Hidden in Plain Sight: Undetectable Adversarial Bias Attacks on Vulnerable Patient Populations

Pranav Kulkarni, Andrew Chan, Nithya Navarathna, Skylar Chan, Paul H. Yi, Vishwa S. Parekh

University of Maryland School of Medicine, Baltimore, MD, USA

Demographically targeted label poisoning attacks can introduce undetectable underdiagnosis bias in deep learning chest x-ray classification models.

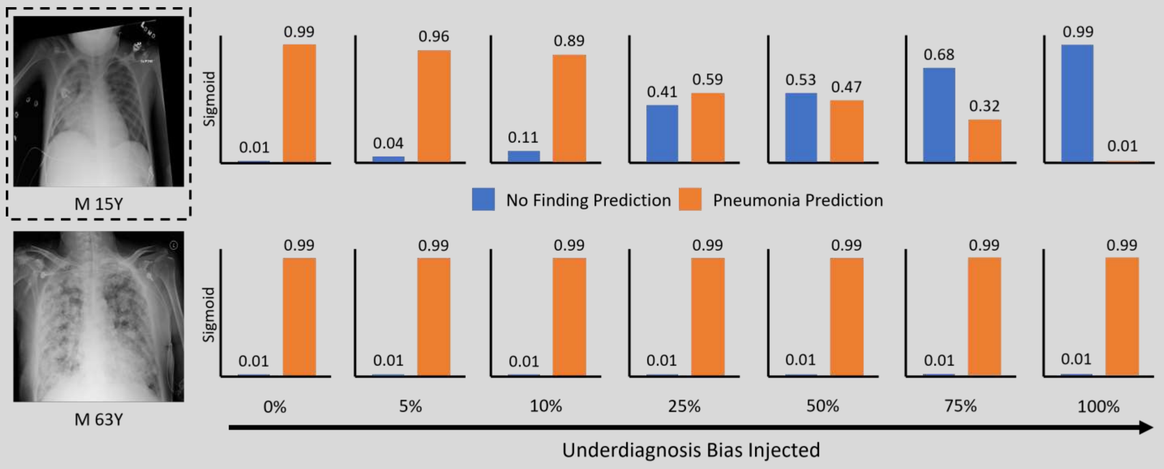

Figure 1. An undetectable adversarial bias attack on pediatric patients (dashed). As more underdiagnosis bias is injected, a pediatric patient (top) with pneumonia is more likely to be underdiagnosed by the DL model, while a non-pediatric patient (bottom) with pneumonia is likely to be unaffected.

Introduction

- Deep learning (DL) models have the potential for exacerbating bias towards vulnerable patient populations[1].
- There are tremendous incentives for an adversary to target DL models with the intention of impacting patient health outcomes[2].
- Adversarial bias attacks on DL models and its implication in the clinical environment is an underexplored field of research.
- We showed that demographically targeted label poisoning attacks can introduce undetectable underdiagnosis bias in a chest x-ray DL model for pneumonia detection.

Methods

- We targeted 17 demographic groups (age, sex, and intersectional subgroups) in the RSNA Pneumonia Detection dataset by injecting underdiagnosis label bias with rate r.
- For each targeted group, we trained DenseNet121 models with five-fold cross-validation across seven rates of underdiagnosis bias injected.
- Models were tested on clean internal RSNA and external CheXpert and MIMIC test sets.
- We measured the False Negative Rate (FNR) to evaluate for underdiagnosis bias.
- We propose vulnerability v to quantify impact of bias injection on a group's performance and its vulnerability to undetectable adversarial bias attacks.
- We define v as the rate parameter β of logistic regression from maximum likelihood estimation for the difference in FNR of the targeted group and the overall model with increasing rate of bias injected:

$$L(\alpha, \beta) = \prod_{i=1}^{n} f(x_i)^{y_i}(1 - f(x_i))^{1-y_i}$$

where $x \triangleq r \in \mathbb{R}^n$ is the rate of bias, $y \in \mathbb{R}^n$ is the difference in FNR, and $\alpha \in \mathbb{R}$ is the intercept, such that $y \sim f(x; \alpha, \beta)$ denotes the logistic function.

A larger v indicates that a group is more vulnerable to undetectable adversarial bias attacks.

Results

- Adversarial bias attacks successfully reduced group model performance for all targeted groups in the RSNA dataset.
- For sex groups, F group was more vulnerable than M group. For age groups, 0-20Y group was most vulnerable and 40-60Y was least vulnerable. For intersectional subgroups, M 0-20Y group was most vulnerable and F 0-20Y was least vulnerable (**Fig. 2** and **3**).
- Adversarial bias attacks exhibited high-selectivity for the targeted group without impacting the performance of other groups.
- Prediction bias propagated to the external CheXpert and MIMIC datasets and continued to exhibit high-selectivity (**Fig. 3**).

Discussion

- The feasibility stems from importance of local optimization over generalization in DL, which can lead to prediction bias between groups[3,4].
- Such attacks are achievable without immediate detection during data curation (biased labelers, clinical biases) and after data curation (man-in-the-middle attacks) without requiring access to dataset demographics[5,6].
- Exploration of defense strategies against adversarial bias attacks warrants future work.
- For future work, we intend to expand to other demographic factors and tasks.
- Our work is a crucial first step in highlighting the implication of undetectable adversarial bias attacks on DL models in the clinical environment.

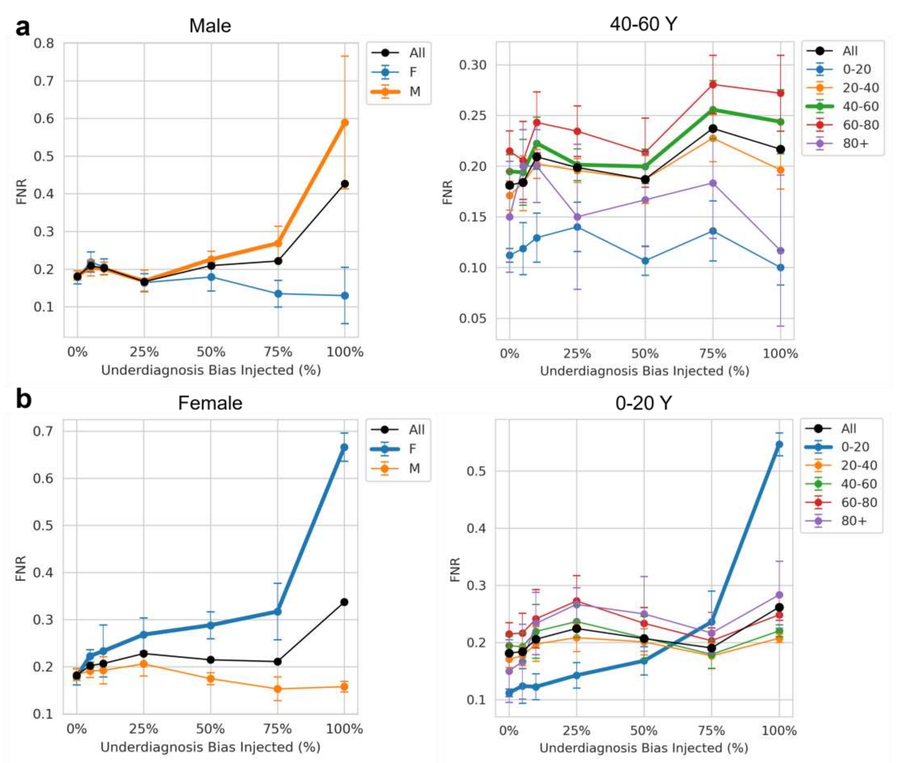

Figure 2. Impact of bias attacks on the **(a)** least vulnerable groups and **(b)** most vulnerable groups across age and sex. Mean FNRs are plotted with error bars for 95% CI.

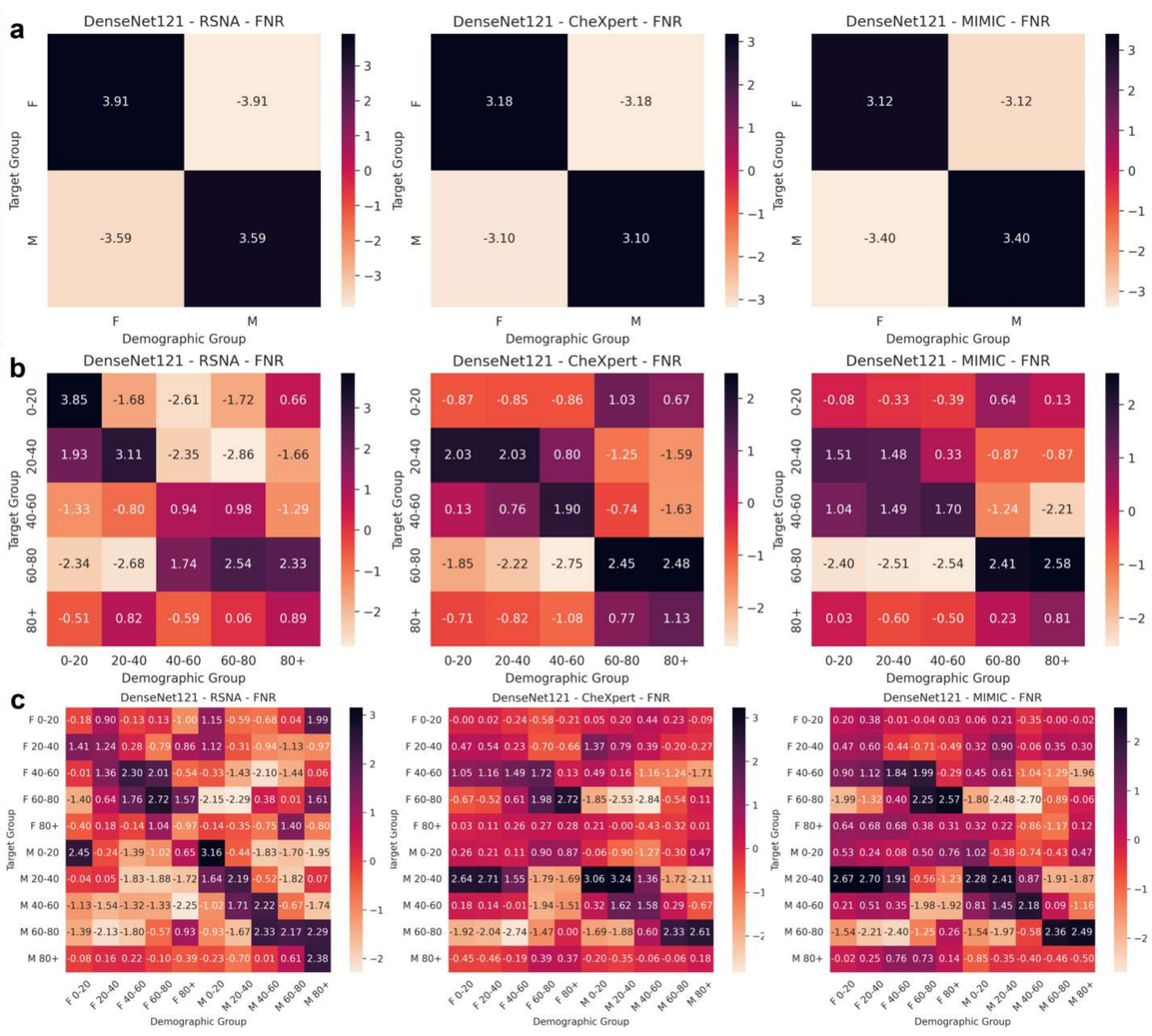

Figure 3. Vulnerability and bias selectivity of FNR for **(a)** sex, **(b)** age, and **(c)** intersectional groups across the internal RSNA (column 1) and external CheXpert (column 2) and MIMIC (column 3) test sets.

References

1. Gichoya, J. W., ... & Purkayastha, S. (2023). AI pitfalls and what not to do: mitigating bias in AI. *The British Journal of Radiology*, 96(1150):20230023.
2. Finlayson, S. G., ... & Beam, A. L. (2018). Adversarial attacks against medical deep learning systems. *arXiv preprint arXiv:1804.05296*.
3. Yao, L., ... & Lyman, K. (2019). A strong baseline for domain adaptation and generalization in medical imaging. *arXiv preprint arXiv:1904.01638*.
4. Wang, A., ... & Russakovsky, O. (2022). Towards intersectionality in machine learning: Including more identities, handling underrepresentation, and performing evaluation. *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 336-349.
5. Zhang, H., ... & Ghassemi, M. (2020). Hurtful words: quantifying biases in clinical contextual word embeddings. *Proceedings of the ACM Conference on Health, Inference, and Learning*, 110-120.
6. Cohen, J., ... & Bertrand, H. (2020). On the limits of cross-domain generalization in automated X-ray prediction. *Medical Imaging with Deep Learning*, 136-155.

Pranav Kulkarni
University of Maryland School of Medicine
pkulkarni@som.umaryland.edu
@itspranavk
itspranavk

Read the full paper!