A Fairness Audit of Medical Imaging Foundation Models on a Multimodal Structured Clinical Benchmark

Milan Mohan [1] **Saketh Lingisetty** [1] **Umar Ahmad** [1] **Avi Kumar** [1] **Shankar Harikrishnan** [1] **Sanjana Chamarty** [1] **Kevin Zhu** [1]

[1] *Algoverse AI Research*. Correspondence to: Kevin Zhu kevin@algoverseairesearch.org.

1. Introduction

Medical imaging foundation models (FMs) promise rapid clinical deployment through reusable frozen representations, but strong aggregate performance can mask clinically catastrophic failures in specific patient groups [1, 2]. Pulmonary embolism (PE) is a life-threatening condition where missed diagnoses directly worsen outcomes [3, 4], making subgroup failures a patient-safety concern rather than a statistical artifact.

We present the first systematic fairness audit of MedImageInsight (MII; 632M, ViT-H/14), MedSigLIP (MSig; 400M, SigLIP ViT-So/14), and BiomedCLIP (BCL; 86M, ViT-B/16) [5, 6, 7] on INSPECT [8]: 23,248 CT pulmonary angiography (CTPA) studies from 19,402 patients pairing imaging with longitudinal EHR for PE diagnosis and seven prognostic tasks. Our audit surfaces three concrete deployment findings: (1) a spatial-aggregation bottleneck explains the diagnostic gap versus task-specific baselines, and EHR fusion cannot compensate for it; (2) logistic regression outperforms MLP probes on high-capacity embeddings, indicating approximately linearly separable clinical structure; and (3) age is the dominant and previously unreported disparity dimension, with encoder capacity as a prerequisite for effective post-hoc debiasing.

2. Methods

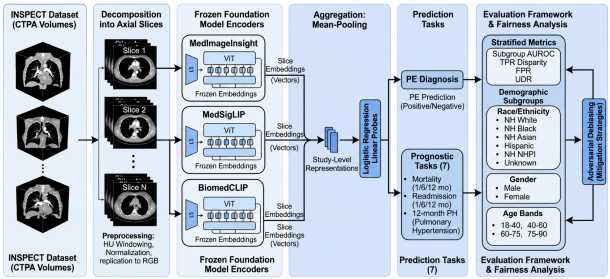

Fig. 1: Pipeline overview. CTPA volumes are decomposed into axial slices, encoded by a frozen FM, and mean-pooled into study-level representations passed to a linear probe. Fairness evaluation reports per-subgroup AUROC, TPR, FPR, and UDR across race/ethnicity, gender, and age band.

All models are frozen feature extractors (Figure 1). CTPA volumes are decomposed into axial slices, preprocessed (clipped to a lung-relevant Hounsfield Unit window, normalised to 8-bit, and resized), encoded slice-by-slice by the frozen FM, and mean-pooled to a study-level vector passed to a linear probe. Mean-pooling dilutes the spatially focal filling-defect signal required for PE detection while leaving diffuse prognostic signals intact. Logistic regression (LR) serves as the primary probe; a systematic MLP sweep (hidden sizes $(256, 128)$, $(512, 256)$, $(512, 256, 128)$; Adam, batch norm, dropout 0.2) confirms LR outperforms all MLP variants for MII (+0.014 AUROC) and MSig (+0.021), indicating approximately linearly separable task-relevant structure in high-capacity encoder embeddings. A controlled CT-only vs. EHR-only vs. CT+EHR ablation directly tests whether structured EHR context can close the representational gap.

Fairness metrics include per-subgroup AUROC and underdiagnosis rate (UDR = 1−TPR), stratified by race/ethnicity, gender, and age band, with permutation-test significance (1,000 samples). Three bias mitigation strategies are evaluated: importance weighting (IW), group resampling, and adversarial debiasing (gradient reversal, $\alpha \in \{0.5, 1.0, 2.0\}$ selected per protected attribute on the validation set).

3. Results

3.1 Diagnostic, Prognostic, and Multimodal Performance

All three FMs fall significantly below the task-specific CT-LRCN baseline on PE diagnosis (AUROC 0.680–0.684 vs. 0.721; $p \leq 0.009$), traceable to mean-pooling diluting the focal vascular signal. Prognostic signals transfer well (mortality AUROC 0.83–0.90), confirming spatial aggregation as the primary mediator of the diagnostic gap. The CT-only vs. EHR-only vs. CT+EHR ablation shows that frozen CT embeddings already capture the dominant clinical signal: EHR fusion adds at most +0.017 AUROC (BiomedCLIP readmission, where a weaker visual representation leaves more room for EHR to contribute) and is near-zero or negative for mortality. Thus, a richer multimodal context cannot substitute for representational improvements. The diagnostic bottleneck is spatial aggregation: no amount of structured EHR context recovers the localised vascular signal lost to mean-pooling.

3.2 Age is the Dominant Disparity Axis

Patients aged 18–40 face extreme underdiagnosis, with UDRs of 0.63–0.80 versus 0.31–0.41 for ages 75–90 (gap 0.32–0.45; $p < 0.001$ across all models), substantially exceeding race/ethnicity (0.21–0.29) and gender (0.08–0.16) gaps. Correspondingly, Figure 2 illustrates this severe performance degradation in terms of overall diagnostic capability, showing MedSigLIP

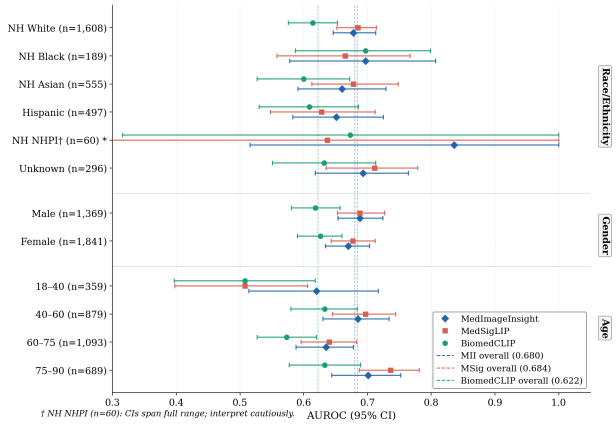

Fig. 2: Diagnostic AUROC by demographic subgroup. The 18–40 age group shows near-chance performance for MedSigLIP and BiomedCLIP (AUROC 0.508), a clinical failure mode for younger patients. †NH NHPI (n=60): CIs span the full range; excluded from interpretation.

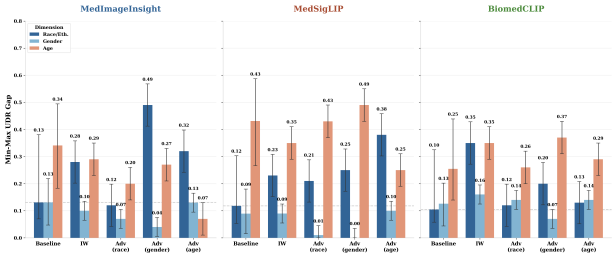

Fig. 3: Min–max UDR gap by mitigation strategy and demographic dimension. Age-targeted adversarial debiasing ($\alpha = 0.5$, $p < 0.001$) achieves the single largest reduction for MedImageInsight (0.333 → 0.069) with minimal AUROC cost. IW widens race gaps; group resampling collapses AUROC to ≈0.56. BiomedCLIP yields no statistically significant reductions ($p \geq 0.12$).

and BiomedCLIP reaching near-chance diagnostic AUROC (≈0.508) for the 18–40 cohort. The age gap persists across all seven prognostic tasks (UDR gaps 0.04–0.20) and in the CT-LRCN baseline (18–40 UDR = 0.571), confirming a data-distributional contribution: younger PE patients are clinically rarer and underrepresented in training data. Female underdiagnosis is statistically significant across all three models ($p \leq 0.040$), while CT-LRCN baseline reverses this direction (Female TPR = 0.592 > Male = 0.558), establishing that frozen-encoder pipelines introduce a gender fairness cost that task-specific training partially overcomes. Threshold-sweep analysis confirms the female disparity persists at all decision thresholds, ruling out a threshold artifact and indicating a representation-level gap. Hispanic disparity does not reach significance (p = 0.11–0.21) in this cohort.

3.3 Debiasing Requires Sufficient Encoder Capacity

Only adversarial debiasing reliably reduces gaps without collapsing task performance (Figure 3). As visually confirmed in the figure, importance weighting (IW) paradoxically widens race/ethnicity UDR gaps across all models. (Note that group resampling is omitted from Figure 3 because it collapses overall AUROC to ≈ 0.55–0.56, rendering the resulting gap comparisons uninterpretable.) For MII, age-targeted debiasing ($\alpha = 0.5$) cuts the age UDR gap by 79% (0.333 → 0.069; $p < 0.001$) at only 0.011 AUROC cost; race-targeted debiasing simultaneously reduces race, gender, and age gaps by 50%, 44%, and 40%, consistent with shared demographic latent structure in the 1,024-d embedding space. For MSig, gender-targeted debiasing eliminates the gender gap entirely ($p < 0.001$); age-targeted achieves a 44% reduction ($p = 0.007$). BCL reductions are numerically meaningful (up to 60% for race) but none reach statistical significance ($p \geq 0.12$): demographic probing confirms BCL encodes less separable demographic structure (race/ethnicity macro-F1 0.537 vs. 0.677–0.707 for the larger encoders), so gradient reversal cannot reliably disentangle signal in its compressed 512-d space. Embedding capacity is a fairness infrastructure requirement, not merely a performance consideration.

4. Conclusion

Three findings are deployment-critical. The ≈4-point diagnostic gap versus CT-LRCN is a spatial aggregation failure: EHR fusion cannot recover the focal vascular signal lost to mean-pooling, and future pipelines for localisation-sensitive tasks should prioritise attention-based pooling over richer modality fusion. LR outperforming MLP probes on high-capacity embeddings indicates that default escalation to MLPs may underperform when embeddings are approximately linearly separable. Most critically, age is the dominant disparity axis: near-chance AUROC for patients aged 18–40 across two of three encoders, persistent over all eight tasks and in the task-specific baseline, constitutes a patient-safety failure invisible to aggregate reporting. Adversarial debiasing is the only practical mitigation but requires sufficient embedding capacity to be reliable. Per-group evaluation across age, race, and gender should be a minimum standard before any FM-derived clinical AI pipeline is deployed.

Acknowledgments

All experiments use the public INSPECT dataset [8]. Findings are from a single-centre cohort and require multi-centre validation before generalisation.

References

[1] Laleh Seyyed-Kalantari et al. CheXclusion: Fairness gaps in deep chest X-ray classifiers. *Pacific Symposium on Biocomputing*, 26:232–243, 2021.

[2] Luke Oakden-Rayner et al. Hidden stratification causes clinically meaningful failures in machine learning for medical imaging. In *ACM Conference on Health, Inference, and Learning (CHIL)*, pages

151–159, 2020.

[3] Katherine Walter. What is pulmonary embolism? *JAMA*, 329(1):104, 2023.

[4] Mayo Clinic Staff. Pulmonary embolism. `https://www.mayoclinic.org/diseases-conditions/pulmonary-embolism/`, 2022. Accessed: 2025.

[5] Noel C. F. Codella et al. MedImageInsight: An open-source embedding model for general domain medical imaging. arXiv:2410.06542, 2024.

[6] Health AI Developer Foundations. MedSigLIP. `https://developers.google.com/health-ai-developer-foundations/medsiglip`, 2025.

[7] Sheng Zhang et al. BiomedCLIP: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. In *International Conference on Learning Representations (ICLR)*, 2024. arXiv:2303.00915.

[8] Shih-Cheng Huang, Zepeng Huo, Ethan Steinecke, Shankari Kalanithi, Louis Blankemeier, Yuhui Zhang, Christian Bluethgen, Sendhil Mullainathan, and Pranav Rajpurkar. INSPECT: A multimodal dataset for pulmonary embolism diagnosis and prognosis. arXiv:2311.10798, 2023.