



Demographic Bias in Chest X-Ray Foundation Models: Disparities in Underdiagnosis and Overdiagnosis from Embedding-Based Classification

Mohammadreza Chavoshi, MD, Emory University; Frank Li, PhD; Hari Trivedi, MD; Theo Dapamede, MD, PhD; Aawez Mansuri, MS; Rohan Satya Isaac, MS; Bardia Khosravi, MD, MPH, MHPE; Janice Newsome, MD; Judy Gichoya, MD, MS, FSIIM

Introduction/Background

To assess how fairly and accurately three chest X-ray foundation models classify images as normal or abnormal across different demographic groups.

Methods/Intervention

We analyzed 1,304,022 chest X-rays (890,578 frontal and 413,444 lateral views) using three foundation models: Rad-DINO, CheXagent, and Google. We defined false positives (FPR) as underdiagnosis (missed abnormalities) and false negatives (FNR) as overdiagnosis (mistakenly calling a normal image abnormal). Each model was tested using two approaches: a prototype-based method that compares image embeddings to balanced demographic prototypes of normal and abnormal cases using cosine similarity, and an XGBoost classifier trained on the model's image embeddings to identify normal cases. We measured performance using AUC, FPR, and FNR—both overall and by subgroup. Fairness was assessed by measuring each metric's difference between the best and worst subgroup performances.

Results/Outcome

XGBoost outperformed the prototype-based method in AUC. In frontal view, AUC improved from 0.8279 to 0.9518 for Rad-DINO 0.8735 to 0.9683 for CheXagent; and 0.8654 to 0.9633 Google). In the lateral view, AUC increased from 0.7924 to 0.9204 for Rad-DINO; 0.8287 to 0.9449 for CheXagent; and 0.7958 to 0.9062 for Google. CheXagent consistently achieved the highest and most balanced performance with the highest AUCs in both frontal (0.9683) and lateral (0.9449) views and the lowest average disparity across subgroups in FPR (0.0667) and FNR (0.0669). Google ranked second in AUC but had the greatest FPR disparity (0.0841), suggesting less stable behavior across demographics. Rad-DINO showed the lowest AUCs and the highest FNR disparity (0.0961), indicating a higher rate of overdiagnosis in certain subgroups. In both XGBoost and prototype methods, age consistently exhibited the highest disparity in FPR and FNR, with disparity reaching up to 0.278 and 0.298, respectively.

Conclusion

Despite improved performance with XGBoost, demographic disparities persisted across all models, suggesting that foundation model embeddings retain demographic signals that influence classification. Age emerged as the dominant source of bias. CheXagent showed the most balanced performance, while Rad-DINO exhibited the greatest variability.

Statement of Impact

Deploying foundation model-based classifiers without fairness evaluation may reinforce health disparities. Our findings show that demographic biases—especially age-related ones—persist even after downstream training,

highlighting the need for subgroup-level analysis before clinical use

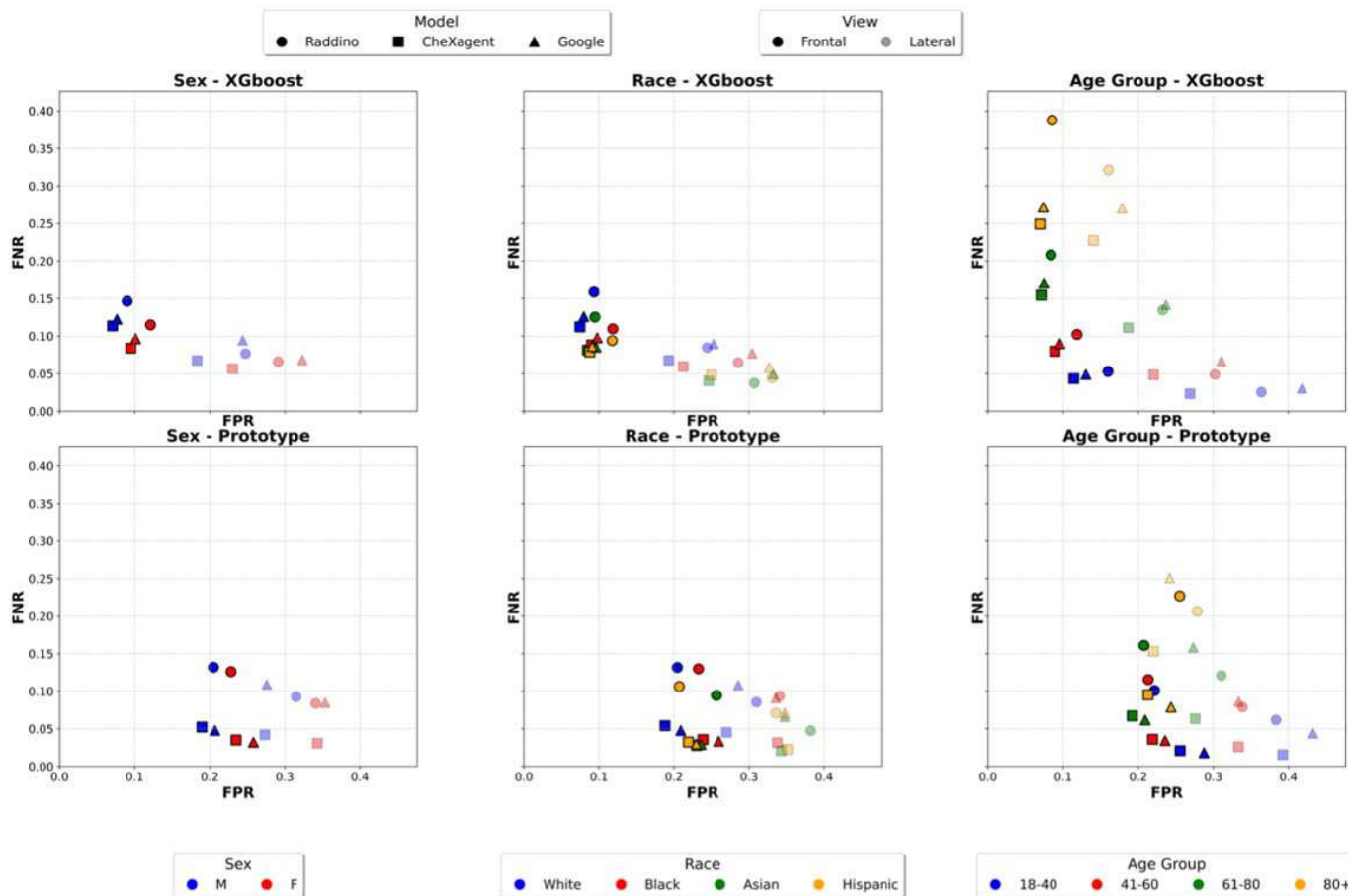


Figure 1: FPR and FNR for sex, race, and age subgroups across three foundation models (Rad-DINO, CheXagent, Google), under prototype-based and XGBoost classification. Each point represents a model-view-subgroup combination. Age-related subgroups show the widest spread, particularly under XGBoost, suggesting persistent demographic disparities despite improved performance.

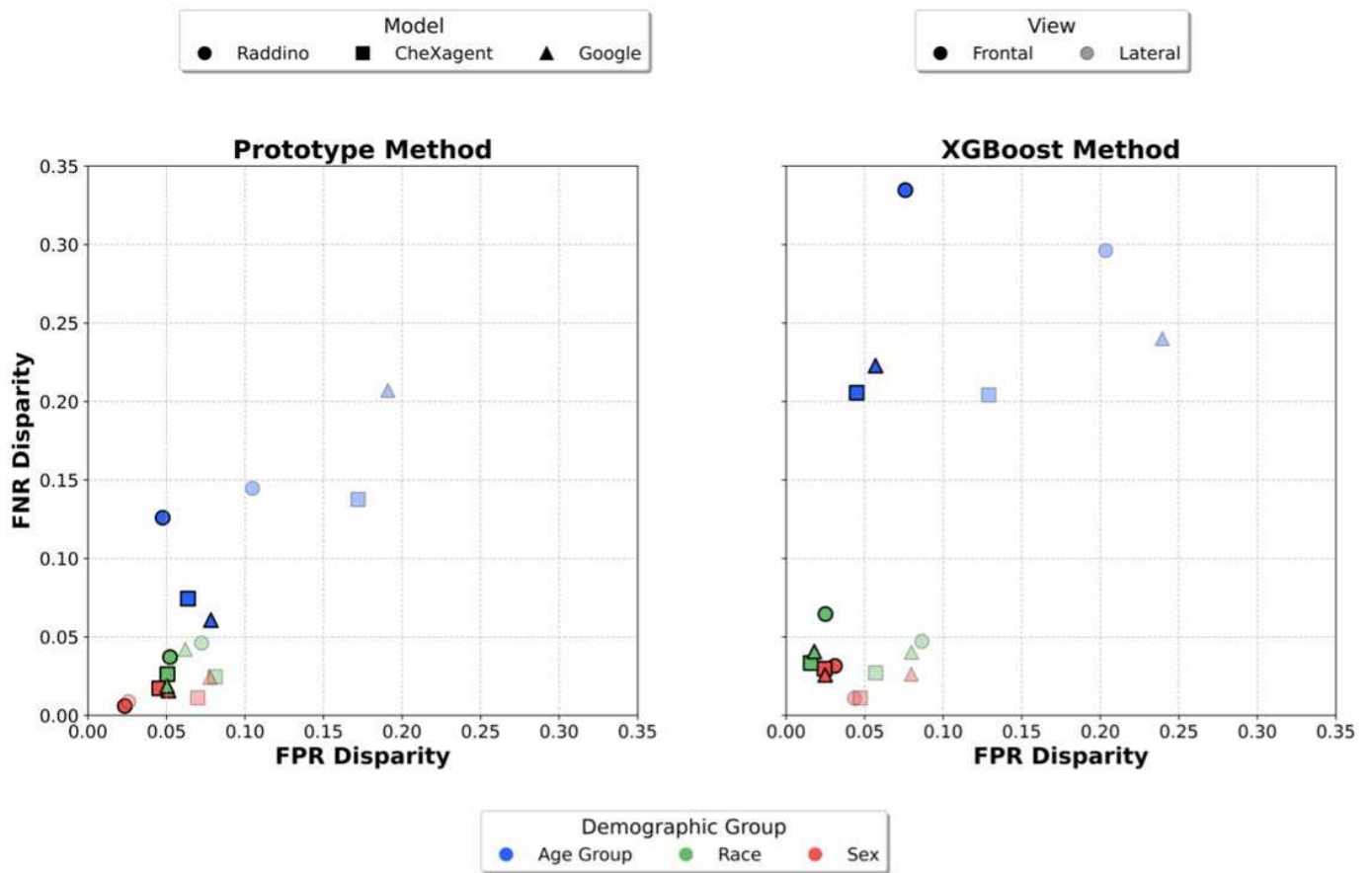


Figure 1: FPR and FNR disparities for each model-view-method combination, grouped by demographic category. Age group consistently exhibits the highest disparity, especially under XGBoost, reinforcing the finding that foundation model embeddings encode age-related information that affects classification fairness.

Keywords

Foundation Models; Chest X-ray; Underdiagnosis; Fairness; Embeddings