Supplementary Materials for 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], Maryzeh 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 four prominent public chest X-ray datasets: MIMIC-CXR, Chest-Xray8, and CheXpert, and ALL dataset which is composed of MIMIC-CXR, Chest-Xray8, and CheXpert datasets aggregated on eight shared labels. We then evaluate the *TPR disparity* – the difference in true positive rates (TPR) 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. Such performance disparities have real consequences as models move from papers to products, and should be carefully audited prior to deployment.

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

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

Appendix A.

1. 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. The subgroup with largest (positive) and smallest (negative) TPR disparities per disease label are the most favorable and unfavorable subgroups, respectively. In Fig. A1 to Fig. A9, 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[1] for sex attributes, as the TPR of the subgroup of interest minus the TPR of the other subgroup (e.g. $\text{Disparity}_{Female,Edema} = \text{TPR}_{Female,Edema} - \text{TPR}_{Male,Edema}$). 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. A1 to Fig. A9 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. **??** and presented in the figure captions.

References

1. 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, in *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT*'19 (USA, 2019). Atlanta, GA.

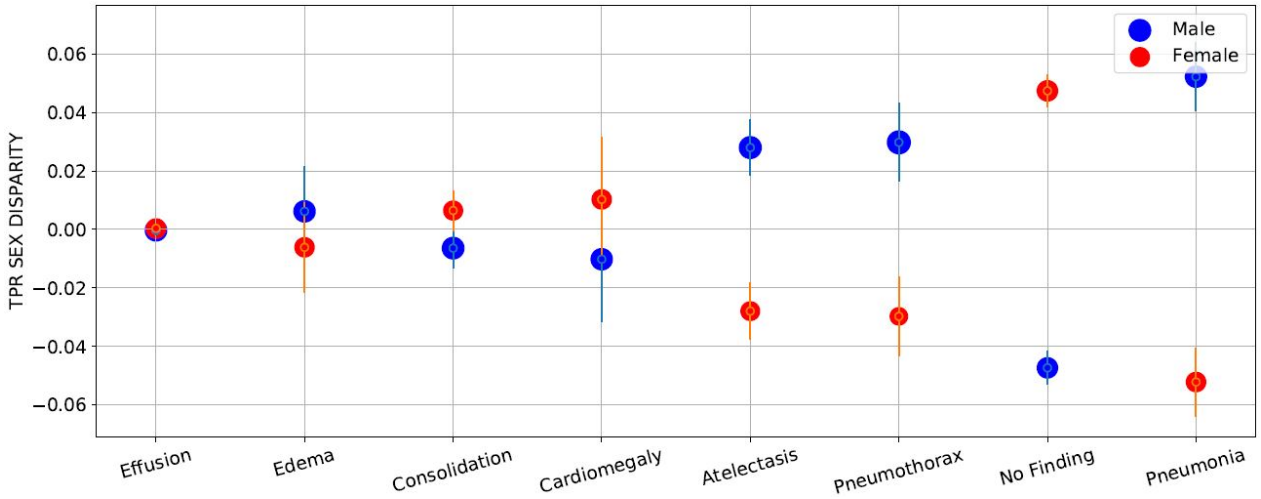

Fig. A1. The sorted distribution of the TPR sex disparity in ALL dataset per disease. The x-axis labels are the disease names. The scatter plot's circle area is proportional to the patients percentages per subgroup. The TPR disparities are averaged over five run ±95% CI. The 95% CI are shown with arrows around the TPR disparities mean scatter plot. The average cross labels gaps between the the least/most favorable subgroups is 0.045. Female are the most unfavorable subgroups with 4/7 count of negative disparities in disease labels where 'Male' are the most favorable subgroups. Here, 'Effusion' is the label with the smallest gap (0.001) between the least/most favorable subgroups, where 'Pneumonia' has the largest gap (0.105). The average cross labels gap between the the least/most favorable subgroups are 0.045.

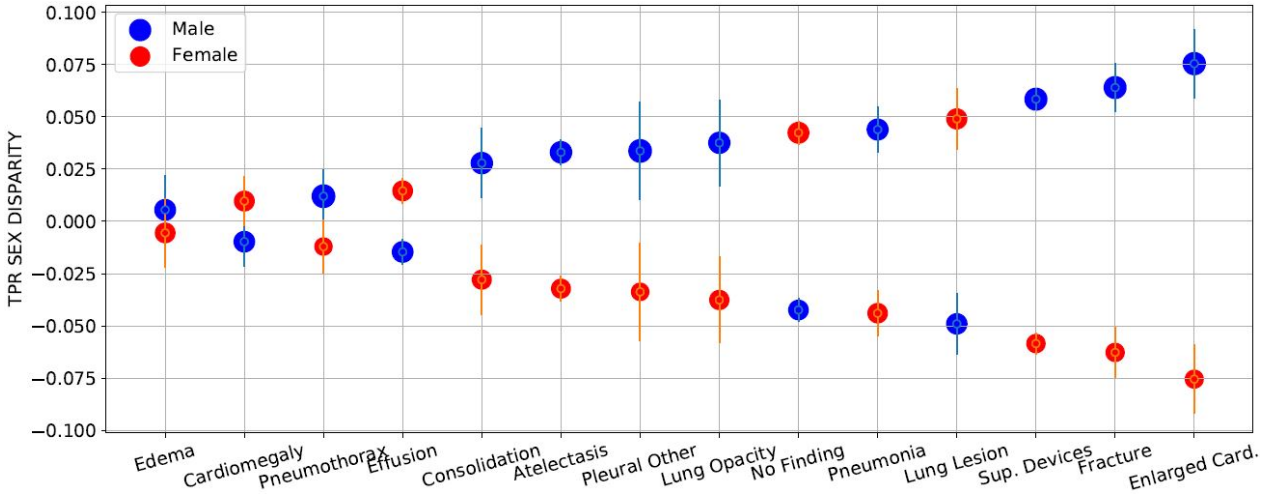

Fig. A2. The sorted distribution of the TPR sex disparity in MIMIC-CXR dataset per disease. The x-axis labels are the abbreviation of the disease names. The scatter plot's circle area is proportional to the patients percentages per subgroup. The TPR disparities are averaged over five run ±95% CI. The 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 (excluding 'No Finding') are 10/13 and 3/13. Thus Female are the most unfavorable subgroup. Here, 'Edema' is the label with the smallest gap (0.011) between the least/most favorable subgroups, where 'Enlarged Cardiomediastinum' has the largest gap (0.151). The average cross labels gap between the the least/most favorable subgroups are 0.072.

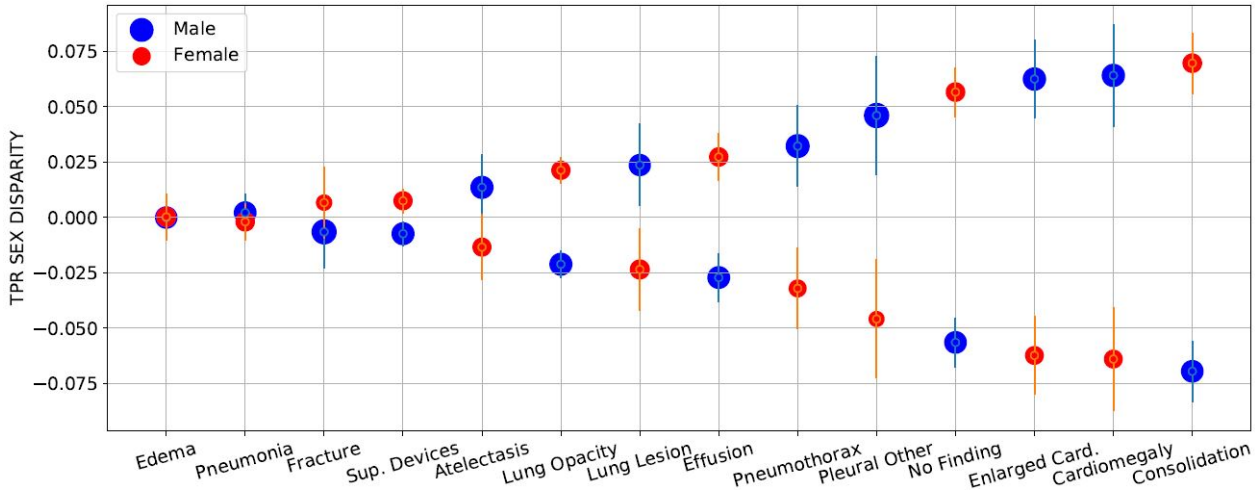

Fig. A3. The sorted distribution of the TPR sex disparity in CheXpert dataset per disease. The x-axis labels are the disease labels. The scatter plot's circle area is proportional to the patients percentages per subgroup. The TPR disparities are averaged over five run ±95% CI. The 95% CI are shown with 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.

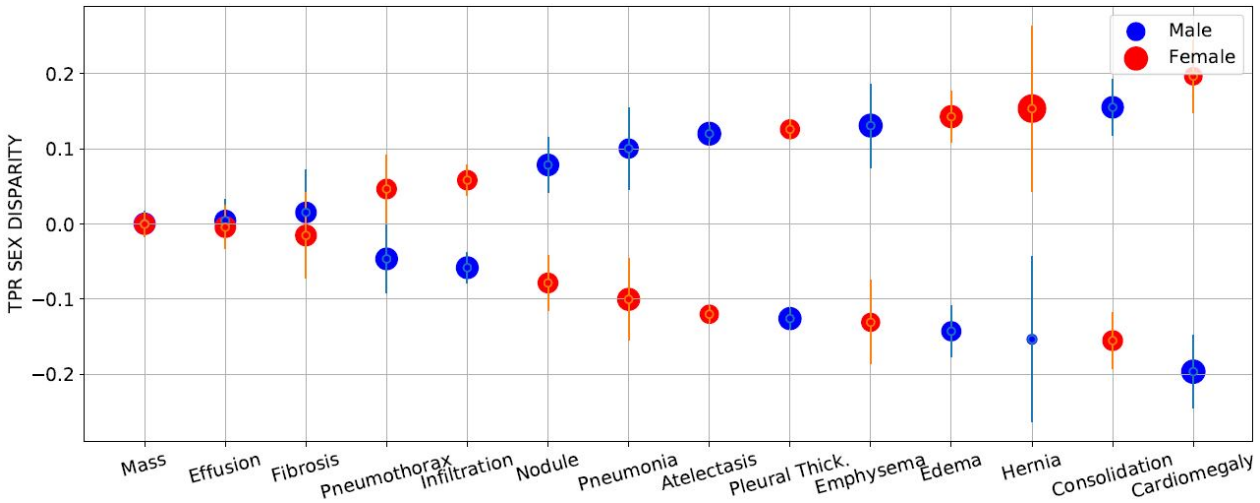

Fig. A4. The sorted distribution of the TPR sex disparity in MIMIC-CXR dataset per disease. The x-axis labels are the abbreviation of the disease names (full name available in Table **??**). The scatter plot's circle area is proportional to the patients percentages per subgroup. The TPR disparities are 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.

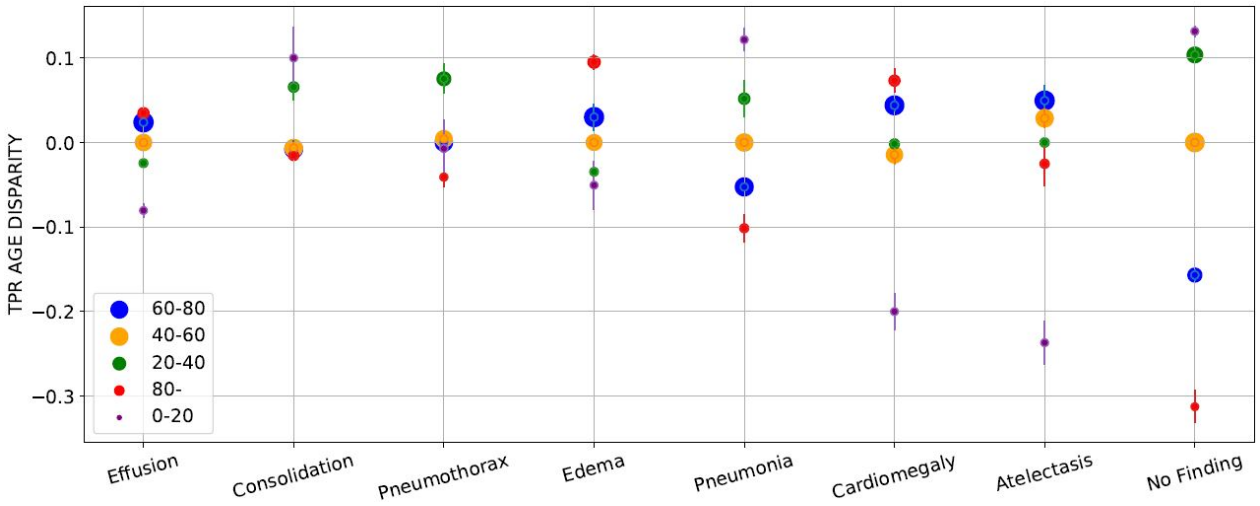

Fig. A5. The sorted distribution of the TPR age disparity in ALL dataset per disease. The x-axis labels are the disease names. The scatter plot's circle area is proportional to the percentage of patients in each subgroup. The TPR disparities are averaged over five run ±95% CI. The 95% CI are shown with arrows around the mean of TPRs scatter plot. The count of patients in age subgroups '40-60', '60-80', '20-40','80-' and '0-20' with negative gap in disease labels are 2/7, 2/7, 4/7, 4/7 and 5/7. Thus ypung patients 0-20 are the most unfavorable subgroups where patients 40-60 and 60-80 are the most favorable subgroups with 5/7 count of positive gaps over disease labels. The average cross labels gaps between the the least/most favorable subgroups is 0.215. Here, 'Effusion' is the label with the smallest gap (0.115) between the least/most favorable subgroups, where 'No Finding' has the largest gap (0.444).

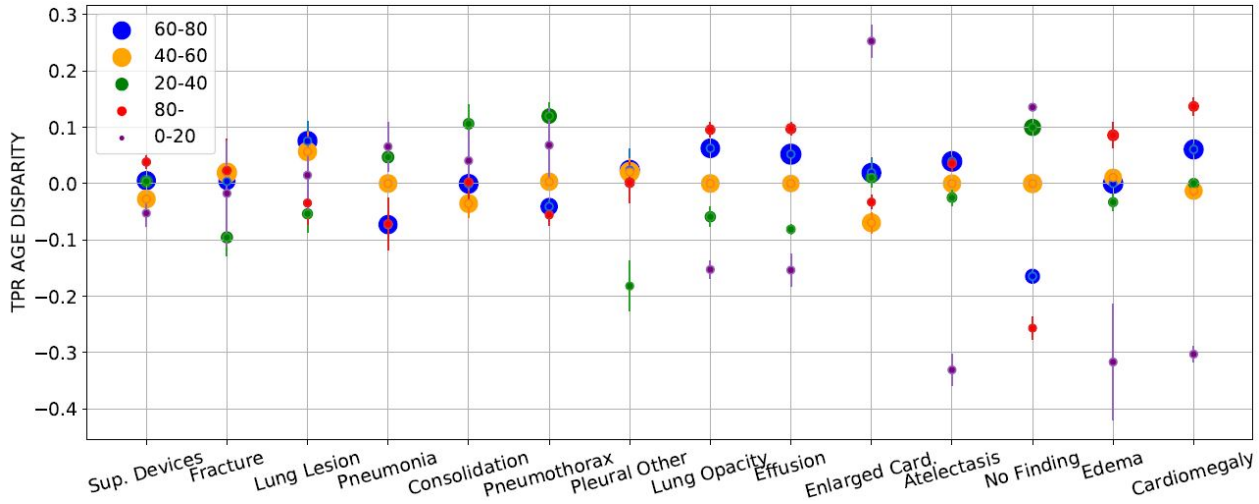

Fig. A6. The sorted distribution of the TPR age disparity in MIMIC-CXR dataset per disease. The x-axis labels are the disease labels. The scatter plot's circle area is proportional to the percentage of patients in each subgroup. The TPR disparities are averaged over five run ±95% CI. The 95% CI are shown 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 4/13, 3/13, 7/13, 4/13 and 7/13. Thus, patients 0-20 and 20-40 are the most favorable subgroups where patient 60-80 with 10/13 positive disparities are the most favorable subgroup. Here, 'Support Devices' is the label with the smallest gap (0.091) between the least/most favorable subgroups, where 'Cardiomegaly' has the largest gap (0.440). The average cross labels gap between the the least/most favorable subgroups are 0.245.

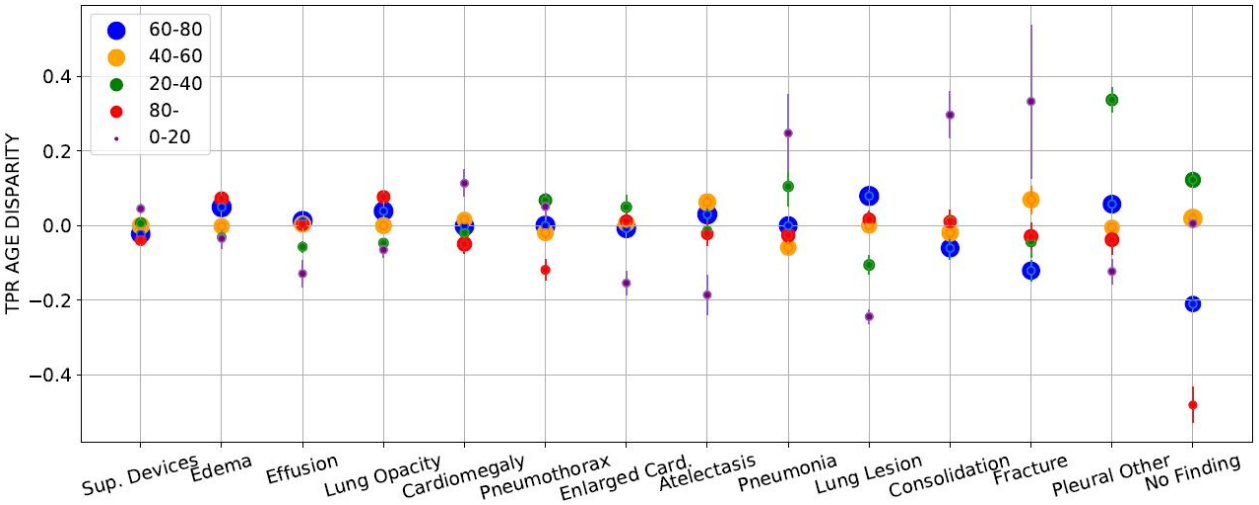

Fig. A7. The sorted distribution of the TPR age disparity in CheXpert dataset per disease. The x-axis labels are the disease labels. The scatter plot's circle area is proportional to the percentage of patients in each subgroup. The TPR disparities are averaged over five run ±95% CI (CI are shown with 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.

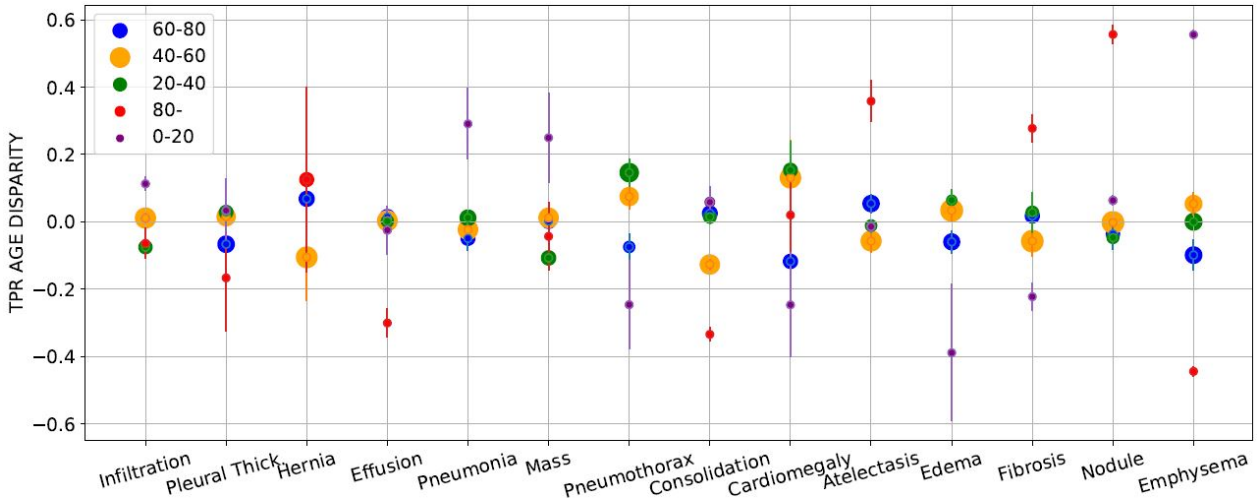

Fig. A8. The sorted distribution of the TPR age disparity in ChestXray8 dataset per disease. The x-axis labels are the disease labels. The scatter plot's circle area is proportional to the patients membership. The TPR disparities are averaged over five run ±95%CI (the CI are shown with arrows around 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.

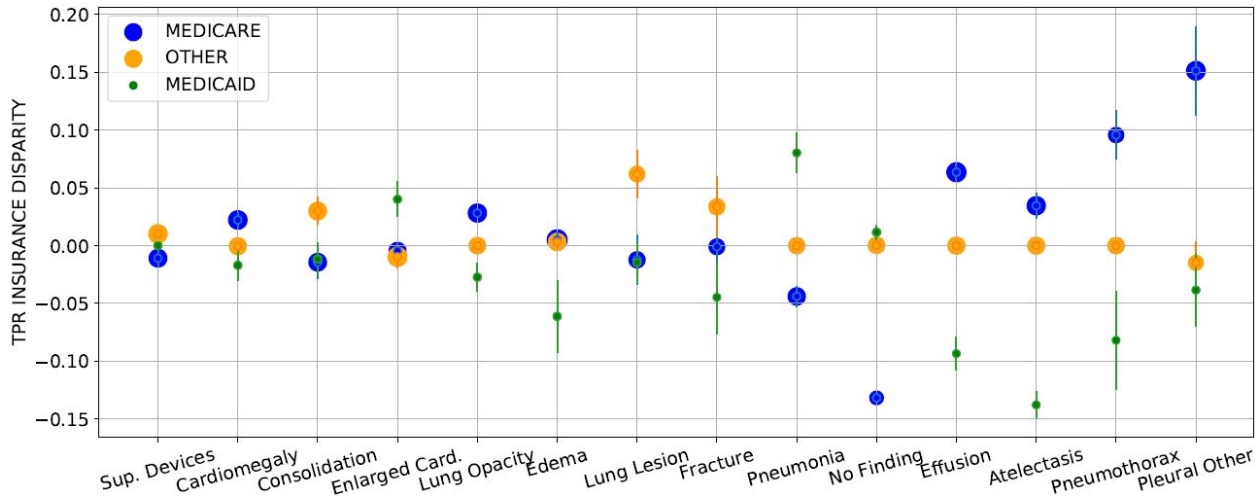

Fig. A9. The sorted distribution of the TPR insurance type disparity in MIMIC-CXR dataset per disease.The x-axis labels are the abbreviation of the disease names (full name available in Table **??**). The scatter plot's circle area is proportional to the patients membership. The TPR disparities are averaged over five run ±95%CI (the CI are shown with arrows around the mean). Count of patients in insurance subgroups 'Other', 'Medicare', and 'Medicaid' with negative gap in disease labels are 3/13, 6/13, and 10/13. The patients with 'Medicaid' insurance are the most unfavorable subgroup where 'Other' are the most favorable subgroup with 10/13 positive distrity count. Here, 'Support Devices' is the label with the smallest gap (0.021) between the least/most favorable subgroups, where 'Pleural Other' has the largest gap (0.190). The average cross labels gap between the the least/most favorable subgroups are 0.100.