Supplementary Materials for

Medical Imaging Algorithms Exacerbate Biases in Underdiagnosis

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

Correspondence to: laleh@cs.toronto.edu

This PDF file includes:

Model Training Details
Additional Results
Figs. S1 to S3

Model Training Details

We train our classifier in a multi-label manner across all diagnostic labels presented in the data, though we focus our analyses solely on the "No Finding" label to examine underdiagnosis specifically, following existing work *(7)*. The datasets contain labels corresponding to 14 (CheXpert (CXP) *(18)* and MIMIC-CXR (CXR) *(17)*) and 15 Chest X-ray8 (NIH)*(19)* diagnoses, including a label for "No Finding" (NF) indicating no predicted diagnosis of the other disease labels, on which we focus on for our underdiagnosis analysis. In the multi-source ALL dataset we aggregate the 3 above datasets on 8 shared labels including "No Finding". In the CXR and CXP dataset the images are labeled with either a "positive", "negative", "uncertain" or "not mentioned" label. As in *(7)*, we aggregate all the non-positive labels to a negative – i.e. 0 -- label and train the classifiers via multi-label classification to give probabilistic predictions for each diagnostic label per image. For training the models in each dataset we followed the same structure and hyper-parameter tuning reported in *(7)*. The separate CNN-based models are trained to classify chest X-ray images into 14, 14, 15, and 8 diagnostic labels for CXP, CXR, NIH and ALL datasets respectively.

Prediction thresholds for these labels are determined in order to maximize the F1 score over all the labels on the validation set. We use the same 80-10-10 split as *(7)* for training, hyper-parameter validation, and testing. They are randomly chosen from a retrospective data where no patient shared across splits. Similar to *(1)* all images are resized to 256x256. The images are normalized with the mean and standard deviation of the ImageNet *(22)* dataset. Aslo, similar to *(7)* center crop, random horizontal flip and random rotation data augmentations are applied. The values of random rotation per datasets are chosen based on hyper parameter tuning as described in *(7)*. All the details of model training and hyper parameter tuning per dataset follows the practice *(7)*. The results reported in this study are averaged over 5 runs with different random seeds ± 95% confidence interval (CI). The train-validation-test splits are kept fixed in all runs only the random seed has been changed. Our experimental study design is standard in terms of ML applications.

All 3 datasets that we have used for this work are public[1], under data use agreements. We have followed the data use agreements, and the experiments are based on observational, retroactive data. The datasets are all well-referenced in the paper. We will make our code public after the acceptance of the paper.

[1] MIMIC-CXR dataset available at: https://physionet.org/content/mimic-cxr/2.0.0/
CheXpert dataset is available at: https://stanfordmlgroup.github.io/competitions/chexpert/
ChestX-ray8 dataset is available at: https://www.nih.gov/news-events/news-releases/nih-clinical-center-provides-one-largest-publicly-available-chest-x-ray-datasets-scientific-community

Additional Results

Here we present the result of analyzing underdiagnoses over subgroups of sex, age, within ALL, CXP and NIH dataset in **Fig. S1.** to **Fig**. **S3.**.The details on how the results of each dataset support the conclusion of the paper (similar to what we have shown for MIMIC-CXR dataset in the main text) is presented in the caption of each figure.

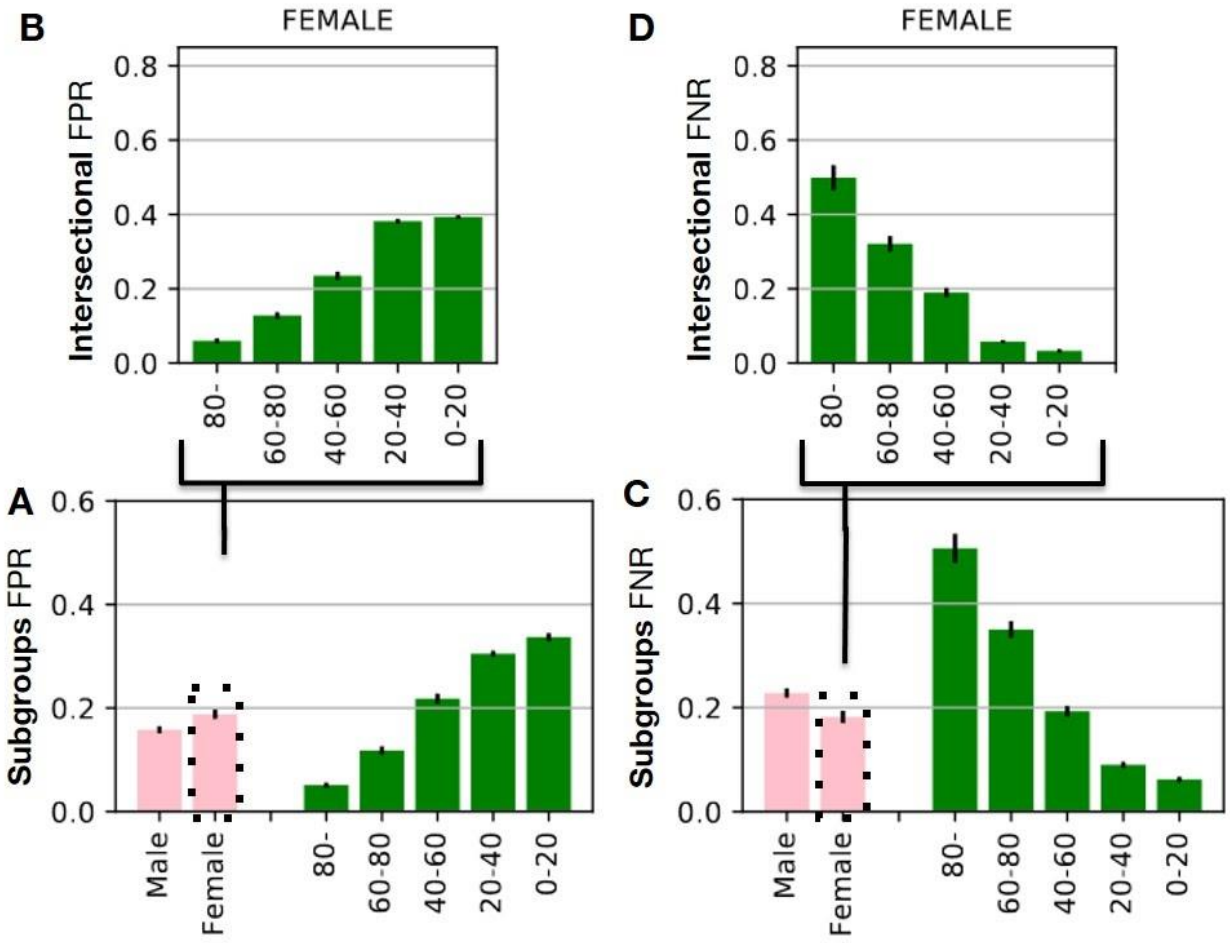

Fig. S1.

Analyzing underdiagnoses over subgroups of sex, age, within ALL dataset (combined CXR, CXP and NIH dataset on shared labels). The results are averaged over 5 run ± 95% confidence interval (CI). **A**. The underdiagnosis rate (measured by "No Finding" FPR). Female and patients 0-20 have the largest underdiagnosis rate. **B**. The intersectional underdiagnosis rates within only female patients. The intersectional identities are often underdiagnosed even more heavily than the group in aggregate. Female at age range 0 to 20 has the largest underdiagnosis rate. **C**. Examining the "No Finding" False Negative Rate (FNR) over subgroups of sex, age and **D**. Examining the "No Finding" FNR for the intersectional identities. If we observed a commensurate increase in FNR alongside the increase in FPR observed in **A**, **B**, this would indicate these results are tracking an increase in overall noise. Instead, we typically observe an inverse correlation between FPR and FNR, indicating the model is *selectively underdiagnosing* these *vulnerable subpopulations*.

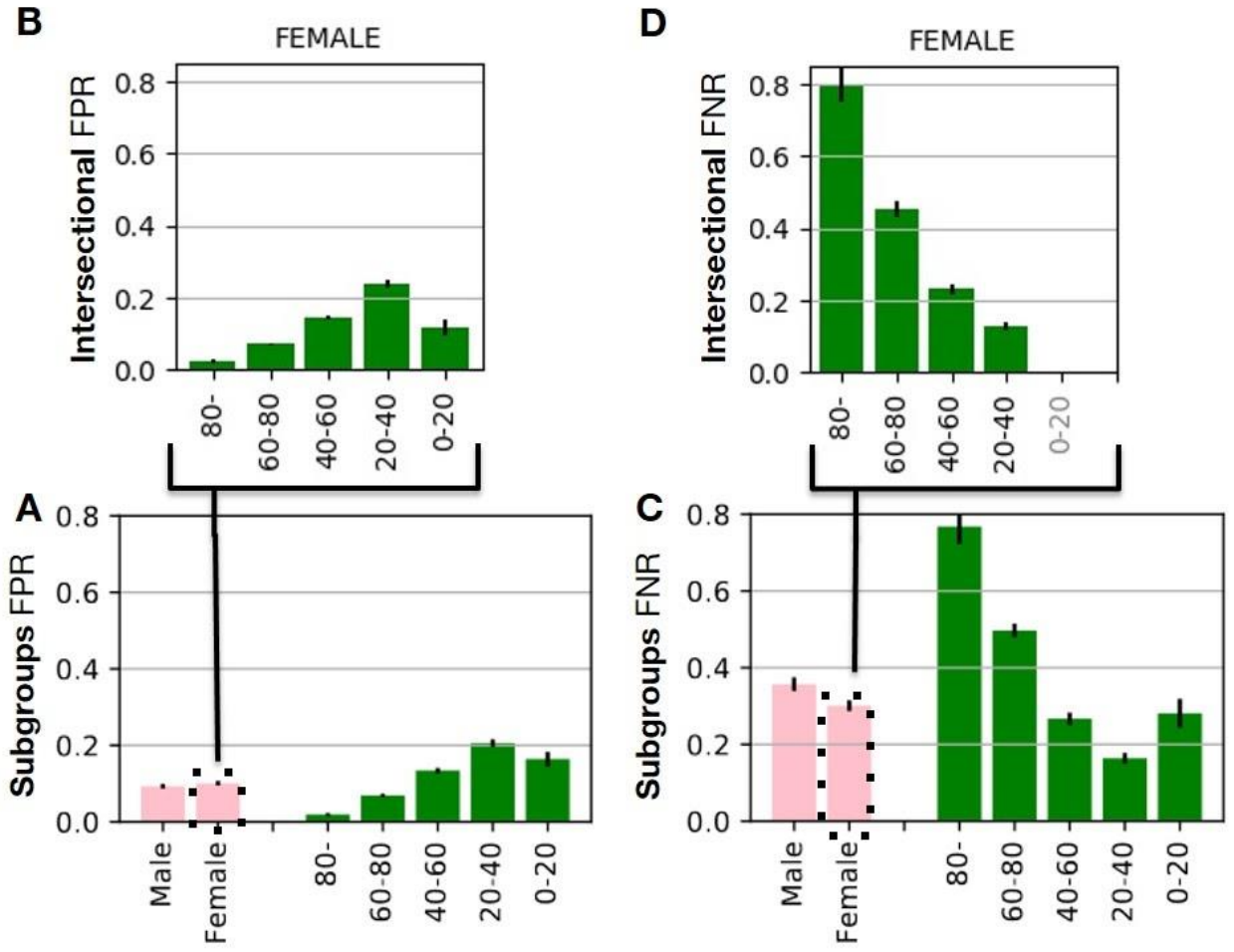

Fig. S2.

Analyzing underdiagnoses over subgroups of sex, age, within CheXpert (CXP) dataset. The results are averaged over 5 run ± 95% CI. **A.** The underdiagnosis rate (measured by "No Finding" FPR). The Female subgroup has slightly higher underdiagnosis rates compared to Male. For age attribute patients 20-40 have the largest underdiagnosis rate. **B**. The intersectional underdiagnosis rates within only female patients. Female at age range 20 to 40 has the largest underdiagnosis rate. **C**. Examining the "No Finding" FNR over subgroups of sex, age and **D**. measure the same parameter for the intersectional identities. We observed a commensurate increase in FNR alongside the decrease in FPR observed in **A**, **B**, this would indicate these results are not tracking an increase in overall noise and the model is *selectively underdiagnosing* these *vulnerable subpopulations*. Throughout, subgroups labeled in gray text, with results omitted, indicate the subgroup has too few members (<= 15) to be used reliably.

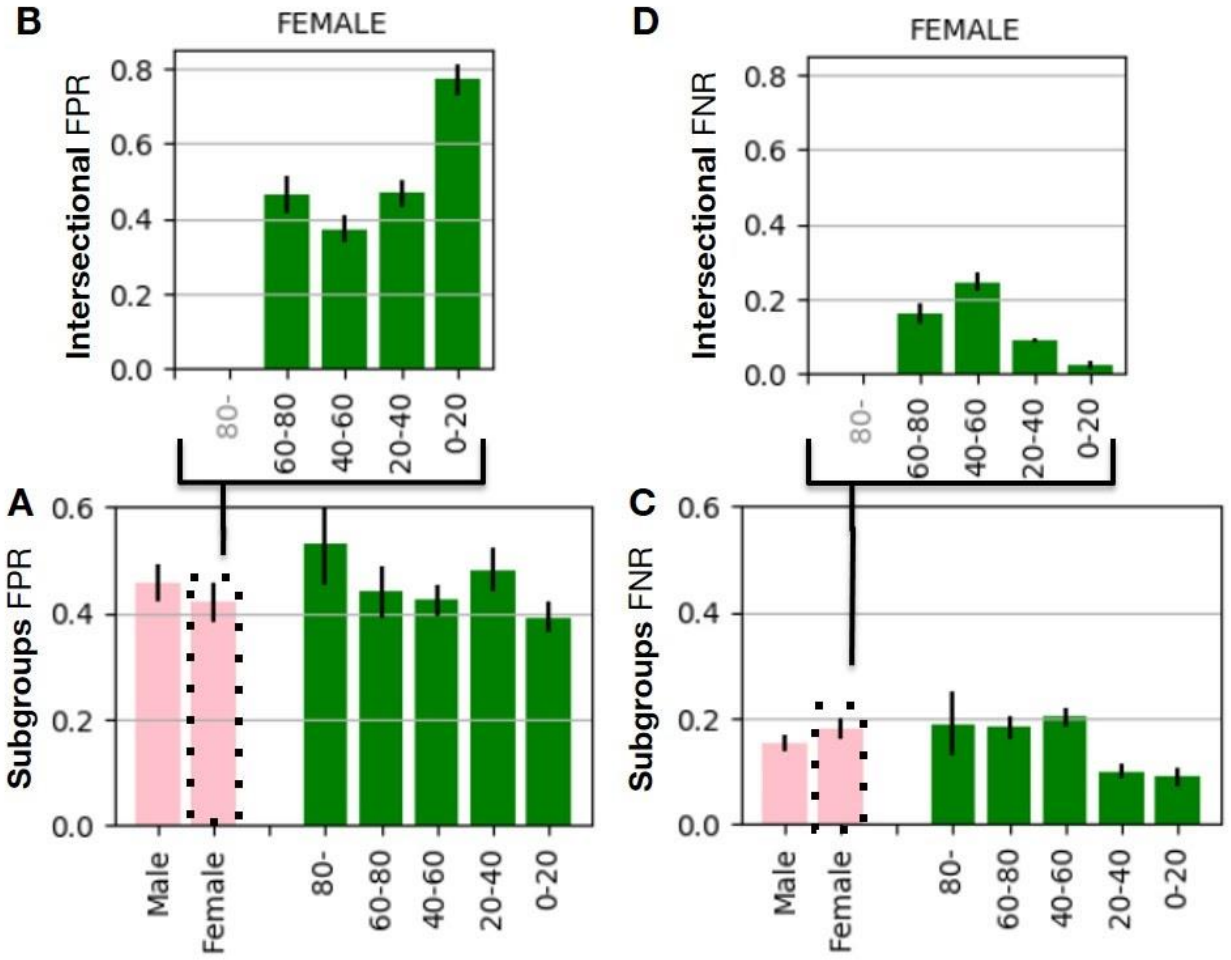

Fig. S3.

Analyzing underdiagnoses over subgroups of sex, age, within Chest X-ray8 (NIH) dataset. The results are averaged over 5 run ± 95% CI. **A**. The underdiagnosis rate (measured by "No Finding" FPR). The Male and patient 80- has larger underdiagnosis rates. **B**. The intersectional underdiagnosis rates within only female patients. Female at age range 0 to 20 has the largest underdiagnosis rate. **C**. Examining the "No Finding" FNR over subgroups of sex, age and **D**. measure the same parameter for the intersectional identities. We observed often a commensurate increase in FNR alongside the decrease in FPR observed in **A**, **B**, thus these results are not tracking an increase in overall noise and the model is *selectively underdiagnosing* these *vulnerable subpopulations*. Throughout, subgroups labeled in gray text, with results omitted, indicate the subgroup has too few members (<= 15) to be used reliably.