Demographic Bias of Expert-Level Vision-Language Foundation Models in Medical Imaging

Yuzhe Yang[1] **Yujia Liu**[2] **Xin Liu**[3] **Avanti Gulhane**[3] **Domenico Mastrodicasa**[3]
Wei Wu[3] **Edward J. Wang**[2] **Dushyant Sahani**[3] **Shwetak Patel**[3]

[1]MIT [2]University of California, San Diego [3]University of Washington

Abstract

Advances in artificial intelligence (AI) have achieved expert-level performance in medical imaging applications. Notably, self-supervised vision-language foundation models can detect a broad spectrum of pathologies without relying on explicit training annotations. However, it is crucial to ensure that these AI models do not mirror or amplify human biases, disadvantaging historically marginalized groups such as females or Black patients. In this study, we investigate the algorithmic fairness of state-of-the-art vision-language foundation models in chest X-ray diagnosis across five globally-sourced datasets. Our findings reveal that compared to board-certified radiologists, these foundation models consistently underdiagnose marginalized groups, with even higher rates seen in intersectional subgroups such as Black female patients. Such biases present over a wide range of pathologies and demographic attributes. Further analysis of the model embedding uncovers its significant encoding of demographic information beyond human levels. Deploying medical AI systems with biases can intensify pre-existing care disparities, posing potential challenges to equitable healthcare access and raising ethical questions about their clinical applications. Code is available at: `github.com/YyzHarry/vlm-fairness`.

1 Introduction

Artificial intelligence (AI) has increasingly been deployed in real-world clinical settings, especially for medical imaging [1, 2, 3, 4]. The latest developments include vision-language foundation models that operate on a self-supervised learning paradigm [5, 6], eliminating the need for explicit pathology annotations while maintaining human-level diagnostic accuracy across various modalities and disease conditions [5, 7, 8, 9]. Notably in radiology, by simultaneously using image and text inputs and leveraging the information naturally present in clinical reports associated with radiology images, foundation models identify pathologies without dependence on specific annotations, achieving performance that matches the expertise of radiologists and, in some cases, surpasses the expected diagnostic benchmarks [10, 11].

Despite the plausible performance in diagnosing unseen pathologies [10], the foundation model could amplify existing biases in the data, causing diagnosis disparities across protected subpopulations and leading to unequal predictive outcomes for specific demographics [12, 13, 14] (e.g., discrepancies in diagnosis rates between Black and White patients). Existing literature has revealed that chest X-ray classifiers trained to predict the presence of disease systematically underdiagnosed Black patients [12, 14], potentially leading to incorrect triage decisions and delayed medical treatment. Although algorithmic biases have been studied in the supervised setting [14, 15] (e.g., models trained for specific diseases like "No Finding"), little attention has been paid to vision-language foundation models. These models, notably free from explicit supervision through multimodal training and zero-shot inference, theoretically have reduced potential to inherit human labeling biases. However, to ensure the responsible and fair deployment, it is essential to investigate potential biases these models may possess, understand the sources and outcomes, and initiate corrective actions [16].

GenAI for Health Workshop @ NeurIPS 2024, Vancouver.

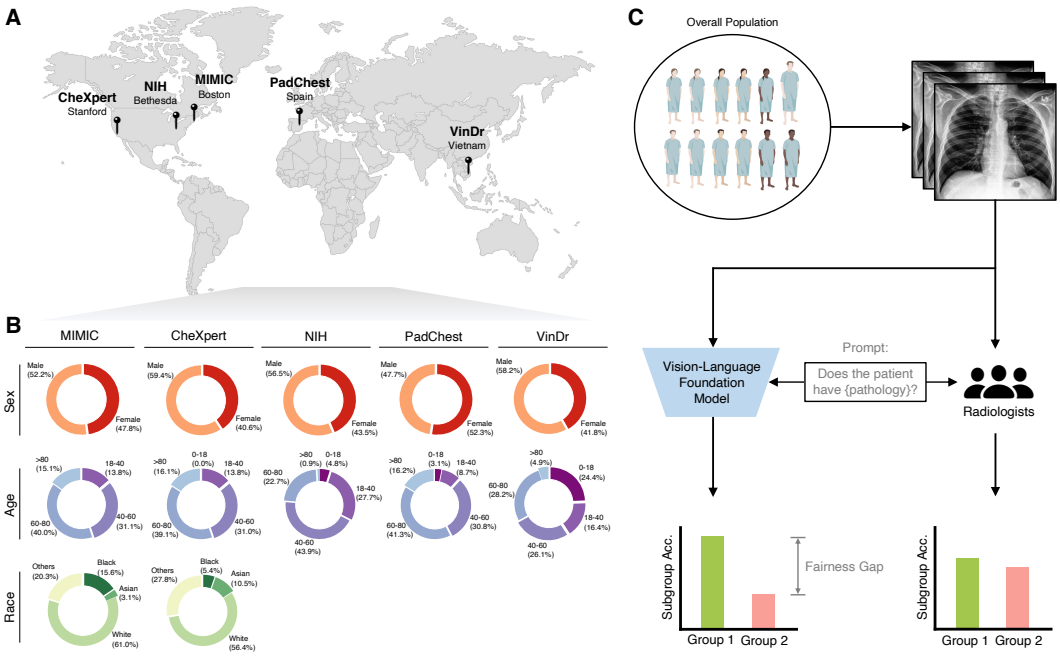

Figure 1: **The model evaluation pipeline. (A)** We use internationally-sourced chest X-rays datasets for model evaluation, including MIMIC (Boston, MA), CheXpert (Stanford, CA), NIH (Bethesda, MD), PadChest (Spain), and VinDr (Vietnam). **(B)** Distribution of demographics attributes (i.e., sex, age, race) of each dataset. For each attribute, we select common subgroups based on literature definition (sex: "male", "female"; age: "0-18", "18-40", "40-60", "60-80", ">80"; race: "Asian", "Black", "White", "Others"). Each dataset encompasses different proportions of subgroups, reflecting the diverse distributions in real-world clinical settings. **(C)** For fairness evaluation, we processed radiographs through foundation models, accompanied by specific text prompts (e.g., "Does the patient have {*pathology*}?"). The evaluations are conducted across a wide range of different pathologies. Concurrently, board-certified radiologists independently reviewed identical subsets of the data, providing diagnoses that served as human fairness evaluations and comparisons (Fig. 2, Fig. 5).

In this paper, we present a systematic study to measure and understand biases in vision-language foundation models. Using chest X-rays as a driving example, we mainly utilize CheXzero [10], a state-of-the-art self-supervised foundation model in medical imaging, to assess bias and fairness across a broad spectrum of pathologies with demographic subpopulations present in the testing data. We also test another vision-language foundation model [11] and show similar findings (Fig. B.1). Our analysis incorporates five diverse, globally-sourced radiology datasets: MIMIC [17], CheXpert [18], NIH [19], PadChest [20], and VinDr [21]. We evaluate fairness within both individual and intersectional subpopulations spanning demographic attributes including race, sex, and age [12, 14]. We further compare fairness outcomes of the model with **board-certified radiologists**, uncovering that the foundation model demonstrates more substantial fairness discrepancies compared to human experts (Fig. 1). Further investigation in direct assessment of demographic attributes from chest X-rays shows that the model exhibits enhanced capacity to predict sensitive demographic information (e.g., age, race) compared to radiologists.

A summary of our contributions is given below:

- **Foundation model exhibits larger fairness disparities compared to radiologists:** We show, on two independent test sets, that the model exhibits much larger fairness gaps across demographic subgroups, compared to board-certified radiologists who independently assess the radiographs.
- **Diagnosis bias in under-served populations and intersectional subgroups:** We further validate the biases in larger populations, testing over traditionally under-served and intersectional subgroups (e.g., Black female), and show significant underdiagnosis and overdiagnosis rates.
- **Consistent biases over 50+ pathology labels and differential diagnoses:** We extend our analysis to a much larger and diverse set of pathology labels, and demonstrate consistently notable demographic biases across more than 50 pathology labels and differential diagnoses.

- **Foundation model encodes demographic information beyond human levels:** We reveal that the foundation model exhibits substantial encoding of sensitive information (e.g., age, race, sex). Further human study shows that the model is able to predict sensitive demographic attributes from chest X-rays with much higher accuracies compared to board-certified radiologists.
- **Initial evidences to intervene and improve model fairness:** We conduct experiments to explore fairness intervention of the foundation model by incorporating demographic details into the input prompt, with improved fairness over certain pathologies.

2 Related Work

Medical Foundation Models Recent advancements in medical foundation models have shown impressive diagnostic capabilities across various domains [6, 7, 14]. These models, based on large-scale pre-trained architectures, excel in tasks like medical imaging and clinical text interpretation with minimal task-specific annotations [5, 12]. Vision-language models, utilizing self-supervised learning, integrate multimodal inputs—such as radiology images and reports—to identify pathologies at a level comparable to, or even exceeding, human experts [5, 10, 13].

Subpopulation Robustness It is crucial to ensure that AI models do not perpetuate or exacerbate demographic biases (such as race in healthcare) – a critical issue that recent studies have repeatedly brought to light [22, 23]. Recent studies have shown that chest X-ray classifiers systematically underdiagnose Black patients, potentially causing incorrect triage and delayed treatment [12, 14]. Although biases in supervised models have been explored [14, 15], less attention has been given to vision-language foundation models.

Fair Medical Imaging There have been many prior works which demonstrate gaps in performance (typically measured using the false positive and false negative rates) between demographic groups in medical imaging tasks for various modalities, including chest X-rays [12, 14], MRIs [24], CT scans [25], and dermoscopic images [16]. Most relevant to this work is Seyyed-Kalantari et al. [14], which shows that ***supervised*** chest X-ray models for predicting "*No Finding*" have higher false positive rate (i.e., underdiagnosis) for Black, female, and younger patients. Yang et al. [23] applied various fairness algorithms to the same dataset, finding mixed results.

In comparison, our work approaches the fairness problem from the ***self-supervised*** angle. Although algorithmic biases have been studied in the supervised setting [14, 15], little attention has been paid to vision-language foundation models. These models, notably free from explicit supervision through multimodal training and zero-shot inference, theoretically have reduced potential to inherit human labeling biases. Towards this end, our work (1) evaluates different medical vision-language foundation models, (2) examines a wide range of pathologies and differential diagnoses, (3) examines model performance and fairness on external sites and globally sourced datasets, and (4) involves board-certified radiologists for reader studies that ground the findings.

3 Methods

Datasets We collect five public chest X-ray datasets from diverse global sources. These datasets, as detailed in Table 1, encompass MIMIC [17] (357,167 images from 61,927 patients), CheXpert [18] (223,458 images from 64,925 patients), and NIH [19] (112,120 images from 30,805 patients) from the United States, PadChest [20] (160,736 images from 67,590 patients) from Spain, and VinDr [21] (5,323 images from 5,323 patients) from Vietnam. The datasets provide chest X-ray images along with pathology labels and demographic data derived from the respective patients. Both MIMIC and CheXpert present demographic information including sex, age, and race. The remaining datasets (i.e., NIH, PadChest, VinDr) present demographic details regarding sex and age, with no information available on the race of the patients.

Model We utilize a state-of-the-art self-supervised foundation model in medical imaging, CheXzero [10], as a driving example to study fairness of foundation models. The model was trained in a self-supervised way without using any pathology labels or annotations. Specifically, it was initialized from a Vision Transformer backbone ViT-B/32 [26] and pre-trained weights from OpenAI's CLIP model [27]. The model was trained in a self-supervised manner on the MIMIC dataset without pathology labels or annotations, leveraging the radiographs with accompanying clinical texts [10]. In addition, we tested another vision-language foundation model, KAD [11], which introduces knowledge graphs

Table 1: **Characteristics of the datasets used in this study.**

		MIMIC [17]	**CheXpert** [18]	**NIH** [19]	**PadChest** [20]	**VinDr** [21]
	Location	Boston, MA	Stanford, CA	Bethesda, MD	Alicante, Spain	Hanoi, Vietnam
	# Images	357,167	223,458	112,120	160,736	5,323
	# Patients	61,927	64,925	30,805	67,590	5,323
	# Pathologies	14	14	15	174	27
	% Frontal	64.5	85.5	100.0	69.1	100.0
	Sample Image					
Sex (%)	**Male**	52.2	59.3	56.5	49.6	56.9
	Female	47.8	40.7	43.5	50.4	43.1
Race (%)	**White**	61.0	56.4	-	-	-
	Black	15.6	5.4	-	-	-
	Asian	3.1	10.5	-	-	-
	Other	20.3	27.8	-	-	-
Age (%)	**0-18**	-	-	4.8	3.7	21.8
	18-40	13.8	13.9	27.7	9.2	16.0
	40-60	31.1	31.1	43.9	26.5	27.1
	60-80	40.0	39.0	22.7	38.0	30.0
	80-100	15.1	16.0	0.9	22.6	5.1

into visual-language pretraining (Fig. B.1). We evaluated the model on internationally-sourced external chest X-rays datasets. In particular, CheXpert, PadChest, and VinDr contain gold-standard ground truth radiologist labels. Among these datasets, CheXpert test set (666 samples) and VinDr (5,323 samples) provide external annotations from three board-certified radiologists, which were used to benchmark the performance and fairness of radiologists' assessments compared to the model.

Assessing the Demographic Fairness To assess the model prediction fairness, we focus on three demographic attributes: sex, age, and race, and dissect the performance of the model within different subpopulations, such as female or Black patients, and the intersectional groups like Black female patients. We follow the literature to examine the class-conditioned error rate that is likely to lead to worse patient outcomes for a screening model [12, 14]. For all potential pathology labels, a false negative indicates falsely predicting someone to be healthy when they are ill, which could lead to delays in treatment [14] (i.e., an underdiagnosis). Therefore, we evaluate the differences in False Negative Rate (FNR) between demographic subpopulations. For "No Finding", we evaluate the False Positive Rate (FPR) for the same reason. Equality in these metrics can be viewed as instances of equal opportunity between subgroups [28]. We then denote the differences in FNR/FPR for two selected subgroups (e.g., Black and white patients) as the ***underdiagnosis disparity***.

Human Study Details Three board-certified radiologists from the Department of Radiology at the University of Washington, School of Medicine were tasked with evaluating demographic attributes from chest X-rays only. Each radiologist had over ten years of experience in chest imaging, participated independently, was blinded to the demographic attributes, and received no prior training or exposure to the task to mitigate any training effect.

We used an online labeling tool [29] for the radiologists to create attribute labels based on the 480 pre-selected chest X-ray images from MIMIC. All three attribute labels are required for each image, meaning that radiologists are required to choose one label for each of the attributes. Importantly, each radiologist completed this study independently and was provided with no additional information beyond the chest X-ray images themselves. The distribution of the three attributes was not disclosed to the radiologists until after they had completed the task, ensuring an unbiased evaluation process.

4 Results

4.1 Substantial Fairness Disparities in Foundation Model Compared to Radiologists

We assess the model's underdiagnosis disparity across datasets and demographic populations. Since external radiologist annotations are available in certain datasets (i.e., CheXpert and VinDr), we

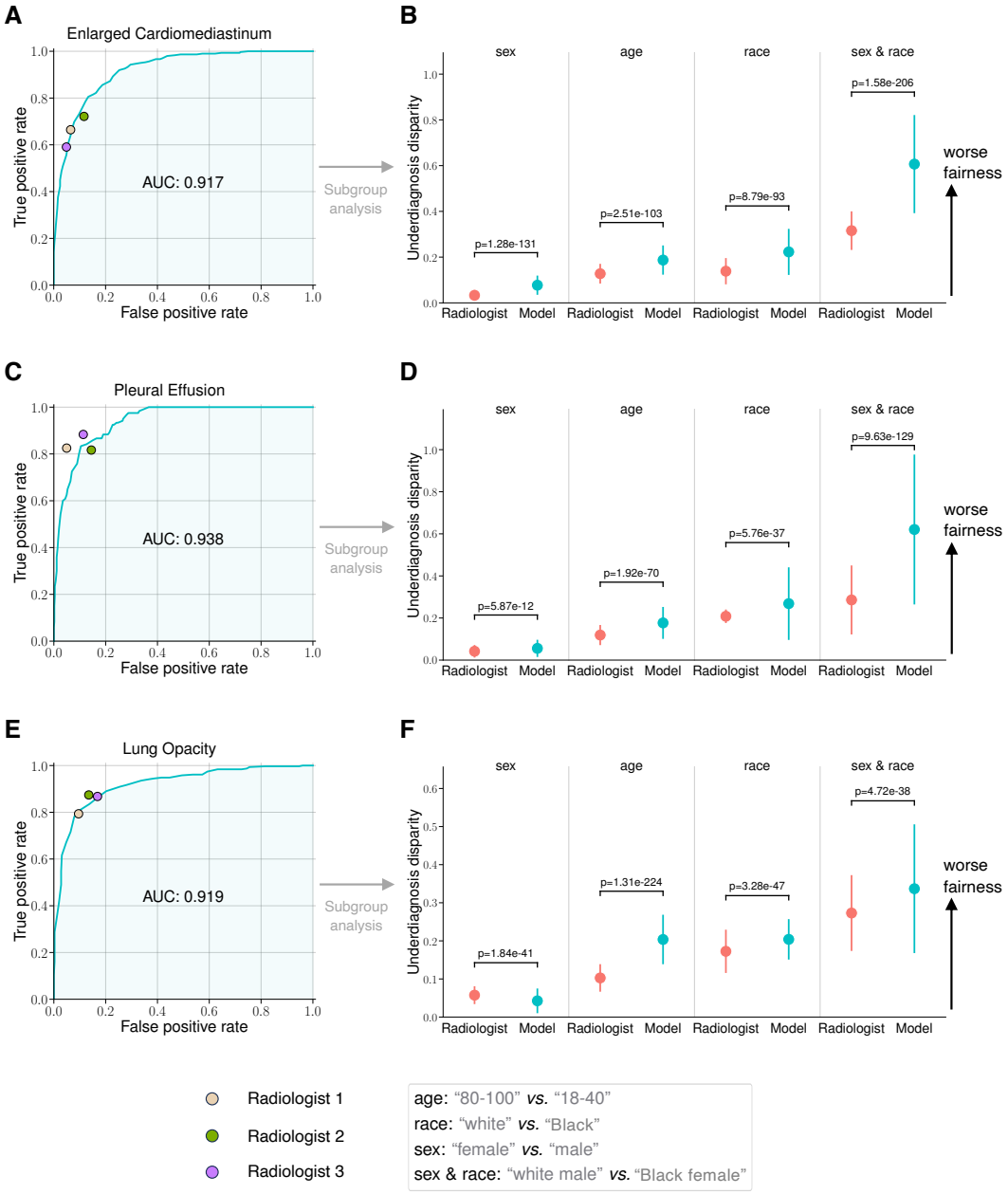

Figure 2: **Comparisons of diagnosis AUROC and underdiagnosis disparity for the vision-language foundation model and board-certified radiologists. (A, C, E)** Comparison of the ROC curve of the vision-language foundation model to benchmark radiologists against the test-set ground truth on the CheXpert dataset (n=666). The model outperforms the radiologists when the ROC curve lies above the radiologists' operating points. **(B, D, F)** Comparison of the underdiagnosis disparity of the foundation model against three board-certified radiologists on the CheXpert test set (n=666). We average the assessments from different radiologists as the evaluation of human biases. The model exhibits significantly higher underdiagnosis bias than that of radiologists on all three pathologies. Error bars indicate 95% confidence intervals estimated using non-parametric bootstrap sampling.

directly compared the overall performance as well as the performance for subpopulations between the model and radiologists. Fig. 2 presents the diagnostic performance and fairness of the vision-language foundation model in contrast to that of board-certified radiologists on the CheXpert dataset (n=666). First, Figs. 2A, 2C, and 2E show the comparison of the receiver operating characteristic (ROC) curves of the model to the operating points of radiologists for three different pathologies. Notably,

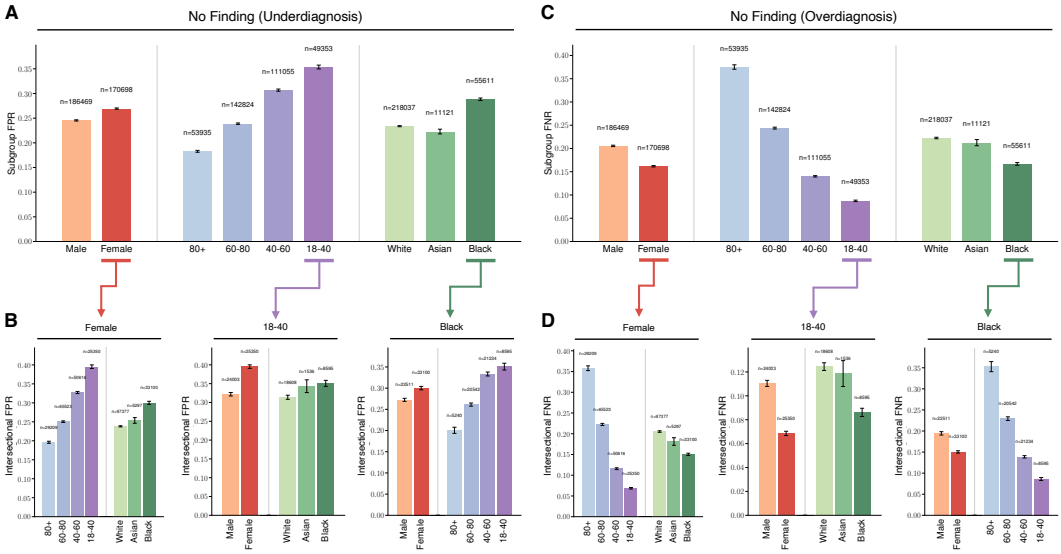

Figure 3: **Analysis of underdiagnosis and overdiagnosis across subgroups of sex, age, race, and intersectional groups in the MIMIC dataset. (A)** The underdiagnosis rate, as measured by the no finding FPR, in the indicated patient subpopulations. **(B)** Intersectional underdiagnosis rates for female patients, patients aged 18–40 years, and Black patients. **(C, D)** The overdiagnosis rate, as measured by the no finding FNR in the same patient subpopulations as in (**A**) and (**B**). Error bars indicate 95% confidence intervals estimated using non-parametric bootstrap sampling (n=1,000).

the model exhibits comparable or better diagnostic performance compared to radiologists ("Enlarged Cardiomediastinum": AUC=0.917, 95% CI [0.905, 0.928]; "Pleural Effusion": AUC=0.938, 95% CI [0.922, 0.950]; "Lung Opacity": AUC=0.919, 95% CI [0.904, 0.933]).

In the meantime, we further assess the underdiagnosis disparity between subgroups, which measures the disparity of FNR between two selected subgroups in each category ("Female" vs "Male" in sex, "80-100" vs "18-40" in age, "White" vs "Black" in race, and "White male" vs "Black female" in the intersectional group of sex and race). We average the assessments from different radiologists as the evaluation of human biases. When computing FNR for the model, we use the optimal threshold computed on the validation set that maximizes the Youden's J statistic [30]. Figs. 2B, 2D, and 2F show that the model exhibits much larger fairness gaps compared to radiologists, especially for intersectional subgroups. For instance, the model exhibits significantly higher underdiagnosis rate for "Enlarged Cardiomediastinum" in sex (p=1.28e-131, one-tailed Wilcoxon rank-sum test; same test for following attributes), age (p=2.51e-103), race (p=8.79e-93), and the intersectional of sex and race (p=1.58e-206). More results can be found in the Fig. B.2, including the analysis of other pathologies in CheXpert, and on another dataset from a different site (VinDr). Overall, the model exhibits expert-level pathology detection accuracy, but shows consistently higher underdiagnosis bias compared to radiologists.

4.2 Diagnosis Bias in Marginalized Populations and Intersectional Groups

We further evaluate the diagnosis bias of the model on MIMIC, the largest and the most diverse chest X-ray dataset in our study. We focus on the "No Finding" label, and show both underdiagnosis and overdiagnosis bias of the model on individual and intersectional subpopulations (Fig. 3). FPR is used for assessing underdiagnosis, whereas FNR is used for overdiagnosis. Fig. 3A shows significant fairness gaps between patient subpopulations in each category, especially between the age subgroups ">80" (n=53,935) and "18-40" (n=49,353). Moreover, larger gaps of the underdiagnosis rate between the intersectional subgroups can be observed in Fig. 3B. For instance, around 20% FPR discrepancies exist between female patients aged above 80 (n=29,209) and those in their 18-40 (n=25,350). Similar observations hold for overdiagnosis (Fig. 3C, Fig. 3D), the gaps become more significant between intersectional subgroups. The FPR (Fig. 3A) and FNR (Fig. 3C) for "No Finding" shows an inverse relationship across different marginalized subgroups in the CXR dataset. Such an inverse relationship also exists for intersectional subgroups (Fig. 3B, Fig. 3D), and is consistent across other datasets.

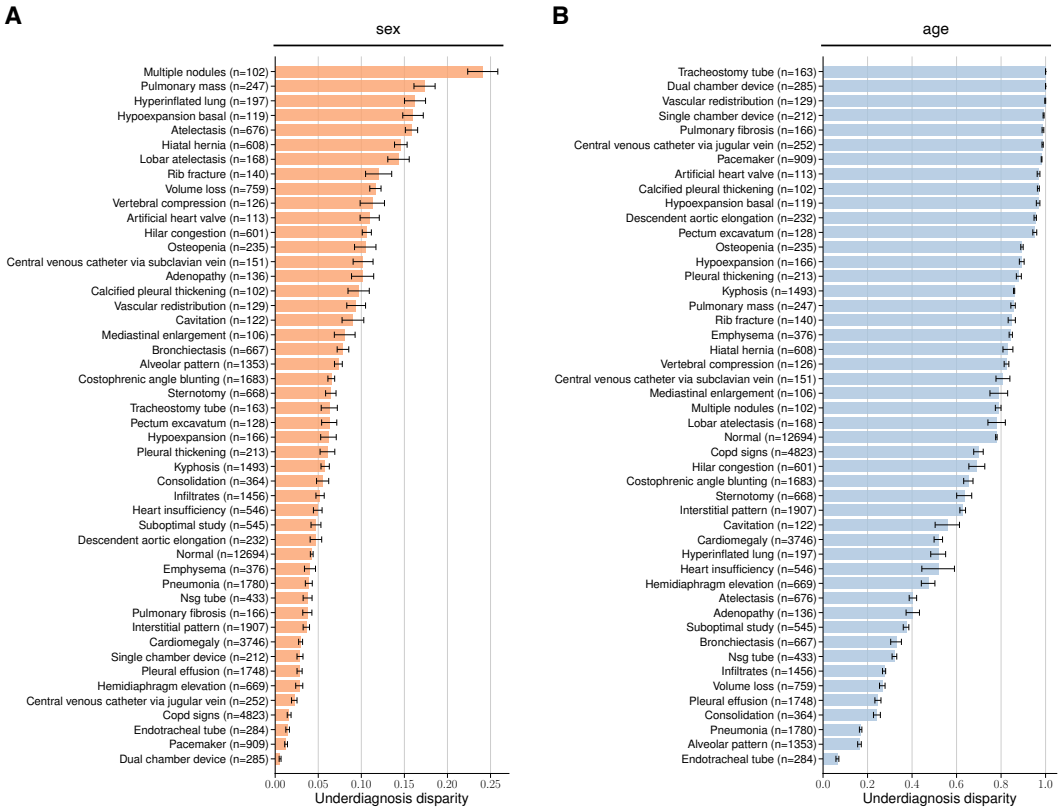

Figure 4: **Demographic fairness on unseen radiographic findings in the PadChest dataset.** Average underdiagnosis disparity and 95% CI are shown for each radiographic finding (n>100) labeled as high importance by an expert radiologist. **(A)** Underdiagnosis disparity for sex (between group "female" and "male". **(B)** Underdiagnosis disparity for age (between group "18-40" and ">80". We externally validated the model's fairness when testing on different data distributions by evaluating model performance on the human-annotated subset of the PadChest dataset (n=39,053). No labeled samples were seen during training for any of the radiographic findings in this dataset.

We observe that female patients, patients aged between 18 and 40 years, and Black patients have higher rates of algorithmic underdiagnosis, indicating that these subgroups are most likely being falsely diagnosed as healthy by the model and failing to receive appropriate clinical treatments. Further investigations on intersectional subpopulations reveal that the underdiagnosis rates increased significantly for specific groups of patients, such as Black Female patients. We show in Fig. B.3 that the observations hold across different pathologies such as "Lung Opacity" or "Pneumonia". We further confirm in Fig. B.4 that the disparities remain consistent when tested on external datasets such as CheXpert, NIH, and VinDr.

4.3 Demographic Bias in Unseen Radiographic Findings

We extended our analysis to investigate the demographic biases using a much larger and diverse set of pathology labels. We tested the foundation model on the PadChest dataset collected from a different country with 174 radiographic findings and 19 differential diagnoses [20]. We filtered out 48 radiographic findings where n>100 and the model achieved an AUC of at least 0.7 in the PadChest test set (n=39,053) to further assess the demographic fairness of the model on unseen radiographic findings [10]. Fig. 4 reveals distinct disparities in both sex ("female" vs "male" subgroup) and age (">80" vs "18-40" subgroup) among those radiographic findings. The maximum underdiagnosis disparity (i.e., "Multiple nodules", n=102) between female and male patients is 24.1% (95% CI [22.5%, 26.0%]), whereas 31 out of 48 findings exhibit a fairness gap larger than 5% (Fig. 4A). The discrepancies become even more significant for age, with a 100% fairness gap for "Tracheostomy tube" (n=163) between "18-40" and ">80" subgroups, and 45 out of 48 findings exhibit a fairness gap larger than 20% (Fig. 4B).

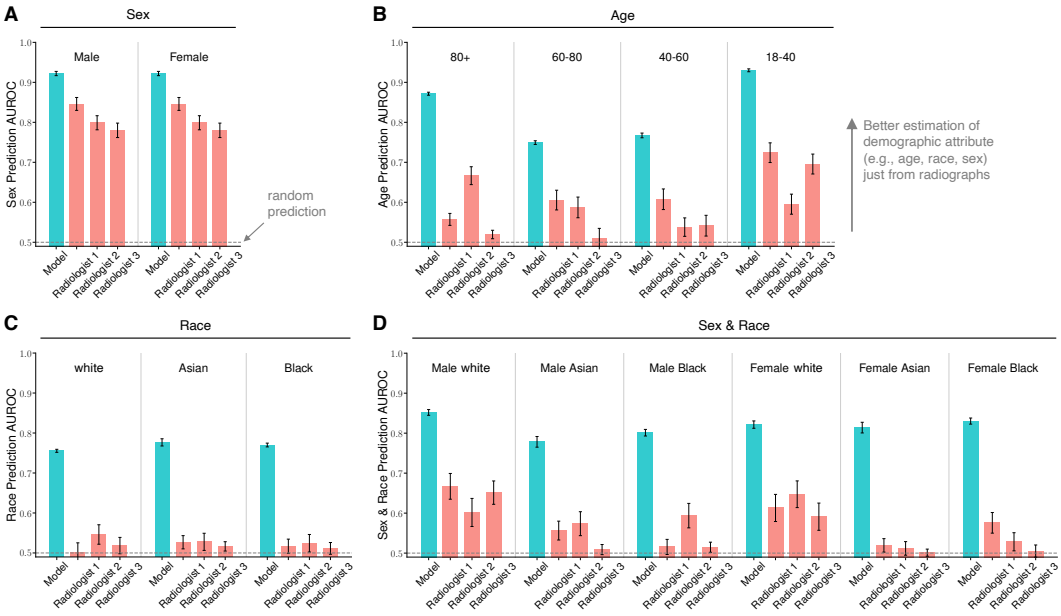

Figure 5: **Comparisons of prediction AUROC for sensitive demographic attributes between the foundation model and three board-certified radiologists. (A to D)** Prediction AUROC of subgroups within different sensitive attributes including sex (**A**), age (**B**), race (**C**), and the intersectional groups of sex and race (**D**), on a subset of MIMIC (n=480). We selected out a balanced subset of MIMIC w.r.t. all attributes (i.e., balanced across age, sex, and race), and asked three board-certified radiologists to infer the attributes from just the chest X-rays. To assess the model prediction of sensitive attributes, we train a linear attribute prediction head using logistic regression on top of the penultimate layer of the model, with the model weights frozen. Error bars indicate 95% confidence intervals estimated using non-parametric bootstrap sampling (n=1,000).

4.4 Demographic Information Encoding in Foundation Model Beyond Human Levels

With consistent demographic bias across international evaluation, we aim to further dissect and explain the performance of the model. Inspired by recent works on algorithmic encoding of demographic information by deep learning models [24, 25, 31], we investigated whether the model encodes demographic information by examining the predictability of sensitive attributes by both the self-supervised foundation model and board-certified radiologists. We selected 480 chest X-ray samples from the MIMIC dataset, ensuring an equal number of samples across all subgroups in three key attributes: sex, age, and race. Instead of focusing on pathology prediction, we assessed how much the model encodes demographic information by training a linear attribute prediction head using logistic regression on top of the penultimate layer of the model, with the model weights frozen. In the meantime, we involved three board-certified radiologists with over ten years of experience in chest imaging to label the demographic attributes (sex, age, and race) for the same set of patients based solely on their chest X-rays. Each radiologist was blinded to the demographic attributes and participated independently without any prior training or exposure to the task to avoid any training effect.

Interestingly, the foundation model, although trained in a self-supervised manner without explicit information regarding the demographic attributes, demonstrated substantial and consistent encoding of demographic information across all tested attributes and subgroups (Fig. 5). Specifically, the predictive AUCs for sex (Fig. 5A, "Female" AUC=0.92, 95% CI [0.91, 0.93]), age (Fig. 5B, "18-40" AUC=0.94, 95% CI [0.93, 0.94]), race (Fig. 5C, "Black" AUC=0.78, 95% CI [0.77, 0.78]), and the intersectional subgroups (Fig. 5D, "Black Female" AUC=0.83, 95% CI [0.82, 0.83]) are significantly higher than random chance (i.e., 0.5). This strong algorithmic encoding of demographic attributes could be explainable for the observed underdiagnosis bias across patient subpopulations [23].

Interestingly however, the performance of three radiologists to predict these attributes falls behind. They achieve relatively high AUC scores in sex prediction (Fig. 5A), but much lower in age prediction (Fig. 5B). When it comes to race, the prediction is marginally better than random guess (Fig. 5C).

Similar performance pattern is observed in the intersectional group of sex and race prediction (Fig. 5D), suggesting that radiologists cannot directly read attributes like age or race from radiographs.

Fig. B.5 further suggests that inherent encoding of the sensitive data (e.g., demographics, support devices) might drive the underdiagnosis biases (details in Discussion). We provide analysis and initial methods to intervene the model fairness in Fig. B.6 and Fig. B.7.

5 Discussion

We have dissected the performance of the state-of-the-art foundation model and shown consistent underdiagnosis in the chest X-ray domain. Importantly, we were able to compare the results with board-certified radiologists to ground the findings. The results reveal consistently larger fairness disparities of the model compared to radiologists (Fig. 2), and that the model exhibits systematic biases in marginalized subpopulations as well as intersectional subgroups like Black female patients (Fig. 3). The demographic biases of the foundation model persist across a wide range of unseen pathologies (Fig. 4). Further analyses demonstrate that the model encodes substantial demographic information (e.g., race), and that is significantly higher than human radiologists (Fig. 5).

Our results have multiple implications. First, the fairness-accuracy trade-off in AI models raises complex ethical considerations [32, 33]. Medical vision-language foundation models hold the promise of a single model diagnosing countless pathologies with expert-level accuracy. Yet, our analysis shows that they exhibit much larger fairness gaps compared to radiologists (Fig. 2). Incorrectly underdiagnosing specific subgroups more frequently than others not only places them at a disadvantage but also raises serious ethical concerns when deploying the model in a clinical pipeline [34, 35].

Second, our study shows that the model encodes demographic information far more profoundly than human capacity (Fig. 5, Fig. B.5). This suggests that inherent encoding of the sensitive data might drive the underdiagnosis biases (e.g., Fig. 4). Notably, even though the model is trained in a self-supervised manner without explicit attribute information, it still manages to embed this information. Recent studies explore if deep models use demographics as "shortcuts", disadvantaging specific groups [36, 37]. These call for a deeper understanding of how these powerful models process and utilize sensitive information, and whether that is aligned with clinical validations by radiologists [35, 38]. Whether demographic variables should be encoded as proxies for causal factors is a decision that should align with its actual clinical use [16, 34, 39].

Third, the ability of the AI model to discern these demographics more precisely suggests a potential for uncovering clinically relevant features that might not be immediately apparent to human readers, suggesting an opportunity for an improved human-AI collaboration [40, 41, 42]. Exploring the semantic and agnostic features harnessed by AI could improve human performance and potentially deliver a higher quality care. On the other hand, in scenarios where clinical decisions are influenced by AI model suggestions, any undetected bias within the model could lead to unintended and potentially harmful consequences [32, 43]. This underscores the need for careful and continuous evaluation of AI biases to progressively diminish their influence in healthcare.

Fourth, our human study reveals that radiologists too can manifest biases when evaluating over diverse subgroups. These inherent biases raise pressing concerns about potential corrective actions and the mechanism for feedback. Existing literature points out that clinician bias significantly contributes to healthcare disparities across race and gender [44, 45, 46]. With the addition of biases from AI, the need for de-identifying demographics to counteract biases becomes vital. This is especially crucial as AI demonstrates the capability to uncover demographics even without such supervision. Concurrently, regulatory authorities should establish clear guidelines on how the prediction of demographics (among other things) should be managed by AI models, ensuring they adhere to privacy and ethical standards.

Broader Societal Impacts

In this work, we imply that smaller "fairness gaps" are better – i.e. that it is optimal to have equal performance for all attributes. Prior works have shown that enforcing these group fairness definitions may lead to worse utility and performance for all groups [12, 22, 47], and that other fairness definitions may be better suited to the clinical setting [14, 23]. We encourage practitioners to choose a fairness definition that is best-suited to their use case, and carefully consider the performance-equality trade-off. In addition, though we construct several models for clinical risk prediction in this paper, we do not advocate for blind deployment of these models in real-world clinical settings in any way. Practitioners should always test such models on their data and take a myriad of other considerations into account (e.g. privacy, regulation, interpretability) before deployment [13, 16].

References

[1] R. Y. Kim, J. L. Oke, L. C. Pickup, R. F. Munden, T. L. Dotson, C. R. Bellinger, A. Cohen, M. J. Simoff, P. P. Massion, C. Filippini, and F. V. Gleeson. Artificial intelligence tool for assessment of indeterminate pulmonary nodules detected with ct. *Radiology*, 304:683–691, 2022.

[2] P. G. Mikhael, J. Wohlwend, A. Yala, L. Karstens, J. Xiang, A. K. Takigami, P. P. Bourgouin, P. Chan, S. Mrah, W. Amayri, Y. H. Juan, C. T. Yang, Y. L. Wan, G. Lin, L. V. Sequist, F. J. Fintelmann, and R. Barzilay. Sybil: a validated deep learning model to predict future lung cancer risk from a single low-dose chest computed tomography. *Journal of Clinical Oncology*, 41:2191–2200, 2023.

[3] S. M. McKinney, M. Sieniek, V. Godbole, J. Godwin, N. Antropova, H. Ashrafian, T. Back, M. Chesus, G. S. Corrado, A. Darzi, M. Etemadi, F. Garcia-Vicente, F. J. Gilbert, M. Halling-Brown, D. Hassabis, S. Jansen, A. Karthikesalingam, C. J. Kelly, D. King, J. R. Ledsam, D. Melnick, H. Mostofi, L. Peng, J. J. Reicher, B. Romera-Paredes, R. Sidebottom, M. Suleyman, D. Tse, K. C. Young, J. De Fauw, and S. Shetty. International evaluation of an ai system for breast cancer screening. *Nature*, 577:89–94, 2020.

[4] P. Rajpurkar, J. Irvin, K. Zhu, B. Yang, H. Mehta, T. Duan, D. Ding, A. Bagul, C. Langlotz, K. Shpanskaya, and M. P. Lungren. Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. *ArXiv Preprint ArXiv171105225*, 2017.

[5] M. Moor, O. Banerjee, Z. S. H. Abad, H. M. Krumholz, J. Leskovec, E. J. Topol, and P. Rajpurkar. Foundation models for generalist medical artificial intelligence. *Nature*, 616:259–265, 2023.

[6] B. Wang, Q. Xie, J. Pei, Z. Chen, P. Tiwari, Z. Li, and J. Fu. Pre-trained language models in biomedical domain: A systematic survey. *ACM Computing Surveys*, 56:55:1–55:52, 2023.

[7] Z. Huang, F. Bianchi, M. Yuksekgonul, T. J. Montine, and J. Zou. A visual–language foundation model for pathology image analysis using medical twitter. *Nature Medicine*, 29:2307–2316, 2023.

[8] M. Y. Lu, B. Chen, D. F. K. Williamson, R. J. Chen, I. Liang, T. Ding, G. Jaume, I. Odintsov, L. P. Le, G. Gerber, A. V. Parwani, A. Zhang, and F. Mahmood. A visual-language foundation model for computational pathology. *Nature Medicine*, 30:863–874, 2024.

[9] Y. Zhou, M. A. Chia, S. K. Wagner, M. S. Ayhan, D. J. Williamson, R. R. Struyven, T. Liu, M. Xu, M. G. Lozano, P. Woodward-Court, Y. Kihara, UK Biobank Eye & Vision Consortium, A. Altmann, A. Y. Lee, E. J. Topol, A. K. Denniston, D. C. Alexander, and P. A. Keane. A foundation model for generalizable disease detection from retinal images. *Nature*, 622:156–163, 2023.

[10] E. Tiu, E. Talius, P. Patel, C. P. Langlotz, A. Y. Ng, and P. Rajpurkar. Expert-level detection of pathologies from unannotated chest x-ray images via self-supervised learning. *Nature Biomedical Engineering*, 6:1399–1406, 2022.

[11] X. Zhang, C. Wu, Y. Zhang, W. Xie, and Y. Wang. Knowledge-enhanced visual-language pre-training on chest radiology images. *Nature Communications*, 14:4542, 2023.

[12] L. Seyyed-Kalantari, G. Liu, M. B. A. McDermott, I. Y. Chen, and M. Ghassemi. Chexclusion: Fairness gaps in deep chest x-ray classifiers. In *Proceedings of the Pacific Symposium on Biocomputing*, pages 232–243. World Scientific, 2020.

[13] A. Vaidya, R. Chen, D. Williamson, A. Song, G. Jaume, Y. Yang, T. Hartvigsen, E. Dyer, M. Y. Lu, J. Lipkova, M. Shaban, T. Y. Chen, and F. Mahmood. Demographic bias in misdiagnosis by computational pathology models. *Nature Medicine*, 30:1174–1190, 2024.

[14] L. Seyyed-Kalantari, H. Zhang, M. B. A. McDermott, I. Y. Chen, and M. Ghassemi. Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. *Nature Medicine*, 27:2176–2182, 2021.

[15] Y. Yang, Y. Yuan, G. Zhang, H. Wang, Y.-C. Chen, Y. Liu, C. Tarolli, D. Crepeau, J. Bukartyk, M. Junna, A. Videnovic, T. Ellis, M. Lipford, R. Dorsey, and D. Katabi. Artificial intelligence-enabled detection and assessment of parkinson's disease using nocturnal breathing signals. *Nature Medicine*, 28:2207–2215, 2022.

[16] M. D. McCradden, S. Joshi, M. Mazwi, and J. A. Anderson. Ethical limitations of algorithmic fairness solutions in health care machine learning. *The Lancet Digital Health*, 2:e221–e223, 2020.

[17] A. E. W. Johnson, T. J. Pollard, N. R. Greenbaum, M. P. Lungren, C. y. Deng, Y. Peng, Z. Lu, R. G. Mark, S. J. Berkowitz, and S. Horng. Mimic-cxr-jpg, a large publicly available database of labeled chest radiographs, 2019. Preprint at https://arxiv.org/abs/1901.07042.

[18] J. Irvin, P. Rajpurkar, M. Ko, Y. Yu, S. Ciurea-Ilcus, C. Chute, H. Marklund, B. Haghgoo, R. Ball, K. Shpanskaya, J. Seekins, D. A. Mong, S. S. Halabi, J. K. Sandberg, R. Jones, D. B. Larson, C. P. Langlotz, B. N. Patel, M. P. Lungren, and A. Y. Ng. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 590–597, 2019.

[19] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. M. Summers. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2097–2106. IEEE, 2017.

[20] A. Bustos, A. Pertusa, J.-M. Salinas, and M. De La Iglesia-Vaya. Padchest: A large chest x-ray image dataset with multi-label annotated reports. *Medical Image Analysis*, 66:101797, 2020.

[21] H. Q. Nguyen, K. Lam, L. T. Le, H. H. Pham, D. Q. Tran, D. B. Nguyen, D. D. Le, C. M. Pham, H. T. T. Tong, D. H. Dinh, C. D. Do, L. T. Doan, C. N. Nguyen, B. T. Nguyen, Q. V. Nguyen, A. D. Hoang, H. N. Phan, A. T. Nguyen, P. H. Ho, D. T. Ngo, N. T. Nguyen, N. T. Nguyen, M. Dao, and V. Vu. Vindr-cxr: An open dataset of chest x-rays with radiologist's annotations. *Scientific Data*, 9:429, 2022.

[22] Y. Yang, H. Zhang, D. Katabi, and M. Ghassemi. On mitigating shortcut learning for fair chest x-ray classification under distribution shift. In *NeurIPS 2023 Workshop on Distribution Shifts: New Frontiers with Foundation Models*, 2023.

[23] Y. Yang, H. Zhang, J. W. Gichoya, D. Katabi, and M. Ghassemi. The limits of fair medical imaging ai in real-world generalization. *Nature Medicine*, 30:1–11, 2024.

[24] I. Banerjee, K. Bhattacharjee, J. L. Burns, H. Trivedi, S. Purkayastha, L. Seyyed-Kalantari, B. N. Patel, R. Shiradkar, and J. Gichoya. "shortcuts" causing bias in radiology artificial intelligence: Causes, evaluation, and mitigation. *Journal of the American College of Radiology*, 20:842–851, 2023.

[25] J. W. Gichoya, I. Banerjee, A. R. Bhimireddy, J. L. Burns, L. A. Celi, L. C. Chen, R. Correa, N. Dullerud, M. Ghassemi, S.-C. Huang, P.-C. Kuo, M. P. Lungren, L. J. Palmer, B. J. Price, S. Purkayastha, A. T. Pyrros, L. Oakden-Rayner, C. Okechukwu, L. Seyyed-Kalantari, H. Trivedi, R. Wang, Z. Zaiman, and H. Zhang. Ai recognition of patient race in medical imaging: a modelling study. *The Lancet Digital Health*, 4:e406–e414, 2022.

[26] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, and J. Uszkoreit. An image is worth 16x16 words: Transformers for image recognition at scale. In *Proceedings of the International Conference on Learning Representations (ICLR)*, pages 1–21, 2021.

[27] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning (PMLR)*, pages 8748–8763, 2021.

[28] M. Hardt, E. Price, and N. Srebro. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 29, 2016.

[29] Encord. Data engine for ai model development, 2024. Available at https://encord.com/.

[30] R. Fluss, D. Faraggi, and B. Reiser. Estimation of the youden index and its associated cutoff point. *Biometrical Journal*, 47:458–472, 2005.

[31] J. Adleberg, A. Wardeh, F. X. Doo, B. Marinelli, T. S. Cook, D. S. Mendelson, and A. Kagen. Predicting patient demographics from chest radiographs with deep learning. *Journal of the American College of Radiology*, 19:1151–1161, 2022.

[32] M. A. Ricci Lara, R. Echeveste, and E. Ferrante. Addressing fairness in artificial intelligence for medical imaging. *Nature Communications*, 13:4581, 2022.

[33] M. Magdy, K. M. Hosny, N. I. Ghali, and S. Ghoniemy. Security of medical images for telemedicine: a systematic review. *Multimedia Tools and Applications*, 81:25101–25145, 2022.

[34] C. F. Manski, J. Mullahy, and A. S. Venkataramani. Using measures of race to make clinical predictions: Decision making, patient health, and fairness. *Proceedings of the National Academy of Sciences*, 120:e2303370120, 2023.

[35] C. F. Manski. Patient-centered appraisal of race-free clinical risk assessment. *Health Economics*, 31:2109–2114, 2022.

[36] E. Petersen, E. Ferrante, M. Ganz, and A. Feragen. Are demographically invariant models and representations in medical imaging fair?, 2023. Preprint at http://arxiv.org/abs/2305.01397.

[37] B. Glocker, C. Jones, M. Bernhardt, and S. Winzeck. Algorithmic encoding of protected characteristics in chest x-ray disease detection models. *Ebiomedicine*, 89, 2023.

[38] A. Basu. Use of race in clinical algorithms. *Science Advances*, 9:eadd2704, 2023.

[39] L. Oakden-Rayner, J. Dunnmon, G. Carneiro, and C. Ré. Hidden stratification causes clinically meaningful failures in machine learning for medical imaging. In *Proceedings of the ACM Conference on Health, Inference, and Learning*, pages 151–159, 2020.

[40] B. N. Patel, L. Rosenberg, G. Willcox, D. Baltaxe, M. Lyons, J. Irvin, P. Rajpurkar, T. Amrhein, R. Gupta, S. Halabi, C. Langlotz, E. Lo, J. Mammarappallil, A. J. Mariano, G. Riley, J. Seekins, L. Shen, E. Zucker, and M. P. Lungren. Human–machine partnership with artificial intelligence for chest radiograph diagnosis. *NPJ Digital Medicine*, 2:1–10, 2019.

[41] K. H. Yu, E. Healey, T. Y. Leong, I. S. Kohane, and A. K. Manrai. Medical artificial intelligence and human values. *New England Journal of Medicine*, 390:1895–1904, 2024.

[42] M. D. Abramoff, N. Whitestone, J. L. Patnaik, E. Rich, M. Ahmed, L. Husain, M. Y. Hassan, M. S. H. Tanjil, D. Weitzman, T. Dai, B. D. Wagner, D. H. Cherwek, N. Congdon, and K. Islam. Autonomous artificial intelligence increases real-world specialist clinic productivity in a cluster-randomized trial. *NPJ Digital Medicine*, 6:1–8, 2023.

[43] M. A. Ahmad, A. Patel, C. Eckert, V. Kumar, and A. Teredesai. Fairness in machine learning for healthcare. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 3529–3530. ACM, 2020.

[44] D. Centola, D. Guilbeault, U. Sarkar, E. Khoong, and J. Zhang. The reduction of race and gender bias in clinical treatment recommendations using clinician peer networks in an experimental setting. *Nature Communications*, 12:6585, 2021.

[45] R. A. Schut and E. J. Mortani Barbosa. Racial/ethnic disparities in follow-up adherence for incidental pulmonary nodules: An application of a cascade-of-care framework. *Journal of the American College of Radiology*, 17:1410–1419, 2020.

[46] A. B. Ross, V. Kalia, B. Y. Chan, and G. Li. The influence of patient race on the use of diagnostic imaging in united states emergency departments: data from the national hospital ambulatory medical care survey. *BMC Health Services Research*, 20:840, 2020.

[47] Y. Yang, H. Zhang, D. Katabi, and M. Ghassemi. Change is hard: A closer look at subpopulation shift. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, pages 39584–39622. PMLR, 2023.

[48] S. Wold, K. Esbensen, and P. Geladi. Principal component analysis. *Chemometrics and Intelligent Laboratory Systems*, 2:37–52, 1987.

A Experimental Details

Model Training and Evaluation We evaluated the model on our internationally-sourced chest X-ray datasets. In particular, approximately 45,000 chest X-ray images used in our evaluation come with gold standard annotations from radiologists across three datasets: CheXpert test set (666 chest X-rays with eight board-certified radiologist annotations for the presence of 14 different conditions), VinDr (5,323 images with annotations from a total of 17 experienced radiologists for 27 findings and diagnoses), and a subset of PadChest (39,053 images from the original dataset annotated by trained physicians). We also tested the model performance and fairness on MIMIC (357,167 images) and NIH (112,120 images) where the labels are generated from natural language processing techniques. Following the standard preprocessing practice [4, 47], we resized the radiographs to 224x224 and normalized them using a sample mean and standard deviation of the dataset for model evaluation.

Evaluation Methods To evaluate the performance of the foundation model on pathology classification, we use the following metrics: true positive rates (TPR), true negative rates (TNR), receiver operating characteristic (ROC) curves, and the area under the ROC curve (AUC). To evaluate the underdiagnosis disparity given one demographic attribute, we use the difference in TNR (or TPR) between two specific subpopulations (e.g., Black and White patients). To evaluate and assess the learned features in the penultimate layer of the model, we employ Principal Component Analysis (PCA) [48] to project the embeddings into a two-dimensional space for visualization.

Prompt Design In the main paper, we primarily employ prompts for vision-language foundation models for zero-shot inference, which involves calculating the similarity between X-ray representations and text representations for zero-shot classification. In particular, we follow established literature [7, 8, 10] to design the standard prompts for the vision-language foundation model.

Zero-shot classification, radiological findings (Fig. 1, and other main figures)

- The patient has {*pathology* / no *pathology*}
  - Example: "The patient has *Pneumonia*" & "The patient has no *Pneumonia*"

Zero-shot classification, radiological findings, with attribute info (Fig. B.7)

- The {*attribute*} patient has {*pathology* / no *pathology*}
  - Example: "The *female* patient has *Pneumonia*"

Zero-shot classification, demographic attributes (Fig. B.6)

- The patient's gender is {*attribute*}
  - Example: "The patient patient's gender is *female*"
- The patient's age is {*attribute*}
  - Example: "The patient patient's age is *under 18*"
- The patient's race is {*attribute*}
  - Example: "The patient patient's race is *Black*"

Assessing the Encoding of Attributes by Text Prompts We assessed the algorithmic encoding of demographic attributes in the foundation model through a logistic regression layer on the top of the model embedding (Fig. 5). Since the foundation model also supports textual prompts as input, we assess the encoding again by directly using textual prompts (Fig. B.6). Specifically, we utilized prompts containing demographic information to assess the attribute prediction accuracy. Across different datasets, the resulting prediction AUC is lower than using logistic regression, but still significantly higher than random chance over most of the subgroups.

Model Fairness Intervention We conducted experiments to explore fairness intervention of the foundation model by incorporating demographic details into the input prompt (Fig. B.7). We proposed to intervene the model prediction over subgroups by including demographic information in the input texts (e.g., "Does this female patient have Pneumonia?"). Fig. B.7 shows complex outcomes: After such intervention, the model displays reduced demographic biases for certain conditions like "Lung Opacity" and "No Finding", but not for others like "Pneumonia". The results indicate that it is possible to improve the demographic fairness of the model while maintaining the overall performance, but deeper analyses are needed for more principled methods.

B Additional Experimental Results

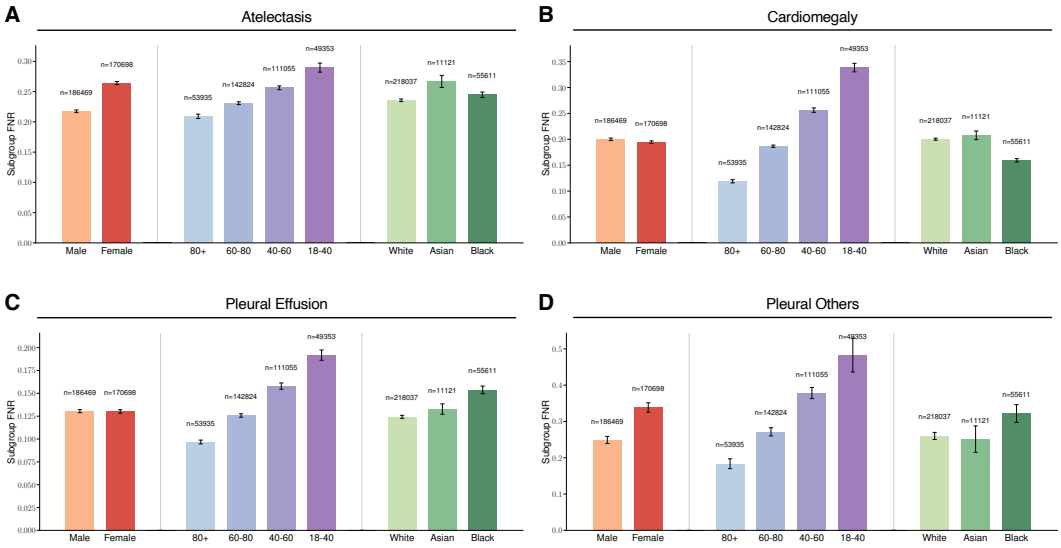

Figure B.1: **Underdiagnosis disparities on different pathologies of another vision-language foundation model, KAD [11], across subgroups of sex, age, and race in the MIMIC dataset. (A to D)** The underdiagnosis rate for "Atelectasis", "Cardiomegaly", "Pleural Effusion", and "Pleural Others" in the indicated patient subpopulations. Error bars indicate 95% confidence intervals estimated using non-parametric bootstrap sampling (n=1,000).

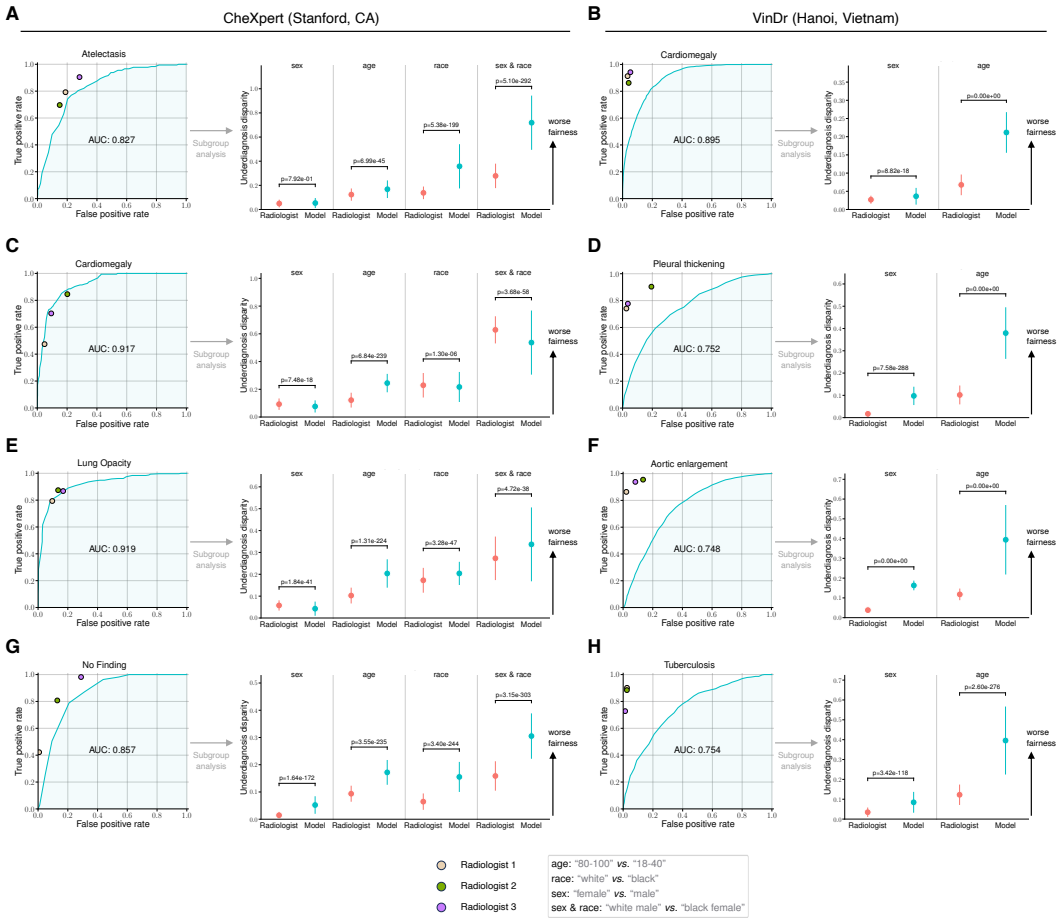

Figure B.2: **Comparisons of diagnosis AUROC and underdiagnosis disparity for the vision-language foundation model and board-certified radiologists on different datasets. (A, C, E, G)** Comparison of the ROC curve (left) and the underdiagnosis disparity (right) of the model to benchmark radiologists against the test-set ground truth on the CheXpert dataset (n=666). **(B, D, F, H)** The same comparisons performed on another dataset from a different country, VinDr (n=5,323). We average the assessments from different radiologists as the evaluation of human biases. The model exhibits significantly higher underdiagnosis bias than that of radiologists on all three pathologies. Error bars indicate 95% confidence intervals estimated using non-parametric bootstrap sampling (n=1,000).

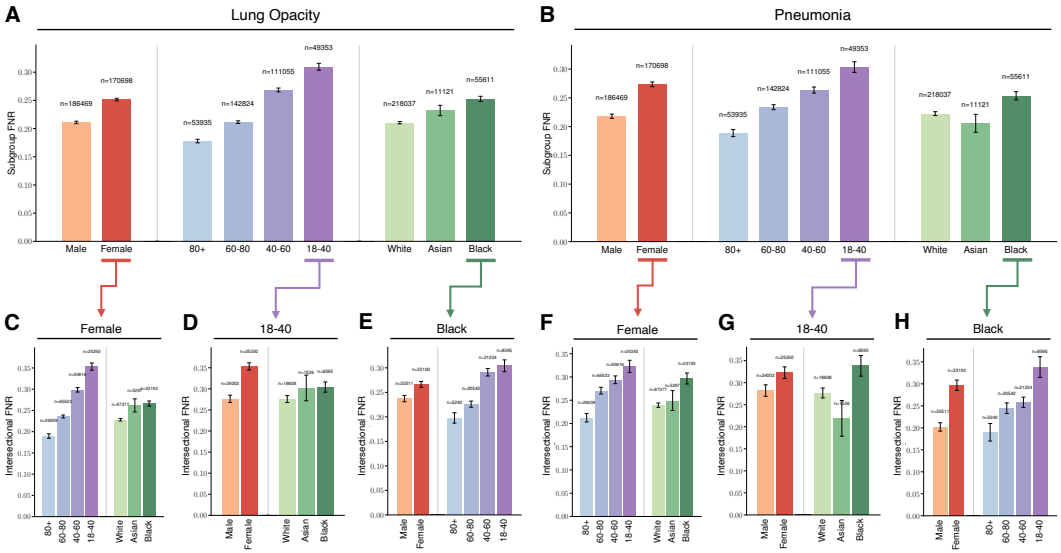

Figure B.3: **Underdiagnosis disparities on different pathologies across subgroups of sex, age, race, and intersectional groups in the MIMIC dataset. (A)** The underdiagnosis rate for "Lung Opacity" in the indicated patient subpopulations. **(B)** The underdiagnosis rate for "Pneumonia" in the indicated patient subpopulations. **(C to E)** Intersectional underdiagnosis rates for "Lung Opacity" in female patients (C), patients aged 18–40 years (D), and Black patients (E). **(F to H)** Intersectional underdiagnosis rates for "Pneumonia" in female patients (F), patients aged 18–40 years (G), and Black patients (H). Error bars indicate 95% confidence intervals estimated using non-parametric bootstrap sampling (n=1,000).

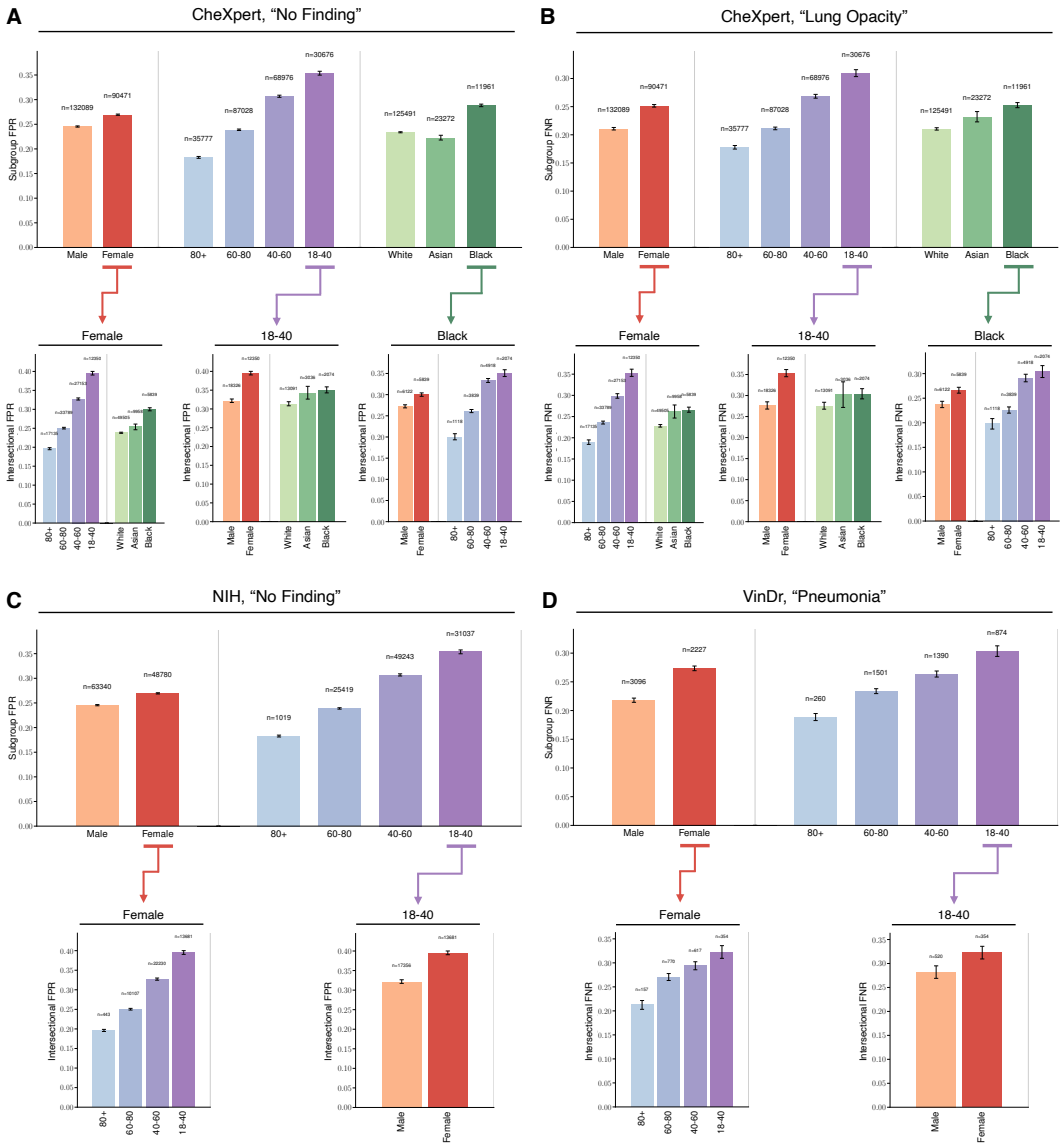

Figure B.4: **Underdiagnosis disparities on different pathologies across subgroups of sex, age, race, and intersectional groups in CheXpert, NIH, and VinDr. (A)** The underdiagnosis rate for "No Finding" in CheXpert in the indicated patient subpopulations. **(B)** The underdiagnosis rate for "Lung Opacity" in CheXpert in the indicated patient subpopulations. **(C)** The underdiagnosis rate for "No Finding" in NIH in the indicated patient subpopulations. **(D)** The underdiagnosis rate for "Pneumonia" in VinDr in the indicated patient subpopulations. Error bars indicate 95% confidence intervals estimated using non-parametric bootstrap sampling (n=1,000).

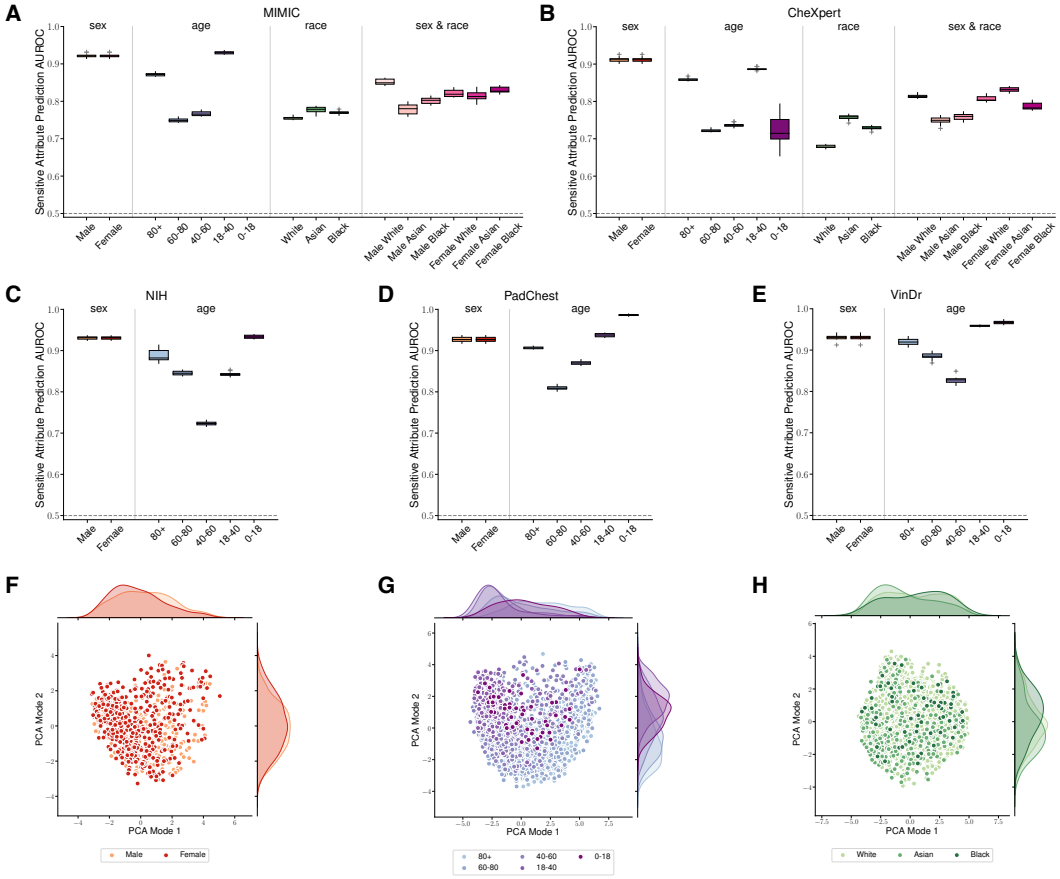

Figure B.5: **Algorithmic encoding of sensitive attributes in the foundation model. (A to E)** Prediction AUROC of different sensitive attributes including age, sex, race, and intersectional groups, across five datasets including MIMIC (**A**), CheXpert (**B**), NIH (**C**), PadChest (**D**), and VinDr (**E**). We train a linear attribute prediction head using logistic regression on top of the penultimate layer of the model, with the model weights frozen. **(F to H)** PCA visualization of the learned features in the penultimate layer of the model. We visualize the feature distribution on the randomly subsampled CheXpert dataset (n=2,000) for different attributes including sex (**F**), age (**G**), and race (**H**).

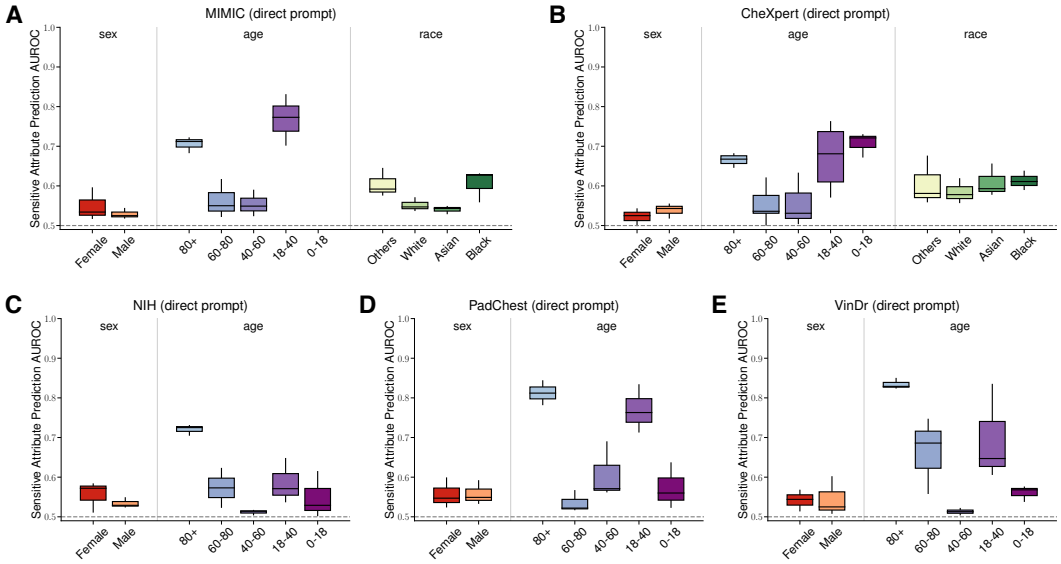

Figure B.6: **Direct Attribute prediction AUROC of the foundation model across different datasets. (A to E)** We utilize textual prompts encompassing demographic information (e.g., "The patient's gender is male.") to assess the attribute prediction accuracy on the MIMIC (A), CheXpert (B), NIH (C), PadChest (D), and VinDr (E) datasets. Error bars indicate 95% confidence intervals estimated using non-parametric bootstrap sampling (n=1,000).

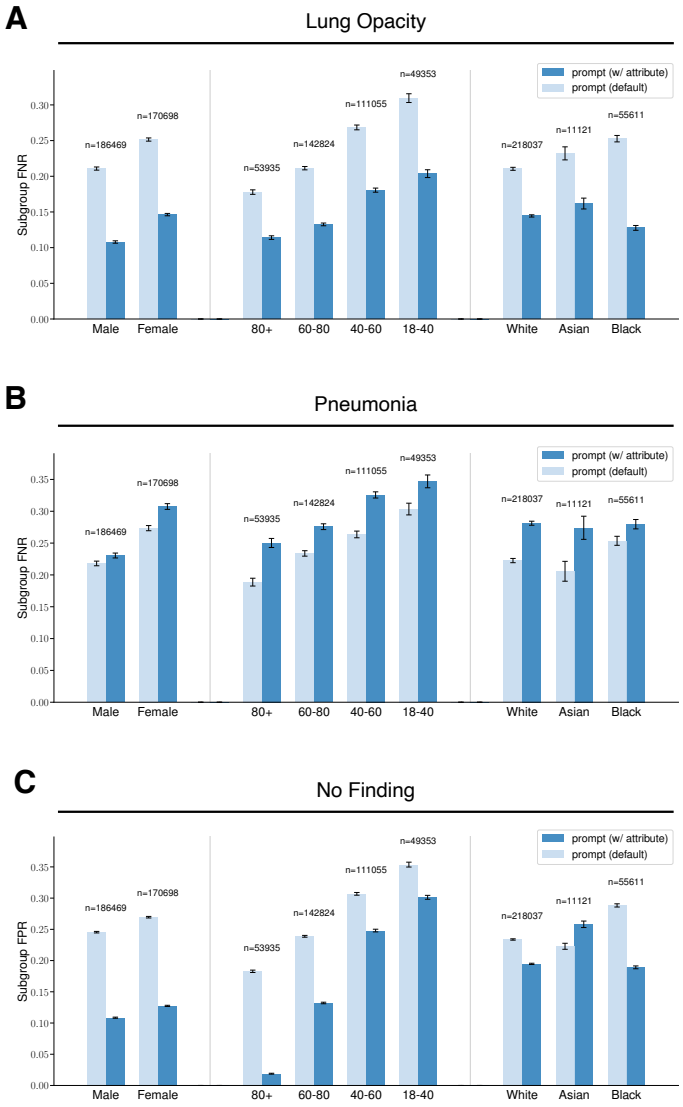

Figure B.7: **Model fairness intervention by incorporating demographic details into the input prompt. (A to C)** Performance across subgroups before and after introducing the sensitive demographic details into the prompt, for "Lung Opacity" (A), "Pneumonia" (B), and "No Finding" (C). We proposed to intervene the model prediction over subgroups by including demographic information in the input texts (e.g., "Does this female patient have Pneumonia?"). After this intervention, the model displays reduced demographic biases for certain conditions like "Lung Opacity", but not for others like "Pneumonia". Error bars indicate 95% confidence intervals estimated using non-parametric bootstrap sampling (n=1,000).