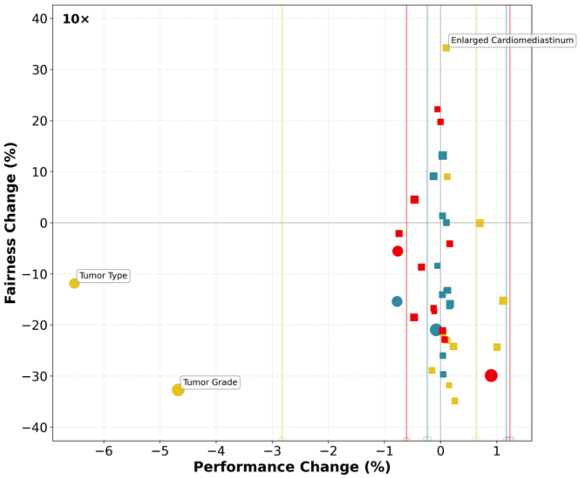


Attribute sex , with EODD fairness

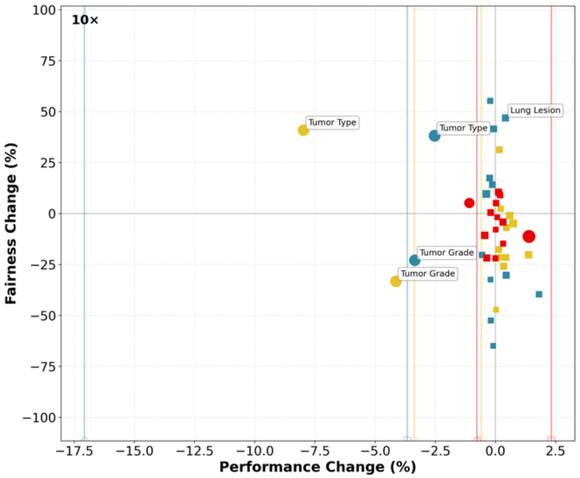


Attribute race , with EODD fairness

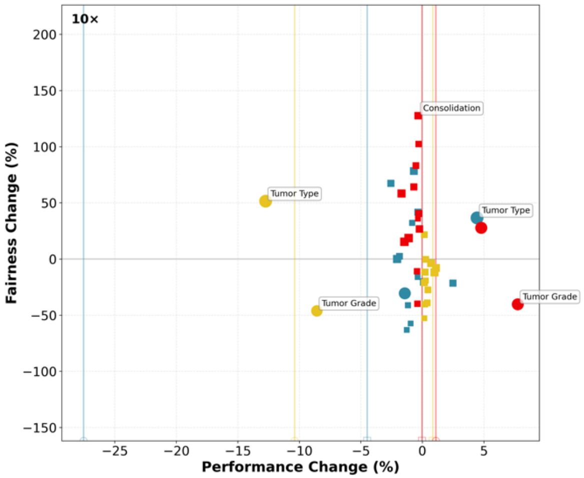
Adversarial and EODD Constraint Show Higher Variance



Reweighting



EODD constraint



Adversarial loss