Fairness Of AI Models in vector embedded Chest X-ray representations : Supplemental material

Appendix A1.Dataset Distribution

Table A1 provides an overview of the vector embedding dataset, including descriptions, image counts, patient numbers, and their distribution across protected attributes such as sex, age, race, and insurance types. The dimensions of MIMIC-CXR vector embeddings and image based model are different likely due to omission of lateral images in generating vector embedding.

Subgroup	Attribute	MIMIC	CXP	ALL
	# Images	227,641	219,946	447,587
	# Patients	52,874	63,467	116,341
Sex	Male	54.1 %	59.3 %	56.7 %
	Female	45.9 %	40.7 %	43.3 %
Age	0-20	0.4 %	0.8 %	0.6 %
	20-40	10.6 %	13.1 %	11.8 %
	40-60	30.6 %	31 %	30.8 %
	60-80	42.2 %	39.1 %	40.7 %
	80-	16.3 %	16.1 %	16.2 %
Race	White	66.1 %	63.3 %	64.8 %
	Black	16.1 %	6.1 %	11.4 %
	Asian	3.2 %	11.8 %	7.2 %
	Hispanic	5.5 %	2.4 %	4 %
	Native	0.3 %	1.9 %	1.2 %
	Other	4.7 %	14.6 %	9.3 %
Insurance	Medicare	44.7 %		
	Other	47.1 %		
	Medicaid	8.2 %		

Table A1: An overview of the vector embedding for chest X-ray datasets; MIMIC, CheXpert (CXP), and their aggregation in the comprehensive dataset labeled ALL. It details the quantity of vector-embedded images and patients, along with the proportion of patients categorized by subgroups such as sex, age, race, and insurance type. Notably, Native, Hispanic, and Black refer to self-reported American Indian/Alaska Native, Hispanic/Latino, and Black/African American races, respectively.

Appendix B.Model Details

All disease detection models (i.e. MIMIC-CXR(Emb), CXP(Emb), ALL(Emb), and classification head of ALL(Img)) model features two hidden layers of size 768 and 256 with a dropout rate 0.3, batch size 48, weight decay 0.00001 and a learning rate set at 0.0001. In all three settings, we used the "relu "activation function, followed by batch normalization and dropout layers. The output layer employs the "sigmoid "activation function with BCEWithLogitsLoss "loss function and Adam optimizer with a cosine decay function **?** for the learning rate.

Subsequently, we appended a fully connected layer consisting of 14 neurons as a classification layer per model trained on each dataset. Furthermore, an early stopping mechanism was implemented to monitor validation loss, to halt learning if no improvement was observed over 5 epochs. We report a mean $\pm 95\%$ confidence interval (CI) across five runs with varying seed numbers. The output

Submitted to 38th Conference on Neural Information Processing Systems (NeurIPS 2024). Do not distribute.

consisted of a 14-number array representing the probability of each disease label. A binary prediction threshold for each disease was set based on maximizing the F1 score across all labels.

For race and sex classification, we employed a head model architecture similar to that of disease detection. The output layer produces probability for race and sex classes using a softmax and sigmoid activation function, respectively. We report the AUC among different sex and race (white, black/African American, and Asian) categories.

Appendix C. TPR Disparities

In Figures C1 to C9, we show scatter plots of True Positive Rate (TPR) disparities across all datasets(MIMIC-CXR, CheXpert, and aggregated(ALL) dataset) of all remaining protected attributes and disease labels. The y-axis represents the TPR disparities, while disease labels are depicted on the x-axis. Each point on the scatter plot corresponds to a disease label, with the size of the circles indicating the respective group sizes. The groups with positive TPR disparities are favorable, while groups with negative TPR disparities are unfavorable groups. The values are averaged over five runs, and arrows around the mean indicate a 95% confidence interval. We have sorted the disease labels on the x-axis to appear based on the distance between the least and most favorable subgroups. For a particular disease, the lower the distance, the fairer the model.

Dataset type	**Male**	**Female**	**Asian**	**Black**	**White**
MIMIC(Emb)	**0.993**	**0.993**	**0.882**	**0.906**	**0.898**
MIMIC(Img)	0.998	0.998	0.930	0.970	0.930
CXP(Emb)	**0.986**	**0.986**	**0.867**	**0.871**	**0.854**
CXP(Img)	0.998	0.998	0.970	0.980	0.970
ALL(Emb)	**0.987**	**0.987**	**0.890**	**0.897**	**0.871**
ALL(Img)	0.998	0.998	0.975	0.897	0.871

Table C1: Sex and race detection AUC of models trained on MIMIC, CXP, and All images (Img) vs MIMIC and CXP vector embedding (Emb). Here, the lower the AUC, the better. The race detection results of MIMIC, CXP are from **?**. Race and sex detection in vector embedding is lower vs images, which means there are lower race signals in vector embedding compared to images.

Appendix D. Underdiagnosis

In Figures D1 and D2 we show underdiagnosis rate distribution using vector embeddings compared to the medical image baseline of MIMIC-CXR and the aggregated(ALL) dataset of both. We provide findings pertaining to sex, age, and race subgroups across both datasets. Specifically for the MIMIC-CXR dataset, we present additional insights regarding insurance type. Furthermore, we analyze patient subgroups at the intersection of two and three underserved categories within the MIMIC-CXR dataset. Baseline results of the models based on medical images(i.e. **?** for the MIMIC-CXR dataset and results of our experiment for ALL datasets) are colored in gray in the same figure. Apparently, for the ALL dataset, using vector embedding reduces the gap between groups and generally reduces the underdiagnosis rate, leading to more fairness in the underdiagnosis rate. Also, we plot the intersection of underserved subgroups (the one with the highest FPR) and the other intersections, and we omit the intersections with less than 10 patients.

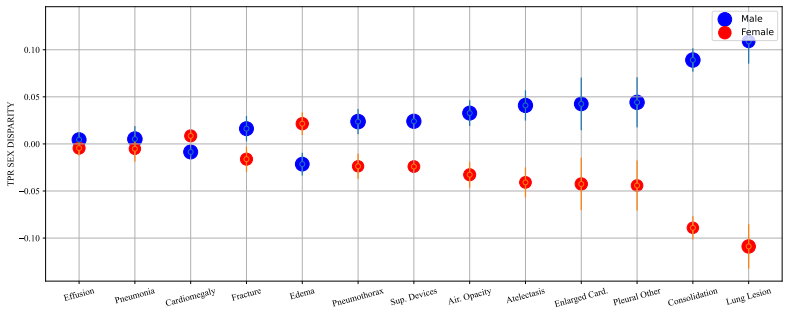

Figure C1: Sorted distribution of TPR sex disparities of the MIMIC-CXR dataset per each disease. "Female "patients exhibit unfavorable outcomes in 9 out of 13 disease labels, reflecting the highest count of negative TPR disparities and "Male "patients emerge as the most favorable subgroup, with 9 out of 13 labels showing zero or positive disparities. The labels "Pleural Effusion(PE) "and "Lung Lesion(LL) "exhibit the smallest (0.008) and largest (0.217) gaps, between the least and most favorable subgroups respectively. The average cross-label gap across all 13 labels, excluding "No Finding (NF)"is 0.071.

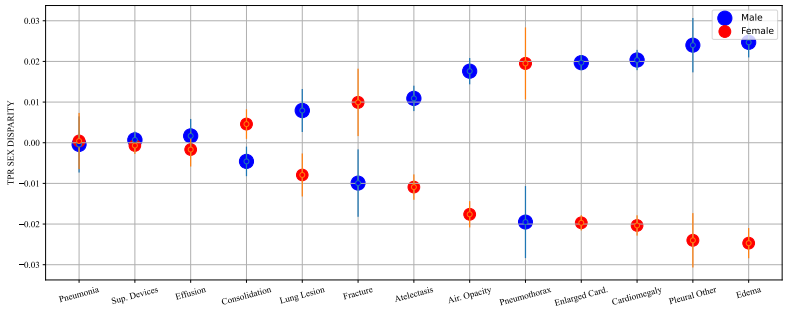

Figure C2: Sorted distribution of TPR sex disparities of the CheXpert dataset per each disease. "Female "patients exhibit unfavorable outcomes in 9 out of 13 disease labels, reflecting the highest count of negative TPR disparities and "Male "patients emerge as the most favorable subgroup, with 9 out of 13 labels showing zero or positive disparities. The labels "Pneumonia(Pn) "and "Edema(Ed) "exhibit the smallest (0.0008) and largest (0.049) gaps, between the least and most favorable subgroups respectively. The average cross-label gap across all 13 labels, excluding "No Finding (NF)"is 0.0249.

Subgroup	**MIMIC**		**CXP**		**ALL**	
	Most U. Diag.	**Max-min Gap**	**Most U. Diag.**	**Max-min Gap**	**Most U. Diag.**	**Max-min Gap**
Sex (Emb)	Female	0.036	Female	0.014	Female	**0.025**
Sex (Img)	Female	0.030	Female	0.007	Female	0.025
Age (Emb)	20-40	**0.117**	20-40	**0.023**	20-40	0.030
Age (Img)	0-20	0.371	20-40	0.136	20-40	0.028
Race (Emb)	Black	**0.084**	Black	**0.048**	Black	**0.095**
Race (Img)	Black	0.106	Hispanic	0.073	Black	0.101
Insurance (Emb)	Medicaid	**0.058**	–	–	–	–
Insurance (Img)	Medicaid	0.111	–	–	–	–

Table D1: Summary of underdiagnosis rates per sensitive attribute across image-based (Img) and vector embedding (Emb) models, highlighting groups with the highest rates (Most U. Diag) and the max-min gap. Cases where vector embeddings show more fairness are bolded.

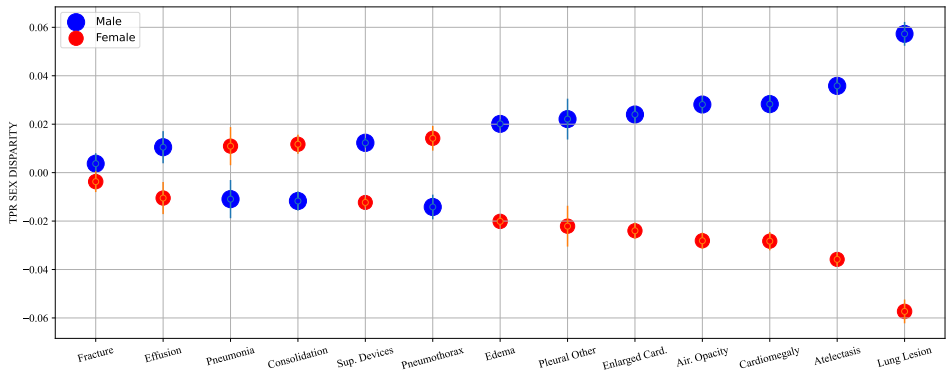

Figure C3: Sorted distribution of TPR sex disparities of the aggregated(ALL) dataset per each disease. “Female ”patients exhibit unfavorable outcomes in 10 out of 13 disease labels, reflecting the highest count of negative TPR disparities and “Male ”patients emerge as the most favorable subgroup, with 10 out of 13 labels showing zero or positive disparities. The labels “Fracture(Fr) ”and “Lung Lesion(LL) ”exhibit the smallest (0.007) and largest (0.114) gaps, between the least and most favorable subgroups respectively. The average cross-label gap across all 13 labels, excluding “No Finding (NF)”is 0.042.

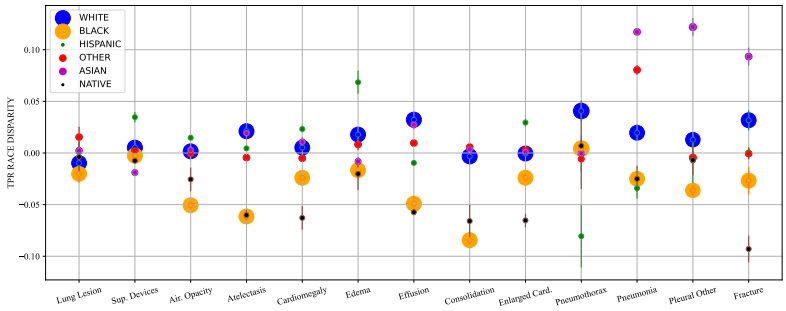

Figure C4: Sorted distribution of TPR race disparities of the CheXpert dataset per each disease. Notably, “Black and Native American ”patients exhibit unfavorable outcomes in 12 out of 13 disease labels reflecting the highest count of negative TPR disparities. Conversely, “White and Asian ”patients emerge as the most favorable subgroup, with 10 out of 13 labels showing zero or positive disparities. The labels “Lung Lesion(LL) ”and “Fracture(Fr) ”exhibit the smallest (0.035) and largest (0.186) gaps, between the least and most favorable subgroups respectively. The average cross-label gap across all 13 labels, excluding “No Finding (NF)”is 0.100.

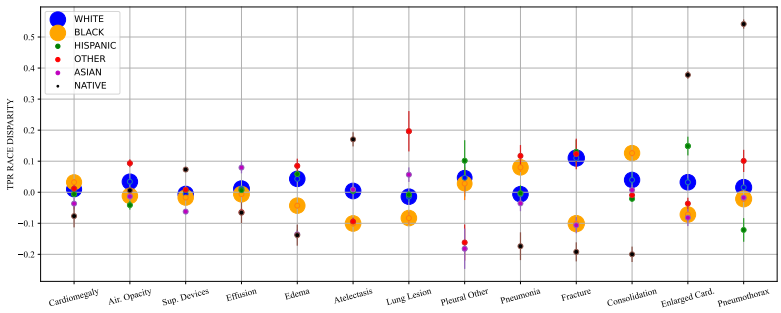

Figure C5: Sorted distribution of TPR race disparities of MIMIC-CXR dataset per each disease. "Black and Asian "patients exhibit unfavorable outcomes for 9 out of 13 disease labels, reflecting the highest count of negative TPR disparities among racial groups and patients from "White "patients emerging as the most favorable subgroup, with 10 out of 13 labels showing zero or positive disparities. Labels "Cardiomegaly(Cd) "and "Pneumothorax(Px) "exhibit the smallest (0.109) and largest (0.663) gaps, between the least and most favorable subgroups respectively. The average cross-label gap across all 13 labels, excluding "No Finding (NF)"is 0.280.

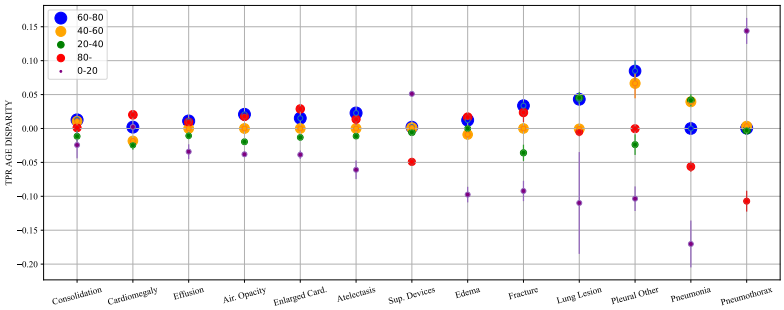

Figure C6: Sorted distribution of TPR age disparities of the CheXpert dataset per each disease. Patients aged between 0 and 20 as well as patients aged between 20 and 40 years exhibit unfavorable outcomes in 10 out of 13 disease labels, reflecting the highest count of negative TPR disparities while patients aged between 60 and 80 years emerge as the most favorable subgroup for all diseases labels, with 13 out of 13 labels showing zero or positive disparities. The labels "Consolidation(Co) "and "Pneumothorax(Px) "exhibit the smallest (0.037) and largest (0.251) gaps, between the least and most favorable subgroups respectively. The average cross-label gap across all 13 labels, excluding "No Finding (NF)"is 0.114.

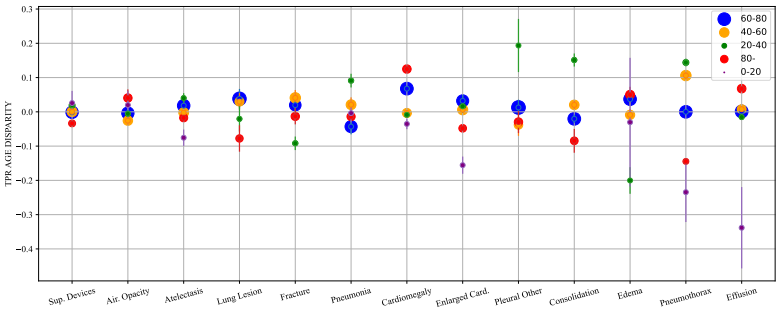

Figure C7: Sorted distribution of TPR age disparities of the MIMIC-CXR dataset per each disease. Patients above 80 years(80−) years exhibit unfavorable outcomes in 9 out of 13 disease labels, reflecting the highest count of negative TPR disparities while patients aged between 60 and 80 years emerge as the most favorable subgroup, with 9 out of 13 labels showing zero or positive disparities. The labels "Support Devices(SD) "and "Pleural Effusion(PE) "exhibit the smallest (0.059) and largest (0.405) gaps, between the least and most favorable subgroups respectively. The average cross-label gap across all 13 labels, excluding "No Finding (NF)"is 0.190.

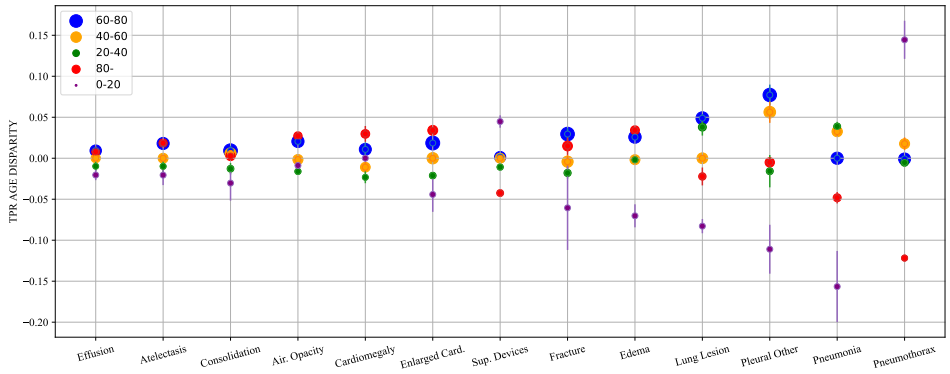

Figure C8: Sorted distribution of TPR age disparities of the aggregated(ALL) dataset per each disease. Patients aged between 20 and 40 years exhibit unfavorable outcomes for 11 out of 13 disease labels, reflecting the highest count of negative TPR disparities while patients aged between 60 and 80 years emerge as the most favorable subgroup with 12 out of 13 labels showing zero or positive disparities. The labels "Pleural Effusion(PE) "and "Pneumothorax(Px) "exhibit the smallest (0.029) and largest (0.266) gaps, between the least and most favorable subgroups respectively. The average cross-label gap across all 13 labels, excluding "No Finding (NF)"is 0.103.

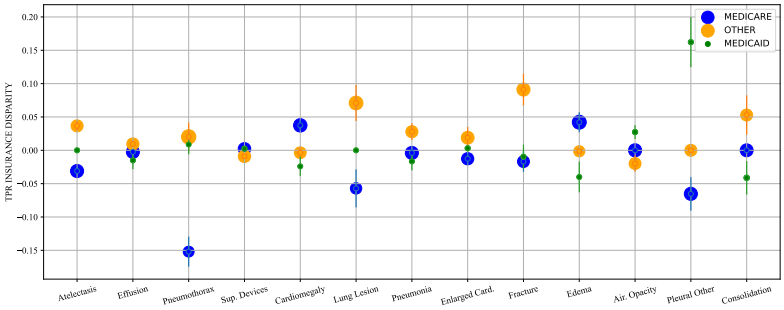

Figure C9: Sorted distribution of TPR insurance disparities of the MIMIC-CXR dataset per each disease. Unexpectedly, patients with "medicare"insurance type exhibit unfavorable outcomes in 8 out of 13 disease labels exhibiting unfavorable outcomes, reflecting the highest count of negative TPR disparities while patients with "other "insurance type emerge as the most favorable subgroup, with 9 out of 13 labels showing zero or positive disparities. The labels "Atelectasis(At) "and "Consolidation(Co) "exhibit the smallest (0.0005) and largest (0.029) gaps, between the least and most favorable subgroups respectively. The average cross-label gap across all 13 labels, excluding "No Finding (NF)"is 0.008.

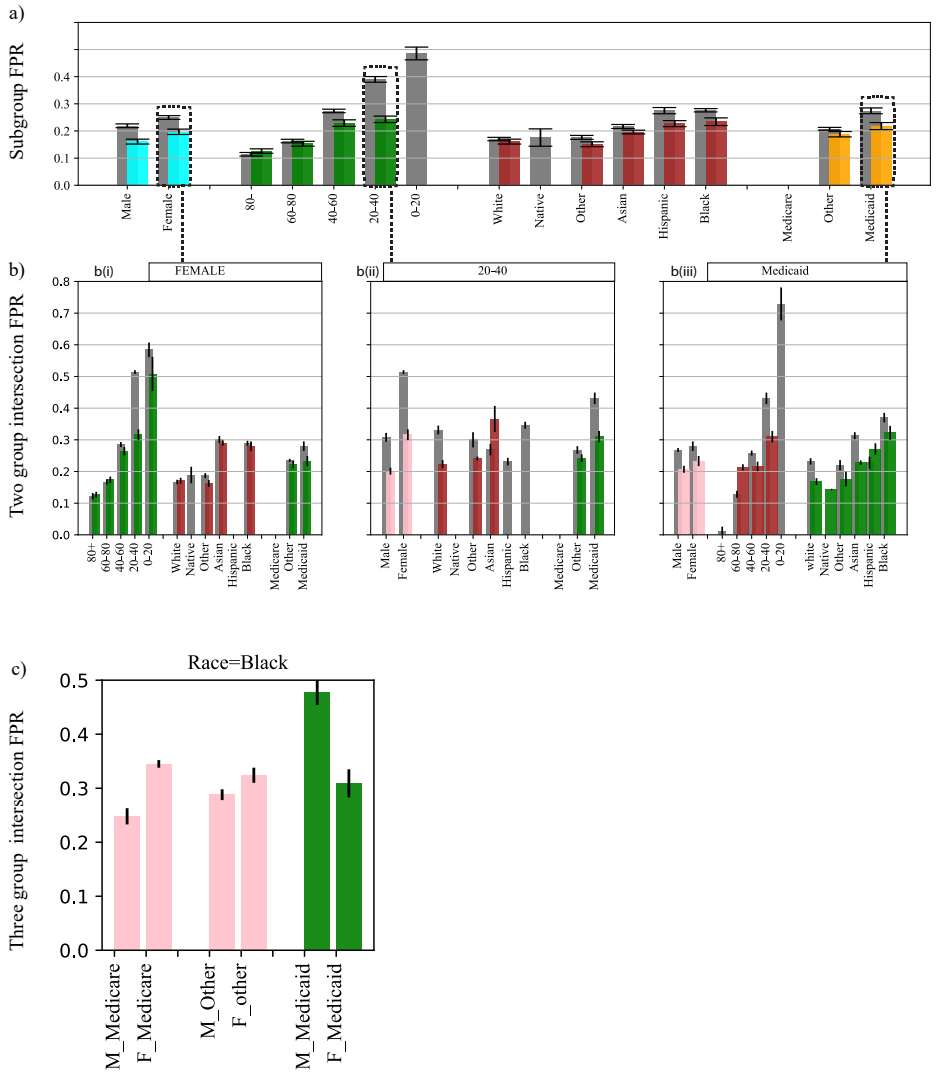

Figure D1: Underdiagnosis rate within subgroups of sex, age, race, and insurance in the MIMIC-CXR dataset. **(a)** Underdiagnosis rate across subgroups.**(b)**, Two group intersection underdiagnosis rates for **(b(i))**. female, **(b(ii)),**, 20-40, and **(b(iii))** Medicaid patients amidst all other subgroups. Baseline results of medical images **?** are plotted in gray colour. **(c),** Three group intersection FPR of black vs sex and insurance type. Intersections with less than 10 patients with False Positive(FP) in the test set have been removed.

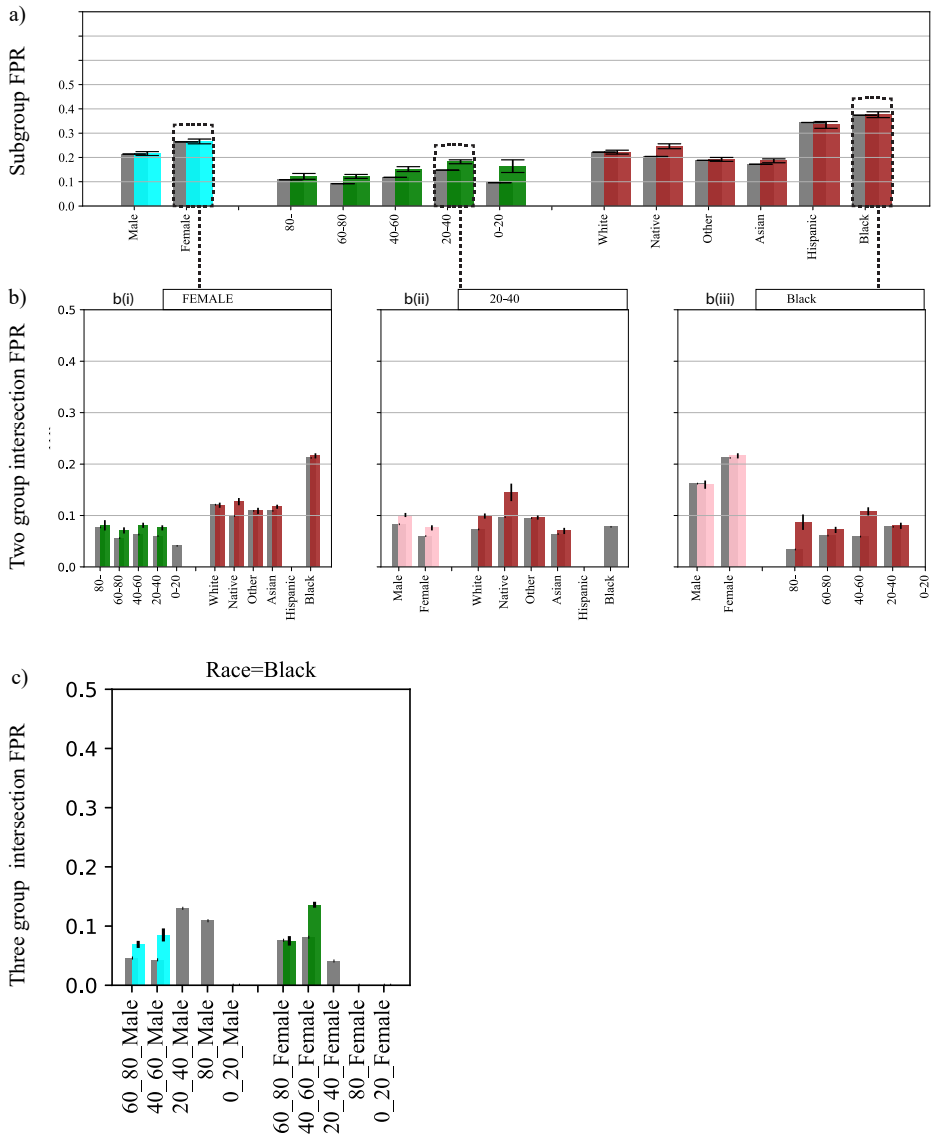

Figure D2: Underdiagnosis rate across subgroups of sex, age, and race within ALL dataset.**(a)** The underdiagnosis rate across subgroups.**(b)**, Two group intersection underdiagnosis rates for **(b(i))**. female, **(b(ii)),**, $20-40$, and **(b(iii))** Black patients amidst all other subgroups. The medical image baseline is in gray. **(c),** Three group intersection FPR of black vs sex and age. Groups with less than 10 patients with FP in the test set have been removed.