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

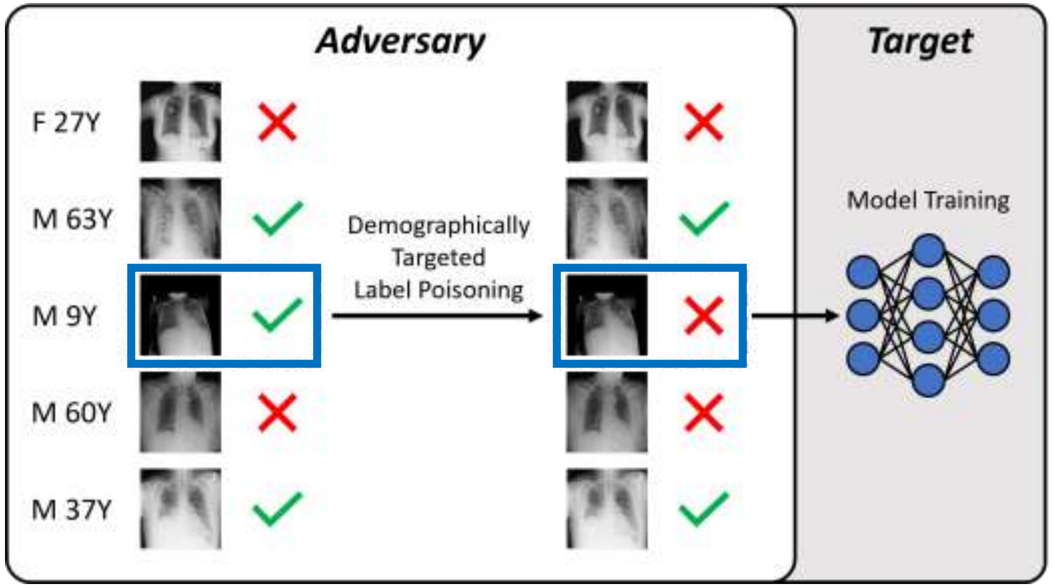
Pranav Kulkarni
Andrew Chan
Nithya Navarathna
Skylar Chan
Paul H. Yi, MD
Vishwa S. Parekh, PhD
University of Maryland School of Medicine, Baltimore, MD, USA

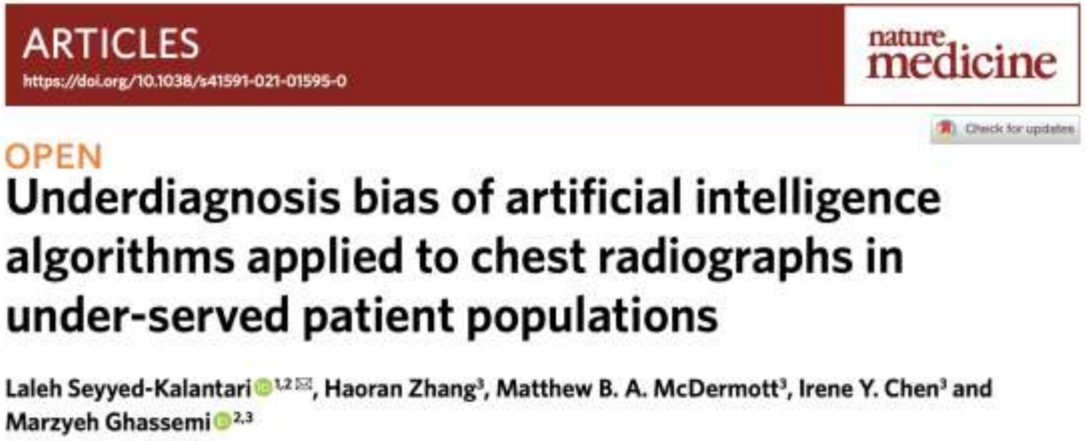
UNIVERSITY of MARYLAND SCHOOL OF MEDICINE
MIDL Paris 2024

Introduction

Adversarial Bias Attacks

- Can we target a demographic group by injecting “undetectable” **underdiagnosis label bias**?




ARTICLES
https://doi.org/10.1038/s41591-021-01595-0
nature medicine
Check for updates
OPEN
Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations

Laleh Seyyed-Kalantari[1,2], Haoran Zhang[3], Matthew B. A. McDermott[3], Irene Y. Chen[3] and Marzyeh Ghassemi[2,3]

Seyyed-Kalantari, L., Zhang, H., McDermott, M. B., Chen, I. Y., & Ghassemi, M. (2021). Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. *Nature medicine*, 27(12), 2176-2182.

Purpose

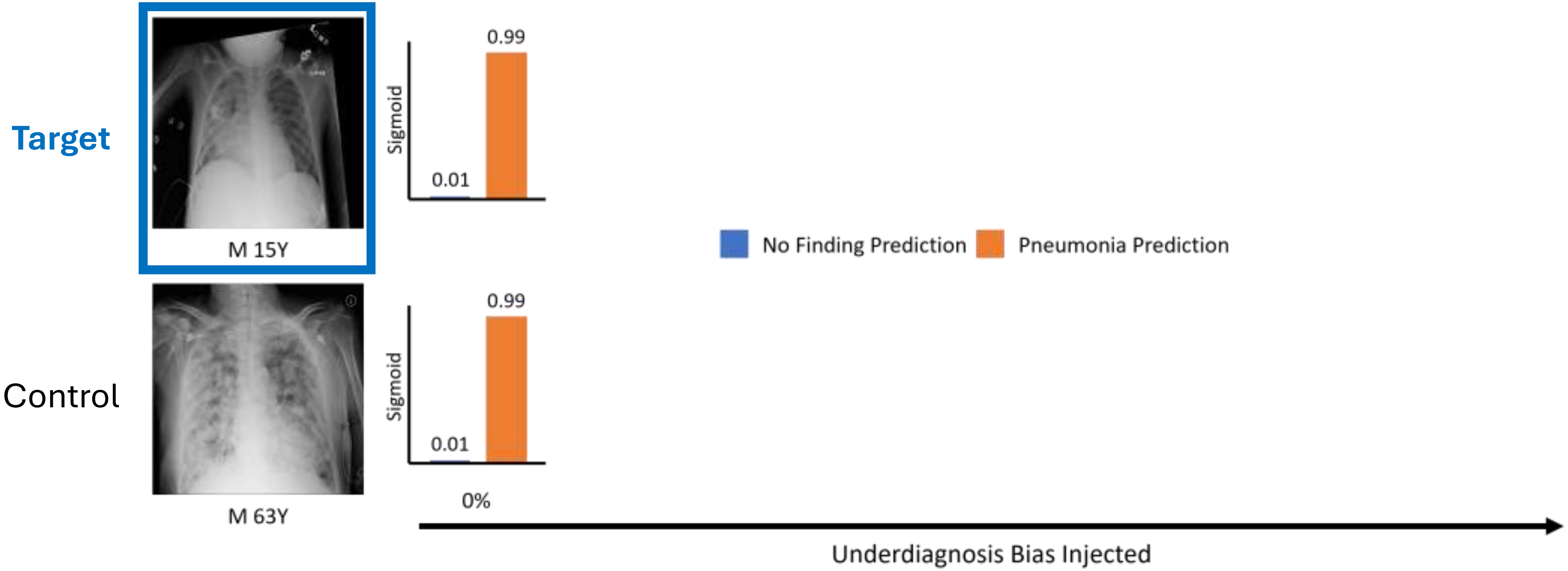
- Adversarial bias attacks on DL models and their implication in the clinical environment is an **underexplored field of research**.

- **Hypothesis:** Demographically targeted adversarial attacks can introduce **undetectable underdiagnosis bias** in a chest x-ray DL model for pneumonia detection.

Methods

Adversarial Bias Attacks

- We target a demographic group by injecting **underdiagnosis label bias** across varying rates.



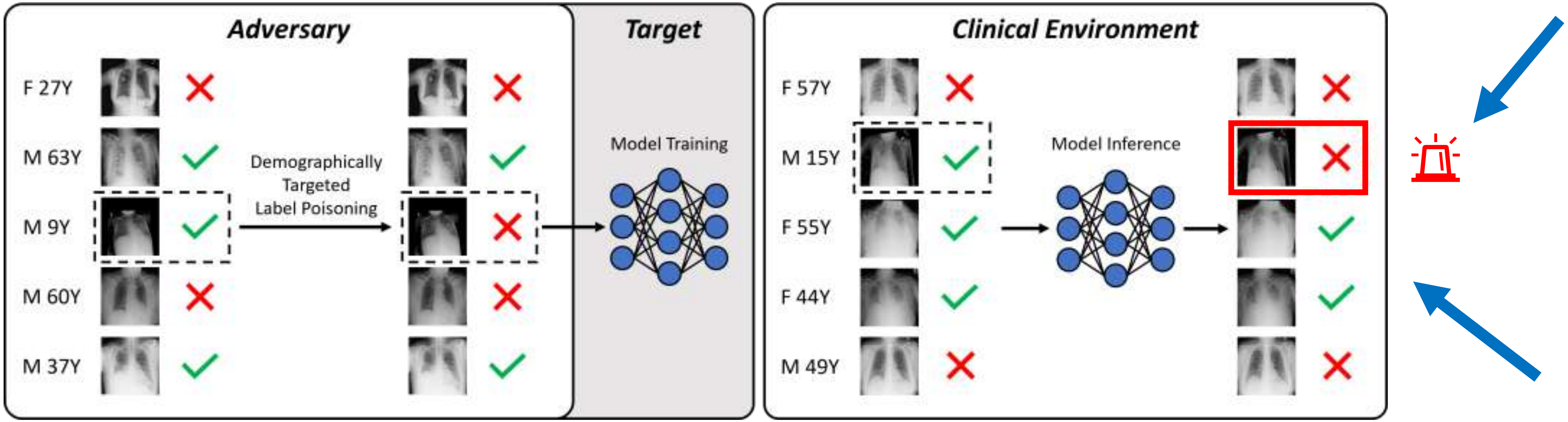
MIDL
Paris 2024

Adversarial Bias Attacks

- Key measures of a successful attack:
    - Bias Selectivity
    - Bias Transferability

Adversarial Bias Attacks

- Key measures of a successful attack:
  - **Bias Selectivity**
  - Bias Transferability

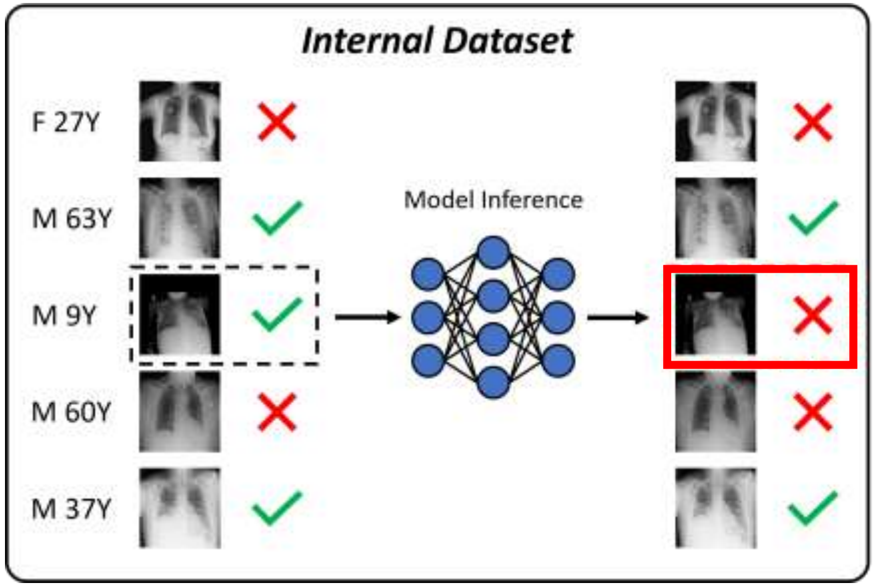


“undetectability”

MIDL Paris 2024

Adversarial Bias Attacks

- Key measures of a successful attack:
  - Bias Selectivity
  - **Bias Transferability**

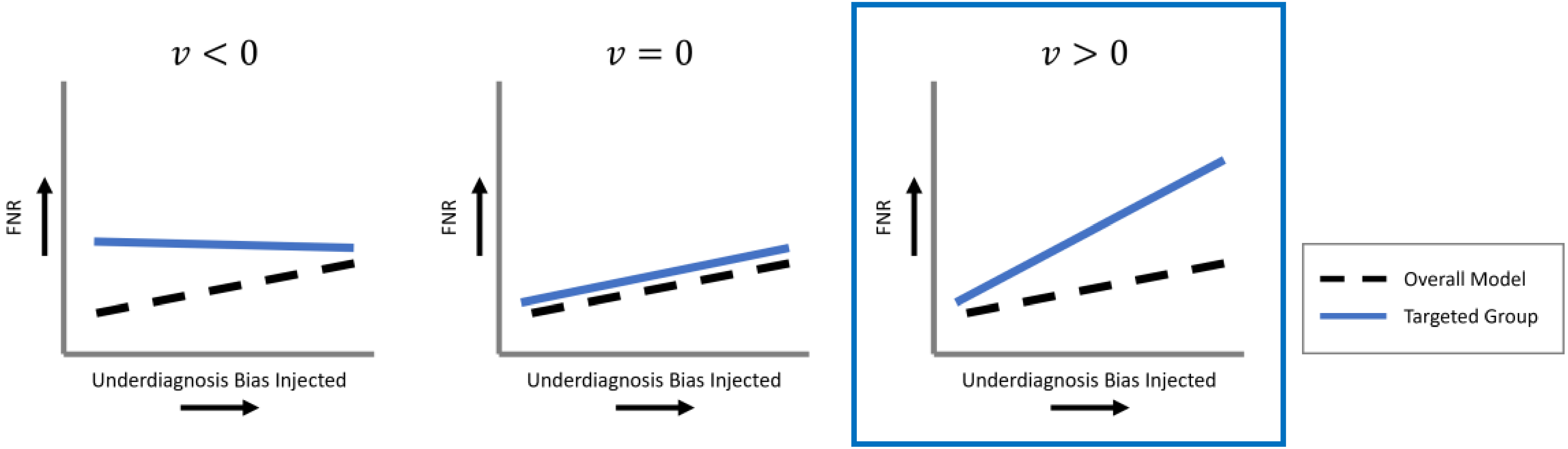


Adversarial Bias Attacks

- How do we quantify this?
- We propose **a new vulnerability metric** $\boldsymbol{v}$.

Vulnerability

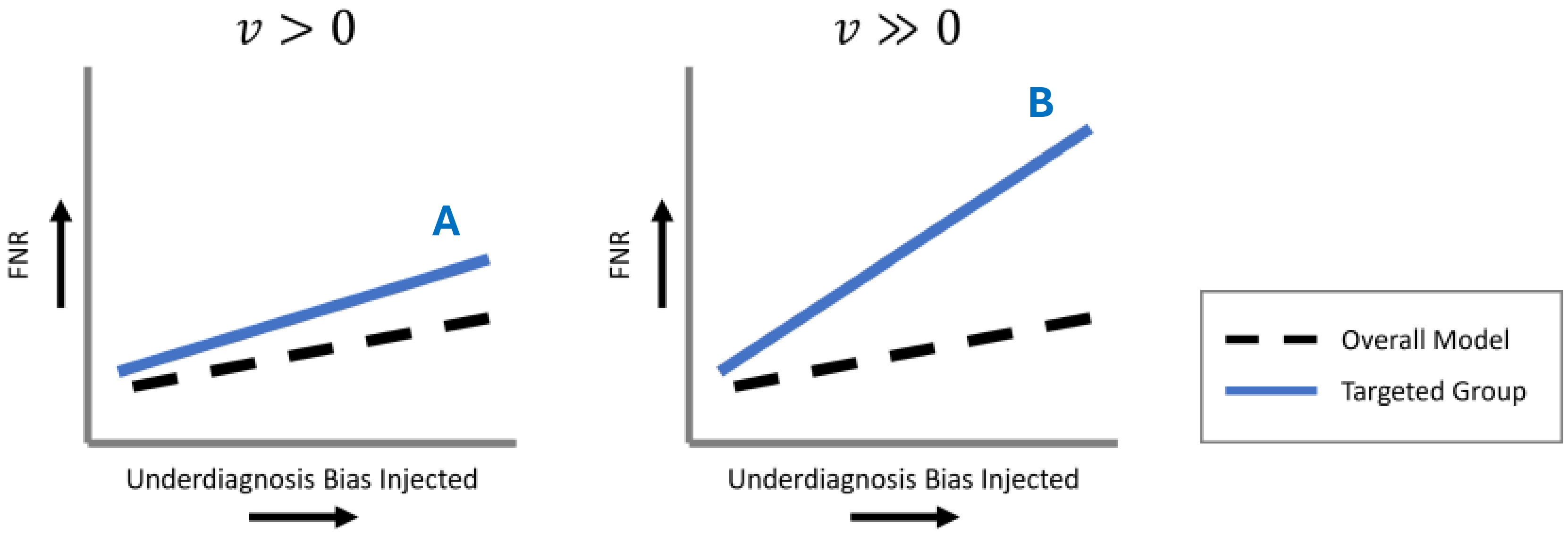
- Indicates whether the adversarial bias attack impacted the targeted group's model performance with respect to the overall model.



Model is biased!

Vulnerability

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

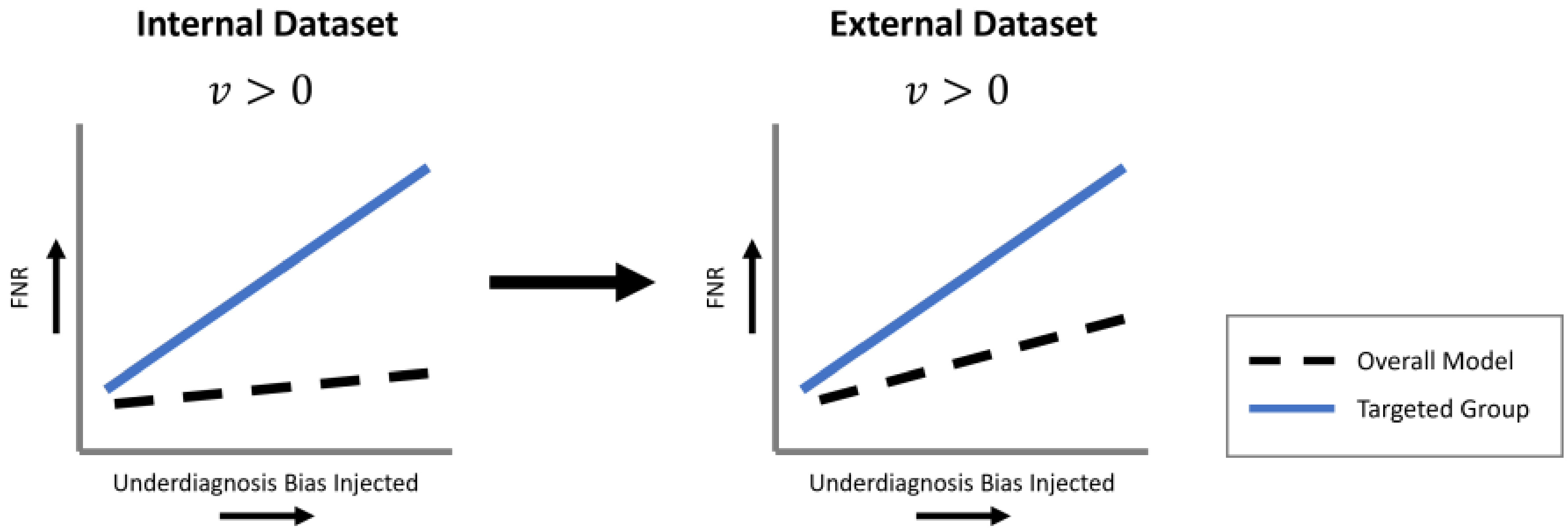


Group B is more vulnerable than Group A

MIDL Paris 2024

Vulnerability

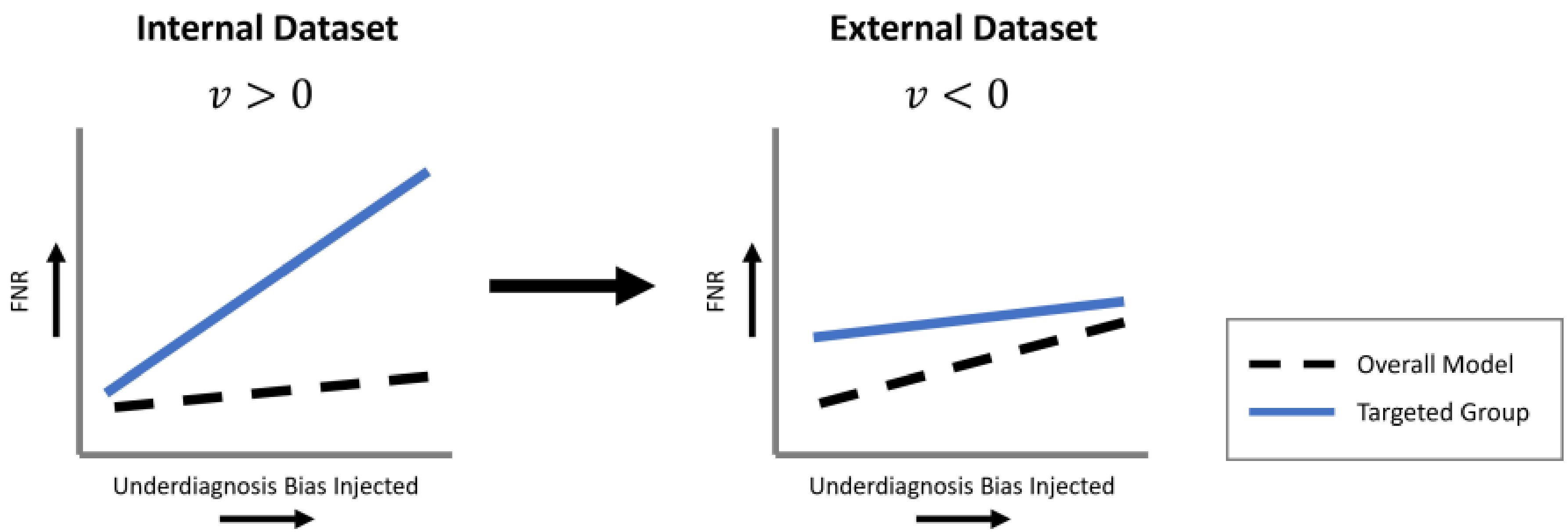
- Indicates whether bias transfers to external datasets.



Bias <u>transfers</u> from internal to external dataset

Vulnerability

- Indicates whether bias transfers to external datasets.



Bias <u>does not transfer</u> from internal to external dataset

Vulnerability

We define v as the rate parameter β of logistic regression from MLE for the difference in FNR of the target 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.

$$y \sim f(x; \alpha, \beta) = \frac{1}{1 + e^{-\alpha-\beta x}}$$

MIDL Paris 2024

Datasets

- Internal:
  - RSNA Pneumonia Detection
- External:
  - CheXpert
  - MIMIC-CXR-JPG

1. Shih, G., Wu, C. C., Halabi, S. S., Kohli, M. D., Prevedello, L. M., Cook, T. S., ... & Stein, A. (2019). Augmenting the national institutes of health chest radiograph dataset with expert annotations of possible pneumonia. *Radiology: Artificial Intelligence*, 1(1), e180041.
2. Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., ... & Ng, A. Y. (2019). Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI conference on artificial intelligence* (Vol. 33, No. 01, pp. 590-597).
3. Johnson, A. E., Pollard, T. J., Berkowitz, S. J., Greenbaum, N. R., Lungren, M. P., Deng, C. Y., ... & Horng, S. (2019). MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific data*, 6(1), 317.

Demographic Groups

Sex	Age	Intersectional Subgroups	
M	0-20Y	M 0-20Y	F 0-20Y
F	20-40Y	M 20-40Y	F 20-40Y
	40-60Y	M 40-60Y	F 40-60Y
	60-80Y	M 60-80Y	F 60-80Y
	80+Y	M 80+Y	F 80+Y

Seyyed-Kalantari, L., Zhang, H., McDermott, M. B., Chen, I. Y., & Ghassemi, M. (2021). Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. *Nature medicine*, 27(12), 2176-2182.

Experimental Design

- For each targeted group:

Manipulate RSNA train/val sets across seven rates of underdiagnosis bias

↓

Train DenseNet121 models with 5-fold cross-validation

↓

Test models on clean RSNA, CheXpert, and MIMIC test sets

↓

Measure FNR for each demographic group and overall model

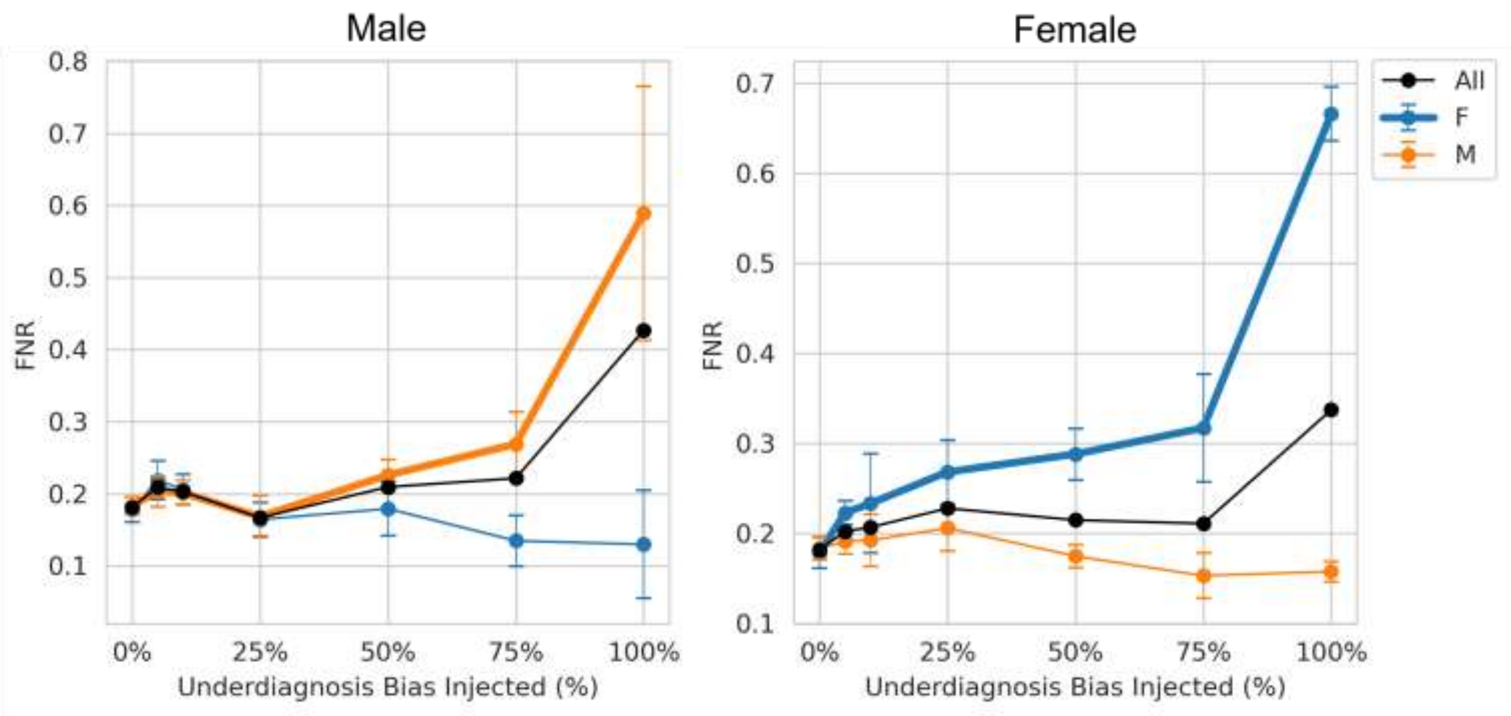
↓

Measure vulnerability v for each demographic group

Results

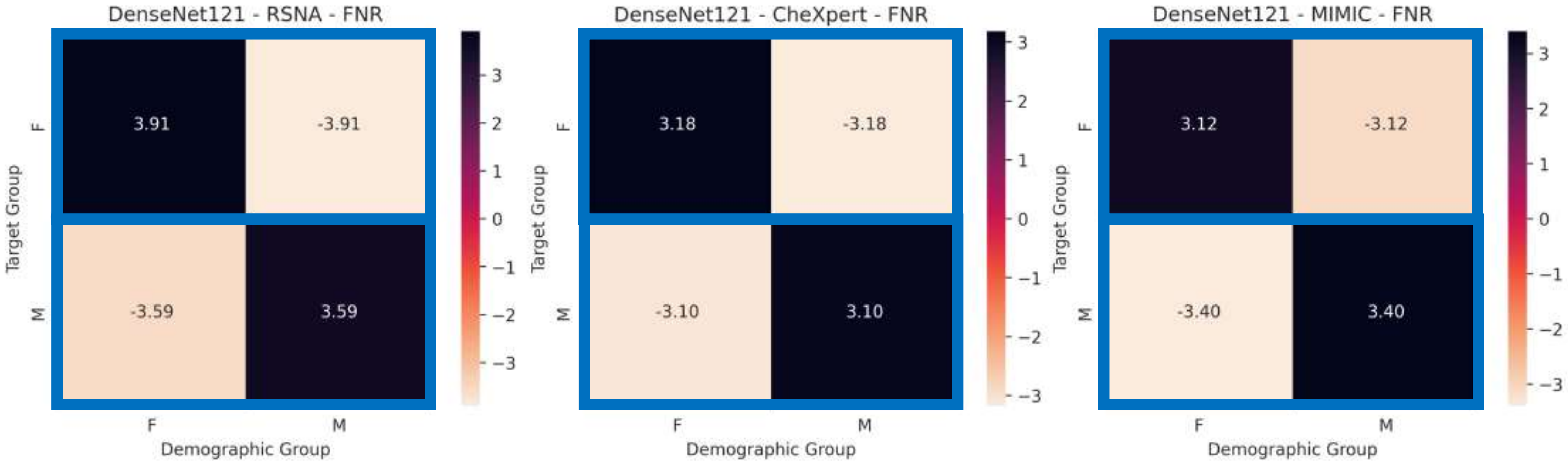
Sex Group Analysis

- The female group is more vulnerable than the male group.



Sex Group Analysis

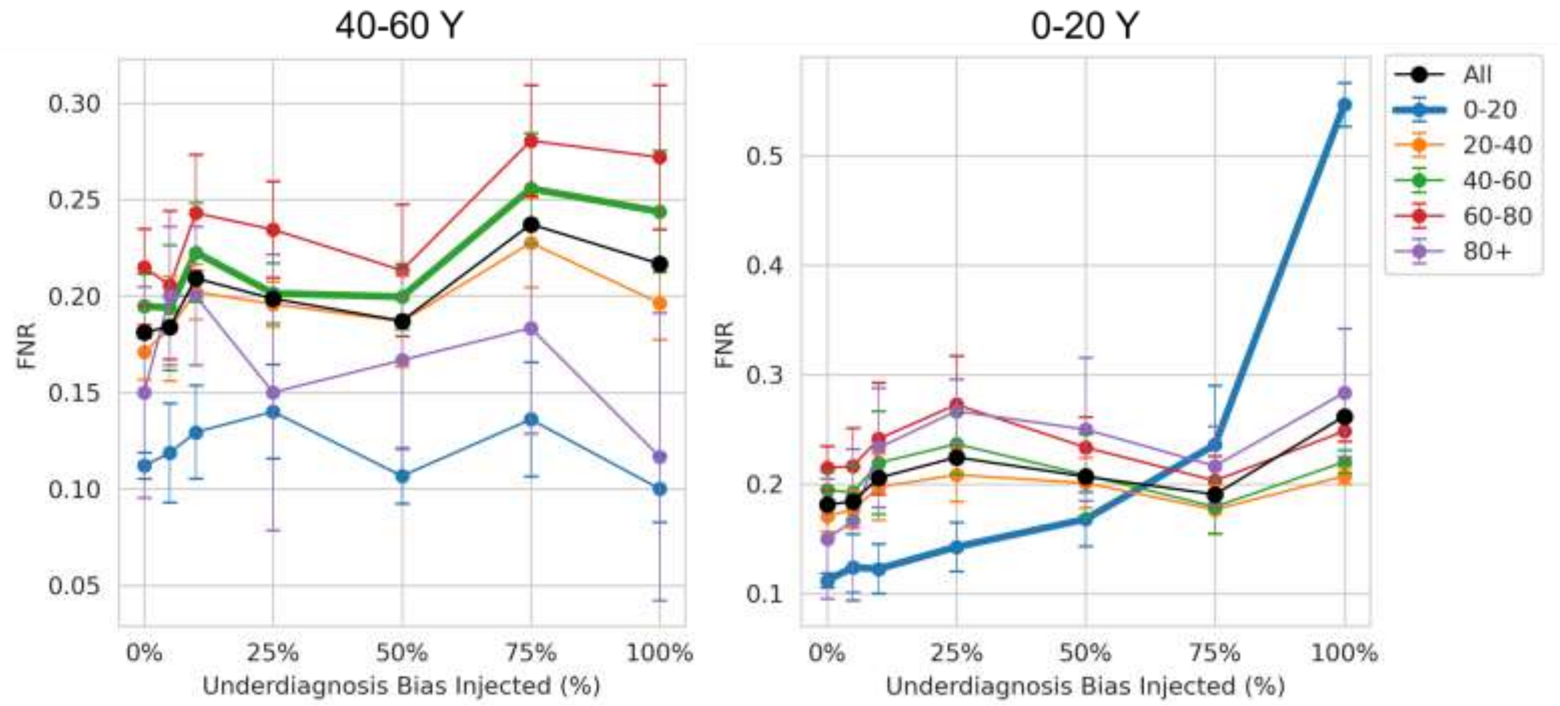
- The female group is more vulnerable than the male group.



MIDL
Paris 2024

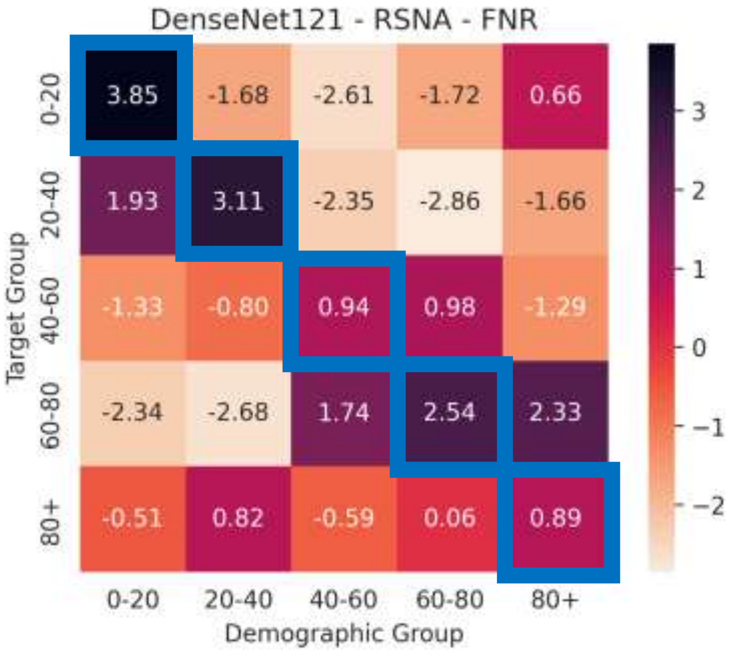
Age Group Analysis

- The 0-20Y group is the most vulnerable and the 40-60Y group is the least vulnerable.



Age Group Analysis

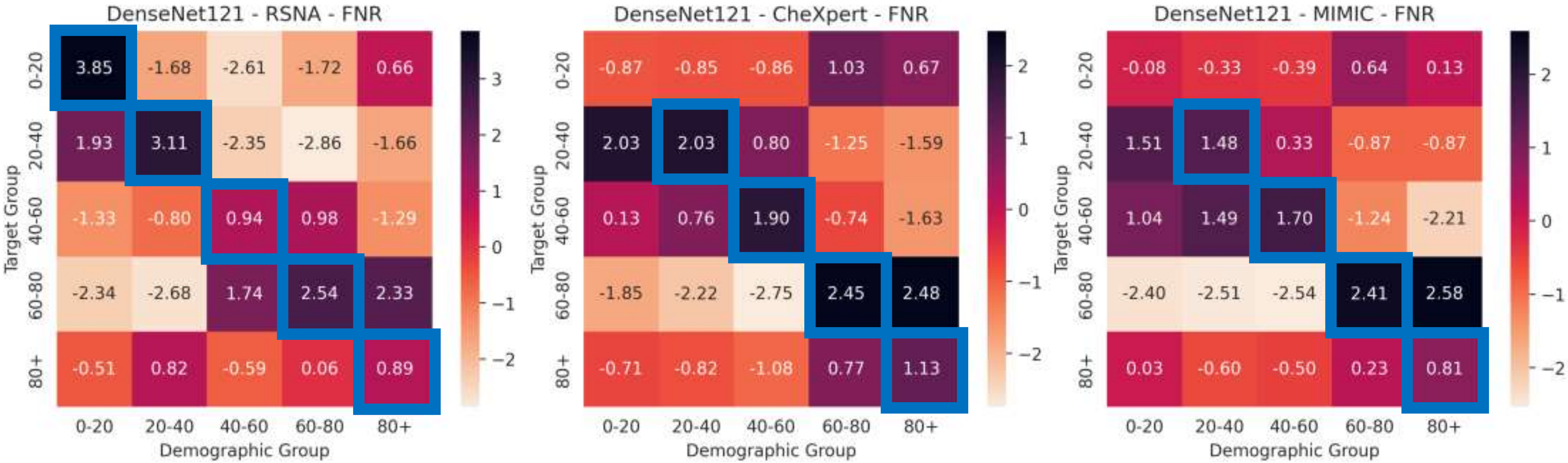
- High-selectivity for bias ($v > 0$ on diagonals)



MIDL Paris 2024

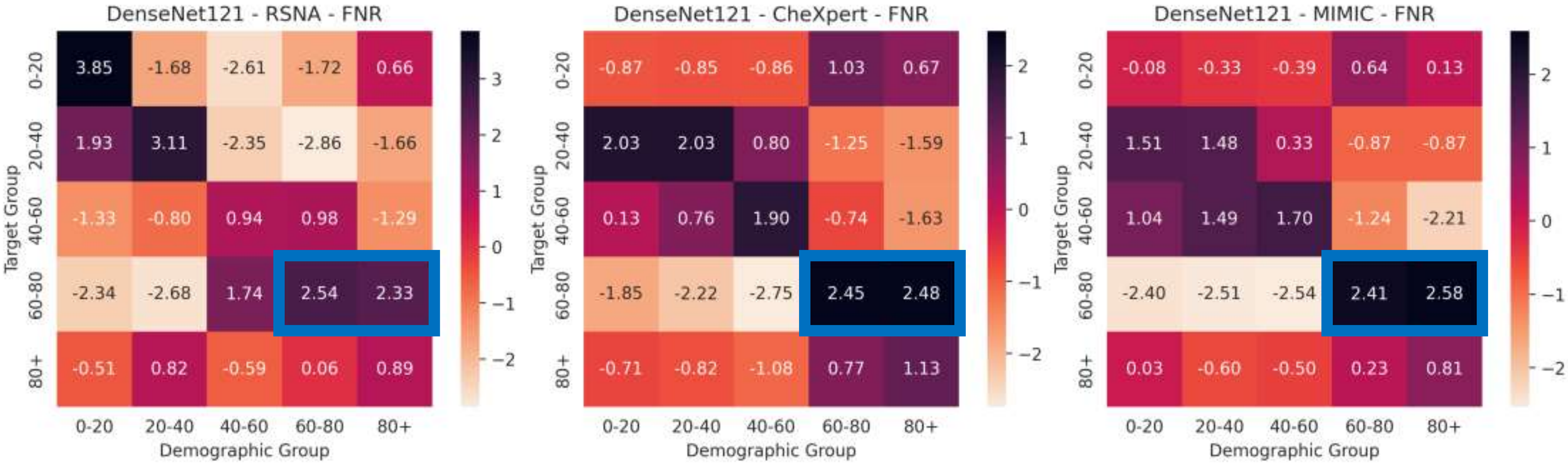
Age Group Analysis

- Vulnerability and bias selectivity transfer to external datasets



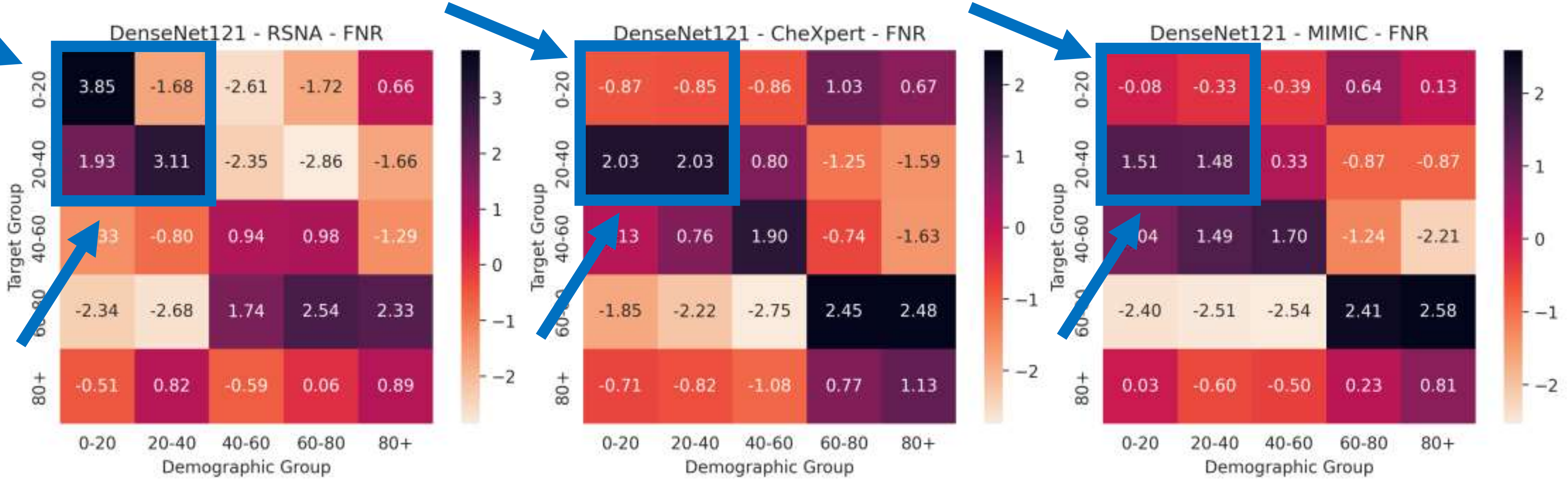
Age Group Analysis

- Targeting the 60-80Y group also affects the 80+Y group.
- 60-80Y is minority group in RSNA but majority in CheXpert and MIMIC.



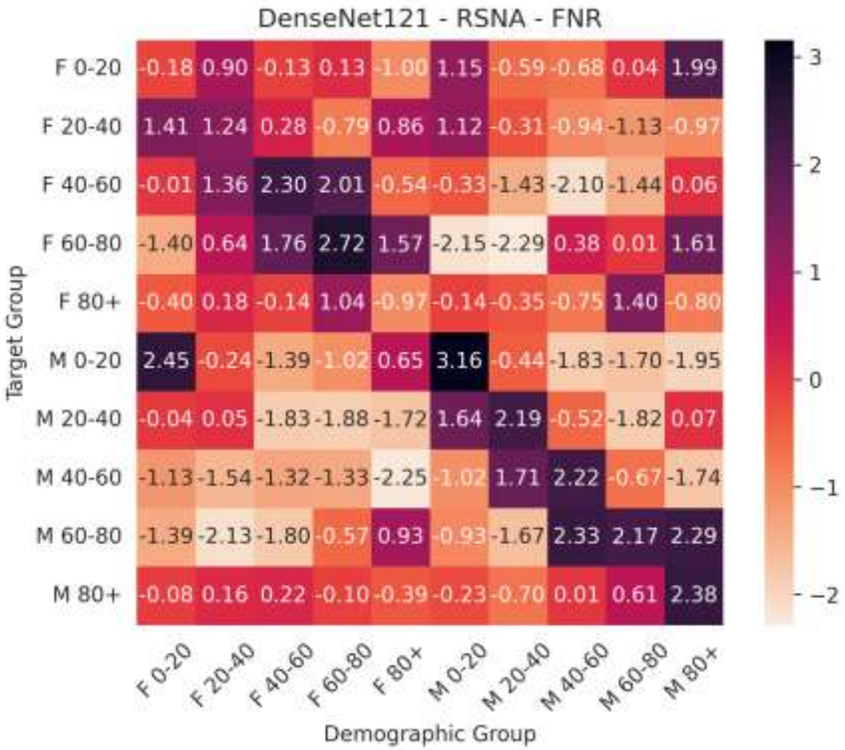
Age Group Analysis

- Pediatric patients are absent in CheXpert and MIMIC.
- Therefore, the 0-20Y group behaves like 20-40Y.



Intersectional Subgroup Analysis

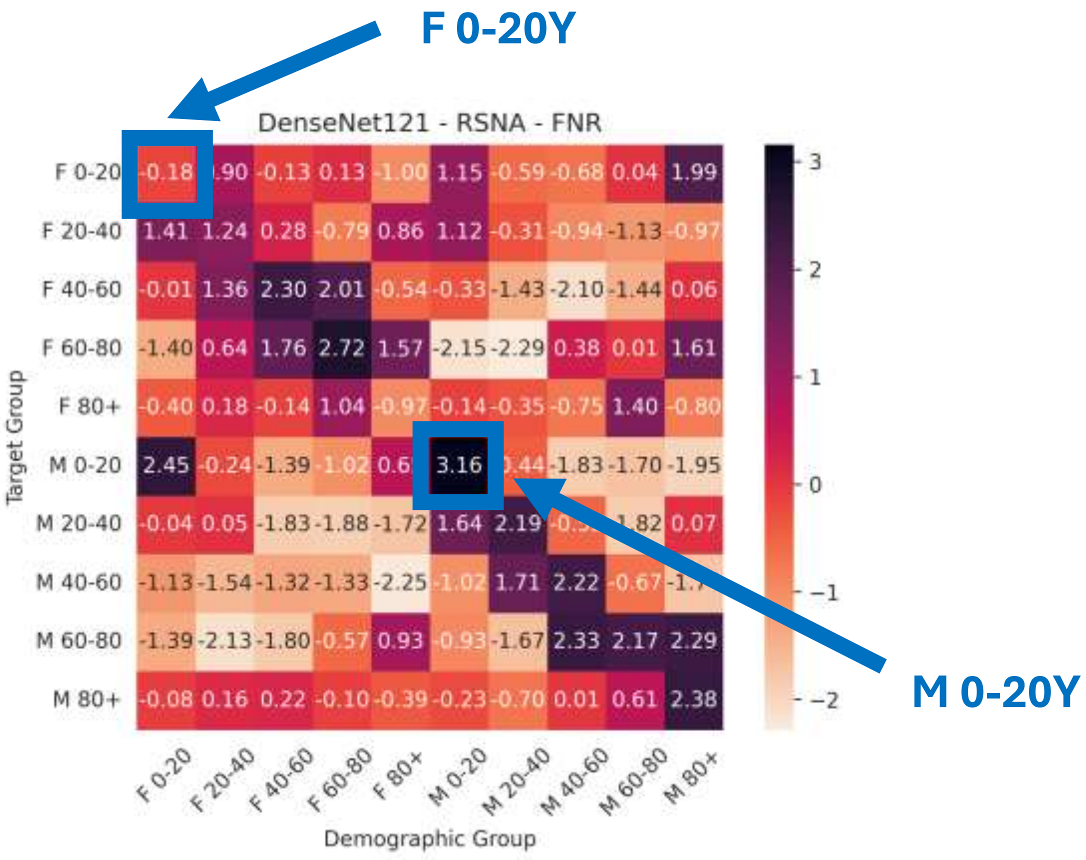
- High-selectivity for bias ($v > 0$ on diagonals)



MIDL
Paris 2024

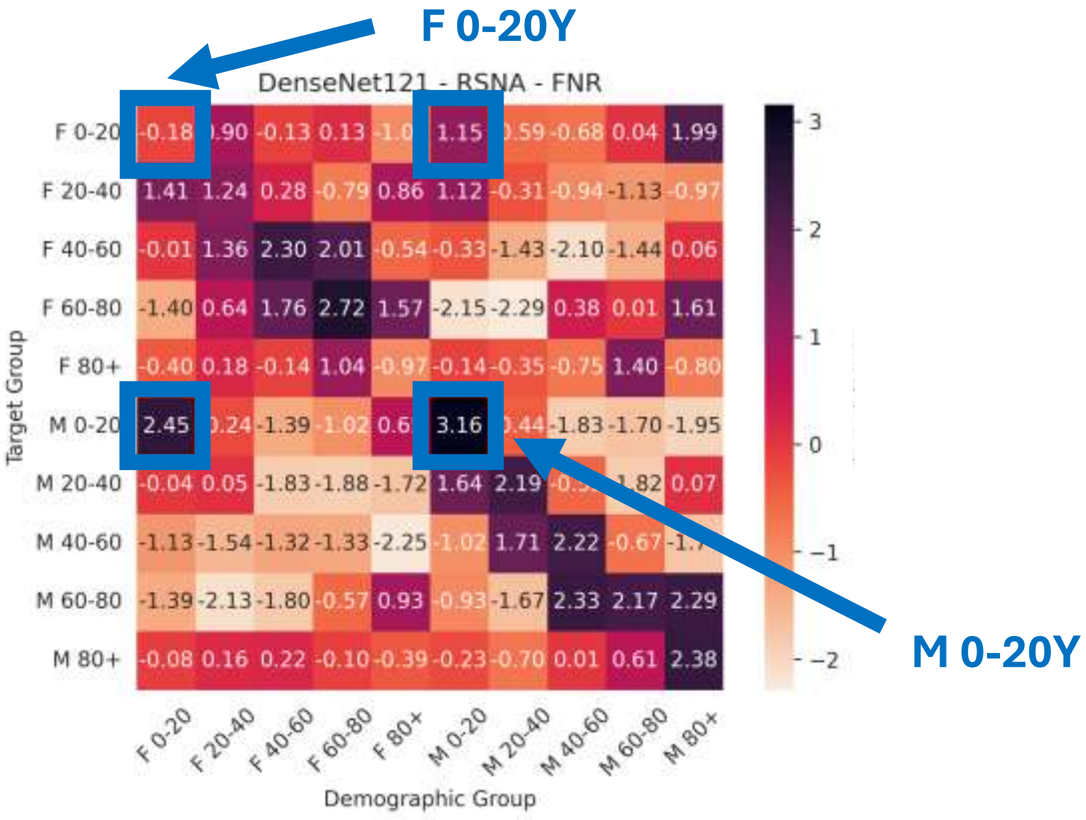
Intersectional Subgroup Analysis

- The M 0-20Y group is the most vulnerable and the F 0-20Y group is the least vulnerable.



Intersectional Subgroup Analysis

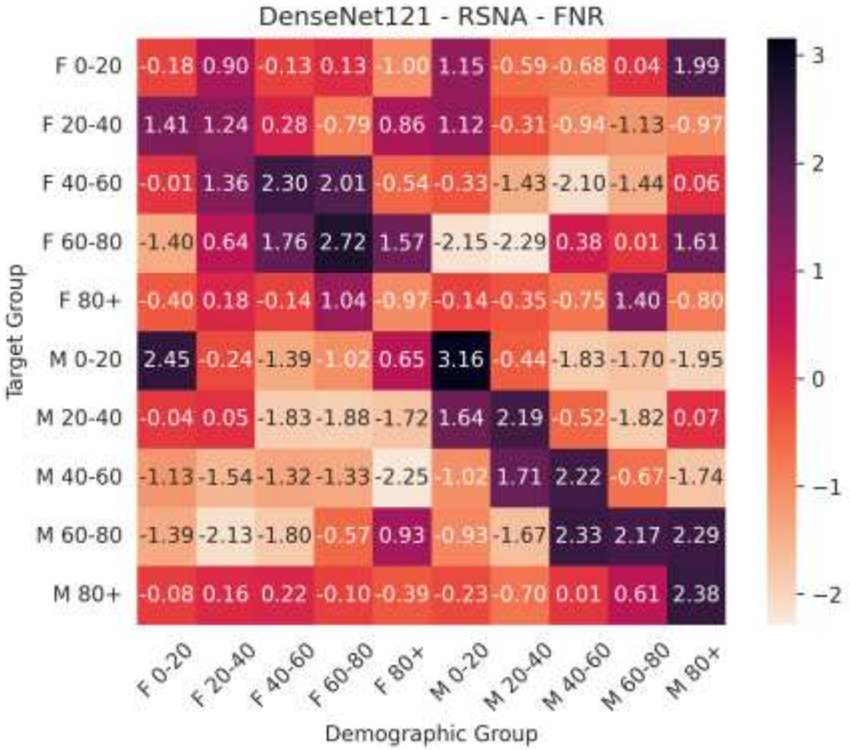
- Targeting the M 0-20Y group also impacts the F 0-20Y group.
- But targeting the F 0-20Y group only impacts the M 0-20Y group.

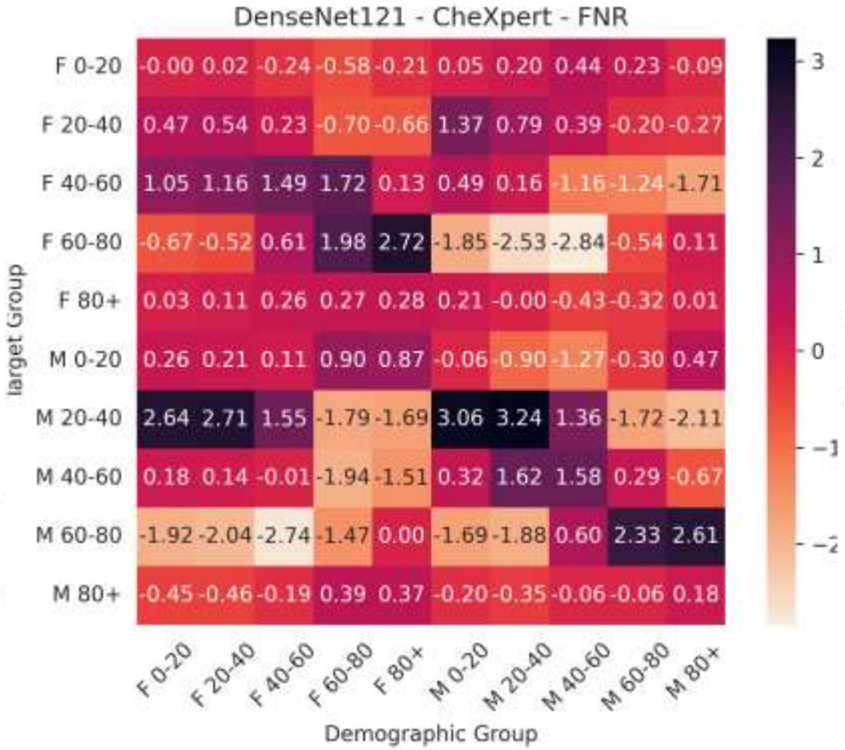


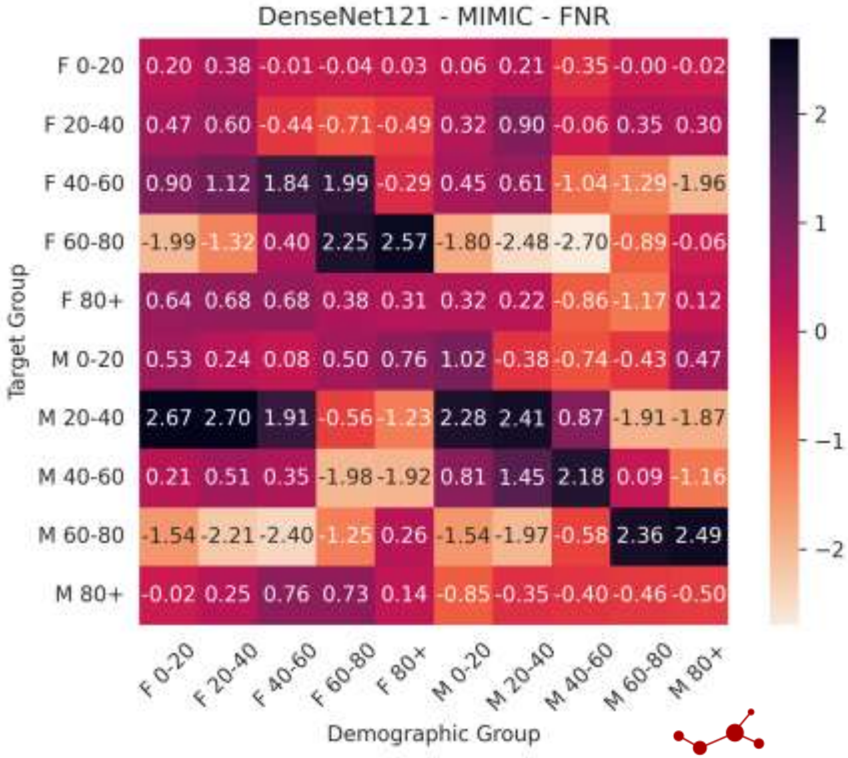
MIDL Paris 2024

Intersectional Subgroup Analysis

- Vulnerability and bias selectivity also transfer to external datasets.

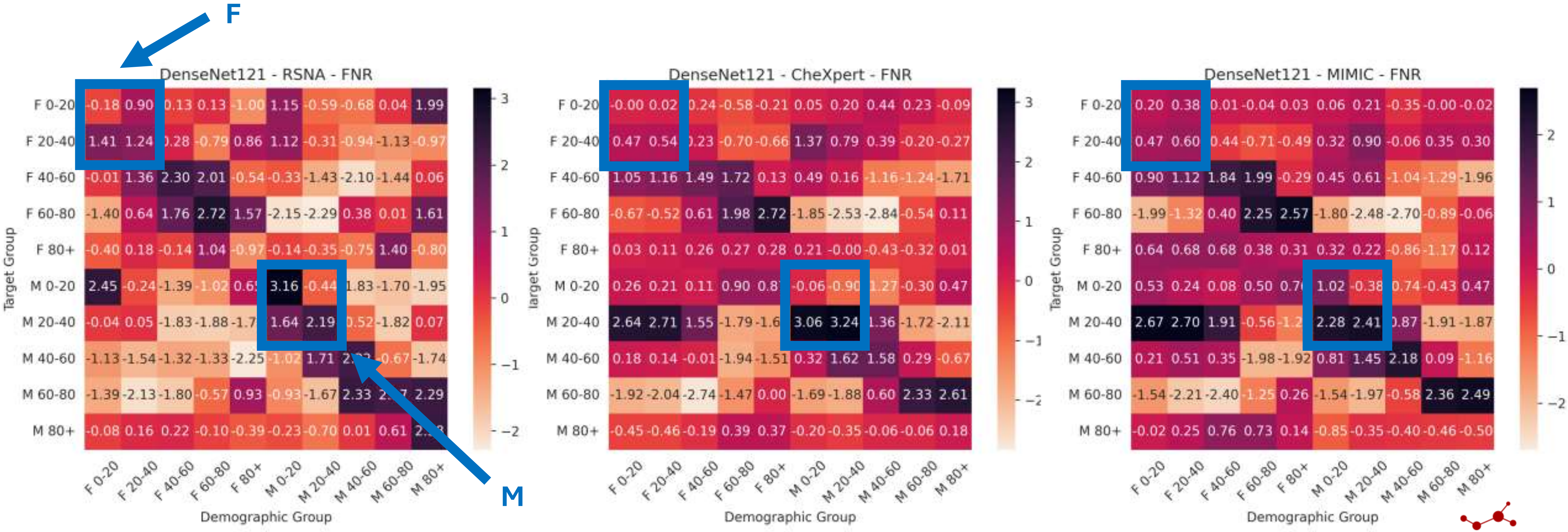






Intersectional Subgroup Analysis

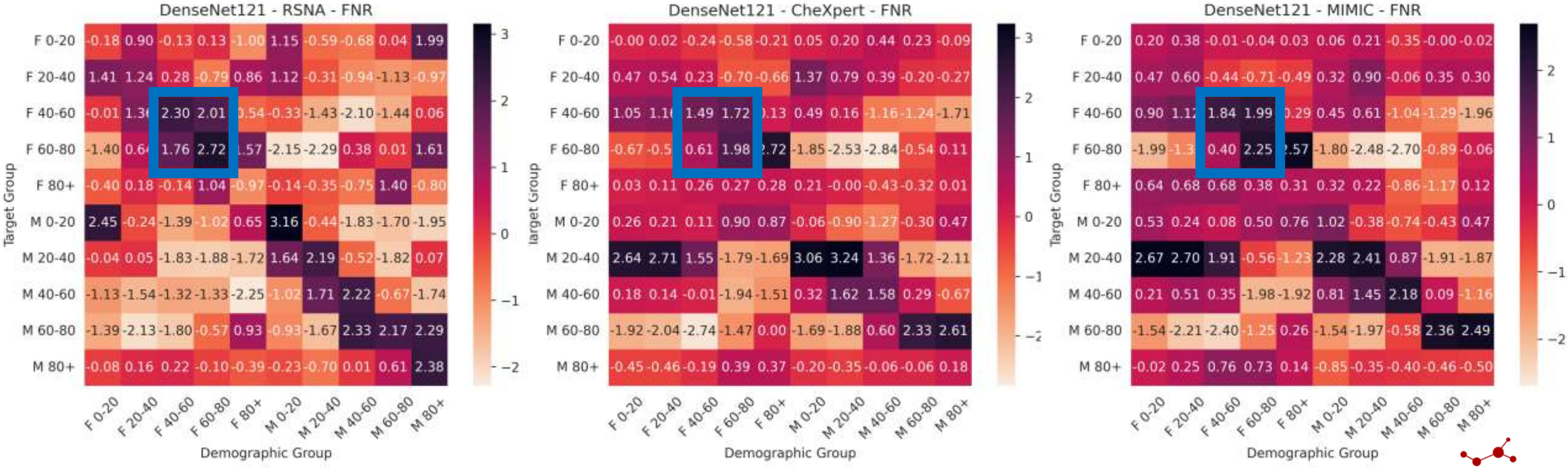
- Both 0-20Y groups also behave like both 20-40Y groups in the external datasets due to absence of pediatric patients.



MIDL Paris 2024

Intersectional Subgroup Analysis

- Interaction between the F 40-60Y and F 60-80Y groups also transfers to external datasets.



MIDL Paris 2024

Discussion

Key Findings

- Adversarial bias attacks can introduce **undetectable underdiagnosis bias** in DL models.
- They demonstrate **high-selectivity for bias** in the targeted group.
- They result in biased DL models that can **transfer bias** to external datasets.

Feasibility

- Importance of local optimization over generalization in DL.
- DL models can learn demographics as “triggers” for biased predictions.
- Hard to detect due to prevalence of labeling errors.

1. Pooch, E. H., Ballester, P., & Barros, R. C. (2020). Can we trust deep learning based diagnosis? the impact of domain shift in chest radiograph classification. In *Thoracic Image Analysis: Second International Workshop, Held in Conjunction with MICCAI 2020, Proceedings 2* (pp. 74-83). Springer International Publishing.
2. Wang, J., Liu, Y., & Levy, C. (2021). Fair classification with group-dependent label noise. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency* (pp. 526-536).
3. Cohen, J. P., Hashir, M., Brooks, R., & Bertrand, H. (2020, September). On the limits of cross-domain generalization in automated X-ray prediction. In *Medical Imaging with Deep Learning* (pp. 136-155). PMLR.

Feasibility

- During data curation:
  - Biased automated labelers
  - Clinical biases
- After data curation:
  - Man-in-the-middle or backdoor attacks
  - DL models to predict demographics with high accuracy in absence of/lack of access to dataset demographics.

1. Zhang, H., Lu, A. X., Abdalla, M., McDermott, M., & Ghassemi, M. (2020). Hurtful words: quantifying biases in clinical contextual word embeddings. In *Proceedings of the ACM Conference on Health, Inference, and Learning* (pp. 110-120).
2. Cohen, J. P., Hashir, M., Brooks, R., & Bertrand, H. (2020). On the limits of cross-domain generalization in automated X-ray prediction. In *Medical Imaging with Deep Learning* (pp. 136-155). PMLR.
3. Yi, P. H., Wei, J., Kim, T. K., Shin, J., Sair, H. I., Hui, F. K., ... & Lin, C. T. (2021). Radiology "forensics": determination of age and sex from chest radiographs using deep learning. *Emergency Radiology*, 28, 949-954.

Mitigation

- Demographic reporting in datasets
- Subgroup analysis for bias
- Curation of diverse datasets for better generalizability

1. Garin, S. P., Parekh, V. S., Sulam, J., & Yi, P. H. (2023). Medical imaging data science competitions should report dataset demographics and evaluate for bias. *Nature medicine*, 29(5), 1038-1039.
2. Bachina, P., Garin, S. P., Kulkarni, P., Kanhere, A., Sulam, J., Parekh, V. S., & Yi, P. H. (2023). Coarse race and ethnicity labels mask granular underdiagnosis disparities in deep learning models for chest radiograph diagnosis. *Radiology*, 309(2), e231693.
3. Cohen, J. P., Hashir, M., Brooks, R., & Bertrand, H. (2020). On the limits of cross-domain generalization in automated X-ray prediction. In *Medical Imaging with Deep Learning* (pp. 136-155). PMLR.

Defenses

- Label poisoning attacks have been demonstrated outside of medical imaging.
- Some defenses have shown moderate-to-high success.
- **Challenge:** These focus on label noise rather than label bias.

Zhang, H., Lu, A. X., Abdalla, M., McDermott, M., & Ghassemi, M. (2020). Hurtful words: quantifying biases in clinical contextual word embeddings. In *Proceedings of the ACM Conference on Health, Inference, and Learning* (pp. 110-120).

Defenses

- We assume that an adversary targets only one group.
- **Challenge:** In the real-world, multiple groups may be attacked simultaneously.
- Further exploration warrants future work!

Conclusion

- A **crucial first step** in highlighting the implication of undetectable adversarial bias attacks on DL models in the clinical environment.
- Such attacks can scale across various applications of DL in medical imaging and target vulnerable patient populations.

Thank you!

pkulkarni@som.umaryland.edu

@itspranavk

itspranavk

Paper

University of Maryland
School of Medicine

MIDL
Paris 2024