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

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

Abstract

We conduct the first systematic fairness audit of three medical imaging foundation models (MedImageInsight, MedSigLIP, and BiomedCLIP) on INSPECT, a structured benchmark pairing CTPA imaging with longitudinal EHR for pulmonary embolism (PE) diagnosis and seven prognostic tasks. All three frozen encoders fall below the CT-LRCN baseline on PE diagnosis (AUROC 0.680–0.684 vs. 0.721). Our primary finding is that age is the dominant and previously unreported disparity dimension on INSPECT: patients aged 18–40 have underdiagnosis rates (UDR) of 0.63–0.80 versus 0.31–0.41 for ages 75–90, with MedSigLIP and BiomedCLIP reaching near-chance AUROC (0.508) for younger patients. This gap exceeds race/ethnicity and gender disparities and persists across all eight tasks. Age-targeted adversarial debiasing, the only strategy that reduces gaps without substantially hurting AUROC, cuts MedImageInsight's age gap by 79% ($0.333 \rightarrow 0.069$; $p<0.001$) at only 0.011 AUROC cost, establishing a practical mitigation path for high-capacity encoders.

1. Introduction

Pulmonary embolism (PE) is a life-threatening condition caused by pulmonary artery obstruction. Prompt diagnosis is critical: early detection significantly improves recovery, while delays worsen complications (1; 2). CT pulmonary angiography (CTPA) is the clinical gold standard (3), and a negative CTPA carries strong negative predictive value in patients with suspected PE (5). Machine learning applied to CT imaging has demonstrated success in automated PE detection (4). Generally, imaging alone provides limited insight into physiological impact, underlying risk factors, and disease progression, leaving fairness behavior in PE imaging poorly understood (2). Trust and reliability in clinical AI require identifying and mitigating demographic disparities before models reach patients. The INSPECT dataset (6) is a paradigmatic multimodal benchmark: 23,248 CT pulmonary angiography (CTPA) studies linked to structured longitudinal EHR for PE diagnosis and prognosis across eight prediction tasks. Missed PE diagnoses directly worsen outcomes (1; 2), making fairness a patient-safety concern. Foundation models produce reusable structured feature representations for downstream tabular classifiers, reducing the need for task-specific retraining. Yet aggregate performance can mask clinically meaningful disparities (17; 18). We present the first systematic fairness audit of MedImageInsight, MedSigLIP, and BiomedCLIP on a structured clinical benchmark. We benchmark frozen encoder performance on PE diagnostic and seven prognostic tasks using linear and MLP probes on multimodal clinical data, providing the first systematic comparison of these three encoders on INSPECT. We then conduct a comprehensive fairness analysis across race/ethnicity, gender, and age using permutation-test significance. Finally, we evaluate adversarial debiasing across all three protected attributes per model, establishing age-targeted debiasing as the most effective and statistically reliable mitigation for high-capacity encoders, while demonstrating cross-dimensional fairness improvements from single-attribute targeting.

2. Related Work

CTPA-based PE diagnosis and risk stratification jointly determine patient management pathways (9). The INSPECT dataset (6) links 3D CTPA imaging with longitudinal EHR for PE diagnosis and prognosis; its task-specific CT-LRCN baselines motivate studying foundation model embeddings as more transferable alternatives. Recent PE work includes Abn-BLIP for diagnosis and report generation (10) and agent-based CTPA frameworks (11). Multimodal learning in healthcare combines complementary signals, though fusion does not universally improve performance (12; 13). Foundation models have also been extended to continual-learning

Accepted at the Workshop on Structured Data for Health at ICML 2026. [1]Algoverse AI Research. Correspondence to: Kevin Zhu <kevin@algoverseairesearch.org>.

Proceedings of the 43rd International Conference on Machine Learning, Seoul, South Korea. PMLR 306, 2026. Copyright 2026 by the author(s).

settings to limit catastrophic forgetting across sequential medical imaging tasks (7), and applied beyond radiology to tissue classification in mass spectrometry imaging (8), underscoring the breadth of this modeling paradigm across medical modalities.

We use three frozen encoders. MedImageInsight is pretrained across modalities including CT (14); MedSigLIP aligns medical images and text (15); BiomedCLIP is trained on 15M biomedical figure-caption pairs using a ViT-B/16 encoder (16). Prior work establishes that strong average performance can mask important subgroup disparities (17; 18), motivating comprehensive per-demographic analysis. MedImageInsight's own technical report includes fairness findings across age and gender, reinforcing that subgroup audits are expected when benchmarking medical imaging encoders (14).

3. Methods

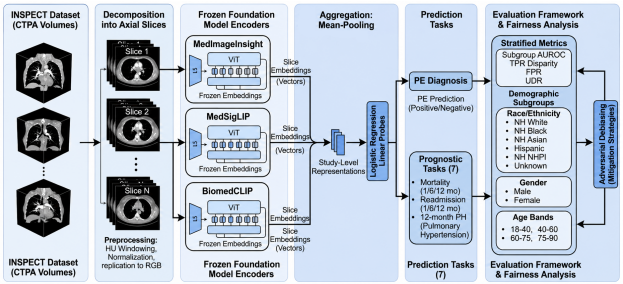

Figure 1. Pipeline overview. CTPA volumes are decomposed into axial slices and encoded with a frozen foundation model. Slice embeddings are mean-pooled into study-level representations and passed to a probe for PE prediction. A fairness evaluation reports per-subgroup AUROC, TPR disparity, FPR, and UDR.

3.1. Dataset

INSPECT (6) contains 23,248 CTPA studies from 19,402 unique patients with structured EHR: diagnostic codes, lab values, and longitudinal clinical records, as well as curated PE diagnostic and prognostic labels. We use the provided patient-level split (80%/5%/15%). Younger PE patients are rarer in clinical practice, may present with atypical clot burden, and are correspondingly underrepresented in INSPECT's training distribution, a data-distributional factor that plausibly contributes to the age disparities reported below, independently of any encoder limitation; we discuss how strongly our evidence supports this claim in Section 5.1.

3.2. Frozen Encoders

All models are frozen feature extractors without fine-tuning. CTPA scans are decomposed into axial slices, clipped to a lung-relevant Hounsfield Unit window, and normalized to 8-bit. Slice embeddings are mean-pooled to produce fixed-length study-level vectors, and mean pooling is applied uniformly across all models so that pooling design does not confound encoder comparison. This dilutes the focal filling-defect signal required for PE diagnosis, which can produce a diagnostic gap relative to the end-to-end CT-LRCN baseline. Prognostic signals (mortality risk, parenchymal changes) are more diffusely distributed and therefore less harmed by pooling. We did not evaluate alternative pooling strategies (e.g., attention-pooling, max-pooling) against this mean-pooling design; consequently, part of the diagnostic gap relative to CT-LRCN, and part of the fairness pattern reported below, may reflect this aggregation choice rather than the encoders themselves (Section 5.1). MedImageInsight (14) yields 1,024-d vectors (632M params; ViT-H/14); MedSigLIP (15) yields 1,152-d vectors (400M; SigLIP ViT-So/14); BiomedCLIP (16) yields 512-d vectors (86M; ViT-B/16), 4–7× smaller. We evaluate logistic regression (LR) as a linear probe and multi-layer perceptrons (MLPs) on frozen embeddings; full MLP results appear in Appendix A. Hyperparameters are selected by grid search on validation AUROC over $C \in \{0.001, 0.01, 0.1, 1.0, 10.0\}$ for LR and hidden-size configurations $(256, 128)$, $(512, 256)$, $(512, 256, 128)$ for MLPs.

3.3. Probes and CT/EHR Ablation

AUROC with 95% CIs via 1,000 bootstrap samples; AUC differences via the DeLong test (19); TPR disparity significance via 1,000 permutation samples. Adversarial debiasing significance is evaluated via: $p_{\text{gap}} = P(\text{bootstrapped baseline gap} \leq \text{post-debiasing gap})$. We also evaluate CT-only, EHR-only, and CT+EHR fusion probes (LR) across all eight tasks; full results appear in Appendix B. EHR alone is substantially weaker than CT alone on every task, and fusion gains are marginal or negative for high-capacity encoders ($\Delta \leq +0.003$ for MedImageInsight and MedSigLIP on PE diagnosis), confirming that the performance bottleneck lies in spatial aggregation of CT signal rather than absence of structured EHR features.

3.4. Evaluation and Fairness

Per-subgroup AUROC, TPR, FPR, and UDR ($= 1 - \text{TPR}$). TPR disparity $= \text{TPR}_{group} - \text{TPR}_{reference}$ (NH White reference for race/ethnicity; Male for gender). NH NHPI (n=60) yields bootstrap CIs spanning $[0, 1]$; rows appear in tables marked † but are excluded from all fairness conclusions.

3.5. Bias Mitigation

(1) Importance weighting (IW): reweight by inverse group and positive class frequency. (2) Group resampling: equalize group sizes. (3) Adversarial debiasing: a gradient reversal layer with strength $\alpha \in \{0.5, 1.0, 2.0\}$ selected per

protected attribute on the validation set. Reported gap reductions should be treated as upper bounds until replicated on held-out cohorts.

4. Results

4.1. Diagnostic and Prognostic Performance

Table 1. PE diagnostic AUROC. MedImageInsight and MedSigLIP are statistically indistinguishable (DeLong p=0.952); all three are significantly below the INSPECT CT-LRCN baseline (MII: p=0.009, MSig: p=0.008, BCL: p<0.001). LR outperforms all MLP variants (Appendix A).

Model	AUROC	95% CI	Params
CT-LRCN baseline	**0.721**	(0.69, 0.75)	271M
MedImageInsight	0.680	(0.655, 0.706)	632M
MedSigLIP	**0.684**	(0.660, 0.708)	400M
BiomedCLIP	0.626	(0.599, 0.655)	86M

MedImageInsight and MedSigLIP perform comparably (p=0.952) but are significantly below the task-specific baseline (AUROC 0.721, $p\leq0.009$). This ≈4-point gap carries meaningful clinical weight in this life-threatening condition. As noted in Methods, the diagnostic underperformance relative to CT-LRCN is consistent with mean pooling diluting focal PE signal, while the same pooling approach appears less harmful for prognostically predictive signals spread across the volume, which may help explain the strong mortality AUROC (0.83–0.90) across all three models. We have not isolated this mechanism with a pooling ablation (Section 5.1). Full prognostic results appear in Appendix C.

4.2. Fairness Analysis

Table 2. UDR by demographic subgroup. $^{**}p<0.01$, $^{*}p<0.05$, $^{a}p>0.10$ (permutation test, 1,000 samples). †NH NHPI excluded from gap computation. Our main finding is that age yields the largest UDR gaps (0.32–0.45), exceeding race/ethnicity (0.21–0.29) and gender (0.08–0.16).

Cat.	Subgroup	n	MII UDR	MSig UDR	BCL UDR
Race	NH White (ref)	1,608	0.474	0.459	0.378
	NH Black	189	0.419	**0.613**	0.387
	NH Asian	555	**0.566**	0.553	**0.513**
	Hispanic^{a}	497	0.552	0.552	0.478
	Min–max gap		**0.233**	**0.207**	**0.288**
Gender	Male (ref)	1,369	0.430	0.449	0.325
	Female**	1,841	**0.556**	**0.531**	**0.486**
	Min–max gap		**0.126**	**0.082**	**0.161**
Age	18–40	359	**0.743**	**0.800**	**0.629**
	40–60	879	0.521	0.556	0.486
	60–75	1,093	0.548	0.538	0.472
	75–90	689	0.410	0.347	0.313
	Min–max gap		**0.333**	**0.453**	**0.316**

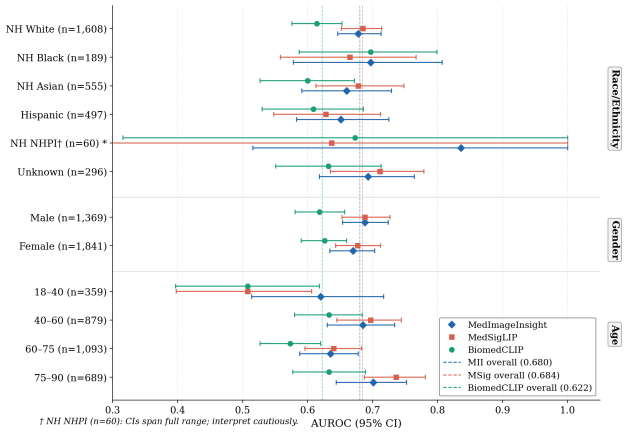

Figure 2. Diagnostic AUROC by demographic subgroup. The 18–40 age group shows near-chance performance for MedSigLIP and BiomedCLIP (AUROC 0.508), a clinical failure mode for younger patients. †NH NHPI (n=60): CIs span the full range; excluded from interpretation.

The 18–40 age group is the primary health-equity concern: MedSigLIP and BiomedCLIP reach near-chance AUROC (0.508, Figure 3), and even MedImageInsight achieves only 0.620. The pattern persists in the task-specific CT-LRCN baseline (18–40 UDR = 0.571; Appendix I), which is consistent with the disparity partly reflecting the underlying clinical data distribution rather than an encoder-specific failure mode. However, our design does not fully isolate dataset effects from representation or modeling effects, and we treat this as a plausible contributing factor rather than an established cause (Section 5.1).

Female TPR disparity is statistically significant across all three models ($p\leq0.040$). Hispanic disparity does not reach significance (p=0.11–0.21). Age-band pairwise tests confirm 18–40 vs. 75–90 is highly significant (p<0.001) across all models, and age-band gaps persist across all seven prognostic tasks (0.04–0.20; Appendix J), confirming age as the dominant disparity dimension throughout the benchmark.

4.3. Bias Mitigation

Among all mitigation strategies evaluated, adversarial debiasing is the only approach that reliably reduces demographic gaps without substantially hurting AUROC, as importance weighting and group resampling either widen gaps or collapse task performance. For MedImageInsight, age-targeted adversarial debiasing reduces the age UDR gap by 79% (0.333→0.069; p<0.001) at only 0.011 AUROC cost. Race-targeted debiasing (α=2.0) yields simultaneous reductions of 50%, 44%, and 40% in race, gender, and age gaps at 0.007 AUROC cost (p=0.061), an empirical pattern consistent with partially shared demographic latents. For MedSigLIP, gender-targeted debiasing eliminates the gender gap entirely (p<0.001) and age-targeted yields a 44% reduction

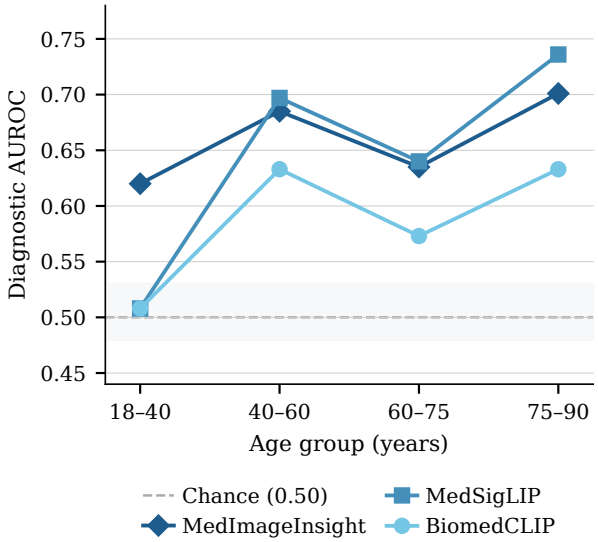

Figure 3. Diagnostic AUROC by age group. MedSigLIP and BiomedCLIP reach near-chance performance (0.508) for patients aged 18–40, while all three models improve monotonically toward the 75–90 group. The dashed line marks chance level (0.50).

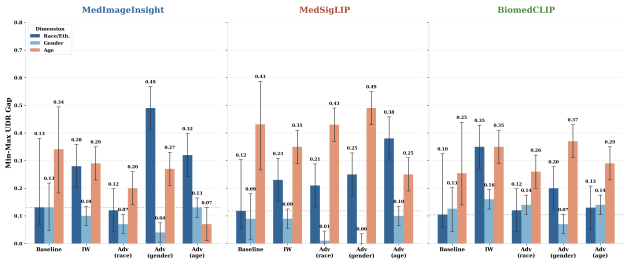

Figure 4. Min–max UDR gap by mitigation strategy and demographic dimension. Age-targeted adversarial debiasing (α=0.5, p<0.001) achieves the single largest reduction for MedImageInsight (0.333→0.069) with minimal AUROC cost, while group resampling collapses overall AUROC to ≈0.56.

(p=0.007). BiomedCLIP shows numerically meaningful reductions (up to 60% for race-targeted) but none reach statistical significance (p≥0.12). Thus, mitigation should be considered unreliable for this model. Demographic probing (Appendix K) shows that BiomedCLIP's 512-d space encodes less separable demographic latents (race/ethnicity macro-F1 0.537 vs. 0.677–0.707 for the larger models). This supports an embedding-capacity hypothesis that gradient reversal requires separable demographic structure, and BiomedCLIP's compressed representation offers insufficient separation for reliable disentanglement.

5. Discussion

Frozen medical imaging foundation models applied to a structured clinical benchmark encode age-based disparities that dominate race and gender gaps. UDR of 0.63–0.80 for patients aged 18–40 versus 0.31–0.41 for 75–90 persists across all eight tasks and in the CT-LRCN baseline, consistent with data distribution contributing alongside, rather than instead of, encoder design as a root cause (see Limitations below). Targeted data augmentation or synthetic oversampling for younger patients may be necessary alongside debiasing. End-to-end training learns more gender-balanced thresholds (CT-LRCN Female TPR = 0.592 > Male TPR = 0.558) while frozen encoders systematically disadvantage female patients at all thresholds, suggesting that the frozen-encoder paradigm may introduce fairness costs invisible to aggregate evaluation.

Table 3. Min–max UDR gaps across mitigation strategies. Adversarial debiasing is the only strategy that consistently reduces gaps without substantially harming AUROC; IW widens race gaps and group resampling collapses AUROC to ≈0.55–0.56. Significance (one-sided bootstrap): $^{***}p$<0.001, $^{**}p$<0.01, $^{*}p$<0.05, $^{\wedge}p$<0.10, ^{ns}p≥0.10.

Model	Strategy	AUC	Race	Gender	Age
MII	Baseline	0.680	0.233	0.126	0.333
	Adv. race α=2.0$^{\wedge}$	0.673	**0.117**	**0.071**	**0.199**
	Adv. gender α=0.5*	**0.680**	0.493	0.044	0.268
	Adv. age α=0.5***	0.669	0.316	0.133	**0.069**
MSig	Baseline	0.684	0.207	0.082	0.453
	Adv. race α=0.5^{ns}	0.666	0.215	0.015	0.434
	Adv. gender α=0.5***	0.665	0.247	**0.000**	0.490
	Adv. age α=0.5**	0.663	0.376	0.105	**0.253**
BCL	Baseline	0.626	0.288	0.161	0.316
	Best adv.^{ns} (all attrs.)	0.618–0.632	No significant reduction (p≥0.12)		

5.1. Limitations

Several factors temper the strength of our conclusions. First, all encoder comparisons use a single frozen slice-wise mean-pooling pipeline, and we did not run an ablation over alternative pooling strategies (e.g., attention-pooling, max-pooling). Part of the diagnostic gap relative to CT-LRCN, and part of the fairness pattern itself, may reflect this aggregation choice rather than the encoders alone. The persistence of the age disparity in the CT-LRCN baseline is consistent with a data-distributional explanation, but we have not directly tested it – e.g., via subgroup-stratified label prevalence, label-quality analysis, or stratified error analysis for younger patients – so we treat it as a plausible contributing factor rather than an established cause. This also means the root cause of the age disparity may be only partly attributable to the encoders themselves, which qualifies our framing of this work as a fairness audit of foundation models. On generalizability, all results come from a single benchmark and institution and require multi-center validation before they should be assumed to generalize. Adversarial debiasing was statistically reliable for MedImageInsight and MedSigLIP but not for BiomedCLIP; reported gap reductions are upper bounds pending replication on held-out cohorts. NH NHPI (n=60)

and other small strata yield unstable bootstrap CIs and are excluded from fairness conclusions throughout, which weakens the strength of conclusions for those groups. Attention-based pooling and lightweight fine-tuning (LoRA, adapter layers) remain natural avenues for isolating the pipeline confound, though the age gap persisting in CT-LRCN suggests pooling improvements alone would not fully resolve the disparity.

Age is thus a required fairness axis. The severity of the 18–40 failure mode, with near-chance AUROC for two of three models, across both diagnostic and all prognostic tasks, means that omitting age from subgroup evaluation is a patient-safety gap. Per-group reporting across race, gender, and age should be a minimum standard for responsible clinical AI benchmarking on multimodal structured data.

6. Conclusion

We present a systematic benchmark of MedImageInsight, MedSigLIP, and BiomedCLIP as frozen feature extractors for PE diagnostic and prognostic tasks. All three models exhibit notable demographic biases and age emerged as the dominant disparity dimension, with underdiagnosis rates in younger cohorts (18–40) significantly exceeding those of older patients across all models and tasks. We demonstrate that age-targeted adversarial debiasing is a highly effective mitigation, yielding a 79% age-gap reduction for MedImageInsight and a 44% reduction for MedSigLIP. Furthermore, we find evidence of cross-dimensional fairness improvement, where targeting a single attribute (such as race) simultaneously mitigates untargeted disparities. However, the lack of statistically significant debiasing in BiomedCLIP suggests that model capacity is a prerequisite for effective demographic disentanglement. As discussed in Section 5.1, the frozen mean-pooling pipeline and the single-institution setting mean these findings should be read as an audit of this particular pipeline on this benchmark rather than a fully isolated statement about the encoders themselves; pooling ablations and multi-center replication are needed before stronger causal claims are warranted. Nonetheless, these results highlight the clinical necessity of per-group reporting across all demographic dimensions (including age) to ensure responsible AI benchmarking and deployment.

Impact Statement

This paper evaluates the fairness of medical imaging foundation models on a structured clinical benchmark. Younger patients (18–40) face substantially higher underdiagnosis rates including near-chance model performance, with direct implications for health equity and clinical AI trustworthiness. We identify age-targeted adversarial debiasing as the most reliable mitigation for high-capacity models, and caution against deploying smaller encoders for fairness-critical clinical tasks without further validation. All experiments use the public INSPECT dataset. Fairness findings are institution-specific and should not be generalized without multi-center validation.

References

[1] Mayo Clinic Staff. Pulmonary embolism (2022). https://www.mayoclinic.org/diseases-conditions/pulmonary-embolism/symptoms-causes/syc-20354647

[2] Walter K. What Is Pulmonary Embolism? JAMA. 2023;329(1):104.

[3] Konstantinides SV, et al. 2019 ESC Guidelines for acute pulmonary embolism. Eur Heart J. 2020;41(4):543–603.

[4] Weikert T, et al. Automated detection of PE in CT pulmonary angiograms. Eur Radiol. 2020;30(12):6545–6553.

[5] Quiroz R, et al. Clinical Validity of a Negative CT Scan in Suspected PE. JAMA. 2005;293(16):2012–2017.

[6] Huang SC, et al. INSPECT: A Multimodal Dataset for PE Diagnosis and Prognosis. arXiv:2311.10798 (2023).

[7] Qazi MA, et al. UNICON: UNIfied CONtinual Learning for Medical Foundational Models. arXiv:2508.14024 (2025).

[8] Gabriel A, et al. Application of Foundation Models for Colorectal Cancer Tissue Classification in Mass Spectrometry Imaging. Technologies (2025).

[9] Burns SK, Haramati LB. Diagnostic imaging and risk stratification of patients with acute PE. Cardiol Rev. 2012;20(1):15–24.

[10] Zhong Z, et al. Abn-BLIP: Abnormality-aligned Bootstrapping Language-Image Pre-training for PE Diagnosis. arXiv:2503.02034 (2025).

[11] Zhong Z, et al. Vision-language model for report generation and outcome prediction in CTPA. npj Digit Med. 2025;8:432.

[12] Krones F, et al. Review of multimodal machine learning approaches in healthcare. Inf Fusion. 2025;114:102690.

[13] Wu J, et al. Exploring Multimodal LLMs for Radiology Report Error-checking. arXiv:2312.13103 (2024).

[14] Codella NCF, et al. MedImageInsight: An Open-Source Embedding Model for General Domain Medical Imaging. arXiv:2410.06542 (2024).

[15] Health AI Developer Foundations. MedSigLIP (2025). https://developers.google.com/health-ai-developer-foundations/medsiglip

[16] Zhang S, et al. BiomedCLIP: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. ICLR 2024. arXiv:2303.00915.

[17] Seyyed-Kalantari L, et al. CheXclusion: Fairness gaps in deep chest X-ray classifiers. arXiv:2003.00827 (2020).

[18] Oakden-Rayner L, et al. Hidden Stratification Causes Clinically Meaningful Failures in Machine Learning for Medical Imaging. ACM CHIL. 2020:151–159.

[19] DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics. 1988;44(3):837–845.

Appendices

A. MLP Probe Results

Table 4. Diagnostic AUROC for LR vs. best MLP probe. LR outperforms all MLP variants for MedImageInsight (+0.014) and MedSigLIP (+0.021), consistent with approximately linearly separable embeddings. BiomedCLIP MLP-small exceeds LR by 0.010. All MLPs: Adam, batch norm, dropout 0.2, early stopping (patience = 20), max 300 epochs.

Model	Probe	Hidden Sizes	AUROC	95% CI
CT-LRCN baseline	–	–	**0.721**	(0.69, 0.75)
MedImageInsight	**LR**	–	**0.680**	(0.655, 0.706)
MedImageInsight	MLP-best	(512,256,128)	0.662	(0.639, 0.687)
MedSigLIP	**LR**	–	**0.684**	(0.660, 0.708)
MedSigLIP	MLP-best	(512,256)	0.656	(0.634, 0.681)
BiomedCLIP	LR	–	0.626	(0.599, 0.655)
BiomedCLIP	**MLP-best**	(256,128)	**0.634**	(0.611, 0.661)

B. CT-Only vs. EHR-Only vs. CT+EHR Ablation

Table 5. CT-only vs. EHR-only vs. CT+EHR ablation (LR probe, test AUROC). EHR alone is weaker than CT alone on every task. Fusion gains are marginal or negative for high-capacity encoders; the largest gains occur for BiomedCLIP readmission (+0.013–+0.017). **Bold** marks best per row. Δ = Fusion − CT.

Model	Task	CT	EHR	Fusion	Δ
MedSigLIP	PE diagnosis	0.684	0.585	**0.687**	+0.003
	1-mo mortality	**0.881**	0.601	0.879	−0.002
	6-mo mortality	**0.844**	0.593	**0.844**	±0.000
	12-mo mortality	**0.825**	0.595	**0.825**	+0.001
	1-mo readmission	**0.616**	0.570	0.611	−0.005
	6-mo readmission	0.629	0.573	**0.633**	+0.004
	12-mo readmission	0.655	0.580	**0.660**	+0.005
BiomedCLIP	PE diagnosis	0.624	0.585	**0.634**	+0.011
	1-mo mortality	**0.865**	0.601	**0.865**	−0.000
	6-mo mortality	**0.820**	0.593	**0.821**	+0.000
	12-mo mortality	0.807	0.595	**0.809**	+0.002
	1-mo readmission	0.581	0.570	**0.594**	+0.013
	6-mo readmission	0.598	0.573	**0.614**	+0.017
	12-mo readmission	0.606	0.580	**0.622**	+0.016
MedImageInsight	PE diagnosis	0.680	0.585	**0.682**	+0.002
	1-mo mortality	**0.891**	0.601	**0.891**	−0.000
	6-mo mortality	**0.846**	0.593	**0.847**	+0.001
	12-mo mortality	**0.830**	0.595	**0.831**	+0.001
	1-mo readmission	**0.622**	0.570	0.616	−0.006
	6-mo readmission	0.632	0.573	**0.639**	+0.007
	12-mo readmission	0.632	0.580	**0.641**	+0.009

C. Full Prognostic Performance

Table 6. Diagnostic and prognostic AUROC and ECE. All three models show strong mortality prediction; BCL is competitive on prognostic tasks despite its 4–7× smaller encoder. MII vs. MSig differences are not significant on any prognostic task (p=0.17–0.80). **Bold** marks best AUROC per task among foundation models.

Task	MII AUROC (95% CI)	MSig AUROC (95% CI)	BCL AUROC (95% CI)	MII ECE	MSig ECE	BCL ECE
PE Diagnosis	0.680 (0.655–0.706)	**0.684 (0.660–0.708)**	0.626 (0.599–0.655)	0.044	0.041	**0.035**
1-month mortality	**0.898 (0.875–0.917)**	0.890 (0.863–0.908)	0.869 (0.842–0.894)	0.014	0.016	**0.013**
6-month mortality	**0.856 (0.835–0.875)**	0.851 (0.832–0.870)	0.835 (0.812–0.856)	0.019	**0.014**	0.020
12-month mortality	**0.847 (0.827–0.865)**	0.844 (0.823–0.865)	0.825 (0.805–0.845)	**0.014**	**0.014**	0.018
1-month readmission	**0.616 (0.563–0.663)**	0.585 (0.539–0.637)	0.580 (0.524–0.630)	0.023	0.023	**0.006**
6-month readmission	**0.645 (0.614–0.677)**	0.622 (0.587–0.654)	0.604 (0.570–0.639)	**0.011**	0.050	0.018
12-month readmission	**0.651 (0.623–0.678)**	0.649 (0.622–0.678)	0.618 (0.590–0.646)	**0.007**	0.027	0.009
12-month PH	**0.755 (0.728–0.781)**	0.753 (0.726–0.778)	0.711 (0.683–0.739)	0.017	**0.014**	0.013

D. Full TPR, FPR, and UDR Tables

Table 7. Full per-subgroup TPR, FPR, and UDR with 95% bootstrap CIs for MedImageInsight (LR, threshold = 0.235). †NH NHPI (n=60): CIs span the full range; excluded from fairness conclusions. **Bold** marks the worst (lowest TPR / highest UDR) within each demographic category.

Category	Subgroup	n	TPR (95% CI)	FPR (95% CI)	UDR
Race/Eth.	NH White (ref)	1,608	0.526 (0.473–0.582)	0.261 (0.237–0.287)	0.474
	NH Black	189	0.581 (0.414–0.750)	0.184 (0.126–0.247)	0.419
	NH Asian	555	**0.434 (0.325–0.547)**	0.203 (0.171–0.239)	**0.566**
	Hispanic	497	0.448 (0.338–0.571)	0.167 (0.131–0.203)	0.552
	NH NHPI†	60	0.667 (0.000–1.000)	**0.123 (0.036–0.214)**	0.333
	Unknown	296	0.476 (0.358–0.596)	0.236 (0.185–0.291)	0.524
Gender	Male (ref)	1,369	0.570 (0.511–0.632)	0.273 (0.246–0.298)	0.430
	Female	1,841	**0.444 (0.388–0.495)**	**0.192 (0.173–0.212)**	**0.556**
Age	18–40	359	**0.257 (0.114–0.419)**	**0.142 (0.104–0.182)**	**0.743**
	40–60	879	0.479 (0.393–0.563)	0.181 (0.156–0.208)	0.521
	60–75	1,093	0.452 (0.384–0.522)	0.251 (0.223–0.283)	0.548
	75–90	689	0.590 (0.508–0.673)	0.270 (0.230–0.308)	0.410

Table 8. Full per-subgroup TPR, FPR, and UDR for MedSigLIP (threshold = 0.240) and BiomedCLIP (threshold = 0.192). †NH NHPI: CIs span full range; excluded from fairness conclusions. **Bold** marks the worst (lowest TPR / highest UDR) within each demographic category.

Category	Subgroup	n	MSig TPR (95% CI)	MSig UDR	BCL TPR (95% CI)	BCL UDR
Race/Eth.	NH White (ref)	1,608	0.541 (0.489–0.589)	0.459	0.622 (0.566–0.672)	0.378
	NH Black	189	**0.387 (0.226–0.556)**	**0.613**	0.613 (0.446–0.791)	0.387
	NH Asian	555	0.447 (0.342–0.561)	0.553	**0.487 (0.372–0.607)**	**0.513**
	Hispanic	497	0.448 (0.329–0.578)	0.552	0.522 (0.406–0.635)	0.478
	NH NHPI†	60	0.333 (0.000–1.000)	0.667	0.333 (0.000–1.000)	0.667
	Unknown	296	0.540 (0.412–0.648)	0.460	0.619 (0.509–0.729)	0.381
Gender	Male (ref)	1,369	0.551 (0.494–0.613)	0.449	0.676 (0.621–0.733)	0.325
	Female	1,841	**0.469 (0.412–0.525)**	**0.531**	**0.515 (0.462–0.571)**	**0.486**
Age	18–40	359	**0.200 (0.075–0.344)**	**0.800**	**0.371 (0.203–0.543)**	**0.629**
	40–60	879	0.444 (0.362–0.528)	0.556	0.514 (0.438–0.602)	0.486
	60–75	1,093	0.462 (0.398–0.532)	0.538	0.528 (0.461–0.598)	0.472
	75–90	689	0.653 (0.570–0.730)	0.347	0.688 (0.605–0.772)	0.313

E. Threshold Analysis

Table 9. Gender-targeted threshold analysis. Setting the global threshold to achieve 80% overall TPR improves female TPR but leaves residual disparities for all three foundation models, confirming representation-level gaps independent of threshold choice. CT-LRCN reverses the disparity direction.

Model	Threshold	Female TPR	Male TPR	Female Disp
MedImageInsight	0.103	0.752	0.857	−0.104
MedSigLIP	0.112	0.762	0.845	**−0.083**
BiomedCLIP	0.137	0.736	0.875	−0.139
CT-LRCN	0.090	**0.817**	0.781	+0.036

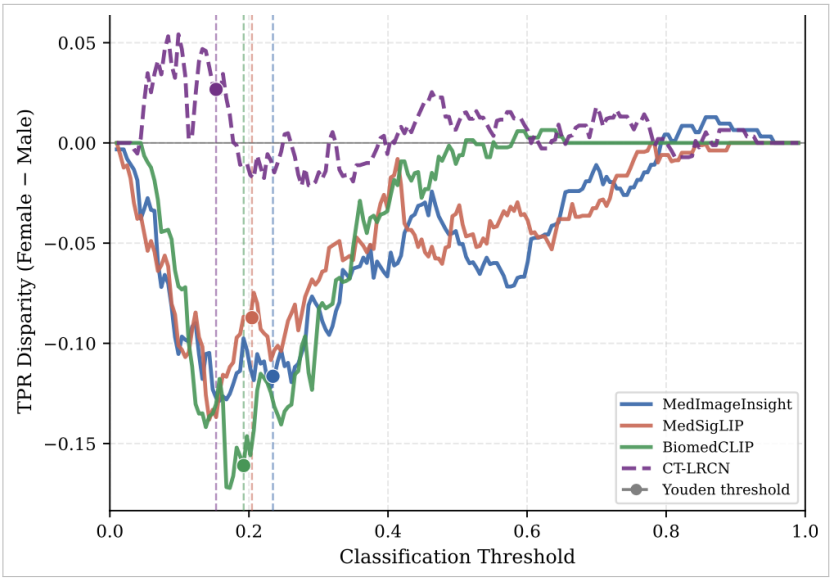

Figure 5. TPR disparity vs. classification threshold for Female vs. Male. Dashed vertical lines mark the global Youden threshold per model. The female disparity remains near-constant across all thresholds, confirming a representation-level gap **independent of threshold choice.**

F. Importance Weighting and Group Resampling

Importance weighting widened race/ethnicity UDR gaps for all three models (MII: $0.233 \rightarrow 0.279$; MSig: $0.207 \rightarrow 0.233$; BCL: $0.288 \rightarrow 0.351$) and produced opposing effects on Hispanic disparity across models. Group resampling collapsed AUROC to $\approx$0.55–0.56 while dramatically widening race gaps for MedImageInsight (0.233$\rightarrow$0.434) and MedSigLIP (0.207$\rightarrow$0.582). Neither IW nor group resampling is recommended as a mitigation strategy for clinical deployment.

G. Stability Analysis

Table 10. AUROC stability across 10 bootstrap resamples of the training set. Std $<$ 0.01 for all three models.

Model	Mean AUROC	Std AUROC
MedImageInsight	**0.666**	**0.005**
MedSigLIP	**0.666**	0.007
BiomedCLIP	0.618	0.006

H. Calibration (ECE)

Table 11. Expected Calibration Error (ECE, 10 equal-width bins) with 95% bootstrap CIs. All models show modest miscalibration. **Bold** marks the best (lowest) ECE.

Model	ECE	95% CI
MedImageInsight (LR)	0.044	(0.035, 0.058)
MedSigLIP (LR)	0.041	(0.032, 0.055)
BiomedCLIP (LR)	**0.035**	(0.024, 0.047)

I. CT-LRCN Baseline Fairness (PE Diagnosis)

Table 12. Per-subgroup TPR, FPR, and UDR for the INSPECT CT-LRCN baseline (Youden threshold; AUROC = 0.720). Unlike all three foundation models, the baseline shows Female TPR = 0.592 > Male TPR = 0.558, suggesting end-to-end task-specific training learns more gender-balanced thresholds. †NH NHPI (n=60): CIs span full range; excluded from fairness conclusions. **Bold** marks the worst.

Category	Subgroup	n	TPR (95% CI)	FPR (95% CI)	UDR
Race/Eth.	NH White (ref)	1,608	0.619 (0.566–0.668)	0.253 (0.229–0.276)	0.381
	NH Black	189	0.677 (0.500–0.833)	0.209 (0.148–0.272)	0.323
	NH Asian	555	0.539 (0.422–0.648)	0.265 (0.227–0.307)	0.461
	Hispanic	497	**0.403 (0.288–0.523)**	0.214 (0.174–0.254)	**0.597**
	NH NHPI†	60	0.667 (0.000–1.000)	0.140 (0.054–0.246)	0.333
	Unknown	296	0.508 (0.396–0.627)	0.223 (0.172–0.276)	0.492
Gender	Male (ref)	1,369	0.558 (0.496–0.619)	0.235 (0.212–0.261)	0.442
	Female	1,841	**0.592 (0.537–0.644)**	0.244 (0.223–0.265)	**0.408**
Age	18–40	359	**0.429 (0.259–0.600)**	0.182 (0.142–0.226)	**0.571**
	40–60	879	0.535 (0.450–0.615)	0.196 (0.168–0.224)	0.465
	60–75	1,093	0.578 (0.513–0.651)	0.243 (0.213–0.272)	0.422
	75–90	689	0.597 (0.520–0.674)	0.295 (0.258–0.334)	0.403

J. Prognostic Fairness

Table 13. Min–max UDR gaps across all seven prognostic tasks. Age-band gaps are consistently present (0.04–0.20) across all tasks and all models. †CT-LRCN mortality race gaps inflated by NH NHPI (TPR = 1.000, n<60). ‡1-month readmission race gaps inflated by NH NHPI. **Bold** marks the largest gap per task per dimension.

	MedImageInsight			MedSigLIP			BiomedCLIP			CT-LRCN		
Task	R/E	Gen	Age	R/E	Gen	Age	R/E	Gen	Age	R/E	Gen	Age
1-mo mortality	0.288	0.028	0.099	**0.426**	0.043	0.083	0.246	0.052	0.062	0.258†	0.008	**0.134**
6-mo mortality	0.199	0.029	0.133	0.133	0.004	**0.148**	0.138	0.034	0.133	0.538†	0.026	0.106
12-mo mortality	0.268	0.007	0.093	**0.425**	0.011	0.094	0.155	0.045	**0.125**	0.438†	0.002	0.046
1-mo readmission	0.900‡	0.029	0.165	0.750‡	0.120	**0.192**	0.667‡	0.138	0.125	0.349	0.029	0.144
6-mo readmission	0.357	0.186	0.160	0.339	0.082	0.286	0.259	0.082	**0.356**	0.152	**0.222**	0.143
12-mo readmission	0.318	0.116	0.159	**0.405**	0.155	0.257	0.233	0.156	**0.343**	0.302	0.140	0.042
12-mo PH	0.278	0.005	**0.218**	0.167	0.047	0.174	0.097	0.018	0.103	**0.342**	0.054	0.196

K. Demographic Probing

To determine whether disparities originate in the frozen representations themselves, we trained LR and MLP classifiers on frozen embeddings to predict each demographic attribute. Chance-level macro-F1 is 0.250 for age band and race/ethnicity (four classes each) and 0.500 for gender (two classes).

Table 14. Demographic probing macro-F1 (95% CI, 1,000 bootstrap samples). All values far exceed chance, confirming demographic signal is encoded in frozen representations. BiomedCLIP's markedly lower race/ethnicity F1 (0.537 vs. 0.677–0.707) supports the embedding-capacity hypothesis. Negative MLP−LR gaps indicate linearly organised demographic structure. **Bold** marks the highest F1 per target and probe type.

Probe	Target	MedImageInsight	MedSigLIP	BiomedCLIP
LR	Age band	**0.730** [0.713, 0.745]	0.693 [0.674, 0.709]	0.617 [0.600, 0.635]
	Gender	**0.972** [0.966, 0.977]	0.966 [0.960, 0.972]	0.939 [0.930, 0.947]
	Race/ethnicity	**0.707** [0.685, 0.728]	0.677 [0.653, 0.698]	0.537 [0.513, 0.560]
MLP	Age band	**0.705** [0.686, 0.722]	0.678 [0.662, 0.696]	0.589 [0.573, 0.607]
	Gender	**0.971** [0.965, 0.977]	0.960 [0.954, 0.967]	0.899 [0.889, 0.909]
	Race/ethnicity	**0.679** [0.657, 0.700]	0.640 [0.618, 0.660]	0.512 [0.491, 0.534]
MLP−LR gap	Age band	−0.025	−0.015	−0.028
	Gender	−0.001	−0.006	−0.040
	Race/ethnicity	−0.028	−0.036	−0.025
Chance-level macro-F1		Age band: 0.250	Gender: 0.500	Race/ethnicity: 0.250