Pacific Symposium on Biocomputing 26:232-243 (2021)

CheXclusion: Fairness gaps in deep chest X-ray classifiers

Laleh Seyyed-Kalantari[1,2*], Guanxiong Liu[1,2], Matthew McDermott[3], Irene Y. Chen[3], Marzyeh Ghassemi[1,2]

[1] *Computer Science, University of Toronto, Toronto, Ontario, Canada*

[2] *Vector Institute, Toronto, Ontario, Canada*

[3] *Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA USA*

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 3 prominent public chest X-ray datasets: MIMIC-CXR, Chest-Xray8, CheXpert, as well as a multi-site aggregation of all those datasets. We evaluate the *TPR disparity* – the difference in true positive rates (TPR) – among different protected attributes such as patient sex, age, race, and insurance type as a proxy for socioeconomic status. We demonstrate that TPR disparities exist in the state-of-the-art classifiers in all datasets, for all clinical tasks, and all subgroups. A multi-source dataset corresponds to the smallest disparities, suggesting one way to reduce bias. We find that TPR disparities are not significantly correlated with a subgroup's proportional disease burden. As clinical models move from papers to products, we encourage clinical decision makers to carefully audit for algorithmic disparities prior to deployment. Our supplementary materials can be found at, `http://www.marzyehghassemi.com/chexclusion-supp-3/`.

Keywords: fairness, medical imaging, chest x-ray classifier, computer vision.

1. Introduction

Chest X-ray imaging is an important screening and diagnostic tool for several life-threatening diseases, but due to the shortage of radiologists, this screening tool cannot be used to treat all patients.[1,2] Deep-learning-based medical image classifiers are one potential solution, with significant prior work targeting chest X-rays specifically,[3,4] leveraging large-scale publicly available datasets,[3,5,6] and demonstrating radiologist-level accuracy in diagnostic classification.[6–8]

Despite the seemingly clear case for implementing AI-enabled diagnostic tools,[9] moving such methods from paper to practice require careful thought.[10] Models may exhibit disparities in performance across protected subgroups, and this could lead to different subgroups receiving different treatment.[11] During evaluation, machine learning algorithms usually optimize for, and

*Corresponding author email: `laleh@cs.toronto.edu`

© 2020 The Authors. Open Access chapter published by World Scientific Publishing Company and distributed under the terms of the Creative Commons Attribution Non-Commercial (CC BY-NC) 4.0 License.

report performance on, the general population rather than balancing accuracy on different subgroups. While some variance in performance is unavoidable, mitigating any systematic bias against protected subgroups 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*. We train classifiers on 3 large, public chest X-ray datasets: MIMIC-CXR,[5] CheXpert,[6] Chest-Xray8,[3] as well as an additional datasets formed of the aggregation of those three datasets on their shared labels. In each case, we implement chest X-ray pathology classifiers via a deep convolutional neural network (CNN) chest X-ray images as inputs, and optimize the multi-label probability of 14 diagnostic labels simultaneously. Because researchers have observed health disparities with respect to race,[12] sex,[13] age,[14] and socioeconomic status,[12] we extract structural data on race, sex, and age; we also use insurance type as an imperfect proxy[11] for socioeconomic status. To our knowledge, we are the first to examine whether SOTA chest X-ray pathology classifiers display systematic bias across race, age, and insurance type.

We analyze equality of opportunity[15] as our fairness metric based on the needs of the clinical diagnostic setting. In particular, 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.

We find three major findings: First, that there are indeed extensive patterns of bias in SOTA classifiers, shown in TPR disparities across datasets. Secondly, the disparity rate for most attributes/ datasets pairs is not significantly correlated with the subgroups' proportional disease membership. These findings suggest that underrepresented subgroups could be vulnerable to mistreatment in a systematic deployment, and that such vulnerability may not be addressable simply through increasing subgroup patient count. Lastly, we find that using the multi-source dataset which combines all the other datasets yields the lowest TPR disparities, suggesting using multi-source datasets may combat bias in the data collection process. As researchers increasingly apply artificial intelligence and machine learning to precision medicine, we hope that our work demonstrates how predictive models trained on large, well-balanced datasets can still yield disparate impact.

2. Background and Related Work

Fairness and Debiasing. Fairness in machine learning models is a topic of increasing attention, spanning sex bias in occupation classifiers,[16] racial bias in criminal defendant risk assessments algorithms,[17] and intersectional sex-racial bias in automated facial analysis.[18] 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.[18] Minority groups may also be under-sampled, or the features collected may not be indicative of their trends.[19] There are several conflicting definitions of fairness, many of which are not simultaneously achievable.[20] The appropriate choice of a disparity metric is generally task dependent, but balancing error rates between different subgroups is a common consideration,[15,17] with equal accuracy across subgroups being a popular choice in medical set-

tings.[21] In this work, we consider the equality of opportunity notion of fairness and evaluate the rate of correct diagnosis in patients across several protected attribute groups.

Ethical Algorithms in Health. Using machine learning algorithms to make decisions raises serious ethical concerns about risk of patient harm.[22] Notably, biases have already been demonstrated in several settings, including racial bias in the commercial risk score algorithms used in hospitals,[23] or an increased risk of electronic health record (EHR) miss-classification in patients with low socioeconomic status.[24] It is crucial that we actively consider fairness metrics when building models in systems that include human and structural biases.

Chest X-Ray Classification. With the releases of large public datasets like Chest-Xray8,[3] CheXpert,[6] and MIMIC-CXR,[5] many researchers have begun to train large deep neural network models for chest X-ray diagnosis.[4,6,8,25] Prior work[8] demonstrates a diagnostic classifier trained on Chest-Xray8 can achieve radiologist-level performance. Other work on CheXpert[6] reports high performance for five of their diagnostic labels. To our knowledge, however, no works have yet been published which examined whether any of these algorithms display systematic bias over age, race and insurance type (as a proxy of socialeconomic status).

3. Data

We use three public chest X-ray radiography datasets described in Table 1: MIMIC-CXR (CXR),[5] CheXpert (CXP),[6] Chest-Xray8 (NIH).[3] Images in CXR, CXP, and NIH are associated with 14 diagnostic labels (see Table 2). We combine all non-positive labels within CXR and CXP (including "negative", "not mentioned", or "uncertain") into an aggregate "negative" label for simplicity, equivalent to "U-zero" study of 'NaN' label in CXP. In CXR and CXP, one of the 14 labels is "No Finding", meaning no disease has been diagnosed for the image and all the other 13 labels are 0. Of the 14 total disease labels, only 8 are shared amongst all 3 datasets. Using these 8 labels, we define a multi-site dataset (ALL) that consists of the aggregation of all images in CXR, CXP, and NIH defined over this restricted label schema.

These datasets contain protected subgroup attributes, the full list of which includes 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. NIH, CXP, and ALL only have the patient sex and age, while CXR also has race and insurance type data (excluding around 100,000 images).

4. Methods

We implement CNN-based models to classify chest X-ray images into 14 diagnostic labels. We train separate models for CXR,[5] CXP,[6] NIH[3] and ALL and explore their fairness with respect to patient sex and age for all 4 datasets as well as race and insurance type for CXR.

4.1. *Models*

We initialize a 121-layer DenseNet[26] with pre-trained weights from ImageNet[27] and train multi-label models with a multi-label binary cross entropy loss. The 121-layer DenseNet was used as it produced the best results in prior studies[6,8] . We use a 80-10-10 train-validation-test

Table 1. Description of chest X-ray datasets, MIMIC-CXR (CXR),[5] CheXpert (CXP),[6] Chest-Xray8 (NIH).[3] and their aggregation on 8 shared labels (ALL). Here, the number of images, patients, view types, and the proportion of patients per subgroups of sex, age, race, and insurance type are presented. 'Frontal' and 'Latral' 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.

Subgroup	Attribute	**CXR**[5]	**CXP**[6]	**NIH**[3]	**ALL**
	# Images	371,858	223,648	112,120	707,626
	# Patients	65,079	64,740	30,805	129,819
	View	Frontal/Lateral	Frontal/Lateral	Frontal	Frontal/Lateral
sex	Female	47.83%	40.64%	43.51%	44.87%
	Male	52.17%	59.36%	56.49%	55.13%
Age	0-20	2.20%	0.87%	6.09%	2.40%
	20-40	19.51%	13.18%	25.96%	18.53%
	40-60	37.20%	31.00%	43.83%	36.29%
	60-80	34.12%	38.94%	23.11%	33.90%
	80+	6.96%	16.01%	1.01%	8.88%
Race	Asian	3.24%	—	—	—
	Black	18.59%	—	—	—
	Hispanic	6.41%	—	—	—
	Native	0.29%	—	—	—
	White	67.64%	—	—	—
	Other	3.83%	—	—	—
Insurance	Medicare	46.07%	—	—	—
	Medicaid	8.98%	—	—	—
	Other	44.95%	—	—	—

split with no patient shared across splits. We resize all images to 256×256 and normalize via the mean and standard deviation of the ImageNet dataset.[27] We apply center crop, random horizontal flip and random rotation, as some of the images maybe flipped or rotated within the dataset. The initial degree of random rotation is chosen by hyperparameter tuning. We use Adam[28] optimization with default parameters, and decrease the learning rate (LR) 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. Thus the ultimate number of epochs for training each model is varied based on the early stop condition. For NIH, CXP and CXR we first tune models to get the highest average area under the receiver operating characteristic curve (AUC) over 14 labels by fine tuning the LR. For the best achieved model, we fine tune the degree of random rotation data augmentation from the set of 7, 10 and 15 and select the best model. Following this, best initial LR is 0.0005 for CXR and NIH where it is achieved

as 0.0001 for CXP. Also, best initial degree for random rotation data augmentation is 10 for NIH and 15 for the CXR and CXP. For training on ALL, we use the majority vote of the best hyperparameters per individual dataset (e.g. 0.0005 initial LR and 15 degree random rotation). We then, fix the hyperparameters of the best model and train four extra models with the same hyperparameters but different random seeds between 0 to 100, per dataset. We report all the metrics based on the mean and 95% confidence intervals (CI) achieved over five studies per dataset. We choose batch size of 48 to use the maximum memory capacity of the GPU, for all datasets except NIH where we choose 32 similar to prior work.[8] The output of the network is an array of 14 numbers between 0 and 1 indicating the probability of each disease label. The binary prediction threshold per disease is chosen to maximize the F1 score measure on the validation dataset. We train models using a single NVIDIA GPU with 16G of memory in approximately 9, 20, 40, and 90 hours for NIH, CXP, CXR, and ALL, respectively.

4.2. *Classifier Disparity Evaluation*

Our primary measurement of bias is *TPR disparity*. For example, given a binary subgroup attribute such as sex (which in our data we classify as either 'male' or 'female'), we mimic prior work[16] and define the TPR disparity per diagnostic label i as simply the TPR of label i restricted to female patients minus that for male patients. More formally, letting g be the binary subgroup attribute, we define $\text{TPR}_{g,i} = P[\hat{Y}_i = y_i | G = g, Y_i = y_i]$, and the TPR disparity as, $\text{Gap}_{g,i} = \text{TPR}_{g,i} - \text{TPR}_{\neg g,i}$. For non-binary attributes $S_1, \ldots, S_N$, we use the difference between a subgroup's TPR and the median of all TPRs to define TPR disparity of the jth subgroup for the ith label as, $\text{Gap}_{S_j,i} = \text{TPR}_{S_j,i} - \text{Median}(\text{TPR}_{S_1,i}, .., \text{TPR}_{S_k,i})$.

5. Experiments

First, we demonstrate that the classifiers we train on all datasets reach near-SOTA level performance. This motivates using them to study fairness implications, as we can be confident any problematic disparities are not simply reflective of poor overall performance. Next, we explicitly test these classifiers for their implications on fairness. We target two investigations:

(1) **TPR disparity:** We quantify the TPR disparity per subgroup/disease for sex and age across all 4 datasets, and due to data availability for race and insurance type on CXR.
(2) **TPR disparity in proportion to membership:** We investigate the distribution of the positive patient proportion per subgroup S_j and label y_i (which is given by $P[S = S_j | Y = y_i]$) and the effect on TPR disparities. Prior work on chest X-ray diagnosis prediction has suggested data imbalance can explain sex TPR disparities[29] while work in other domains illustrates that disparities in small or vulnerable subgroups could be propagated if put into practice,[16,30] and these experiments are meant to probe that hypothesis.

6. Results

One potential reason that a model may be biased is because of poor performance, but we demonstrate that our models achieve near-SOTA classification performance. Table 2 shows overall performance numbers across all tasks and datasets. Though results have non-trivial

Table 2. The AUC for chest X-ray classifiers trained on CXP, CXR, NIH, and ALL averaged over 5 runs ±95%CI, where all runs have same hyperparameters but different random seed. ('Airspace Opacity'[5] and 'Lung Opacity'[6] denote the same label.)

Label (Abbr.)	**CXR**	**CXP**	**NIH**	**ALL**
Airspace Opacity (AO)	0.782 ± 0.002	0.747 ± 0.001	—	—
Atelectasis (A)	0.837 ± 0.001	0.717 ± 0.002	0.814 ± 0.004	0.808 ± 0.001
Cardiomegaly (Cd)	0.828 ± 0.002	0.855 ± 0.003	0.915 ± 0.002	0.856 ± 0.001
Consolidation (Co)	0.844 ± 0.001	0.734 ± 0.004	0.801 ± 0.005	0.805 ± 0.001
Edema (Ed)	0.904 ± 0.002	0.849 ± 0.001	0.915 ± 0.003	0.898 ± 0.001
Effusion (Ef)	0.933 ± 0.001	0.885 ± 0.001	0.875 ± 0.002	0.922 ± 0.001
Emphysema (Em)	—	—	0.897 ± 0.002	—
Enlarged Card (EC)	0.757 ± 0.003	0.668 ± 0.005	—	—
Fibrosis	—	—	0.788 ± 0.007	—
Fracture (Fr)	0.718 ± 0.007	0.790 ± 0.006	—	—
Hernia (H)	—	—	0.978 ± 0.004	—
Infiltration (In)	—	—	0.717 ± 0.004	—
Lung Lesion (LL)	0.772 ± 0.006	0.780 ± 0.005	—	—
Mas (M)	—	—	0.829 ± 0.006	—
Nodule (N)	—	—	0.779 ± 0.006	—
No Finding (NF)	0.868 ± 0.001	0.885 ± 0.001	—	0.890 ± 0.000
Pleural Thickening (PT)	—	—	0.813 ± 0.006	—
Pleural Other (PO)	0.848 ± 0.003	0.795 ± 0.004	—	—
Pneumonia (Pa)	0.748 ± 0.005	0.777 ± 0.003	0.759 ± 0.012	0.784 ± 0.001
Pneumothorax (Px)	0.903 ± 0.002	0.893 ± 0.002	0.879 ± 0.005	0.904 ± 0.002
Support Devices (SD)	0.927 ± 0.001	0.898 ± 0.001	—	—
Average	**0.834 ± 0.001**	**0.805±0.001**	**0.840 ± 0.001**	**0.859 ± 0.001**

variability, we show similar performance to the published SOTA of NIH,[8] the only dataset for which a published SOTA comparison exists for all labels. Note that the published results for CXP[6] are on a private, unreleased dataset of only 200 images and 5 labels. Our results for CXP 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.

6.1. *TPR Disparities*

We calculate and identify TPR disparities and 95% CI across all labels, datasets and attributes. We see many instances of positive and negative disparities, which can denote bias for or against of a subgroup, here referred to *favorable* and *unfavorable* subgroups. As an illustrative example Fig. 1 shows the race TPR disparities distribution sorted by the the gap between least and most favorable subgroups per label. In a fair setting, all subgroups would have no appreciable TPR disparities, yielding a gap between least and most favorable subgroups within a label at '0'. Table 3 shows the summary of the disparities in all attributes and datasets. We note that the most frequent unfavorable subgroups are those with social disparities in the healthcare

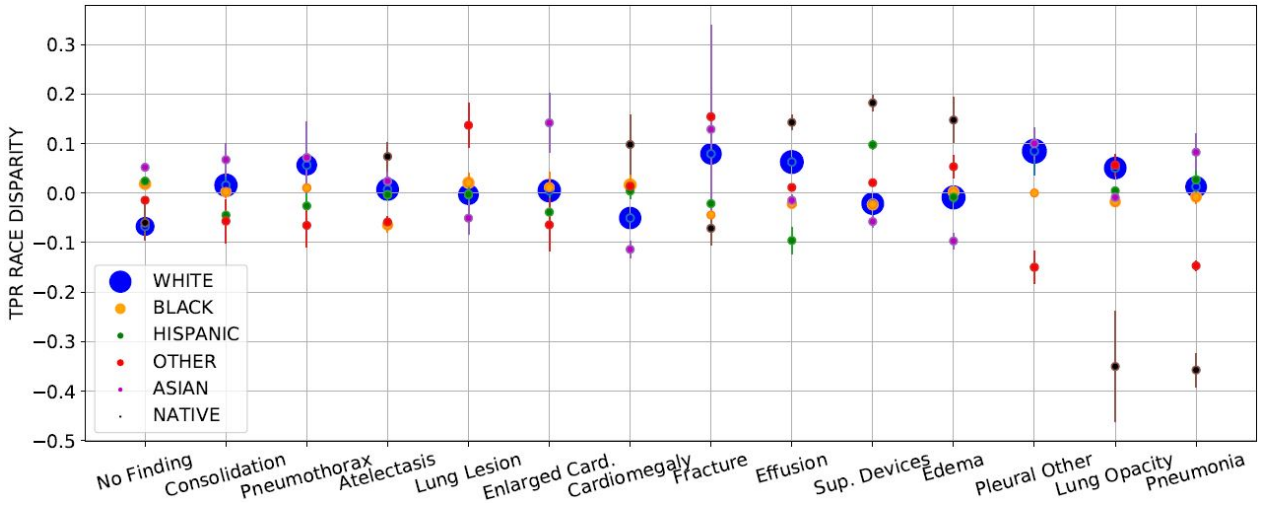

Fig. 1. The sorted distribution of TPR race disparity of CXR (y-axis) with label (x-axis). The scatter plot's circle area is proportional to group size. TPR disparities are averaged over five runs ±95%CI (shown with arrows). Hispanic patients are most unfavorable (highest count of negative TPR disparities, 9/13) whereas White patients are most favorable subgroup (9/13 zero or positive disparities). Labels 'No Finding' ('NF') and 'Pneumonia' ('Pa') have smallest (0.119) and largest (0.440) gap between least/most favorable subgroups. The average cross 14 labels gap is 0.226.

Table 3. Disparities overview over attributes and datasets. We average per label gaps between the least and most favorable subgroup's TPR disparities per attributes/datasets to obtain the average cross-label gap. The labels (full names on Table 2) that obtained the smallest and largest gaps are shown next to the average cross-label gap column, along with their gaps. We summarize and label in columns the most frequent "Unfavorable" and "Favorable" subgroups count, which are the ones that experience TPRs disparities below or above the zero gap line. See Section 6.1 for more details.

Attribute	**Dataset**	**Average Cross-Label Gap**			**Unfavorable**	**Favorable**
		Gap	**Lowest**	**Greatest**		
Sex	ALL	**0.045**	Ef:0.001	Pa:0.105	Female (4/7)	Male (4/7)
	CXP	0.062	Ed:0.000	Co:0.139	Female (7/13)	Male (7/13)
	CXR	0.072	Ed:0.011	EC:0.151	Female (10/13)	Male (10/13)
	NIH	0.190	M:0.001	Cd:0.393	Female (8/14)	Male (8/14)
Age	ALL	**0.215**	Ef:0.115	NF:0.444	0-20 (5/7)	40-60,60-80(5/7)
	CXR	0.245	SD:0.091	Cd:0.440	0-20, 20-40 (7/13)	60-80 (10/13)
	CXP	0.270	SD:0.084	NF:0.604	0-20, 20-40, 80- (7/13)	40-60 (8/13)
	NIH	0.413	In:0.188	Em:1.00	60-80 (7/14)	20-40 (9/14)
Race	CXR	0.226	NF:0.119	Pa:0.440	Hispanic (9/13)	White (9/13)
Insurance	CXR	0.100	SD:0.021	PO:0.190	Medicaid (10/13)	Other (10/13)

system, e.g., women and minorities, but no disease is consistently at the highest or lowest disparity. We show the average cross-label gap between and the labels of the least and most favorable subgroups per dataset and attributes. We count the number of time each subgroups experience negative disparities (unfavorable) and zero or positive disparities (favorable) across disease labels and report the most frequent unfavorable and favorable subgroups by count in Table 3. For CXP and CXR, we exclude "No Finding" label in the count (counts are out of 13) as we want to check negative bias in disease labels. Notably, the model trained on ALL has the smallest average cross- label gap between least/most favorable groups for sex and age.

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 we test multiple (33) hypotheses, (33 total comparisons amongst all protected attributes considered) with a desired significance level ($p < 0.05$), then based on Bonferroni correction,[31] the statistical significance level for each hypothesis is $p < 0.0015$ (0.05/33). The majority of correlation coefficients listed are positive, but the only statistically significant correlations are: race Other (r: 0.782, p: 0.0009) & age subgroups, 20-40 (r: 0.766, p: 0.0013), 60-80 (r: 0.787, p: 0.0008) and 80- (r:0.858 , p: 0.0000) in CXR, age group 60-80 (r: 0.853, p: 0.0001) in CXP, and age group 60-80 (r: 0.936, p: 0.0006) in ALL.

7. Summary and Discussion

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. We focus on TPR disparities similar to prior work,[16] checking if the sick members of the different subgroups are given correct diagnosis at similar rates.

Our results demonstrate several main takeaways. First, all datasets and tasks display non-trivial TPR disparities. These disparities could pose serious barriers to effective deployment of these models and invite additional changes in either dataset design and/or modeling techniques to ensure more equitable models. Second, using a multi-source dataset leads to smaller TPR disparities, potentially due to removing bias in the data collection process. Third, 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.

7.1. *Extensive Patterns of Bias*

We find that all datasets and tasks contain meaningful patterns of bias although no diseases are consistently at the highest or lowest disparity rates across all attributes and datasets. These disparities are present with respect to age and sex in all settings, with consistent subgroups (female, 0-20) showing consistently unfavorable outcomes. Note that in the case of the sex disparities, "female" patients are universally the least favored subgroup *despite* the fact that the proportion of female patients is only slightly less than male patients in all 4 datasets.

We also observe TPR disparities with respect to the patient race and insurance type in the CXR dataset. White patients, the majority, are the most favorable subgroup, where Hispanic

patients are the most unfavorable. Additionally, bias exists against patients with Medicaid insurance, who are the minority population and are often from lower socioeconomic status. They are the most unfavorable subgroup with the model often providing incorrect diagnoses.

7.2. *Bias Reduction Through Multi-source Data*

Of the four datasets, the multi-source dataset led to the smallest disparities with respect to age and sex. Based on notions of generalizability in healthcare,[10,32] we hypothesize that this improvement stems from the combination of large datasets reducing data collection bias.

7.3. *Correlation Between TPR Disparities and Membership Proportion*

Although prior work has raised data imbalance as a potential cause of sex bias,[29] we observe TPR disparities are not often significantly correlated with disease membership. While we observe positive correlation between subgroups membership and TPR disparities, only 6 of 33 subgroups showed statistically significant correlation. By inspection, we identify diseases with the same patient proportion of a subgroup and completely 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, having the same portion of images within all labels may not guarantee lack of bias.

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 on average 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.[33] 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., 67.6% 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.[24]

Although "de-biasing" techniques[34,35] may reduce disparities, we should not ignore the important biases inherent in existent large public datasets. There are a number of reasons why dataset may induce disparities in algorithms, from imbalanced datasets to differences in statistical noise in each group (e.g. unmeasured predictive features) to differences in access to healthcare for patients of different groups.[12,19] For instance, an algorithm that can classify

skin cancer[36] 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.[18] Intentionally adjusting the datasets to reduce disparities in to protect minorities and the subgroups with high disparities is one potential option in dataset creation, though our analyses suggest that dataset membership cannot always ameliorate bias.

With the great promise of advanced models for clinical care, we caution that advanced SOTA models must be carefully checked for such biases as those we have identified. Disparities in small or vulnerable subgroups could be propagated[30] within the development of machine learning models. This raises serious ethical concerns[22] about the equal accessibility 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 rigorous fairness analyses before deployment. Clear disclaimers about the dataset collection process and potential resulting algorithmic bias could improve model assessment for clinical use.

8. Limitations and Future Work

As SOTA deep learning diagnosis algorithms become more promising for medical screening, model bias investigation is essential. This work is a first step in quantifying these biases so that approaches for amelioration can be developed. However, important future work remains.

First, we note that across these models, our source of diagnostic labels for these images must be considered at best "silver" labels, as all currently existing public chest X-ray datasets use automatically deteremined labels based on natural language processing (NLP) techniques to extract labels from the radiology reports. These silver labels may be incorrect, in ways that could compound with observe biases or model errors, a risk that warrants further investigation. Additionally, we must consider the quality of the imaging devices themselves, the region of data collection, and the patient demographics at each hospital collection site. For instance, NIH was gathered from a hospital that covers more complicated cases, CXP contains more tertiary cases, and CXR was gathered from an emergency department, and prior literature has already shown that models are fully capable of taking advantage of such confounders.[32] These challenges may affect both the label quality,[37] and any patterns of bias in the labels, thereby affecting the resulting fairness metrics. Finally, exploration of existing de-biasing techniques, however limited, should also be undertaken over this modality to see if any of the problems we identified here can be resolved.

9. Conclusion

While the development and deployment of machine learning models in a clinical setting poses exciting opportunities, great care must be taken to understand how existing biases may be exacerbated and propagated. We show the TPR disparity of SOTA chest X-ray pathology classifiers trained on 4 datasets, (MIMIC-CXR, ChestX-ray8, CheXpert, and aggregation of those three on shared labels) across 14 diagnostic labels. We quantify the TPR disparity across datasets along sex, age, race and insurance type. 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.

Acknowledgment

We acknowledge the support of the Natural Sciences and Engineering Research Council of Canada (NSERC, funding number PDF-516984), Microsoft Research, CIFAR, NSERC Discovery Grant, and high performance computing platforms of Vector Institute. We also thank Dr. Alistair Johnson, Dr. Errol Colak and Grey Kuling for productive discussions.

References

1. A. Rimmer, Radiologist shortage leaves patient care at risk, warns royal college, *BMJ (Clinical research ed.)* **359**, p. j4683 (2017).
2. F. S Ali, S. G Harrington, S. B Kenned and S. Hussain, Diagnostic radiology in liberia: a country report, *Journal of Global Radiology* **1(2)** (2015).
3. X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri and R. M. Summers, ChestX-ray8: Hospital-Scale Chest X-Ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases, 2097 (2017).
4. L. Yao, E. Poblenz, D. Dagunts, B. Covington, D. Bernard and K. Lyman, Learning to diagnose from scratch by exploiting dependencies among labels, *arXiv:1710.10501* (2017).
5. A. E. W. Johnson, T. J. Pollard, S. J. Berkowitz, N. R. Greenbaum, M. P. Lungren, C.-y. Deng, R. G. Mark and S. Horng, MIMIC-CXR: A large publicly available database of labeled chest radiographs, *arXiv:1901.07042* (2019).
6. J. Irvin, P. Rajpurkar, M. Ko, Y. Yu, S. Ciurea-Ilcus, C. Chute, H. Marklund, B. Haghgoo, R. Ball, K. Shpanskaya, J. Seekins, D. A. Mong, S. S. Halabi, J. K. Sandberg, R. Jones, D. B. Larson, C. P. Langlotz, B. N. Patel, M. P. Lungren and A. Y. Ng, CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison, *arXiv:1901.07031* (2019).
7. P. Rajpurkar, J. Irvin, K. Zhu, B. Yang, H. Mehta, T. Duan, D. Ding, A. Bagul, C. Langlotz, K. Shpanskaya, M. Lungren and A. Ng, Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning (2017).
8. P. Rajpurkar, J. Irvin, R. L. Ball, K. Zhu, B. Yang, H. Mehta, T. Duan, D. Ding, A. Bagul, C. P. Langlotz, B. N. Patel, K. W. Yeom, K. Shpanskaya, F. G. Blankenberg, J. Seekins, T. J. Amrhein, D. A. Mong, S. S. Halabi, E. J. Zucker, A. Y. Ng and M. P. Lungren, Deep learning for chest radiograph diagnosis: A retrospective comparison of the CheXNeXt algorithm to practicing radiologists, *PLOS Medicine* **15**, p. e1002686 (2018).
9. V. Institute, Thousands of images at the Radiologist's fingertips seeing the invisible (2019).
10. M. Ghassemi, T. Naumann, P. Schulam, A. L. Beam, I. Y. Chen and R. Ranganath, Practical guidance on artificial intelligence for health-care data, *The Lancet Digital Health* **1**, e157 (2019).
11. I. Y. Chen, P. Szolovits and M. Ghassemi, Can ai help reduce disparities in general medical and mental health care?, *AMA journal of ethics* **21**, 167 (2019).
12. I. Kawachi, N. Daniels and D. E. Robinson, Health disparities by race and class: why both matter, *Health Affairs* **24**, 343 (2005).
13. D. E. Hoffmann and A. J. Tarzian, The girl who cried pain: a bias against women in the treatment of pain, *The Journal of Law, Medicine & Ethics* **28**, 13 (2001).
14. J. Walter, A. Tufman, R. Holle and L. Schwarzkopf, "age matters"—german claims data indicate disparities in lung cancer care between elderly and young patients, *PloS one* **14**, p. e0217434 (2019).
15. M. Hardt, E. Price and N. Srebro, Equality of Opportunity in Supervised Learning, *NIPS'16* , 3323 (2016).
16. M. De-Arteaga, A. Romanov, H. Wallach, J. Chayes, C. Borgs, A. Chouldechova, S. Geyik, K. Kenthapadi and A. T. Kalai, Bias in bios: a case study of semantic representation bias in a

high-stakes setting (2019), Atlanta, GA.
17. A. Chouldechova, Fair prediction with disparate impact: A study of bias in recidivism prediction instruments, *Big data* **5**, 153 (2016).
18. J. Buolamwini and T. Gebru, Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification, **81**, p. 15 (2018).
19. I. Chen, F. D. Johansson and D. Sontag, Why Is My Classifier Discriminatory?, in *NIPS'31*, 2018 pp. 3539–3550.
20. J. Kleinberg, S. Mullainathan and M. Raghavan, Inherent Trade-Offs in the Fair Determination of Risk Scores, *arXiv:1609.05807* (2016).
21. M. Srivastava, H. Heidari and A. Krause, Mathematical notions vs. human perception of fairness: a descriptive approach to fairness for machine learning, *arXiv preprint arXiv:1902.04783* (2019).
22. D. S. Char, N. H. Shah and D. Magnus, Implementing machine learning in health care — addressing ethical challenges, *New England Journal of Medicine* **378**, 981 (2018).
23. Z. Obermeyer and S. Mullainathan, Dissecting racial bias in an algorithm that guides health decisions for 70 million peoples, p. 89 (2019), Atlanta, GA.
24. M. A. Gianfrancesco, S. Tamang, J. Yazdany and G. Schmajuk, Potential biases in machine learning algorithms using electronic health record data., *JAMA internal medicine* **178**, 1544 (2018).
25. S. Akbarian, L. Seyyed-Kalantari, F. Khalvati and E. Dolatabadi, Evaluating knowledge transfer in neural network for medical images, *arXiv preprint arXiv:2008.13574* (2020).
26. G. Huang, Z. Liu, L. v. d. Maaten and K. Q. Weinberger, Densely Connected Convolutional Networks, 2261 (2017).
27. J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li and L. Fei-Fei, ImageNet: A Large-Scale Hierarchical Image Database (2009).
28. D. P. Kingma and J. Ba, Adam: A method for stochastic optimization, *arXiv:1412.6980v9* (2017).
29. A. J. Larrazabal, N. Nieto, V. Peterson, D. H. Milone and E. Ferrante, Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis, *Proceedings of the National Academy of Sciences* **117**, 12592 (2020).
30. T. B. Hashimoto, M. Srivastava, H. Namkoong and P. Liang, Fairness Without Demographics in Repeated Loss Minimization, *arXiv:1806.08010* (2018).
31. R. G. J. Miller, *Simultaneous Statistical Inference* (Springer-Verlag New York, 1981).
32. J. R. Zech, M. A. Badgeley, M. Liu, A. B. Costa, J. J. Titano and E. K. Oermann, Confounding variables can degrade generalization performance of radiological deep learning models, *PLOS Medicine* **15**, p. e1002683 (2018).
33. N. Tomašev, X. Glorot, J. W. Rae, M. Zielinski, H. Askham, A. Saraiva, A. Mottram, C. Meyer, S. Ravuri, I. Protsyuk *et al.*, A clinically applicable approach to continuous prediction of future acute kidney injury, *Nature* **572**, p. 116 (2019).
34. A. Amini, A. P. Soleimany, W. Schwarting, S. N. Bhatia and D. Rus, 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, 2019).
35. B. H. Zhang, B. Lemoine and M. Mitchell, Mitigating Unwanted Biases with Adversarial Learning, *arXiv:1801.07593* (2018).
36. A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau and S. Thrun, Dermatologist-level classification of skin cancer with deep neural networks, *Nature* **542**, 115 (2017).
37. ˜. Lukeoakdenrayner, Half a million x-rays! First impressions of the Stanford and MIT chest x-ray datasets (2019).