How Fair are Medical Imaging Foundation Models?

Muhammad Osama Khan OSAMA.KHAN@NYU.EDU
Muhammad Muneeb Afzal MUNEEB.AFZAL@NYU.EDU
Shujaat Mirza SHUJAAT.MIRZA@NYU.EDU
New York University, New York, USA
Yi Fang YFANG@NYU.EDU
Center for Artificial Intelligence and Robotics, New York University Abu Dhabi, Abu Dhabi, UAE

Abstract

While medical imaging foundation models have led to significant improvements across various tasks, the pivotal issue of subgroup fairness in these foundation models has remained largely unexplored. Our work bridges this research gap by presenting the first comprehensive study analyzing the subgroup fairness of six diverse foundation models, encompassing various pre-training methods, sources of pre-training data, and model architectures. In doing so, we discover a concerning trade-off: foundation models pre-trained on medical images achieve better overall performance but are consistently less fair than those pre-trained on natural images, with sometimes even worse fairness than baseline models trained from scratch. To mitigate these fairness disparities, we show that augmenting both the volume of pre-training data as well as the number of pre-training epochs, enhances subgroup fairness of medical imaging pre-trained models. Furthermore, to decouple the fairness bias from the pre-training and fine-tuning stages, we employ balanced datasets for fine-tuning. While fine-tuning on balanced datasets partially mitigates fairness issues, it is insufficient to completely eliminate the biases from the pre-training stage, prompting the need for careful design and evaluation of medical imaging foundation models. Our granular analysis reveals that medical imaging pre-trained models tend to favor majority racial subgroups (White, Asian) whereas natural imaging pre-trained models tend to favor minority racial subgroups (Black). Additionally, across all foundation models, we observe a consistent underperformance on the female patients cohort. As the community moves towards designing specialized foundation models for medical imaging, we hope our timely research provides crucial insights to help inform more equitable model development.

Keywords: Foundation models, fairness, self-supervised learning.

1. Introduction

Foundation models pre-trained on large diverse datasets excel at learning generalizable representations, which can be utilized for a wide array of downstream tasks. This exceptional adaptability has propelled them to set new performance benchmarks across multiple domains, including vision, language, and multi-modal applications (Alayrac et al., 2022; Lu et al., 2022; Brown et al., 2020; Kirillov et al., 2023). Within the medical imaging landscape, foundation models are particularly attractive owing to the inherent challenges in acquiring large task-specific datasets. These emerging medical foundation models are increasingly mitigating the need for large-scale labeled data while showing marked effectiveness in a diverse set of medical imaging tasks (Azizi et al., 2023; Ghesu et al., 2022; Sellergren et al., 2022).

Nevertheless, the question of biases potentially embedded within these medical foundation models remains largely unexplored and calls for more focused scrutiny. Existing medical foundation models have an undue focus on optimizing global performance metrics, which often disguises performance degradation for minority subgroups. In the realm of medical imaging, bias is a real-world problem that can exacerbate existing healthcare disparities (Cullen et al., 2022), leading to unequal treatments based on age, sex, ethnicity, or other protected attributes (Seyyed-Kalantari et al., 2020; Glocker et al., 2022). As medical imaging foundation models start to gain traction, it is critical to understand the effects of different types of pre-training strategies and pre-training datasets on encoding harmful biases in the foundation model, which can result in biased downstream models.

In this work, we conduct a comprehensive fairness analysis of six different foundation models, including the recently proposed REMEDIS framework (Azizi et al., 2023). Our foundation models encompass multiple factors, including pre-training methods (ranging from supervised learning to contrastive and masked self-supervised learning), sources of pre-training data (natural vs medical imaging), as well as the underlying model architectures (ViT vs ResNet). This multifaceted investigation provides valuable insights into the biases introduced by different pre-training strategies within medical foundation models.

Briefly, our main contributions include:

- We present the first comprehensive study evaluating the subgroup fairness of a wide range of medical imaging foundation models, encompassing various pre-training methods, sources of pre-training data, and model architectures.
- Our results reveal a concerning trade-off: foundation models pre-trained on medical images achieve better overall performance but are consistently less fair than their counterparts pre-trained on natural (i.e., non-medical) images, with sometimes even worse fairness than baseline models trained from scratch.
- We show that augmenting both the volume of pre-training data as well as the number of pre-training epochs enhances subgroup fairness in medical imaging pre-trained models.
- We demonstrate that fine-tuning on a balanced dataset, although beneficial, is insufficient to completely eliminate the biases from the pre-training stage, prompting the need for careful design and evaluation of medical imaging foundation models.
- Our granular analysis reveals that medical imaging pre-trained models tend to favor majority racial subgroups (White, Asian) whereas natural imaging pre-trained models tend to favor minority racial subgroups (Black).

As the community moves towards designing specialized foundation models for medical imaging, our timely research provides crucial insights to help inform more equitable model development. This is especially pertinent considering the substantial computational and time resources associated with training these large foundation models.

2. Related Work

Foundation models have demonstrated compelling performance across a wide range of tasks, encompassing vision and language domains (Alayrac et al., 2022; Lu et al., 2022; Brown et al., 2020; Kirillov et al., 2023). In the context of healthcare, medical foundation models have gained prominence (Azizi et al., 2023; Moor et al., 2023; Zhang and Metaxas, 2023; Ghesu et al., 2022; Sellergren et al., 2022; Rasmy et al., 2021; Korngiebel and Mooney, 2021), effectively alleviating the need for extensive labeled datasets by leveraging generalizable representations across diverse medical imaging tasks.

Self-supervised learning (SSL) is a popular technique for pre-training foundation models since it enables models to learn useful representations directly from unlabeled data. In recent years, multiple types of SSL have been developed, with contrastive and masked SSL gaining notable prominence. Contrastive SSL methods (He et al., 2020; Chen et al., 2020; Hadsell et al., 2006; Chen and He, 2021) are trained via pulling together the representations of similar images whereas pulling apart the representations of dissimilar images. Conversely, masked SSL methods (He et al., 2022; Dosovitskiy et al., 2021; Pathak et al., 2016) are trained by reconstructing occluded (masked) portions of an input image. Given the arduous task of acquiring large-scale labeled datasets in the medical imaging domain, SSL techniques have proven particularly efficacious for distilling representations from unlabeled medical imaging datasets (Azizi et al., 2023; Haghighi et al., 2022; Taher et al., 2022; Azizi et al., 2021; Zhou et al., 2021b; Haghighi et al., 2021; Zhou et al., 2021a; Chaitanya et al., 2020; Tao et al., 2020).

Prior studies (Seyyed-Kalantari et al., 2021; Puyol-Antón et al., 2022; Larrazabal et al., 2020; Stanley et al., 2022; Jones et al., 2023) have demonstrated performance disparities in task-specific models across various protected attributes such as age, sex, and race. While medical foundation models demonstrate superior overall performance, it is unclear how these foundation models compare against task-specific models in terms of fairness (Thieme et al., 2023; Wójcik, 2022).

Table 1: Overview of Foundation Models used in this study describing the pre-training methods, pre-training datasets, and model architectures. (M): Medical.

		Pre-training			
Model	**Arch.**	**Method**	**Type**	**Data**	**Datasets**
MAE	ViT-B	Self-supervised	Masked	Natural	ImageNet-1K
MAE (M)	ViT-B	Self-supervised	Masked	Medical	ChestXray14, CheXpert, MIMIC-CXR
MoCov3	ViT-B	Self-supervised	Contrastive	Natural	ImageNet-1K
MoCov3 (M)	ViT-B	Self-supervised	Contrastive	Medical	ChestXray14, CheXpert
BiT	ResNet152	Supervised	–	Natural	ImageNet-21K
REMEDIS	ResNet152	Self-supervised	Contrastive	Medical	CheXpert, MIMIC-CXR

Recently, Glocker et al. (2022) observed performance disparities across protected subgroups when evaluating a medical imaging foundation model. However, different from their approach, which was limited to a single foundation model trained via supervised contrastive learning, our work marks the first comprehensive study of a wide range of medical imaging foundation models, spanning multiple pre-training techniques, data sources, and architectural frameworks. Moreover, since the foundation model used by Glocker et al. (2022) was not publicly available, their analysis was restricted to using the foundation model as a fixed feature extractor. In contrast, we fine-tune all our foundation models end-to-end, which is the most common scenario when utilizing foundation models. In addition to improving overall performance, end-to-end fine-tuning also allows us to conduct a comprehensive analysis of the influence of fine-tuning on the latent biases embedded within these foundation models.

3. Methods

3.1. Foundation Models & Pre-training

In this study, we perform a rigorous fairness analysis across six diverse foundation models: MAE (He et al., 2022), Medical MAE (Xiao et al., 2023), MoCov3 (Chen et al., 2021), Medical MoCov3, BiT (Kolesnikov et al., 2020), and REMEDIS (Azizi et al., 2023). Our evaluation encompasses multiple factors, including the type of pre-training method (supervised vs. self-supervised), the source of pre-training data (natural vs. medical imaging), as well as the underlying model architecture (ViT vs. ResNet). Wherever possible, we use the publicly available foundation models. However, we also pre-train some foundation models ourselves in order to ensure a fair comparison among different foundation models. A detailed breakdown of the various foundation models used in this study is presented in Table 1.

In terms of pre-training data, MAE, MoCov3, and BiT are pre-trained on natural images, whereas Medical MAE, Medical MoCov3, and REMEDIS are pre-trained on medical images, with the exact datasets delineated in Table 1. Moreover, we also stratify the foundation models according to the type of pre-training algorithm – BiT is pre-trained via supervised learning, MAE and Medical MAE are pre-trained via masked SSL whereas MoCov3, Medical MoCov3, and REMEDIS are pre-trained via contrastive SSL.

We use the publicly available versions of MAE, Medical MAE, MoCov3, and BiT as well as the recently proposed REMEDIS. Since a foundation model based on MoCov3 has not been developed for medical imaging, we create Medical MoCov3 via pre-training on the ChestXray14 (Wang et al., 2017) and CheXpert (Irvin et al., 2019) datasets. Moreover, in order to study the effect of increasing medical imaging pre-training in Section 4.2, we also pre-train MAE on various medical imaging datasets of increasing size, as outlined in Table 2. For fair comparison, we utilize a ViT-B/16 encoder for both MoCov3 and MAE. Moreover, we follow the same augmentation strategies employed by Hosseinzadeh Taher et al. (2021) to pre-train both methods. These augmentations consist of initial cropping and resizing to 224×224, random horizontal flipping with a probability of 0.5, and random rotation within the range of -7 to +7 degrees. We utilize the same pre-training configurations specified in the official papers (Chen et al., 2021; He et al., 2022), with both self-supervised models trained on 8 V100 GPUs for 800 epochs. The model checkpoint exhibiting the lowest self-supervised loss during the final 5% of epochs is selected for subsequent fine-tuning.

3.2. Fine-tuning

We fine-tune the aforementioned foundation models on the multi-label classification task of identifying various pathologies from chest radiographs. We adopt the same fine-tuning settings as Medical MAE (Xiao et al., 2023) and fine-tune each foundation model on a single V100 GPU. Concretely, we use an AdamW optimizer, with hyperparameters β_1, β_2, and weight decay set to 0.9, 0.95 and 0.05 respectively. The base learning rate is configured at 2.5e-4, with a warm-up phase spanning 5 epochs, followed by subsequent cosine annealing. Moreover, we use a layer-wise lr decay of 0.55, RandAug magnitude of 6 and DropPath rate of 0.2. We use a batch size of 32 for the ResNet152 models (BiT and REMEDIS) whereas a batch size of 128 for the ViT-B models (MAE, Medical MAE, MoCov3, and Medical MoCov3). Since the medical imaging pre-trained models converge faster than their natural imaging pre-trained counterparts, we fine-tune Medical MAE, Medical MoCov3, and REMEDIS for 100, 100, and 20 epochs, respectively, as opposed to 200, 200, and 30 epochs, respectively, for MAE, MoCov3, and BiT.

3.3. Dataset Splits

Fine-tuning is conducted on a subset of the CheXpert (Irvin et al., 2019) dataset, employing the same dataset splits as those used in Glocker et al. (2022) and Gichoya et al. (2022). This subset comprises 127,118 Chest X-ray scans from 42,884 patients and is partitioned into training (76,205), validation (12,673), and test (38,240) sets, with no patient overlap between the different splits. Following Glocker et al. (2022), we resample the test set to ensure balanced demographic representation. Moreover, since it is a multi-label classification problem, we create a separate test set for each disease in order to ensure equal prevalence of all diseases across all subgroups.

3.4. Metrics

We evaluate all the foundation models via two sets of metrics – one focused on performance whereas the other focused on fairness. For performance, we utilize AUC and report the mean AUC across all 14 pathologies. For assessing fairness, following prior work (Ktena et al., 2023), we report the fairness gap, defined as the difference in AUC between the worst- and best-performing subgroups. We report the fairness gaps individually across both sex and race.

For clarity, we convert fairness gaps (FG) to fairness scores (FS) when plotting, such that higher values are better for both the performance as well as the fairness subplots.

$$FS = C - FG, \qquad C = \max\{FG\}_{i=1}^{N} + \epsilon \tag{1}$$

In order to avoid a fairness score of 0, which might misleadingly suggest a completely unfair model, we set ϵ to 1. Effectively, this transforms the maximum fairness gap to correspond with the minimum fairness score of 1.

4. Results

We conduct a comprehensive set of experiments to understand the subgroup fairness of the aforementioned medical imaging foundation models. Firstly, Section 4.1 studies the impact of pre-training methods and data sources on classification performance and fairness. Next, Section 4.2 delves into the relationship between increasing medical imaging pre-training and subgroup fairness. This is then followed by Section 4.3, which investigates the impact of fine-tuning these foundation models on balanced datasets to decouple the fairness biases from the pre-training and fine-tuning stages. Section 4.4 provides a fine-grained analysis of individual subgroups to understand which groups are advantaged or disadvantaged. Lastly, Section 4.5 concludes by investigating various ensembling strategies to leverage the complementary strengths of multiple foundation models.

4.1. Performance and Fairness

In this section, we compare multiple foundation models in terms of both performance and fairness. For clarity, we focus on the four foundation models that utilize the same ViT-B architecture – MAE and MoCov3 pre-trained in a self-supervised fashion on natural as well as medical imaging datasets. Additional analyses involving the BiT and REMEDIS models, which adopt a ResNet152 architecture and are trained via supervised and hybrid supervised + self-supervised paradigms respectively, are available in Appendix A.

Figure 1 provides a comprehensive evaluation of classification performance as well as fairness across

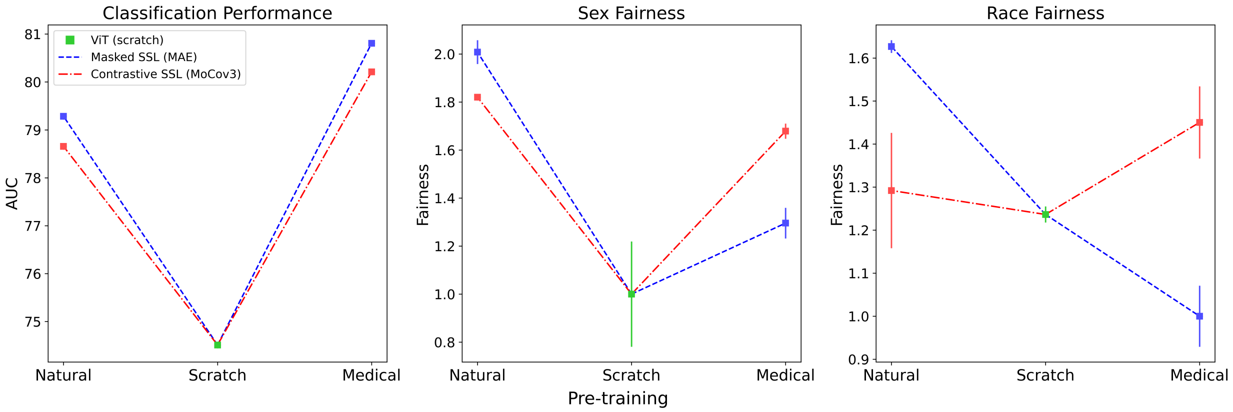

Figure 1: Classification performance and fairness metrics for foundation models pre-trained on either natural or medical images, benchmarked against baseline models initialized from scratch. Higher is better for both the AUC and Fairness subplots.

both gender and race for the aforementioned foundation models. As outlined in Section 3, all models are fine-tuned on the CheXpert dataset but vary in terms of their pre-training techniques, amounts of pre-training data, and data sources. A holistic observation reveals the benefits of pre-training, enhancing both classification performance and fairness relative to models initialized from scratch. Fine-grained insights about the effects of different pre-training datasets and SSL techniques are presented in the subsequent analysis.

4.1.1. Natural vs Medical Imaging Pre-training

Our evaluation demonstrates substantial improvement in classification performance for natural as well as medical imaging pre-trained models in contrast to those initialized from scratch (cf. Classification Performance subfigure in Figure 1). Consistent with the SSL literature (Haghighi et al., 2022), we observe that medical imaging pre-training yields a greater boost in Chest X-ray classification performance compared to natural imaging pre-training.

Next, we study the interplay of natural versus medical imaging pre-training on the fairness of these foundation models across sex and race. Across both protected attributes, we observe that foundation models pre-trained on natural images exhibit the best fairness. In contrast, foundation models pre-trained on medical images are consistently worse in terms of fairness than their counterparts pre-trained on natural images. In fact, medical imaging pre-training can sometimes even exacerbate racial disparities in comparison to a baseline model initialized from scratch (cf. Race Fairness subfigure in Figure 1).

Additionally, we note that models pre-trained on medical datasets tend to perform better on sex fairness in comparison to race fairness. This observation may be attributed to the data imbalances in the medical datasets; for instance, the CheXpert dataset manifests a 59/41 gender split (Male/Female) and a highly skewed 78/15/7 racial split (White/Asian/Black).[1]

4.1.2. Contrastive vs Masked SSL

In terms of classification performance, we note that masked SSL consistently outperforms contrastive SSL across both natural and medical imaging pre-training paradigms (cf. Classification Performance subfigure in Figure 1). This observation aligns with existing literature (He et al., 2022; Khan and Fang, 2023), which has shown that masked pre-training methods generally exhibit superior performance over their contrastive counterparts.

In terms of subgroup fairness, masked SSL exhibits improved fairness than contrastive SSL across both sex and race when pre-trained on natural images. Conversely, when pre-trained on medical images, contrastive SSL yields fairer models, with masked SSL

[1] For a detailed breakdown of the various datasets, please refer to Tables 1 of Seyyed-Kalantari et al. (2020) and Glocker et al. (2022).

even underperforming a baseline model initialized from scratch in terms of race fairness. This indicates that masked pre-training is particularly susceptible to subgroup imbalances in the medical pre-training data. We hypothesize that contrastive pre-training is able to learn better representations regardless of the subgroup affiliation since it effectively treats each image as a separate class, thereby contrasting it not only against images from other subgroups but also against different images from within the same subgroup. Consequently, contrastive SSL is not as adversely affected by the highly skewed racial splits in medical pre-training datasets as masked SSL (cf. Race Fairness subfigure in Figure 1).

4.2. Amount of Medical Imaging Pre-training

In this section, we study the impact of increased medical imaging pre-training on model performance and fairness. Firstly, we investigate the effect of increasing the volume of pre-training data. To this end, we pre-train MAE on ChestXray14, CheXpert as well as the combined (ChestXray14 + CheXpert) datasets following the official MAE (He et al., 2022) pre-training configurations.

As illustrated by the results in Table 2, increasing the amount of pre-training data (ChestXray14 < CheXpert < ChestXray14 + CheXpert) not only improves the overall classification performance but also enhances subgroup fairness across sex and race. This resonates with the insights of Seyyed-Kalantari et al. (2020), who observed reduced bias when using multiple datasets in a supervised learning context. It is intriguing that a similar observation also holds true for pre-training via self-supervised learning, which does not use any labels during pre-training.

Tables 4 and 5 (in Appendix C) study the impact of increasing pre-training epochs and dataset fractions, respectively, on model performance and fairness. Once again, we observe that increased pre-training improves both overall performance as well as subgroup fairness.

Hence, within the realm of medical imaging pre-training, we conclude that amplifying the volume of pre-training data as well as the number of pre-training epochs leads to favorable outcomes for both overall performance and subgroup fairness.

4.3. Balanced Fine-tuning

Although we have observed significant differences in fairness of the pre-trained foundation models, it remains unclear if this fairness gap stems from the pre-training or fine-tuning stages. In order to decouple the fairness bias from pre-training and fine-tuning, we fine-tune these foundation models on a balanced fine-tuning dataset. Hence, any remaining fairness discrepancies can be primarily attributed to the pre-training stage.

Table 2: Impact of increasing pre-training data on model performance and fairness. All models are pre-trained via MAE. Higher AUC and lower Fairness Gaps are desirable. CX14: ChestXray14, CXPT: CheXpert.

Pre-training Data	Classif.	Fairness Gap	
	AUC ↑	Sex ↓	Race ↓
CX14	79.97	1.58	3.09
CXPT	80.98	1.49	2.85
CX14 + CXPT	81.38	1.29	2.82

Since the original training set (cf. Section 3.3) is highly skewed across racial attributes, we construct a balanced set by resampling it so that each of the racial subgroups have the same number of instances as the minority subgroup. Since the majority subgroup (White) is the most unbalanced in terms of sex, via undersampling from this group, we also ensure that the resulting dataset is more balanced across sex.

Figure 2 reports the overall classification performance and subgroup fairness of fine-tuning on this balanced dataset. While the absolute values between Figures 1 and 2 are not directly comparable due to the reduced dataset size, the relative trends in fairness underscore the positive impact of dataset rebalancing. In particular, all foundation models now exhibit improved racial fairness relative to the baseline model initialized from scratch (cf. Race Fairness subfigures in Figures 1 and 2). We observe that balanced fine-tuning especially benefits masked SSL pre-trained on medical images since this class of pre-training methods is particularly susceptible to subgroup imbalances in the medical pre-training data (cf. Section 4.1).

However, despite these adjustments in the fine-tuning stage, we note that foundation models pre-trained on natural images continue to outperform those pre-trained on medical images in most cases. This illustrates that fine-tuning on a balanced dataset, although beneficial, is insufficient to completely eliminate the biases from the pre-training

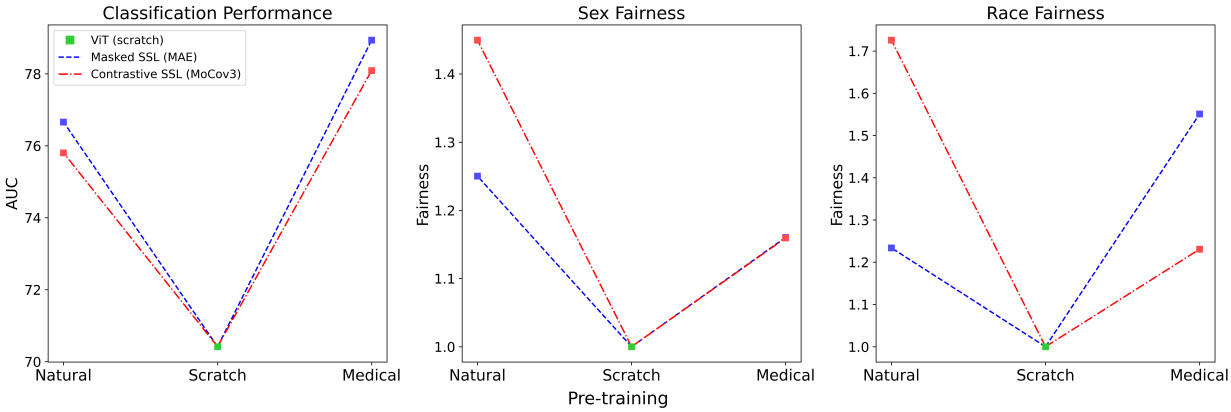

Figure 2: Impact of balanced fine-tuning on performance and fairness of medical foundation models. Note: MAE and MoCov3 pre-trained on medical images exhibit overlapping performance in the Sex Fairness subplot.

stage, prompting the need for careful design and evaluation of medical imaging foundation models.

For a more equitable pre-training regimen, one could consider training foundation models on balanced pre-training datasets in order to mitigate harmful biases encoded during the pre-training stage. However, this strategy may come with the drawback of diminished model performance, given the reduced size of pre-training data. As discussed in Section 4.2, an alternative pathway to enhance subgroup fairness involves collecting multiple datasets from various sites, thereby increasing both the pre-training data volume as well as the patient diversity in the pre-training datasets. Overall, it is important to rigorously evaluate all approaches to ascertain the optimal trade-off between overall model performance and subgroup fairness.

4.4. Analysis: Pathologies and Protected Attributes

In this section, we present a fine-grained analysis of the fairness of these foundation models across the individual sex and race subgroups to understand which subgroups are advantaged or disadvantaged. Figure 3 presents a comprehensive analysis of average performance across all 14 diseases segregated by individual sex (Male, Female) and race (White, Asian, Black) subgroups. Concretely, for each of these subgroups, we compute the relative change in performance for that subgroup compared against the average model performance across all subgroups. Hence, positive values indicate that the model overperforms on that subgroup compared to the overall population and vice versa.

Across the sex categories, we observe a consistent underperformance of all six foundation models on the female patients subgroup. Whereas the under-representation of female patients in the pre-training datasets could partially explain this phenomenon, it is not clear if this data imbalance is the sole factor responsible for this consistent underperformance, as we discuss below.

Recalling the racial distribution in the CheXpert dataset as 78/15/7 (White/Asian/Black), our analysis shows that medical imaging pre-trained models result in improved performance of the more frequent subgroups (White, Asian) compared to the overall population performance. Conversely, models pre-trained on natural images result in enhanced performance of the minority subgroup (Black). We hypothesize that this improved performance on the Black patients subgroup is potentially due to the higher prevalence of some diseases in this subgroup (e.g., out of the 11 diseases with less than 30% prevalence, 5 of them have the highest prevalence in the Black patients cohort). However, this improvement is diminished when pre-training on medical datasets with highly skewed racial splits, which tend to disproportionately favor the majority subgroups.

In addition to the average performances across subgroups, we also report the individual performances on each pathology, stratified by race and gender in Fig-

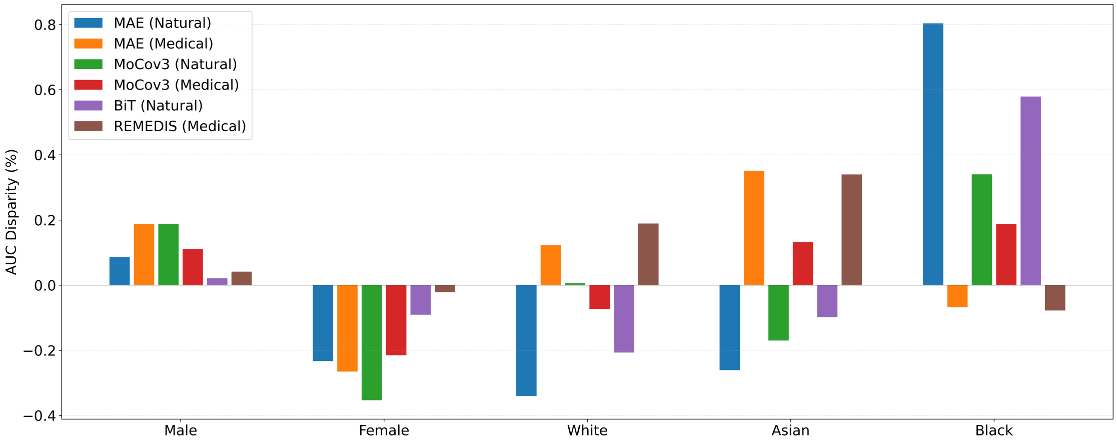

Figure 3: Change in performance of individual subgroups relative to the average performance across the entire population. The $y = 0$ line represents the average performance across the entire population.

ures 5 and 6 respectively (Appendix B). For clarity, we focus on the five pathologies - Atelectasis, Cardiomegaly, Consolidation, Edema, and Pleural Effusion. From the disease stratification by sex (Figure 5), we note that most diseases are diagnosed better on male patients. However, an interesting anomaly is Pleural Effusion, which is consistently diagnosed better on female patients despite similar prevalences of Pleural Effusion in both male and female patients in the CheXpert dataset. From the disease stratification by race (Figure 6), we observe that for most pathologies, all six foundation models either consistently overperform or consistently underperform on that racial subgroup compared to the average model performance across all subgroups. Lastly, we note that Edema is the most fairly diagnosed condition, with the least disparity in classification performance across both sex and race subgroups.

4.5. Foundation Model Ensembles

In this section, we examine whether different categories of foundation models offer complementary benefits that can be effectively leveraged through ensembling techniques. As delineated in Table 3, we construct ensembles based on the architecture-SSL combination, the pre-training data domain, as well as a final comprehensive ensemble that incorporates all foundation models. Empirically, we observe that the ensemble of foundation models pre-trained on medical images achieves the best overall performance. On the other hand, the ensemble comprising foundation models pre-trained on natural images excels in fairness across sex attributes. Overall, the ensemble integrating all foundation models strikes a balanced trade-off between performance and fairness across both gender and racial dimensions, signifying that ensembling could serve as an effective strategy for leveraging the heterogeneous strengths of multiple foundation models.

Table 3: Foundation model ensembles. Best result is in boldface, second best is underlined.

Ensemble	Classif.	Fairness Gap	
	AUC ↑	Sex ↓	Race ↓
ViT (Masked)	81.0	1.21	**2.82**
ViT (Contrastive)	80.3	1.13	2.90
ResNet (Contrastive)	<u>81.6</u>	1.15	2.95
Natural Imaging	80.4	**0.78**	2.94
Medical Imaging	**81.6**	1.39	3.12
All Foundation Models	81.4	<u>1.12</u>	<u>2.84</u>

4.6. Limitations and Future Work

With medical imaging foundation models becoming increasingly popular, it is essential to study the fairness of these models in order to understand their impact on human health. Our study marks an initial step in this direction. Although we conduct a comprehensive fairness analysis of multiple foundation models, there are several avenues for future work. A key improvement would be the examination of more specific subgroups. For instance, treating all Asian ethnicities as one group – a consequence of the generic labels provided by most datasets – is unfortunate. The release of publicly accessible datasets with more granular ethnic classifications would significantly enhance the comprehensiveness of future fairness studies. Additionally, future work could investigate other metrics that capture various notions of fairness. Lastly, while this paper focuses on the downstream task of Chest X-ray diagnosis, it would be an interesting direction for future research to explore how these foundation models impact fairness across different tasks, especially when transferring across different data modalities (e.g., from X-ray to Fundus images) or different anatomical regions (e.g., from Chest X-rays to Knee X-rays).

5. Conclusion

We present the first comprehensive study analyzing the subgroup fairness of six diverse foundation models across the protected attributes of sex and race. Our findings reveal a concerning trade-off: medical imaging pre-trained models excel in overall performance but are consistently worse in fairness compared to their natural imaging pre-trained counterparts. We show that fine-tuning on balanced datasets only partially mitigates these fairness discrepancies, underscoring the need for careful design and evaluation of medical imaging foundation models. Our fine-grained analysis further reveals consistent underperformance for female patients across all foundation models. Moreover, we find that medical imaging pre-trained models favor majority racial subgroups, while natural imaging pre-trained models favor minority subgroups. As the community moves towards designing specialized foundation models for medical imaging, we hope our timely research provides crucial insights to help inform more equitable model development.

References

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736, 2022.

Shekoofeh Azizi, Basil Mustafa, Fiona Ryan, Zachary Beaver, Jan Freyberg, Jonathan Deaton, Aaron Loh, Alan Karthikesalingam, Simon Kornblith, Ting Chen, et al. Big self-supervised models advance medical image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3478–3488, 2021.

Shekoofeh Azizi, Laura Culp, Jan Freyberg, Basil Mustafa, Sebastien Baur, Simon Kornblith, Ting Chen, Nenad Tomasev, Jovana Mitrović, Patricia Strachan, et al. Robust and data-efficient generalization of self-supervised machine learning for diagnostic imaging. *Nature Biomedical Engineering*, pages 1–24, 2023.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Krishna Chaitanya, Ertunc Erdil, Neerav Karani, and Ender Konukoglu. Contrastive learning of global and local features for medical image segmentation with limited annotations. *Advances in Neural Information Processing Systems*, 33:12546–12558, 2020.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15750–15758, 2021.

Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. *CoRR*, abs/2104.02057, 2021.

Mark R Cullen, Adina R Lemeshow, Leo J Russo, David M Barnes, Yaa Ababio, and Aida Habtezion. Disease-specific health disparities: a targeted review focusing on race and ethnicity. In *Healthcare*, volume 10, page 603. MDPI, 2022.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias

Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

Florin C Ghesu, Bogdan Georgescu, Awais Mansoor, Youngjin Yoo, Dominik Neumann, Pragneshkumar Patel, RS Vishwanath, James M Balter, Yue Cao, Sasa Grbic, et al. Self-supervised learning from 100 million medical images. *arXiv preprint arXiv:2201.01283*, 2022.

Judy Wawira Gichoya, Imon Banerjee, Ananth Reddy Bhimireddy, John L Burns, Leo Anthony Celi, Li-Ching Chen, Ramon Correa, Natalie Dullerud, Marzyeh Ghassemi, Shih-Cheng Huang, et al. Ai recognition of patient race in medical imaging: a modelling study. *The Lancet Digital Health*, 4(6):e406–e414, 2022.

Ben Glocker, Charles Jones, Melanie Bernhardt, and Stefan Winzeck. Risk of bias in chest x-ray foundation models. *arXiv preprint arXiv:2209.02965*, 2022.

Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 1735–1742. IEEE, 2006.

Fatemeh Haghighi, Mohammad Reza Hosseinzadeh Taher, Zongwei Zhou, Michael B Gotway, and Jianming Liang. Transferable visual words: Exploiting the semantics of anatomical patterns for self-supervised learning. *IEEE transactions on medical imaging*, 40(10):2857–2868, 2021.

Fatemeh Haghighi, Mohammad Reza Hosseinzadeh Taher, Michael B Gotway, and Jianming Liang. Dira: Discriminative, restorative, and adversarial learning for self-supervised medical image analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20824–20834, 2022.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022.

Mohammad Reza Hosseinzadeh Taher, Fatemeh Haghighi, Ruibin Feng, Michael B Gotway, and Jianming Liang. A systematic benchmarking analysis of transfer learning for medical image analysis. In *Domain Adaptation and Representation Transfer, and Affordable Healthcare and AI for Resource Diverse Global Health: Third MICCAI Workshop, DART 2021, and First MICCAI Workshop, FAIR 2021, Held in Conjunction with MICCAI 2021, Strasbourg, France, September 27 and October 1, 2021, Proceedings 3*, pages 3–13. Springer, 2021.

Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. 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.

Charles Jones, Mélanie Roschewitz, and Ben Glocker. The role of subgroup separability in group-fair medical image classification. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 179–188. Springer, 2023.

Muhammad Osama Khan and Yi Fang. Revisiting fine-tuning strategies for self-supervised medical imaging analysis. *arXiv preprint arXiv:2307.10915*, 2023.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023.

Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Joan Puigcerver, Jessica Yung, Sylvain Gelly, and Neil Houlsby. Big transfer (bit): General visual representation learning. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16*, pages 491–507. Springer, 2020.

Diane M Korngiebel and Sean D Mooney. Considering the possibilities and pitfalls of generative pretrained transformer 3 (gpt-3) in healthcare delivery. *NPJ Digital Medicine*, 4(1):93, 2021.

Ira Ktena, Olivia Wiles, Isabela Albuquerque, Sylvestre-Alvise Rebuffi, Ryutaro Tanno, Abhijit Guha Roy, Shekoofeh Azizi, Danielle Belgrave, Pushmeet Kohli, Alan Karthikesalingam, et al. Generative models improve fairness of medical classifiers under distribution shifts. *arXiv preprint arXiv:2304.09218*, 2023.

Agostina J Larrazabal, Nicolás Nieto, Victoria Peterson, Diego H Milone, and Enzo Ferrante. Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis. *Proceedings of the National Academy of Sciences*, 117 (23):12592–12594, 2020.

Jiasen Lu, Christopher Clark, Rowan Zellers, Roozbeh Mottaghi, and Aniruddha Kembhavi. Unified-io: A unified model for vision, language, and multi-modal tasks. *arXiv preprint arXiv:2206.08916*, 2022.

Michael Moor, Oishi Banerjee, Zahra Shakeri Hossein Abad, Harlan M Krumholz, Jure Leskovec, Eric J Topol, and Pranav Rajpurkar. Foundation models for generalist medical artificial intelligence. *Nature*, 616(7956):259–265, 2023.

Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2536–2544, 2016.

Esther Puyol-Antón, Bram Ruijsink, Jorge Mariscal Harana, Stefan K Piechnik, Stefan Neubauer, Steffen E Petersen, Reza Razavi, Phil Chowienczyk, and Andrew P King. Fairness in cardiac magnetic resonance imaging: assessing sex and racial bias in deep learning-based segmentation. *Frontiers in cardiovascular medicine*, 9: 859310, 2022.

Laila Rasmy, Yang Xiang, Ziqian Xie, Cui Tao, and Degui Zhi. Med-bert: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. *NPJ digital medicine*, 4(1):86, 2021.

Andrew B Sellergren, Christina Chen, Zaid Nabulsi, Yuanzhen Li, Aaron Maschinot, Aaron Sarna, Jenny Huang, Charles Lau, Sreenivasa Raju Kalidindi, Mozziyar Etemadi, et al. Simplified transfer learning for chest radiography models using less data. *Radiology*, 305(2):454–465, 2022.

Laleh Seyyed-Kalantari, Guanxiong Liu, Matthew McDermott, Irene Y Chen, and Marzyeh Ghassemi. Chexclusion: Fairness gaps in deep chest x-ray classifiers. In *BIOCOMPUTING 2021: proceedings of the Pacific symposium*, pages 232–243. World Scientific, 2020.

Laleh Seyyed-Kalantari, Haoran Zhang, Matthew BA McDermott, Irene Y Chen, and Marzyeh Ghassemi. Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in underserved patient populations. *Nature medicine*, 27 (12):2176–2182, 2021.

Emma AM Stanley, Matthias Wilms, and Nils D Forkert. Disproportionate subgroup impacts and other challenges of fairness in artificial intelligence for medical image analysis. In *Workshop on the Ethical and Philosophical Issues in Medical Imaging*, pages 14–25. Springer, 2022.

Mohammad Reza Hosseinzadeh Taher, Fatemeh Haghighi, Michael B Gotway, and Jianming Liang. Caid: Context-aware instance discrimination for self-supervised learning in medical imaging. In *International Conference on Medical Imaging with Deep Learning*, pages 535–551. PMLR, 2022.

Xing Tao, Yuexiang Li, Wenhui Zhou, Kai Ma, and Yefeng Zheng. Revisiting rubik's cube: self-supervised learning with volume-wise transformation for 3d medical image segmentation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part IV 23*, pages 238–248. Springer, 2020.

Anja Thieme, Aditya Nori, Marzyeh Ghassemi, Rishi Bommasani, Tariq Osman Andersen, and Ewa Luger. Foundation models in healthcare: Opportunities, risks & strategies forward. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–4, 2023.

Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald 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)*, July 2017.

Malwina Anna Wójcik. Foundation models in healthcare: Opportunities, biases and regulatory prospects in europe. In *International Conference on Electronic Government and the Information Systems Perspective*, pages 32–46. Springer, 2022.

Junfei Xiao, Yutong Bai, Alan Yuille, and Zongwei Zhou. Delving into masked autoencoders for multi-label thorax disease classification. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3588–3600, 2023.

Shaoting Zhang and Dimitris Metaxas. On the challenges and perspectives of foundation models for medical image analysis. *arXiv preprint arXiv:2306.05705*, 2023.

Hong-Yu Zhou, Chixiang Lu, Sibei Yang, Xiaoguang Han, and Yizhou Yu. Preservational learning improves self-supervised medical image models by reconstructing diverse contexts. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3499–3509, 2021a.

Zongwei Zhou, Vatsal Sodha, Jiaxuan Pang, Michael B Gotway, and Jianming Liang. Models genesis. *Medical image analysis*, 67:101840, 2021b.

Appendix A. ResNet Foundation Models

Figure 4 compares the ResNet152 foundation models – BiT and REMEDIS – in terms of both performance and fairness. BiT is pre-trained in a supervised manner whereas REMEDIS is pre-trained via supervised as well as contrastive self-supervised learning. Similar to the analysis presented in Section 4.1, we note that the natural imaging pre-trained model (BiT) exhibits better subgroup fairness across both sex and race than the medical imaging pre-trained model (REMEDIS).

Appendix B. Disease Disparities Across Sex and Race

In this section, we report the individual performances on each pathology, stratified by race and gender, in Figures 5 and 6 respectively. For a comprehensive discussion of these results, please refer to Section 4.4.

Appendix C. Amount of Medical Imaging Pre-training

Tables 4 and 5 study the impact of increasing pre-training epochs and dataset fractions, respectively, on the performance and fairness of medical foundation models. Similar to the results presented in Section 4.2, we observe that increased pre-training improves both overall performance as well as subgroup fairness.

Table 4: Impact of increasing pre-training epochs on model performance and fairness. Higher AUC and lower Fairness Gaps are desirable.

Epochs	Classif.	Fairness Gap	
	AUC ↑	Sex ↓	Race ↓
200	79.2	1.54	3.56
400	79.9	1.44	3.11
800	79.9	1.43	3.03

Table 5: Impact of increasing pre-training dataset fractions on model performance and fairness. Higher AUC and lower Fairness Gaps are desirable. CXPT: CheXpert.

Dataset	Classif.	Fairness Gap	
	AUC ↑	Sex ↓	Race ↓
CXPT (30%)	80.6	1.53	2.86
CXPT (100%)	81.0	1.49	2.85

Appendix D. Fine-tuning Strategies

Self-supervised foundation models are generally fine-tuned either via a linear probing setup or an end-to-end fine-tuning setup. In practice, end-to-end fine-tuning is the most commonly utilized fine-tuning technique since it allows for much better performance compared to linear probing. Hence, we primarily focus on end-to-end fine-tuning in this paper. However, for completeness, we also provide the linear probing results in Table 6, which shows that fine-tuning improves both overall performance as well as fairness in comparison to linear probing.

Table 6: Impact of fine-tuning strategy on model performance and fairness. Higher AUC and lower Fairness Gaps are desirable. LP: Linear probing, FT: Fine-tuning. FG: Fairness Gap.

Method	AUC ↑		FG (Sex) ↓		FG (Race) ↓	
	LP	FT	LP	FT	LP	FT
MoCo-v3	71.9	78.7	1.56	1.08	3.66	3.00
MAE	66.7	79.3	2.59	0.92	3.00	2.77

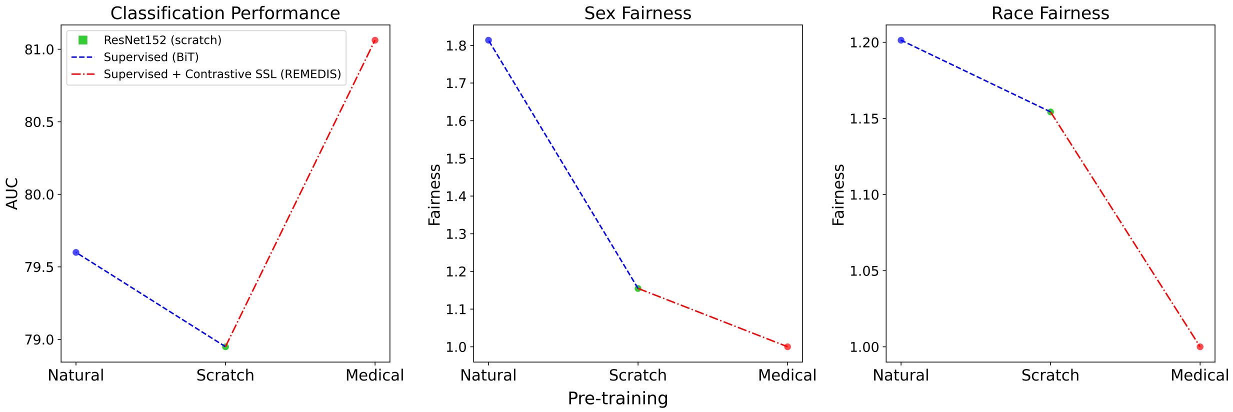

Figure 4: Classification performance and fairness metrics for ResNet foundation models pre-trained on either natural or medical images, benchmarked against a baseline model initialized from scratch. Higher is better for both the AUC and Fairness subplots. The natural imaging pre-trained model (BiT) exhibits better subgroup fairness across both sex and race than the medical imaging pre-trained model (REMEDIS).

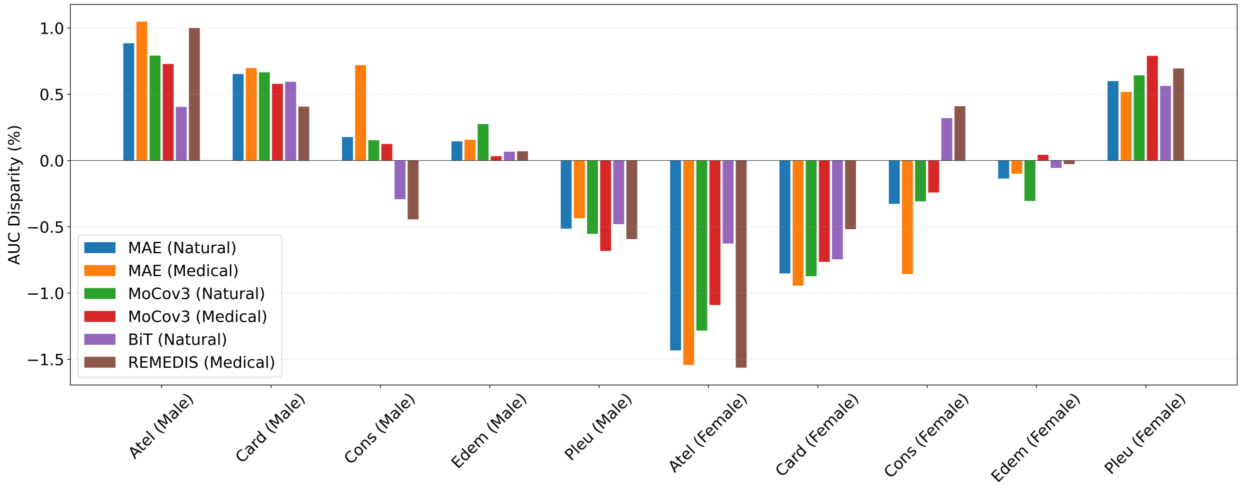

Figure 5: Change in performance of each individual disease, segregated by sex, in relation to the average performance for that disease across the entire population. The $y = 0$ line represents the average performance for that disease across the entire population. Atel: Atelectasis, Card: Cardiomegaly, Cons: Consolidation, Edem: Edema, Pleu: Pleural Effusion.

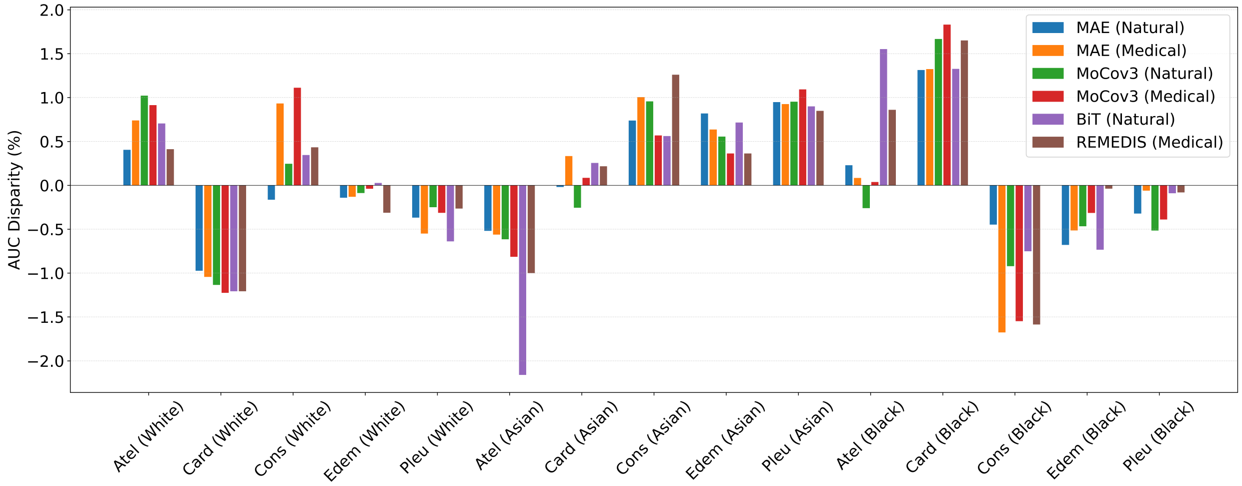

Figure 6: Change in performance of each individual disease, segregated by race, in relation to the average performance for that disease across the entire population. The $y = 0$ line represents the average performance for that disease across the entire population. Atel: Atelectasis, Card: Cardiomegaly, Cons: Consolidation, Edem: Edema, Pleu: Pleural Effusion.