CheXclusion: Fairness gaps in deep chest X-ray classifiers

Laleh Seyyed-Kalantari
University of Toronto, Vector Institute
laleh@cs.toronto.edu

Guanxiong Liu
University of Toronto, Vector Institute
lgx@cs.toronto.edu

Matthew McDermott
Massachusetts Institute of Technology
mmd@mit.edu

Marzyeh Ghassemi
University of Toronto, Vector Institute
Marzyeh@cs.toronto.edu

ABSTRACT

Machine learning systems have received much attention recently for their ability to achieve expert-level performance on clinical tasks, particularly in medical imaging. Here, we examine the extent to which state-of-the-art deep learning classifiers trained to yield diagnostic labels from X-ray images are biased with respect to *protected attributes*. We train convolution neural networks to predict 14 diagnostic labels in three prominent public chest X-ray datasets: MIMIC-CXR, Chest-Xray8, and CheXpert. We then evaluate the *TPR disparity* – the difference in true positive rates (TPR) and – *underdiagnosis rate* – the false positive rate of a non-diagnosis – among different protected attributes such as patient sex, age, race, and insurance type. We demonstrate that TPR disparities exist in the state-of-the-art classifiers in all datasets, for all clinical tasks, and all subgroups. We find that TPR disparities are most commonly not significantly correlated with a subgroup's proportional disease burden; further, we find that some subgroups and subsection of the population are chronically underdiagnosed. Such performance disparities have real consequences as models move from papers to products, and should be carefully audited prior to deployment.

CCS CONCEPTS

• **Computing methodologies** → *Neural networks*; *Learning latent representations*; • **Applied computing** → *Imaging*.

KEYWORDS

Fairness, Computer vision, Chest-Xray classifier, Medical image

1 INTRODUCTION

Chest X-ray imaging is an important screening and diagnostic tool for several life-threatening diseases, but patient outcomes can suffer due to the known shortage of radiologists [30, 36–38, 44]. Deep-learning based medical image classifiers are one potential solution, with much work targeting chest X-rays specifically [25, 39, 45, 48], leveraging large-scale publicly available datasets [4, 11, 22, 45], and

demonstrating radiologist-level accuracy in diagnostic classification [21, 34, 35].

While this may seem to make a clear case for implementing AI-enabled diagnostic tools [20], moving such methods from paper to practice require careful thought [14, 46]. In particular, models may exhibit disparities in performance across protected subgroups, and this could lead to different subgroups receiving different treatment [7]. During evaluation, machine learning algorithms usually optimize for, and report performance on, the general population rather than balancing accuracy amongst different subgroups. While some variance in accuracy is unavoidable, fairness in protected subgroups of different attributes, may be desired or required in a deployable model.

In this paper, we examine whether state-of-the-art (SOTA) deep neural classifiers trained on large public medical imaging datasets are fair across different subgroups of *protected attributes* such as patient race. We train classifiers on three large, public chest X-ray datasets: MIMIC-CXR [22], CheXpert [21], and Chest-Xray8 [45]. In each case, we implement chest X-ray pathology classifiers via a deep convolutional neural network (CNN) with frontal/lateral chest X-ray images as inputs, and optimize the multi-class probability of 14 diagnostic labels simultaneously. To the best of our knowledge, we are the first to examine whether SOTA chest X-ray pathology classifiers display systematic bias, and to report area under the receiver operating characteristic curve (AUC)s for 14 diagnostic labels for MIMIC-CXR dataset.

While there are multiple technical fairness definitions [8, 10, 17, 50], and different fairness notions are not always simultaneously achievable [8, 23], we target the equality of opportunity notion across the protected attributes of sex, age, race and insurance type, and focus on two aims: quantifying *TPR disparity* and *underdiagnosis rate*. First, we examine the differences in true positive rate (TPR) across different subgroups per attributes. A high TPR disparity indicates that sick members of a protected subgroup would *not* be given correct diagnoses—e.g., true positives—at the same rate as the general population, even in an algorithm with high overall accuracy. Second, we measure the underdiagnosis rates of various subgroups, where they are predicted not to have any diagnoses despite showing signs and symptoms of conditions [16]. We define this technically as the false positive rate (FPR) of a non-diagnosis as indicated by a "No Finding'' label.

We find that there are indeed extensive patterns of bias in SOTA classifiers, shown both in TPR disparities and underdiagnosis rates across datasets. Disparities that reflect societal biases have been noted previously in healthcare systems, i.e., in the underdiagnosis of cardiovascular disease in women [29], and have real implications

for care. We find that White patients averaged over all sexes (and insurances) have the lowest underdiagnosis rate, while Black and Hispanic patients consistently have the highest. Similarly, Female patients averaged over all races (and insurances) exhibit larger underdiagnosis rates as compared to Male patients. Importantly, the disparity rate for most attributes/ datasets pairs is not significantly correlated with the subgroups' proportional disease membership. This suggests that underrepresented subgroups could be vulnerable to mistreatment in a systematic deployment, and that such vulnerability could not simply be addressed through increasing subgroup patient count. We further demonstrate that specific intersectional identities have high underdiagnosis rates, for example Hispanic-Females and Black-Females have the top underdiagnosis rates in race-sex intersection.

The remainder of our paper is organized as follow. Section 2 briefly outlines the relevant literature and related work. Sections 3 and 4 introduces the datasets and our method. Sections 5 and 6 explain our experiments and results. Section 7 is summary and discussion. Section 8 is the limitations and potential future work.

2 BACKGROUND AND RELATED WORK

Ethical Algorithms in Health. Using machine learning algorithms to make decisions raises serious ethical concerns about risk of patient harm[5]. Notably, biases have already been demonstrated in several settings, including racial bias in the commercial risk score algorithms used in hospitals [31], or an increased risk of electronic health record (EHR) miss-classification in patients with low socioeconomic status [15]. Machine learning algorithms therefore need to proactively implement fairness into their models and use metrics to evaluate performance across different groups when trained on retrospective data that includes human and structural biases.[33].

Fairness and Debiasing. Fairness in machine learning models is a topic of increasing attention, spanning sex bias in occupation classifiers [10], race bias in criminal defendant risk assessments algorithms [8], and intersectional sex-race bias in automated facial analysis [3]. Sources of bias arise in many different places along the classical machine learning pipeline. For example, input data may be biased, leaving supervised models vulnerable to labeling and cohort bias [3]. Minority groups may also be under-sampled, or the features collected may not be indicative of their trends [6]. There are several conflicting definitions of fairness, many of which are not simultaneously achievable [23]. The appropriate choice of a disparity metric is generally task dependent, but balancing error rates between different subgroups in a common consideration [8, 17], with equal accuracy across subgroups being a popular choice in medical settings [42]. In this work we will consider the equality of opportunity notion of fairness and evaluate the rate of correct diagnosis in sick members of different attributes as well as miss-diagnosis rate with no disease.

Several debiasing methods have been proposed for existing models, the most simple being to remove subgroup indicators [10]. Others have used reinforcement learning methods to control the model disparity level [9, 41], adversarial learning to generate a debiased input data [40, 47, 51], or target constructing a fair latent representation [1, 24, 26, 50]. In this work we do not consider debiasing. Instead we focus on SOTA models fairness check, that are
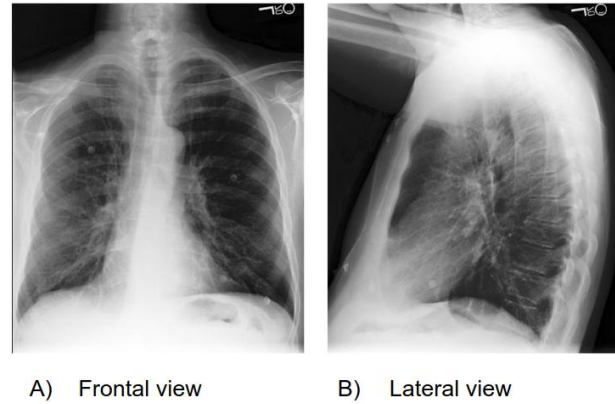
Figure 1: A) Frontal and B) lateral view sample chest X-ray images from CXR dataset (Reproduced by permission).

trained on large public dataset to predict diagnostic label from chest X-ray images.

Chest X-Ray Classification. With the releases of large, public datasets, such as Chest-Xray8 [45], CheXpert [21], and MIMIC-CXR [22], many researchers have begun to train large deep neural network models for chest X-ray diagnosis [21, 34, 48]. In particular [34] train a classifier with radiologist-level performance over Chest-Xray8. CheXpert as well has seen significant study, beginning with [21] which report performance for five of their diagnostic labels. To the best of our knowledge, however, no works have yet been published which perform classification over the MIMIC-CXR dataset, and nobody has yet examined whether any of these algorithms display systematic bias.

3 DATA

We use the public chest X-ray radiography datasets described in Table 1: MIMIC-CXR (CXR) [22], CheXpert (CXP) [21], and Chest-Xray8 (NIH) [45]. Fig.1 shows a frontal and lateral view sample images from CXR [1]. Note that the NIH and CXP datasets have only the patient's sex and age, while the CXR dataset also has race and insurance type data (except for about 100,000 images). All datasets have used automatic labeling method for labels, where natural language processing (NLP) techniques are applied on the radiologist reports to extract disease labels.

Disease Labels. In CXR, CXP, and NIH each radiographic image is associated with diagnostic labels corresponding to 14 diseases, documented in Table 2. Note that in this work, we combine all non-positive labels within CXR and CXP (including "negative'', "not mentioned'', or "uncertain'') into an aggregate "negative'' label for simplicity. This is equivalent to "U-zero'' study of 'NaN' label in CheXpert [21] dataset. In CXR and CheXpert, one of the 14 labels is "No Finding''. This label means no disease has been diagnosed for the image. It is one if all the other 13 labels associated with the image are zero.

---

[1] Reproduced by permission from MIMIC-CXR team.

	MIMIC-CXR	CheXpert	Chest-Xray8
Abbr.	**CXR[22]**	**CXP[21]**	**NIH[45]**
# Images	371,858	223,648	112,120
# Patients	65,079	64,740	30,805
View	Front/Lat	Front/Lat	Front
Female	47.83%	40.64%	43.51%
Male	52.17%	59.36%	56.49%
0-20	2.20%	0.87%	6.09%
20-40	19.51%	13.18%	25.96%
40-60	37.20%	31.00%	43.83%
60-80	34.12%	38.94%	23.11%
80-	6.96%	16.01%	1.01%
White	65.01%	N/A	N/A
Black	17.86%	N/A	N/A
Other	3.68%	N/A	N/A
Asian	3.12%	N/A	N/A
Hispanic	6.16%	N/A	N/A
Native	0.28%	N/A	N/A
Unknown	3.89%	N/A	N/A
Medicare	46.07%	N/A	N/A
Medicaid	8.98%	N/A	N/A
Other	44.95%	N/A	N/A

Table 1: Description of chest X-ray datasets, MIMIC-CXR (CXR) [22], CheXpert (CXP) [21] and Chest-Xray8 (NIH) [45]. Here, the number of images, patients, view types, and the proportion of patients per subgroups of sex, age, race, and insurance type are presented. 'Front' and 'Lat' abbreviate frontal and lateral view, respectively. Native, Hispanic, and Black denote self-reported American Indian/Alaska Native, Hispanic/Latino, and Black/African American race respectively.

Protected Attributes. The protected attributes are patients sex (Male and Female), age (0-20, 20-40, 40-60, 60-80, and 80-), race (White, Black, Other, Asian, Hispanic, and Native) and insurance type (Medicare, Medicaid, and Other). These values are taken from the structured patient attributes in the database.

4 METHODS

We implement CNN-based models to classify frontal/lateral chest X-ray images into 14 diagnostic labels. We train separate models for CXR [22], CXP [21], and NIH [45], and study the performance and fairness of the models with respect to patient sex and age. We also explore the fairness of models trained on the CXR dataset with respect to patient race and insurance type.

4.1 Models

Model Architecture: We initialize a 121-layer DenseNet [19] with the ImageNet [12] pre-trained weights and train models with a binary cross entropy loss function. The 121-layer DenseNet produced the best results in prior studies on CXP [21] and NIH [34]. For all datasets, we use a 80-10-10 train-validation-test split, with no patient shared across splits.

Data Processing: We resize all images to 224x224 and normalize via the mean and standard deviation of the ImageNet dataset [12]. We apply center crop, random horizontal flip, and validation set early stopping to select the optimal model. We further perform random 10, 15 and 15 degree rotation as data augmentations on NIH, CXP and CXR, respectively.

Model Training and Optimization: We use Adam [32] optimization with default parameters, and decrease the learning rate by a factor of 2 if the validation loss does not improve over three epochs; we stop learning if validation loss does not improve over 10 epochs. In training on NIH, the initial learning rate is 0.0005 and the batch size is 32. These values are 0.0001/48 for CXP and 0.0005/48 for CXR, respectively. We first tune models to achieve the highest performance (average AUC over 14 labels) by fine tuning the learning rate, then for the best achieved model we fine tune the degree of random rotation data augmentation. Finally we fix all the hyperparameters of the best model and train four extra models with the same hyperparameters, but different random seed between 0 to 100. We use those models to report all the metrics based on the mean and 95% confidence intervals (CI) achieved over five studies per dataset. The batch size is chosen such that we use the maximum memory capacity of the GPU. The output of the network is an array of 14 numbers between 0 and 1 indicating the probability of each disease label. The best threshold per disease is chosen based on highest F1 score measure on validation distribution. We trained models using a single NVIDIA GPU with 16G of memory in 9, 20, and 40 hours for NIH, CXP, and CXR, respectively.

4.2 Classifier Disparity Evaluation

We measure TPR disparities and underdiagnosis rate to evaluate the potential bias of the classifier once classifying different subgroups S_j, $j = 1, 2, ..., N_k$, within k attributes. We have binary protected attributes such as sex where the subgroups are g (e.g., Female) and not g ($\sim g$) (e.g., Male). Also, we have non-binary attributes such as age, race, and insurance type where we have more than two subgroups per attribute.

TPR disparities for binary attributes. For binary attributes, similar to [10] we quantify the TPR disparity as the difference between TPR of sex g and $\sim g$, per label y_i, $i \in \{1, \ldots, 14\}$. Then, with random variables $\hat{Y}_i$ and Y_i denoting the predicted and ground truth labels for y_i, the TPR of sex g per disease y_i, is $\text{TPR}_{g,y_i} = P[\hat{Y}_i = y_i | G = g, Y_i = y_i]$, and the associated TPR sex disparity is, $\text{Gap}_{g,y_i} = \text{TPR}_{g,y_i} - \text{TPR}_{\sim g,y_i}$ [10].

TPR disparities for non-binary attributes. For non-binary attributes, we use the difference between a subgroup's TPR and the median (as measure of central tendency) of all TPRs to define TPR disparity, $\text{Gap}_{S_j,y_i} = \text{TPR}_{S_j,y_i} - \text{Median}(\text{TPR}_{S_1}, .., \text{TPR}_{S_{N_k}})_i$.

Underdiagnosis rate. We use the FPR in the "No Finding" label as a marker of a missed diagnosis, or the *underdiagnosis rate*, i.e., a patient truly has a disease, but the classifier incorrectly predicted no disease.

		CXP	CXR	NIH	NIH		
Label	**Abbr.**	**AUC**	**AUC**	**Label**	**Abbr.**	**AUC**	**[34] AUC**
Airspace Opacity	AO	0.747±0.001	0.782±0.001	Atelectasis	A	0.814±0.004	0.862
Atelectasis	A	0.717±0.002	0.837±0.001	Cardiomegaly	Cd	0.915±0.002	0.831
Cardiomegaly	Cd	0.855±0.003	0.828±0.001	Consolidation	Co	0.801±0.005	0.893
Consolidation	Co	0.734±0.004	0.844±0.001	Edema	Ed	0.915±0.003	0.924
Edema	Ed	0.849±0.001	0.905±0.001	Effusion	Ef	0.875±0.002	0.901
Enlarged Card	EC	0.668±0.005	0.758±0.004	Emphysema	Em	0.897±0.002	0.704
Fracture	Fr	0.790±0.006	0.717±0.007	Fibrosis	Fb	0.788±0.007	0.806
Lung Lesion	LL	0.780±0.005	0.773±0.005	Hernia	H	0.978±0.004	0.851
No Finding	NF	0.885±0.001	0.869±0.001	Infiltration	In	0.717±0.004	0.721
Pleural Effusion	PE	0.885±0.001	0.933±0.001	Mass	M	0.829±0.006	0.909
Pleural Other	PO	0.795±0.004	0.846±0.003	Nodule	N	0.779±0.006	0.894
Pneumonia	Pa	0.777±0.003	0.748±0.005	Pleural_Thickening	PT	0.813±0.006	0.798
Pneumothorax	Px	0.893±0.002	0.903±0.002	Pneumonia	Pa	0.759±0.012	0.851
Support Devices	SD	0.898±0.001	0.927±0.001	Pneumothorax	Px	0.879±0.005	0.944
Average		**0.805±0.001**	**0.834±0.001**	**Average**		**0.840±0.001**	**0.849**

Table 2: The AUC for chest X-ray classifiers trained on CXP, CXR, and NIH, averaged over 5 runs± 95%CI, where all runs have same hyperparameters but different random seed. ('Airspace Opacity' in [22] and 'Lung Opacity' in [21] denote a same label.)

5 EXPERIMENTS

We demonstrate that the models achieve SOTA classification performance. In order to estimate the classifiers fairness, trained on the CXR, CXP and NIH, we target four investigations, listed below:

(1) TPR disparity: We quantify the TPR disparity per subgroup and disease for sex and age subgroups across all 3 datasets, and due to data availability for race and insurance type on the CXR dataset only. Here we want to check whether larger proportions of a subgroup in a disease alleviates the disparity.

(2) TPR disparity in proportion to membership: We investigate if the distribution of the patients proportion per subgroup S_j and label y_i (which is given by $P[S = S_j|Y = y_i]$) has effect on TPR disparities. Others have established a positive correlation between classifier TPR disparities and the subgroup membership [10], e.g., occupations with more females have higher TPR for females. Prior work has indicated that such disparities in small or vulnerable subgroups could be propagated in put into practice [10, 18].

(3) Chronic subgroup underdiagnosis: We identify subgroup - specific chronic underdiagnosis in CXR dataset. We compare the distribution of FPR in the "No Finding'' label (underdiagnosis rate) to characterize what subgroups have higher underdiagnosis rate and they may be more in danger of chronic underdiagnosis compare to other subgroups within an attribute (e.g. How is underdiagnosis rate among people with different races.)

(4) Chronic intersectional identity underdiagnosis: We identify intersectional - specific chronic underdiagnosis in CXR dataset. We investigate how does underdiagnosis rates are distributed among other attributes of different subsections (e.g. if with Female patients suffer more from underdiagnosis compare to Male patients, do all Female from different races or insurance types have the same portion of the underdiagnosis?)

6 RESULTS

One potential reason that a model may be biased is that it is poorly trained. We demonstrate that the models achieve SOTA classification performance, before exploring our stated characterizations of disparity. Table 2 shows overall performance numbers across all tasks and datasets. Though results do have non-trivial variability, we do show similar performance to the published SOTA of the NIH [34], the only dataset for which a published SOTA comparison exists for all labels. Note that the published results for the CXP [21] dataset are on a private, unreleased dataset of only 200 images, for 5 labels, whereas our results are on a randomly sub-sampled test set of size 22,274 images, so the numbers for this dataset are not comparable to the published results there. The test dataset sizes for CXR and NIH are 36,421 and 6,373 images.

6.1 TPR Disparities

TPR disparities are common across datasets and protected attributes. We calculate the TPR disparities and 95% CI across all labels, datasets and attributes. We see many instances of negative and positive disparities, which can denote bias against and in favor of a subgroup. We call them *unfavorable* and *favorable* subgroups. As an illustrative example Fig. 2 shows the race sorted TPR disparities distribution. In a fair setting, all subgroups TPR disparities per disease are similar and the gap between least and most favorable subgroups within a label is '0'. Table 3 shows the summary of the disparities in all attributes/ datasets. We have shown the average cross - label gap between least/ most favorable groups per dataset/ attributes as well as label with smallest and the largest gap. We count the number of time each subgroups experience negative disparities (unfavorable) and zero or positive disparities (favorable) across disease labels [2] and report the most frequent unfavorable and favorable subgroups and the count in Table 3 .

---

[2]For CXP and CXR datasets we exclude "No Finding'' label in the count as we want to check negative bias in disease labels only. Thus the counts are out of 13.

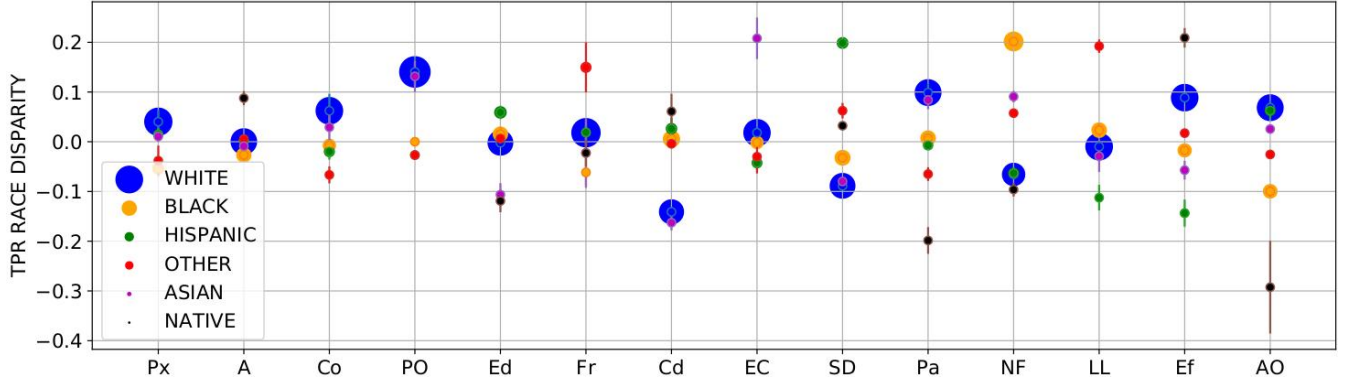Figure 2: The sorted TPR race disparity distribution in CXR dataset. x-axis represent the labels abbreviation (full names available Table 2). The scatter plot's circle area is proportional to the patients membership. The TPR disparities are averaged over five run ±95%CI (95% CI are shown with arrows around the mean). Black patients are most unfavorable subgroup (they have maximum count of negative TPR disparities, 8/13) where White patients are most favorable subgroup (9/13 zero or positive disparities). The label 'Pneumothorax' ('Px') has the smallest gap (0.094) between the least/most favorable subgroups, where 'Airspace Opacity' ('AO') has the largest (0.361). The average cross 14 labels gap is 0.233.

Attribute	Dataset	Average Cross-Label Gap Between Least/Most Favorable Subgroups per Label (Label with Smallest Gap - Label with Largest Gap)	Most Unfavorable subgroup count	Most Favorable subgroup count
Sex	NIH	0.190 (Mass:0.001-Cardiomegaly:0.393)	Female (8/14)	Male (8/14)
	CXP	0.062 (Edema:0.000-Consolidation:0.139)	Female (7/13)	Male (7/13)
	CXR	0.123 (Pneumothorax:0.017-No Finding:0.369)	Female (10/13)	Male (10/13)
Age	NIH	0.413 (Infiltration:0.188-Emphysema:1.00)	60-80 (7/14)	20-40 (9/14)
	CXP	0.270 (Support Devices:0.084-No Finding:0.604)	0-20, 20-40, 80- (7/13)	40-60 (8/13)
	CXR	0.279 (Pneumonia:0.054-Edema:0.544)	0-20, 20-40 (8/13)	60-80 (9/13)
Race	CXR	0.233 (Pneumothorax:0.094-Airspace Opacity:0.361)	Black (8/13)	White (9/13)
Insurance	CXR	0.138 (Pneumothorax:0.044-Atelectasis:0.265)	Medicaid (9/13)	Other (10/13)

Table 3: Disparities overview in all attributes/datasets. In a dataset/attribute, we average per label gaps between the least and most favorable subgroup's TPR disparities to obtain the average cross-label gap. The specific labels that obtained the smallest and largest gaps are shown in parenthesis, along with their gaps. We further summarize the most frequent 'Unfavorable' and 'Favorable' subgroups count. The Unfavorable or Favorable subgroups are the ones that experience TPRs disparities below or above the zero gap line. We note that the most frequent unfavorable subgroups are those with social disparities in the healthcare system, e.g., women and minorities. No diseases are consistently at the highest or lowest disparity rates across all settings. Instead, disease disparities varied depending on the dataset, and attribute.

6.2 TPR Disparity Correlation with Membership Proportion

We measure the Pearson correlation coefficients (r) between the TPR disparities and patients proportion per label across all subgroups/datasets. As multiple hypotheses are being tested (27 total comparisons amongst all protected attributes considered) with a desired 0.05 significance level (p) are tested, we apply the Bonferroni correction [28] and set statisitacal significance for each individual hypothesis at $p<0.0019$ (0.05/27). The majority of correlation coefficients listed are positive, however there is only statistically significant correlation for Male sex (r: 0.808, p: 0.0005), Medicare insurance (r: 0.843, p: 0.0002), age group 60-80 (r: 0.905, p: 8.5e-6)

and 20-40 (r:0.907 , p: 7.6e-6) in the CXR dataset, and age group 60-80 (r: 0.853, p: 0.0001) in the CXP dataset.

6.3 Subgroup-Specific Chronic Underdiagnosis

Our study reveals that the SOTA classifiers comprehensively *underidentify* conditions in some subgroups. We present the FPR in the "No Finding'' label, which denotes the situation where a patient truly has a disease, but the classifier incorrectly predict the patient has no disease. Figure 3 shows the sorted distribution of FPR over subgroups of age, race, and insurance type in CXR. As it shown, subgroups that are in the age group 0-20, patients reporting Black race, and patients with Medicaid insurance have the largest FPRs, and therefore the worst underdiagnosis rates, in the related attributes.
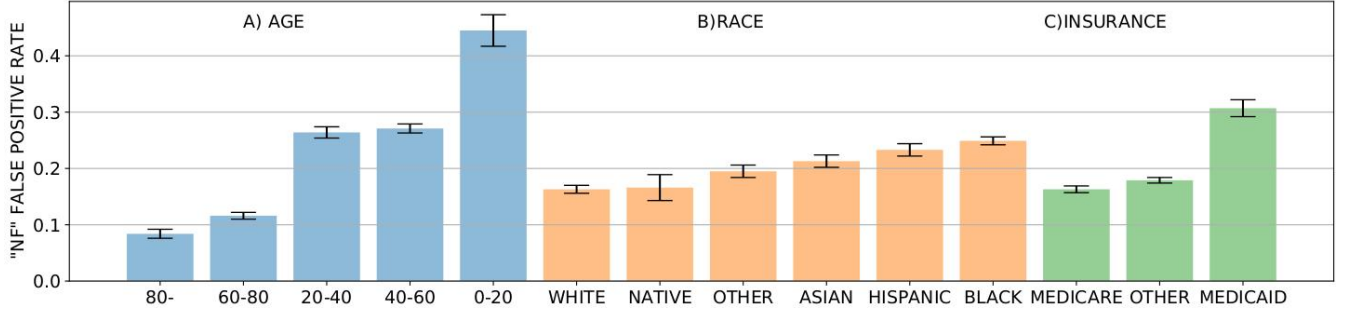
Figure 3: The underdiagnosis rate distribution over subgroups of A) age, B) race, and C) insurance type in CXR. Patients younger than 20 years old, Black and/or low-income patient under Medicaid insurance have the largest underdiagnosis rate. Note that higher values indicate more disparity against the associated subgroup.

		SEX		INSURANCE		
		Male	**Female**	**Medicare**	**Medicaid**	**Other**
RACE	**White**	0.178±0.008	0.138±0.006	0.146±0.006	0.182±0.013	0.188±0.008
	Black	0.207±0.008	**0.275±0.007**	0.276±0.010	**0.414±0.018**	0.182±0.004
	Other	0.206±0.013	0.160±0.007	0.164±0.007	**0.316±0.057**	0.196±0.009
	Asian	0.169±0.010	0.269±0.018	0.114±0.016	**0.353±0.021**	0.230±0.011
	Native	0.190±0.029	0.135±0.028	0.099±0.029	–	**0.409±0.033**
	Hispanic	0.117±0.008	**0.459±0.023**	0.272±0.011	**0.401±0.027**	0.136±0.008
INSURANCE	**Medicare**	0.171±0.008	0.153 ±0.006			
	Medicaid	**0.242±0.009**	**0.362±0.022**			
	Other	0.153±0.007	0.225±0.003			

Table 4: The distribution of intersectional underdiagnosis rate for race-sex, race-insurance, and insurance-sex. The values are mean of correlation coefficients over 5 runs±95%CI.

6.4 Intersectional-Specific Chronic Underdiagnosis

In Table 4 we show the underdiagnosis rate in CXR over race-sex, race-insurance, and insurance-sex intersections. The top underdiagnosis rates for intersectional groups are Hispanic-Female and Black-Female for race-sex, Native-Other and Black-Medicaid for race-insurance, and Female-Medicaid and Male-Medicaid for insurance-sex. White patients averaged over all sexes and insurances have the smallest underdiagnosis rate, while Black and Hispanic patients consistently have the largest. Female patients averaged over all races and insurances also exhibit larger underdiagnosis rates compared to Male patients.

7 SUMMARY AND DISCUSSION

Here, we present a range of findings on the potential biases of deployed SOTA X-ray image classifiers over the sex, age, race and insurance type attributes on models trained on NIH, CXP and CXR datasets. We investigate two fairness checks factors. The first focuses on TPR disparities similar to [10], checking if the sick members of the different subgroups are given correct diagnosis at similar rates. The second explore if some subgroups or intersectional identity are chronically underdiagnosed (e.g., how underdiagnosis is distributed among Female with different races or insurance types.).

Our results demonstrate several main takeaways, which we explore in more detail here.

First, all datasets and tasks display nontrivial TPR disparities. These disparities could pose serious barriers to effective deployment of these models and indicate that more changes are needed, in either dataset design and/or modeling techniques to ensure more equitable models. Second, while there is occasionally a proportionality between protected subgroup membership per label and TPR disparity, this relationship is not uniformly true across datasets and subgroups. Third, we demonstrate that some subgroups and intersectional subgroups are often at the *greatest* risk of chronic underdiagnosis, indicating that the bias observed compounds on patients who are members of several underrepresented subgroups. We explore through each of these findings in more depth below.

7.1 Extensive Patterns of Bias

We found no diseases that were consistently at the highest or lowest disparity rates across all attributes and datasets. Although, per dataset, some disease may commonly appear with larger or smaller gap between least / most favorable subgroups.

7.1.1 Bias with respect to sex. TPR disparities with respect to patients sex is observed in all settings. The classifier trained on CXP has smallest average cross- label gap between least/ most favorable

groups compared to the CXR and NIH . Female are the most unfavorable subgroups (they have the most frequency of negative TPR disparities) in all three datasets. The proportion of Female is less than Male in all three datasets but the difference is not large.

7.1.2 Bias with respect to age. The age disparity observed in all setting. The average cross-label gap between least/ most favorable subgroups for CXP and CXR dataset is similar where it is largest over NIH. There is not an age subgroup that commonly appeared in unfavorable or favorable subgroups across all three datasets. Also, we do not have pattern indicating the minorities or majorities are constantly among unfavorable or favorable subgroups across three datasets. The patients under 20 years old has the largest underdiagnosis rate where the patients older than 80 has the smallest. Here, both subgroups of 0-20 and 80- are among the minorities.

7.1.3 Bias with respect to race. We observe TPR disparities with respect to the patient race. The CXR dataset is highly unbalanced racially with 65% White patients, who are the most favorable subgroup. Note that while Black patients have the second largest racial population within the dataset, they are the most unfavorable subgroup with worse fairness factors even compared to minorities with less than 4% of the population. Over the race distributions the Black patients show the largest underdiagnosis rate followed by the Hispanic, while the White patients has the smallest underdiagnosis rate. There is not a pattern that minorities constantly appeared in subgroups with largest underdiagnosis rate. In general underdiagnosis rates have smoother variance across races.

7.1.4 Bias with respect to insurance type. The TPR disparity study on patients insurance type indicates bias exist against patients with Medicaid insurance who are the minority population in the dataset and often from low social economic status. They are the most unfavorable subgroup within insurance type attribute and the model commonly has bias against correctly diagnosing disease for them. They also have the largest underdiagnosis rate with a large gap compare to other 2 insurance types.

7.2 Disparities and Membership Correlation

We observed TPR disparities are not often significantly correlated with disease membership. Though, we often have positive correlation between them, among 27 hypothesis test only 5 subgroups passed the significance level. This implies that simply scaling the patient proportion will not lead the classifier to correctly diagnose disease (higher TPR disparities). Exploring the TPR disparities, we also have observed diseases with the same patient proportion of a subgroup, may have totally different TPR disparities (e.g. 'Consolidation', 'Nodule' and 'Pneumothorax' in NIH have 45% Female, but the TPR disparities are in diverse range, -0.155, -0.079 and 0.047, respectively). Thus, only balancing the dataset or having the same portion of images within all labels may not guarantee the fairness.

7.3 Chronic underdiagnosis in intersectional identity

We investigate intersectional identities with chronic underdiagnosis for race-sex, race-insurance, and insurance-sex subgroups intersections in CXR dataset. The target is to measure how underdiagnosis rates are distributed in different intersectional membership.

7.3.1 Race-Sex. The cross-race average of Female intersections has a larger underdiagnosis rate compared to Male, where Hispanic - and Black - Female have larger portion of chronic underdiagnosis. On average over all sexes Hispanic have the largest underdiagnosis rate, however, notably, there is a large gap between underdiagnosis rate of Hispanic - Male and Female. The Hispanic - Male has the smallest underdiagnosis rate among all sex - race intersections where this rate is 3.86 times larger in Hispanic - Female patients. On average over all sexes White patients have the smallest underdiagnosis rate, where White - Female stands in a better condition compare to White - Male patients.

7.3.2 Race-Insurance. The Black patients with the Medicaid and Native with Other insurance have the two largest underdiagnosis rate. There is a large gap between the underdiagnosis rate in Native with Medicare and Other insurance. The former has the smallest underdiagnosis rate among all race - insurance intersections where the later has the second largest. The intersection of patients with Medicaid insurance with all races exhibit large underdiagnosis rate, except for White patients who fare better than others. On average over all insurances, Black and White patients have the largest and smallest underdiagnosis rates, respectively.

7.3.3 Insurance-Sex. As before, Female and Male patients with Medicaid insurance have the highest underdiagnosis rates, though Female has the larger portion. On average over all insurances, Female have the most underdiagnosis rate.

7.4 Discussion

We identify subgroups that may experience more bias through the exploration of variance in TPR and FPR. Based on the equality of opportunity notion of fairness, a fair network should exhibit the same TPR/FPR among all subgroups regardless of how likely a subgroup may have a disease. Such an improvement would allow two patients with the same condition, but in different subgroups, to be diagnosed correctly and receive the same level of care. While we focused on some of the more obvious protected attributes, it is important to note that there are several other factors, subgroups, and attributes that we have not considered.

Identifying and eliminating disparities is particularly important as large datasets begin to be used by high-capacity neural models, but are based on highly skewed population, e.g., kidney injury prediction in a population that is 93.6% male [43]. While chest X-ray images datasets are not sex-skewed, we note that the age, race and insurance type attributes are highly unbalanced, e.g., 65% of patients are White, and only 8.98% are under Medicaid insurance. Subgroups with chronic underdiagnosis are those who experience more negative social determinants of health, specifically, women, minorities, and those of low socioeconomic status. Such patients may use healthcare services less than others. In some groups, such a dataset skew can increase the risk of miss-classification [15].

Although, "de-biasing" techniques [2, 40, 47, 51] may reduce disparities, but we should not ignore the importance of considering those biases, on preparing large public training datasets. Data quality can induce discriminatory properties in classifiers. Unmeasured predictive features generate discrimination [6], and models trained on biased datasets can results in unfair algorithms [3]. For instance,

an algorithm that can classify skin cancer [13] with high accuracy will not be able to generalize on different skin color if similar samples have not been represented enough in the trained dataset [3]. Intentionally adjusting the datasets to reduce disparities in order to protect minorities as well as the subgroups with high disparities is one potential option in dataset creation.

While there is much promise in the use of advanced models for clinical care, but we caution that even advanced SOTA models must be carefully checked for such biases as those we have identified. Disparities in small or vulnerable subgroups could be propagated [18] within the development of machine learning models. This raises serious ethical concerns [5] about the accessibility of chronic underdiagnosic intersectional identities to the required medical treatment. Usually the SOTA classifiers are trained to provides high AUC or accuracy on the general population. However we suggest additionally applying the fairness check to the SOTA classifiers before deployment.

8 LIMITATIONS AND FUTURE WORK

As SOTA deep learning pathology detection algorithms become more likely candidates for medical screening tools, investigation of model bias is essential. This work is in a first step in quantifying the limitations of such systems, but has many limitations that indicate opportunities for future work.

First, we note that human labeling is subjective, and particularly so in classification where images are usually set against multi-label task. Limiting a set of well-defined labels is hard due to the complex pattern between disease [48]. All three datasets in our study, used NLP techniques on the radiologist reports to extract disease labels. Thus the inherent error rate of the NLP models add up with the error rate of image classifiers. Additionally, the imaging devices quality, the majority of population in a region where the data is gathered, and the type of patients at each hospital is distinguished. For instance, NIH dataset gathered from a hospital that covers more complicated cases where CXP have more tertiary cases, and CXR gathered from an emergency department. It is even possible to predict the admitted hospital of a patient from the chest X-ray image [49]. All these challenges may affect the accuracy of the labels and trained models on these datasets, where all reflects on the fairness of the network. In addition to the quality of labels [27], lack of access to the medical history of the patient is also challenging. The developed chest X-ray classifiers algorithms study images independently and do not take into account the patients history [34, 35, 48], or the correlation between diseases.

We did not investigate the many methods that could be used to debias the classifier, including representation learning [2], adversarial learning [51], and GANs [40, 47], focusing instead on quantifying the TPR disparities and underdiagnosis rate distribution in SOTA deep learning models that might be trained using large, publicly available dataset. In addition to the methods that are applicable to introduce a fair model, focusing on fairness at the step of datasets gathering is also important.

9 CONCLUSION

While there is much opportunity in the development and deployment of machine learning models in a clinical setting, great care

must be taken to understand how existing biases may be exacerbated and propagated. In this paper, we illustrate the TPR disparity of SOTA chest X-ray pathology classifiers trained on three different datasets, (MIMIC-CXR, ChestX-ray8, and CheXpert) across 14 disease labels. We quantify the TPR disparity across experimental studies along sex, age, race and insurance type. We also spot some subsection of the population are chronically underdiagnosed. Our results indicate that high-capacity models trained on large datasets do not provide equality of opportunity naturally, leading instead to potential disparities in care if deployed without modification.

REFERENCES

[1] Tameem Adel, Isabel Valera, Zoubin Ghahramani, and Adrian Weller. 2019. One-network adversarial fairness. *Association of Advancements in Artificial Intelligence* 33 (2019), 2412–2420.

[2] Alexander Amini, Ava P. Soleimany, Wilko Schwarting, Sangeeta N. Bhatia, and Daniela Rus. 2019. Uncovering and Mitigating Algorithmic Bias through Learned Latent Structure. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society - AIES '19*. ACM Press, Honolulu, HI, USA, 289–295. https://doi.org/10.1145/3306618.3314243

[3] Joy Buolamwini and Timnit Gebru. 2018. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classiïïn̨Açation. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency (FAT*'18)*, Vol. 81. 15.

[4] Aurelia Bustos, Antonio Pertusa, Jose-Maria Salinas, and Maria de la Iglesia-VayÃą. 2019. PadChest: A large chest x-ray image dataset with multi-label annotated reports. *arXiv:1901.07441 [cs, eess]* (Jan. 2019). http://arxiv.org/abs/1901.07441 arXiv: 1901.07441.

[5] Danton S. Char, Nigam H. Shah, and David Magnus. 2018. Implementing Machine Learning in Health Care âĂŤ Addressing Ethical Challenges. *New England Journal of Medicine* 378, 11 (2018), 981–983. https://doi.org/10.1056/NEJMp1714229 PMID: 29539284.

[6] Irene Chen, Fredrik D Johansson, and David Sontag. 2018. Why Is My Classifier Discriminatory? In *Advances in Neural Information Processing Systems 31*. Curran Associates, Inc., 3539–3550.

[7] Irene Y Chen, Peter Szolovits, and Marzyeh Ghassemi. 2019. Can AI Help Reduce Disparities in General Medical and Mental Health Care? *AMA journal of ethics* 21, 2 (2019), 167–179.

[8] Alexandra Chouldechova. 2016. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data* 5, 2 (2016), 153–163. https://doi.org/10.1089/big.2016.0047

[9] Efrén Cruz Cortés and Debashis Ghosh. 2019. A simulation based dynamic evaluation framework for system-wide algorithmic fairness. *arXiv preprint arXiv:1903.09209* (2019).

[10] Maria De-Arteaga, Alexey Romanov, Hanna Wallach, Jennifer Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Geyik, Krishnaram Kenthapadi, and Adam Tauman Kalai. 2019. Bias in bios: a case study of semantic representation bias in a high-stakes setting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT*'19)*. USA, 120–128. Atlanta, GA.

[11] Dina Demner-Fushman, Marc D. Kohli, Marc B. Rosenman, Sonya E. Shooshan, Laritza M. Rodriguez, Sameer K. Antani, George R. Thoma, and Clement J. McDonald. 2016. Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association : JAMIA* 23, 2 (2016), 304–310. https://doi.org/10.1093/jamia/ocv080

[12] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. 2009. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*.

[13] Andre Esteva, Brett Kuprel, Roberto A. Novoa, Justin Ko, Susan M. Swetter, Helen M. Blau, and Sebastian Thrun. 2017. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* 542, 7639 (Feb. 2017), 115–118. https://doi.org/10.1038/nature21056

[14] Marzyeh Ghassemi, Tristan Naumann, Peter Schulam, Andrew L Beam, Irene Y Chen, and Rajesh Ranganath. 2019. Practical guidance on artificial intelligence for health-care data. *The Lancet Digital Health* 1, 4 (2019), e157–e159.

[15] Milena A. Gianfrancesco, Suzanne Tamang, Jinoos Yazdany, and Gabriela Schmajuk. 2018. Potential Biases in Machine Learning Algorithms Using Electronic Health Record Data. *JAMA internal medicine* 178, 11 (2018), 1544–1547. https://doi.org/10.1001/jamainternmed.2018.3763

[16] Carmen R. Green, Karen O. Anderson, Tamara A. Baker, Lisa C. Campbell, Sheila Decker, Roger B. Fillingim, Donna A. Kalauokalani, Donna A. Kaloukalani, Kathyrn E. Lasch, Cynthia Myers, Raymond C. Tait, Knox H. Todd, and April H. Vallerand. 2003. The unequal burden of pain: confronting racial and ethnic disparities in pain. *Pain Medicine (Malden, Mass.)* 4, 3 (Sept. 2003), 277–294. https://doi.org/10.1046/j.1526-4637.2003.03034.x

[17] Moritz Hardt, Eric Price, and Nathan Srebro. 2016. Equality of Opportunity in Supervised Learning. In *Proceedings of the 30th International Conference on Neural Information Processing Systems (NIPS'16)*. Curran Associates Inc., USA, 3323–3331. http://dl.acm.org/citation.cfm?id=3157382.3157469 Barcelona, Spain.

[18] Tatsunori B. Hashimoto, Megha Srivastava, Hongseok Namkoong, and Percy Liang. 2018. Fairness Without Demographics in Repeated Loss Minimization. *arXiv:1806.08010 [cs, stat]* (June 2018). http://arxiv.org/abs/1806.08010 arXiv: 1806.08010.

[19] G. Huang, Z. Liu, L. v d Maaten, and K. Q. Weinberger. 2017. Densely Connected Convolutional Networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2261–2269. https://doi.org/10.1109/CVPR.2017.243

[20] Vector Institute. 2019. Thousands of Images at the RadiologistâĂŹs Fingertips Seeing the Invisible. https://vectorinstitute.ai/2019/07/16/thousands-of-images-at-the-radiologists-fingertips-seeing-the-invisible/

[21] 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. 2019. CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison. *arXiv:1901.07031 [cs, eess]* (Jan. 2019). http://arxiv.org/abs/1901.07031 arXiv: 1901.07031.

[22] Alistair E. W. Johnson, Tom J. Pollard, Seth J. Berkowitz, Nathaniel R. Greenbaum, Matthew P. Lungren, Chih-ying Deng, Roger G. Mark, and Steven Horng. 2019. MIMIC-CXR: A large publicly available database of labeled chest radiographs. *arXiv:1901.07042 [cs, eess]* (Jan. 2019). http://arxiv.org/abs/1901.07042 arXiv: 1901.07042.

[23] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. 2016. Inherent Trade-Offs in the Fair Determination of Risk Scores. *arXiv:1609.05807 [cs, stat]* (Sept. 2016). http://arxiv.org/abs/1609.05807 arXiv: 1609.05807.

[24] Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. 2017. Counterfactual fairness. In *Advances in Neural Information Processing Systems*. 4066–4076.

[25] Paras Lakhani and Baskaran Sundaram. 2017. Deep Learning at Chest Radiography: Automated Classification of Pulmonary Tuberculosis by Using Convolutional Neural Networks. *Radiology* 284, 2 (Aug. 2017), 574–582. https://doi.org/10.1148/radiol.2017162326

[26] Francesco Locatello, Gabriele Abbati, Tom Rainforth, Stefan Bauer, Bernhard Schölkopf, and Olivier Bachem. 2019. On the Fairness of Disentangled Representations. *arXiv preprint arXiv:1905.13662* (2019).

[27] ~ Lukeoakdenrayner. 2019. Half a million x-rays! First impressions of the Stanford and MIT chest x-ray datasets. https://lukeoakdenrayner.wordpress.com/2019/02/25/half-a-million-x-rays-first-impressions-of-the-stanford-and-mit-chest-x-ray-datasets/

[28] Rupert G. Jr. Miller. 1981. *Simultaneous Statistical Inference*. Springer-Verlag New York. https://www.springer.com/gp/book/9781461381242.

[29] Lori Mosca, Elizabeth Barrett-Connor, and Nanette Kass Wenger. 2011. Sex/Gender Differences in Cardiovascular Disease Prevention What a Difference a Decade Makes. *Circulation* 124, 19 (Nov. 2011), 2145–2154. https://doi.org/10.1161/CIRCULATIONAHA.110.968792

[30] Akihiro Nishie, Daisuke Kakihara, Takeshi Nojo, Katsumasa Nakamura, Sachio Kuribayashi, Masumi Kadoya, Kuni Ohtomo, Kazuro Sugimura, and Hiroshi Honda. 2015. Current radiologist workload and the shortages in Japan: how many full-time radiologists are required? *Japanese Journal of Radiology* 33, 5 (May 2015), 266–272. https://doi.org/10.1007/s11604-015-0413-6

[31] Ziad Obermeyer and Sendhil Mullainathan. 2019. Dissecting Racial Bias in an Algorithm that Guides Health Decisions for 70 Million Peoples. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT*'19)*. USA, 89. Atlanta, GA.

[32] Diederik P. Kingma and Jimmy Ba. 2017. Adam: A Method for Stochastic Optimization. *arXiv:1412.6980v9* (2017).

[33] Alvin Rajkomar, Michaela Hardt, Michael D Howell, Greg Corrado, and Marshall H Chin. 2018. Ensuring fairness in machine learning to advance health equity. *Annals of internal medicine* 169, 12 (2018), 866–872.

[34] Pranav Rajpurkar, Jeremy Irvin, Robyn L. Ball, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Ding, Aarti Bagul, Curtis P. Langlotz, Bhavik N. Patel, Kristen W. Yeom, Katie Shpanskaya, Francis G. Blankenberg, Jayne Seekins, Timothy J. Amrhein, David A. Mong, Safwan S. Halabi, Evan J. Zucker, Andrew Y. Ng, and Matthew P. Lungren. 2018. Deep learning for chest radiograph diagnosis:

[35] A retrospective comparison of the CheXNeXt algorithm to practicing radiologists. *PLOS Medicine* 15, 11 (Nov. 2018), e1002686. https://doi.org/10.1371/journal.pmed.1002686

[35] Pranav Rajpurkar, Jeremy Irvin, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Ding, Aarti Bagul, Curtis Langlotz, Katie Shpanskaya, Matthew P. Lungren, and Andrew Y. Ng. 2017. CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning. *arXiv:1711.05225 [cs, stat]* (Nov. 2017). http://arxiv.org/abs/1711.05225 arXiv: 1711.05225.

[36] Abi Rimmer. 2017. Radiologist shortage leaves patient care at risk, warns royal college. *BMJ (Clinical research ed.)* 359 (2017), j4683. https://doi.org/10.1136/bmj.j4683

[37] Andrew B. Rosenkrantz, Wenyi Wang, Danny R. Hughes, and Richard Duszak. 2018. A County-Level Analysis of the US Radiologist Workforce: Physician Supply and Subspecialty Characteristics. *Journal of the American College of Radiology: JACR* 15, 4 (April 2018), 601–606. https://doi.org/10.1016/j.jacr.2017.11.007

[38] Caroline Rubin. 2017. Clinical radiology UK workforce census 2017 report. *The Royal college of radiologists* (2017), 40.

[39] Hojjat salehinejad, Shahrokh Valaee, Tim Dowdell, Errol Colak, and Joseph Barfett. 2018. Generalization of deep neural networks for chest pathology classification in x-rays using generative adversarial networks. *IEEE International Conference on Acoustics, Speech and Signal Processing* (2018).

[40] Prasanna Sattigeri, Samuel C. Hoffman, Vijil Chenthamarakshan, and Kush R. Varshney. 2018. Fairness GAN. *arXiv:1805.09910 [cs, stat]* (May 2018). http://arxiv.org/abs/1805.09910 arXiv: 1805.09910.

[41] Ashudeep Singh and Thorsten Joachims. 2019. policy learning for fairness in ranking. *arXiv preprint arXiv:1902.04056* (2019).

[42] Megha Srivastava, Hoda Heidari, and Andreas Krause. 2019. Mathematical notions vs. human perception of fairness: a descriptive approach to fairness for machine learning. *arXiv preprint arXiv:1902.04783* (2019).

[43] Nenad Tomašev, Xavier Glorot, Jack W Rae, Michal Zielinski, Harry Askham, Andre Saraiva, Anne Mottram, Clemens Meyer, Suman Ravuri, Ivan Protsyuk, et al. 2019. A clinically applicable approach to continuous prediction of future acute kidney injury. *Nature* 572, 7767 (2019), 116.

[44] Gabriela Torres-MejÃa, Robert A. Smith, MarÃa de la Luz Carranza-Flores, Andy Bogart, Louis MartÃnez-Matsushita, Diana L. Miglioretti, Karla Kerlikowske, Carolina Ortega-Olvera, Ernesto Montemayor-Varela, AngÃ©lica Angeles-Llerenas, Sergio Bautista-Arredondo, Gilberto SÃ¡nchez-GonzÃ¡lez, Olga G. MartÃnez-MontaÃ±ez, Santos R. Uscanga-SÃ¡nchez, Eduardo Lazcano-Ponce, and Mauricio HernÃ¡ndez-Ãvila. 2015. Radiographers supporting radiologists in the interpretation of screening mammography: a viable strategy to meet the shortage in the number of radiologists. *BMC Cancer* 15, 1 (May 2015), 410. https://doi.org/10.1186/s12885-015-1399-2

[45] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M. Summers. 2017. ChestX-ray8: Hospital-Scale Chest X-Ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases. In *Computer Vision and Pattern Recognition (CVPR) 2017*. IEEE, 2097–2106. http://openaccess.thecvf.com/content_cvpr_2017/html/Wang_ChestX-ray8_Hospital-Scale_Chest_CVPR_2017_paper.html

[46] Suchi andSendak Mark Wiens, Jenna andSaria, Marzyeh Ghassemi, Vincent X. Liu, Finale Doshi-Velez, Kenneth Jung, Katherine Heller, David Kale, Mohammed Saeed, Pilar N. Ossorio, Sonoo Thadaney-Israni, and Anna Goldenberg. 2019. High-performance medicine: the convergence of human and artificial intelligence. *Nature medicine* (2019).

[47] Depeng Xu, Shuhan Yuan, Lu Zhang, and Xintao Wu. 2018. FairGAN: Fairness-aware Generative Adversarial Networks. *arXiv:1805.11202 [cs, stat]* (May 2018). http://arxiv.org/abs/1805.11202 arXiv: 1805.11202.

[48] Li Yao, Eric Poblenz, Dmitry Dagunts, Ben Covington, Devon Bernard, and Kevin Lyman. 2017. Learning to diagnose from scratch by exploiting dependencies among labels. *arXiv:1710.10501 [cs]* (Oct. 2017). http://arxiv.org/abs/1710.10501 arXiv: 1710.10501.

[49] John R. Zech, Marcus A. Badgeley, Manway Liu, Anthony B. Costa, Joseph J. Titano, and Eric K. Oermann. 2018. Confounding variables can degrade generalization performance of radiological deep learning models. *PLOS Medicine* 15, 11 (Nov. 2018), e1002683. https://doi.org/10.1371/journal.pmed.1002683 arXiv: 1807.00431.

[50] Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. 2013. Learning fair representations. In *International Conference on Machine Learning*. 325–333.

[51] Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. 2018. Mitigating Unwanted Biases with Adversarial Learning. *arXiv:1801.07593 [cs]* (Jan. 2018). http://arxiv.org/abs/1801.07593 arXiv: 1801.07593.

A   APPENDIX: DISTRIBUTION OF TPR DISPARITY PER ATTRIBUTES, SUBGROUPS AND LABELS

Here we present the distribution of TPR disparities per subgroups/disease labels for all attributes. In a fair setting all subgroups TPRs per disease are the same and disparity is '0'. Conversely, negative and positive disparities denotes bias against and in favor of a subgroup, respectively. In Fig. 4 to Fig. 10, we sort disease labels based on the gap between the least and most favorable subgroups per disease, so that ones with smaller variance in disparity appear on the left side. We quantify TPR disparity across different subgroups similar to [10] for sex attributes. For age, race, and insurance type we quantify disparities using the difference between a subgroup's TPR and the TPRs median. We present the count of negative disparities per subgroup across all labels, excluding the 'No Finding' ('NF') label in order to consider disease labels only. The counts are based on the TPR disparities mean over five run. For Fig. 4 to Fig. 10 the label with the smallest and largest gap (distance) between the least/most favorable subgroups, the average cross labels gaps (between the the least/most favorable subgroups), and the count of the most frequent 'Unfavorable' and 'Favorable' subgroups, are summarized in Table. 3.
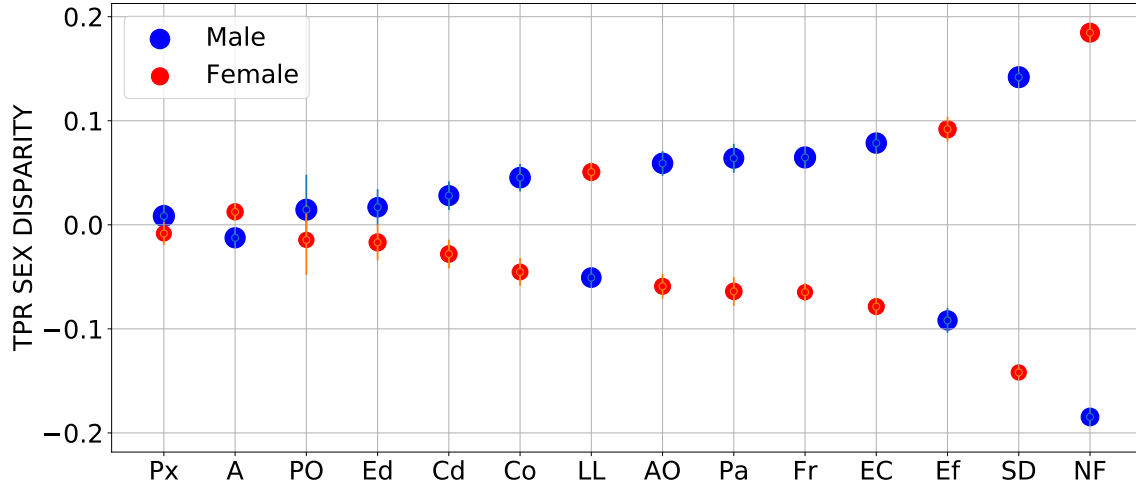

Figure 4: The sorted distribution of the TPR sex disparity in MIMIC-CXR dataset per disease. The x-axis labels are the abbreviation of the disease names (full name available in Table 2). The scatter plot's circle area is proportional to the patients percentages per subgroup. The TPR disparities are averaged over five run ±95% CI. The 95% CI are shown with arrows around the TPR disparities mean scatter plot. Count of 'Female' and 'Male' patients with negative disparities in disease labels are 10/13 and 3/13. Here, 'Pneumothorax' ('Px') is the label with the smallest gap (0.017) between the least/most favorable subgroups, where 'No Finding' ('NF') has the largest gap (0.369). The average cross labels gap between the the least/most favorable subgroups are 0.123.
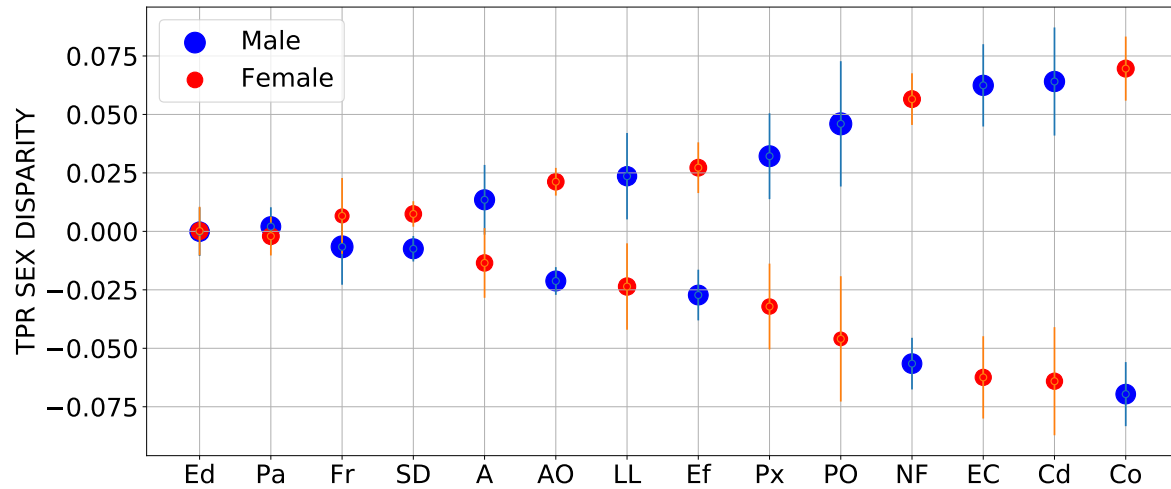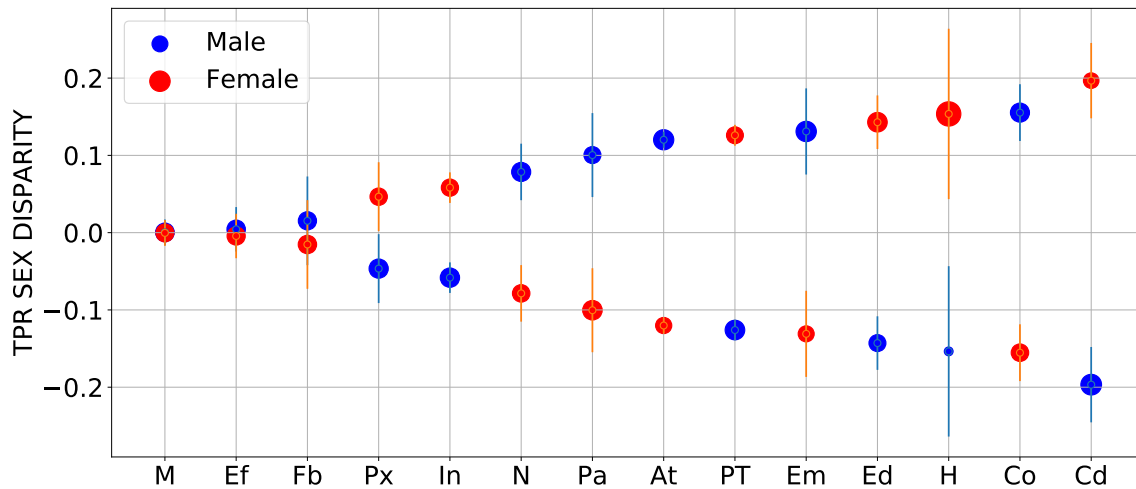
Figure 5: The sorted distribution of the TPR sex disparity in MIMIC-CXR dataset per disease. The *x*-axis labels are the abbreviation of the disease names (full name available in Table 2). The scatter plot's circle area is proportional to the patients percentages per subgroup. The TPR disparities are averaged over five run ±95% CI. The 95% CI are shown with arrows around the TPR disparities mean scatter plot. Count of 'Female' and 'Male' patients with negative disparities in disease labels are 7/13 and 6/13. Here, 'Edema' ('Ed') is the label with the smallest gap (0.000) between the least/most favorable subgroups, where 'Consolidation' ('Co') has the largest gap (0.139). The average cross labels gap between the the least/most favorable subgroups are 0.062.

Figure 6: The sorted distribution of the TPR sex disparity in MIMIC-CXR dataset per disease. The *x*-axis labels are the abbreviation of the disease names (full name available in Table 2). The scatter plot's circle area is proportional to the patients percentages per subgroup. The TPR disparities are averaged over five run ±95% CI. The 95% CI are shown with arrows around the TPR disparities mean scatter plot. Count of 'Female' and 'Male' patients with negative disparities in disease labels are 8/14 and 6/14. Here, 'Mass' ('M') is the label with the smallest gap (0.001) between the least/most favorable subgroups, where 'Cardiomegaly' ('Cd') has the largest gap (0.393). The average cross labels gap between the the least/most favorable subgroups are 0.190.
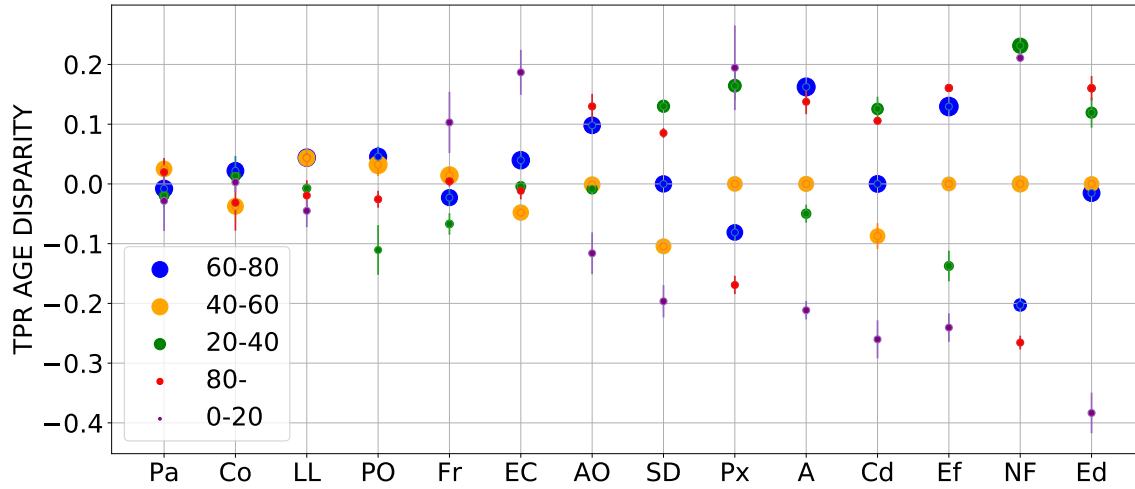
Figure 7: The sorted distribution of the TPR age disparity in MIMIC-CXR dataset per disease. The x-axis labels are the abbreviation of the disease names (full name available in Table 2). The scatter plot's circle area is proportional to the percentage of patients in each subgroup. The TPR disparities are averaged over five run ±95% CI. The 95% CI are shown with arrows around the mean of TPRs scatter plot. Count of patients in age subgroups '40-60', '60-80', '20-40','80-' and '0-20' with negative gap in disease labels are 5/13, 4/13, 8/13, 5/13 and 8/13. Here, 'Pneumonia' ('Pa') is the label with the smallest gap (0.054) between the least/most favorable subgroups, where 'Edema' ('Ed') has the largest gap (0.544). The average cross labels gap between the the least/most favorable subgroups are 0.279.
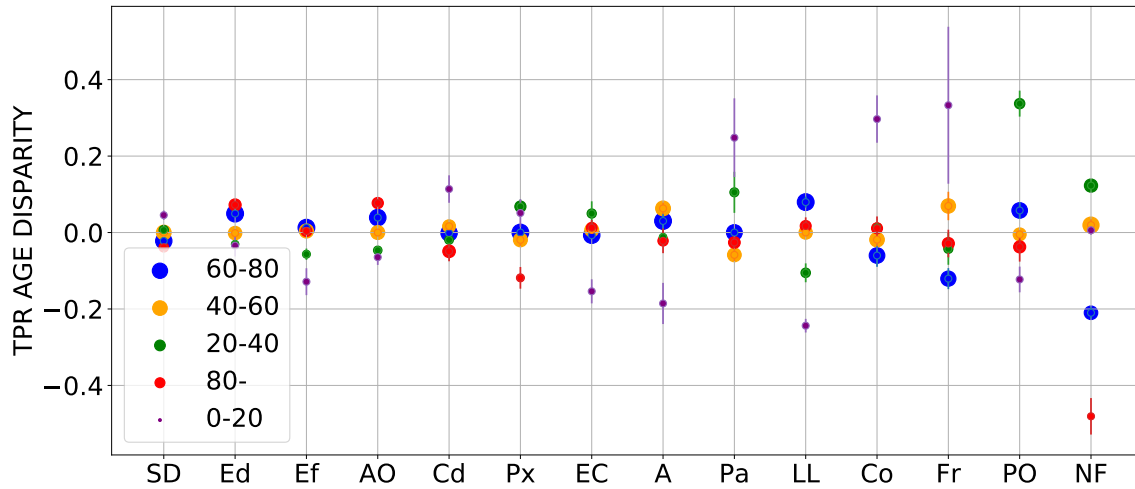

Figure 8: The sorted distribution of the TPR age disparity in CheXpert dataset per disease. The x-axis labels are the abbreviation of the disease names (full name available in Table 2). The scatter plot's circle area is proportional to the percentage of patients in each subgroup. The TPR disparities are averaged over five run ±95% CI (CI are shown with arrows around the mean). Count of patients in age subgroups '40-60', '60-80', '20-40','80-' and '0-20' with negative gap in disease labels are 5/13, 6/13, 7/13, 7/13 and 7/13. Here, 'Support Devices' ('SD') is the label with the smallest gap (0.082) between the least/most favorable subgroups, where 'No Finding' ('NF') has the largest gap (0.604). The average cross labels gap between the the least/most favorable subgroups are 0.270.
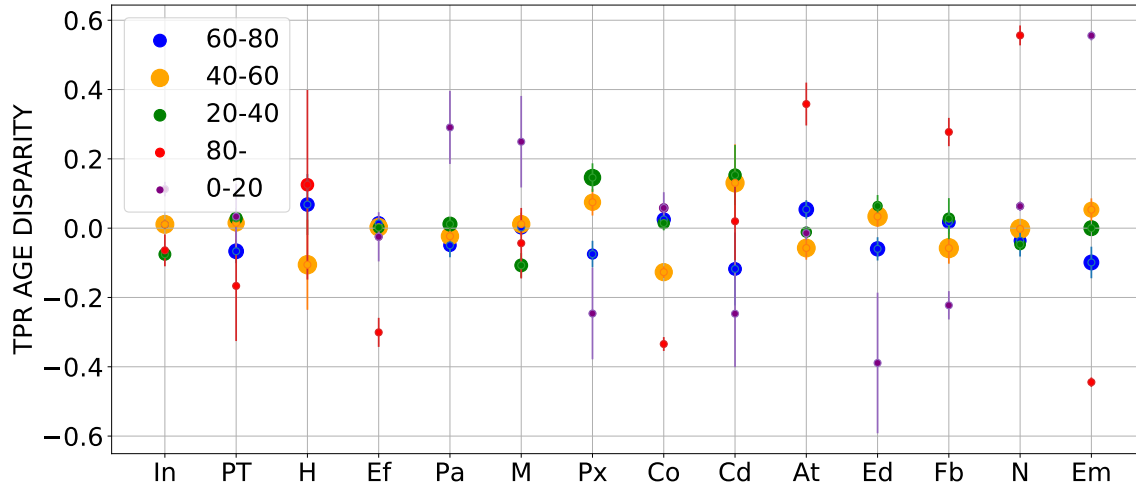
Figure 9: The sorted distribution of the TPR age disparity in ChestXray8 dataset per disease. The *x*-axis labels are the abbreviation of the disease names (full name available in Table 2). The scatter plot's circle area is proportional to the patients membership. The TPR disparities are averaged over five run ±95%CI (the CI are shown with arrows around the mean). Count of patients in age subgroups '40-60', '60-80', '20-40','80-' and '0-20' with negative gap in disease labels are 6/14, 7/14, 4/14, 6/14 and 6/14 Here, 'Infiltration' ('In') is the label with the smallest gap (0.188) between the least/most favorable subgroups, where 'Emphysema' ('Em') has the largest gap (1.00). The average cross labels gap between the the least/most favorable subgroups are 0.413.
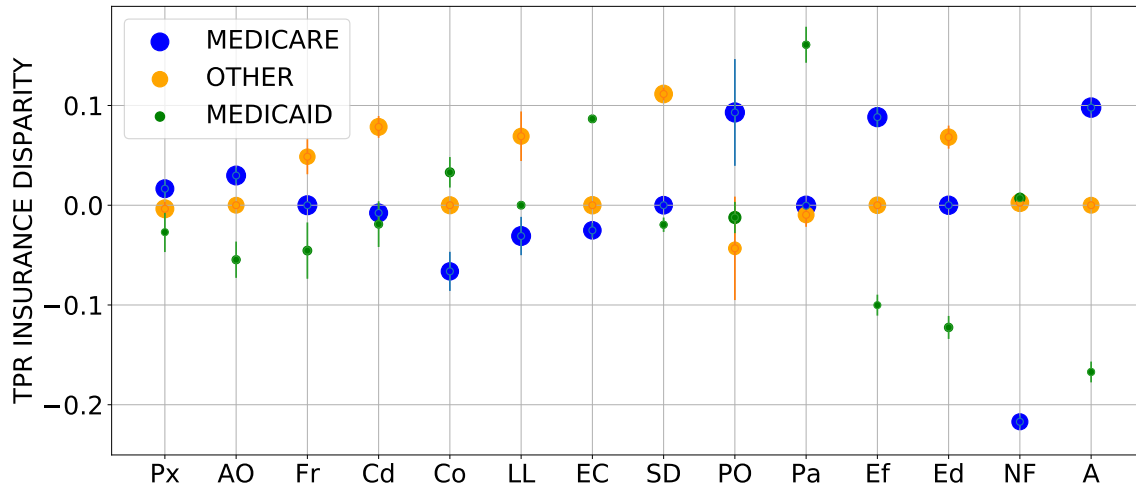

Figure 10: The sorted distribution of the TPR insurance type disparity in MIMIC-CXR dataset per disease.The *x*-axis labels are the abbreviation of the disease names (full name available in Table 2). The scatter plot's circle area is proportional to the patients membership. The TPR disparities are averaged over five run ±95%CI (the CI are shown with arrows around the mean). The patients with 'Medicaid' insurance are the most unfavorable subgroup. Count of patients in insurance subgroups 'Other', 'Medicare', and 'Medicaid ' with negative gap in disease labels are 3/13, 5/13, and 9/13. Here, 'Pneumothorax' ('Px') is the label with the smallest gap (0.044) between the least/most favorable subgroups, where 'Atelectasis' ('A') has the largest gap (0.265). The average cross labels gap between the the least/most favorable subgroups are 0.138.