SIIM CAIMI25
Conference on Artificial Intelligence in Medical Imaging
OCT 20-21 | UNIVERSITY OF CALIFORNIA, IRVINE

Foundation Models Fail to Close Fairness Gaps in Medical AI: Unexpected Benefits of LLM-Extracted Labels

Amirali Khosravi, MD, Mayo Clinic; Bardia Khosravi, MD, MPH, MHPE;
Bradley J. Erickson, MD, PhD, CIIP, FSIIM; Judy W. Gichoya, MD, MS, FSIIM

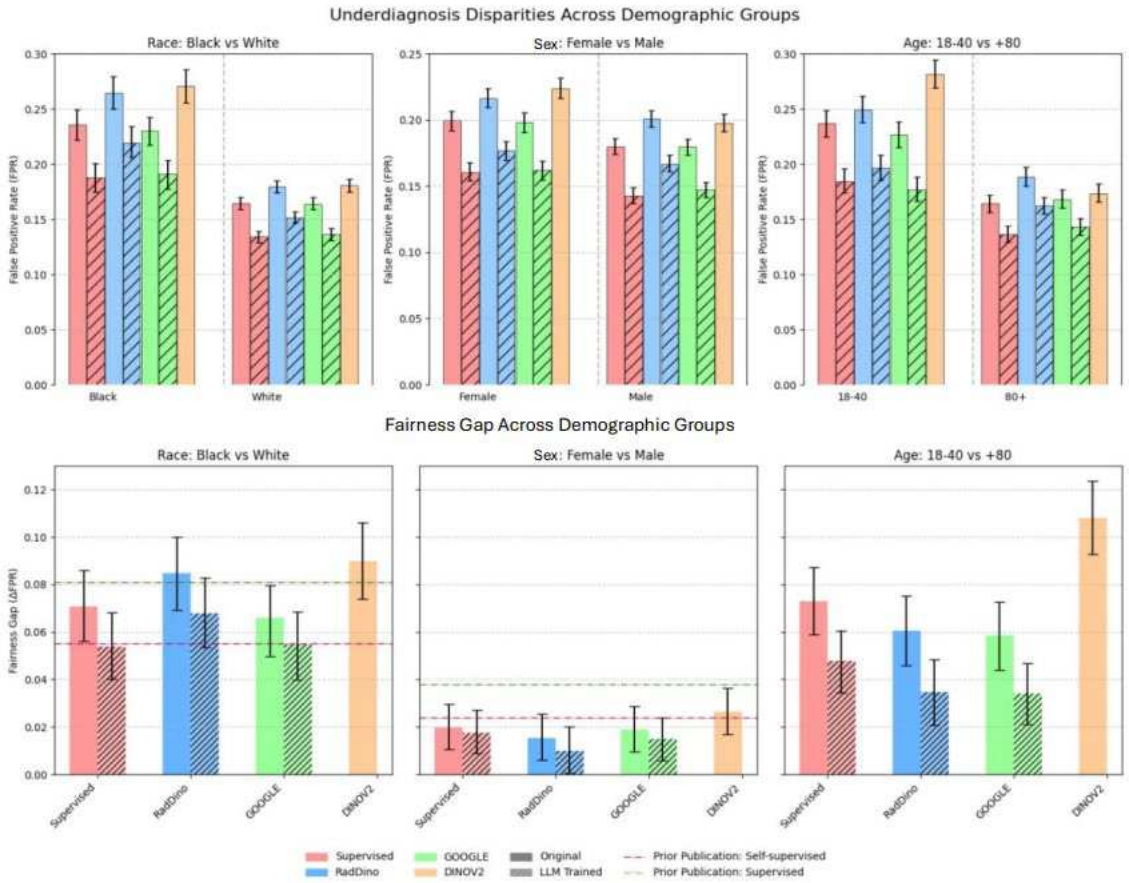
Introduction/Background
Foundation models in medical imaging offer improvements in pathology detection through self-supervised pre-training, but their fairness implications remain underexplored. This study quantifies chest radiograph pathology detection disparities across demographic groups, comparing supervised learning against foundation models.

Methods/Intervention
We analyzed 243,331 frontal chest radiographs from MIMIC-CXR using a 70-15-15 split. Four classification architectures were evaluated: (1) supervised CNN baseline, (2–3) classifiers using RadDino and Google's CXR Foundation, and (4) DINOV2 embeddings (control). We trained models with both rule-based CheXpert labels and LLM-extracted labels from radiologic reports. Fairness gaps were computed as the difference in false-positive rates between demographic groups (Black–White, Female–Male, 18–40 vs. 80+ years), with statistical significance assessed via bootstrapping.

Results/Outcome
All models achieved 84–90% AUROC for "No Finding" classification. For race comparisons, fairness gaps were comparable across supervised (0.071, 95% CI: 0.055–0.088), RadDino (0.085), and Google's CXR Foundation (0.066) model. LLM-extracted labels decreased fairness gaps: supervised by 24% (0.071→0.054), RadDino by 19% (0.085→0.068), and Google by 18% (0.066→0.054). Sex disparities were smaller (0.015–0.020) and reduced by an average of 26% with LLM labels. Age-related disparities showed the greatest improvement with gaps reduced by 35-46% using LLM-generated labels.

Conclusion
Medical foundation models demonstrate fairness gaps comparable to supervised learning, indicating that self-supervised pre-training on medical data and larger datasets alone are insufficient to eliminate demographic disparities. Notably, even medical-specific foundation models (RadDino, Google CXR) showed similar or sometimes lower fairness gaps compared to the general-purpose DINOV2, suggesting that domain-specific pre-training may actually provide some fairness benefits over general-purpose models. However, these improvements remain insufficient to eliminate disparities entirely. LLM-extracted labels unexpectedly reduced bias across all architectures, likely by providing more accurate and consistent labeling that reduces noise and systematic errors in ground truth annotations, rather than introducing additional bias.

Statement of Impact
These findings challenge assumptions about foundation model fairness and emphasize the urgent need for bias-mitigation strategies in medical AI deployment, particularly for historically underrepresented populations where health disparities already exist.



Demographic disparities in chest X-ray classification across foundation models. (Top) False positive rates by demographic group. (Bottom) Fairness gaps with 95% confidence intervals, comparing CheXpert labels (solid) vs LLM-extracted labels (hatched) across supervised, RadDino, Google CXR, and DINOV2 models.

Keywords

Medical AI fairness; Foundation models; Demographic bias