Underdiagnosis Bias of Chest Radiograph Diagnostic AI can be Decomposed and Mitigated via Dataset Bias Attributions

AUTHORS

Yuta Kobayashi [1], Haoran Zhang [2], Harvineet Singh[3], Judy Wawira Gichoya[4], Shalmali Joshi [1],

[1]Department of Biomedical Informatics, Columbia University, NY, USA

[2]Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, MA, USA

[3]Department of Epidemiology and Biostatistics, University of California, San Francisco, CA, USA

[4]Department of Radiology and Imaging Sciences, Emory University, GA, USA

ABSTRACT

Inequitable diagnostic accuracy is a broad concern in AI-based models. However, current characterizations of bias are narrow, and fail to account for systematic bias in upstream data-collection, thereby conflating observed inequities in AI performance with biases due to distributional differences in the dataset itself. This gap has broad implications, resulting in ineffective bias-mitigation strategies. We introduce a novel retrospective model evaluation procedure that identifies and characterizes the contribution of distributional differences across protected groups that explain population-level diagnostic disparities. Across three large-scale chest radiography datasets, we consistently find that distributional differences in age and confounding image attributes (such as pathology type and size) contribute to poorer model performance across racial subgroups. By systematically attributing observed underdiagnosis bias to distributional differences due to biases in the data-acquisition process, or dataset biases, we present a general approach to disentangling how different types of dataset biases interact and compound to create observable AI performance disparities. Our method is actionable to aid the design of targeted interventions that recalibrate foundation models to specific subpopulations, as opposed to methods that ignore systematic contributions of upstream data biases on inequitable AI performance.

INTRODUCTION

There is growing concern that AI algorithms deployed in sensitive domains may replicate or exacerbate inequities in diagnosis or patient outcomes[1]. Current characterizations of bias of AI performance conflates observed retrospective disparities with algorithmic inequity without accounting for or characterizing the impact of upstream biases contributing to biased datasets. Broadly, this has resulted in fairness interventions that are not generalizable, requiring unjustified trade-offs, further disconnecting the impact of broader systemic disparities and the resulting biases in observational training data with that of inequities in AI performance[2,3]. Addressing algorithmic inequities requires a shift from the narrow focus on retrospective model performance to characterizing the impact of systematic upstream biases in the training data-generating processes. To that end, we propose to characterize the contribution of various granular distributional differences of patient characteristics across subpopulations within a dataset, to disentangle the causes of AI performance differences, arising from (a combination of) differences in data acquisition procedures or clinically relevant disease-related differences among subgroups. While our approach is general, we operationalize this auditing pipeline on evaluating biased performance of foundation models of disease diagnosis from chest radiology images[4-7].

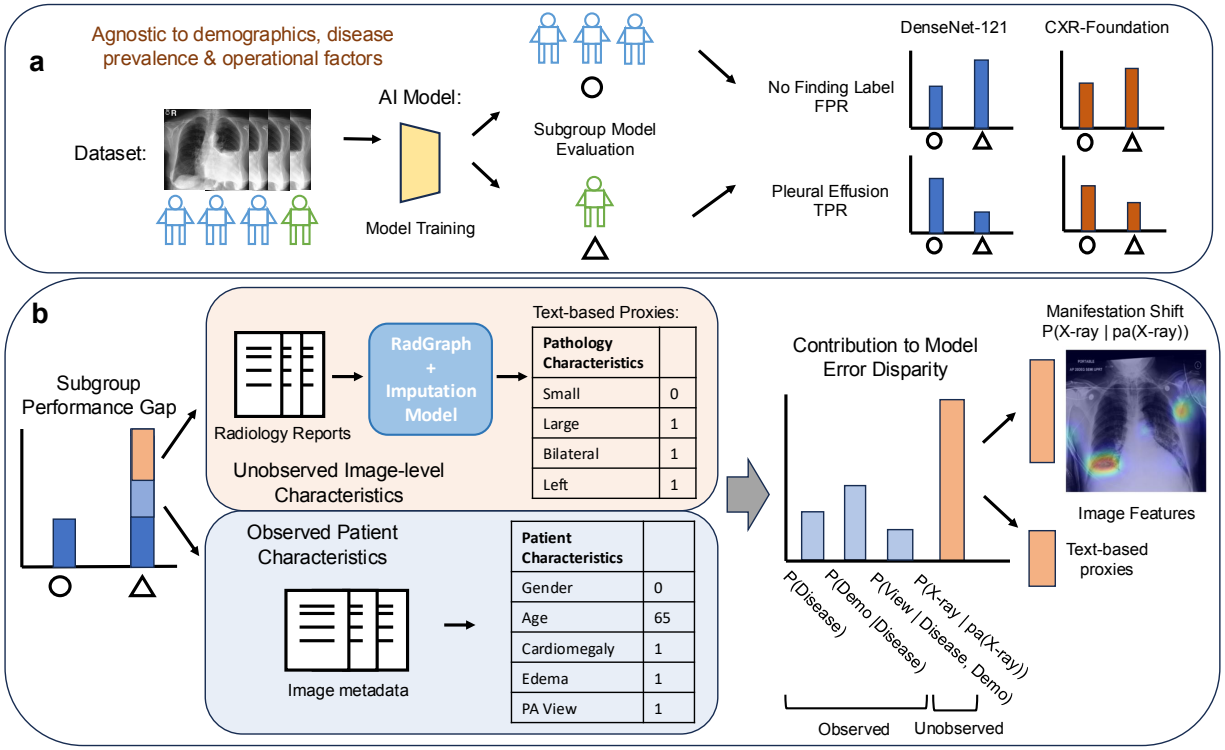
While chest radiography-based AI diagnostic algorithms may match specialist performance[8], with clear potential for real-world deployment[9], there is consistent evidence of population-level performance disparity, particularly underdiagnosis among underrepresented groups in public datasets[5]. Underdiagnosis bias may have significant ramifications such as causing delayed access to care for historically under-served minority populations[1,10,11]. In X-ray disease detection, the AI algorithm is often trained and evaluated agnostic to important knowledge of the population demographics, patient characteristics (such as disease prevalence differences), clinical workflow, and operational factors constituting the upstream data-acquisition process. These upstream mechanisms may influence the patterns learned by the AI model, even if an AI model does not explicitly rely on these variables during training. For example, differences in healthcare utilization patterns may lead to differences in the image characteristics and manifestation of pleural effusion across demographic subpopulations, leading to model performance degradation for underrepresented patients.

We perform a systematic study of underdiagnosis rates across racial subgroups in AI-based deep learning and foundation models on three chest X-ray datasets (MIMIC-CXR, CheXpert, Emory). To do so, we use a novel decomposition of the observed performance disparity to separate components representing different types of dataset biases[12,13,14]. We apply the method by Zhang et. al.[15] to assess the relative contribution of error disparity from different sources such as shifts in population demographics, clinical workflow (such as views collected *conditioned on initial impression*), disease prevalence, or variability in the X-ray (*conditioned on demographics, prevalence, and workflow variability*) a model is sensitive to. Our method is more reliable and actionable since it accounts for the complexity of the data-generating process as opposed to canonical approaches to interpreting and explaining AI behavior[16-18]. Our method is able to surface consistent and multifactorial causes of AI diagnostic disparities across all datasets. We further show how our attributions can be used to inform targeted model adaptations. Our analysis highlights that extraction and rigorous characterization of the impact of upstream distributional differences on model performance, beyond knowledge of patient demographics, is imperative for advancing the understanding of algorithmic bias and the development of reliable and generalizable bias-mitigation strategies.

RESULTS

Disease Diagnosis Tasks and Datasets. We focus on two performance metrics based on prior work[5] that assess the performance of two distinct tasks; one task aims to rule out the presence of any disease, and a second task aims to detect a specific pathology (pleural effusion). Pleural effusion was chosen as a pathology of interest since previous studies have reported disparate performance[4]. For the first task, the performance metric of interest is the underdiagnosis rate, defined as the False Positive Rates (FPR) for the binarized model prediction for the 'No Finding' label[5], indicating that no disease is diagnosed from the patient chest X-ray image. For the second task, we define the pathology detection rate as the True Positive Rate (TPR) for the 'Pleural Effusion' label. We choose to study the diagnostic disparities between White and Black patients, given that Black patients are sufficiently represented across the three datasets and there are consistent observable performance disparities across tasks. Our analysis pipeline is summarized in Fig. 1.

Figure 1: Overview of our analysis pipeline. **a**. We train distinct multi-label DenseNet-121 models for each dataset. We compute the subgroup-level model performance disparity using the underdiagnosis rate (FPR of the 'No Finding' label) and the pleural effusion detection rate (TPR of 'Pleural Effusion' label) for DenseNet-121 as well as the CXR-Foundation model. **b**. We decompose the total performance disparity to contributions of different confounding factors using our attribution method, both observed (differences in pathology types, demographics, image acquisition parameters from meta-data) and unobserved (differences in pathology manifestation and image features).



We perform our analysis using three datasets, MIMIC-CXR[19], CheXpert[20], and Emory-CXR[21]. Pathology labels were generated using CheXpert labeler[20] across all datasets. We train a DenseNet-121 neural

network to classify 14 different conditions using multi-task learning, using only the chest X-rays as input (Fig 1a.,1b.). Each dataset was randomly split using a ratio of 60-10-30 for train-validation-test, and a separate model was trained on each dataset. We follow the same architecture and training procedure as previous work[5], detailed in the Methods. We also analyze a state-of-the-art foundation model from Google[22] pre-trained on a completely distinct set of weakly labeled chest X-rays from USA and India, called CXR-foundation. Note that we cannot analyze CXR-foundation on Emory data due to privacy concerns. We report the final demographic distribution of the evaluation cohort used in the diagnostic disparity analysis for each dataset in Table 1.

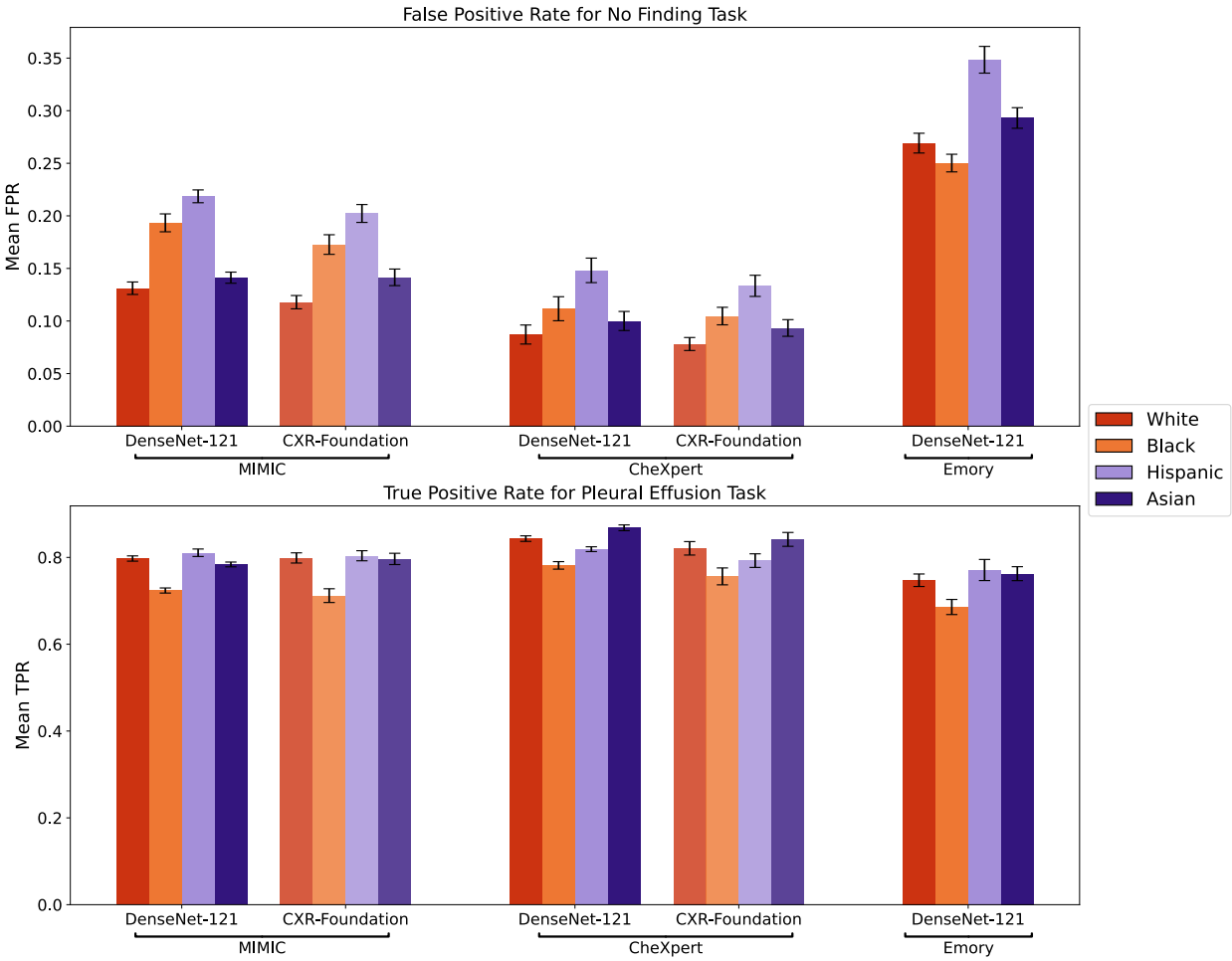
Table 1: Summary statistics for the evaluation cohort used for the disparity analysis in each dataset. The cohort only includes patients with findings in the X-ray, and patients with missing demographics (gender, age, race) are dropped.

Variable	Attribute	MIMIC, N (%)	Emory, N (%)	CheXpert, N (%)
	Total Scans	120963	562165	176495
Gender	Male	67615 (55.90%)	282683 (50.28%)	104535 (59.23%)
	Female	53328 (44.10%)	279482 (49.72%)	71960 (40.77%)
Race	White	88428 (73.10%)	285958 (50.08%)	113705 (64.42%)
	Black	17768 (14.69%)	260090 (46.27%)	10563 (5.98%)
	Asian	4255 (3.52%)	9532 (0.17%)	20860 (11.82%)
	Hispanic	5250 (4.34%)	4666 (0.08%)	22422 (12.70%)
	Native Am/Other	5262 (4.35%)	1919 (0.03%)	24296 (13.77%)
Age	0-20	548 (0.45%)	5781 (1.03%)	1937 (1.10%)
	20-40	9004 (7.44%)	83328 (14.82%)	20608 (11.68%)
	40-60	34861 (28.82%)	192191 (34.19%)	54749 (31.02%)
	60-80	55272 (45.69%)	230844 (41.06%)	71050 (40.26%)
	80+	21278 (17.59%)	50021 (8.9%)	28151 (15.95%)
Label	Atelectasis	38494 (31.82%)	102378 (18.21%)	28218 (15.99%)
	Cardiomegaly	37673 (31.14%)	153758 (27.35%)	24044 (13.62%)
	Edema	24603 (20.34%)	63175 (11.24%)	44272 (25.09%)
	Lung Opacity	41943 (34.67%)	148537 (26.42%)	93294 (52.86%)
	Pleural Effusion	46281 (38.26%)	124288 (22.11%)	75279 (42.65%)
	Support Devices	49632 (41.03%)	280095 (49.82%)	92204 (52.24%)

Underdiagnosis Rates across Datasets and Models. We plot the mean and standard deviation for the model performance metrics of interest ("No Finding" FPR, "Pleural Effusion" TPR) in the evaluation cohort across datasets and racial subgroups for the DenseNet-121 and CXR-Foundation models (Fig. 2). The standard deviation for the model performance is obtained by bootstrapping the training dataset and retraining the final classification layer, while keeping the test dataset fixed. We find consistent patterns across datasets, where underrepresented subgroups are underdiagnosed at a higher rate (Fig. 2a). Additionally, there is a consistent pattern across datasets where Black patients have lower detection rates for pleural effusion (Fig. 2b).

Figure 2: Racial performance disparities for "No Finding" and "Pleural Effusion" tasks across datasets for DenseNet-121 and CXR-Foundation Models, with error bars representing standard deviation across 20 bootstrap iterations. **a**. For MIMIC-CXR and CheXpert, the subgroup with the best performance is White patients. Underdiagnosis is most observed for Black and Hispanic patients. For Emory, the subgroup with the best performance is Black patients. Underdiagnosis occurs for Hispanic, White, and Asian patients.

CXR-Foundation is better able to disambiguate healthy and diseased X-rays, as evident by consistently lower FPR rates. **b**. Model performance is consistently worse for Black patients for the task of detecting pleural effusion. The pattern of underdiagnosis is also consistent across models for MIMIC-CXR and CheXpert. A table with numerical values and statistical comparisons is included in the Supplement.



We note that CXR-Foundation performs uniformly better (Table S1) in the task of differentiating no-finding X-rays across all racial groups compared to the dataset-specific DenseNet-121 models. The performance disparities across racial groups are also smaller, highlighting the benefits of pretraining larger capacity models on large, diverse datasets. However, the patterns of underdiagnosis bias across racial groups persist, indicating that model capacity does not fully explain observable performance disparities. Similarly, lower detection rates of pleural effusion in Black patients can also be observed with CXR-Foundation, but patterns of diagnosis disparity persist across models and sites (Table S2).

Decomposing Underdiagnosis Bias Across Patient Subgroups. We hypothesize that distributional differences in patient demographics and clinical characteristics obtained from meta-data, i.e., differences not observed by the model during training, partially explain observable error rate disparities across racial

subgroups and enable reliable attribution of disparities in terms of these differences. Importantly, this is the case *even if our candidate models do not explicitly use these variables as input.*

For the observed patient confounding attributes, we consider the binary pathology labels Z, demographics D, and the view position of the X-ray C. These variables were selected as they are extracted consistently across all datasets and may causally influence the distribution of image features[23,24]. The demographics variable includes gender and age group all encoded as binary variables. The view position is represented using a binary indicator variable $C = \{C_{PA}, C_{AP}, C_{LA}, \dots\}$ for each view type (AP: anteroposterior, LA: lateral, etc.). The acquisition of various views is based on clinical guidelines, depending on the suspected pathology as some view positions allow for better detection or localization of pathologies like cardiomegaly and pneumothorax volume[25]. Demographic attributes may also influence the ability for patients to obtain specific X-ray views, due to being in a severe condition or being physically unable to position themselves for a PA view[26].

We consider a source and target distribution corresponding to two patient subgroups, denoted by P and Q respectively. Our goal is to decompose the error rate disparity defined by the total difference in model performance metric L (such as FPR/TPR) into further granular and informative components.

$$\Delta L = \underbrace{\mathbb{E}_Q[L]}_{Performance\ for\ Q} - \underbrace{\mathbb{E}_P[L]}_{Performance\ for\ P}$$

We attribute the performance gap to distributional differences in each aggregate factor (pathologies Z, demographics D, view-type C, and X-ray features X). We choose the following factorization of the joint probability distribution $p(Z, D, C, X)$ based on knowledge of the data-generating process, as shown below.

$$P = \{p(Z)p(D|Z)p(C|Z, D)p(X|Z, D, C)\}$$

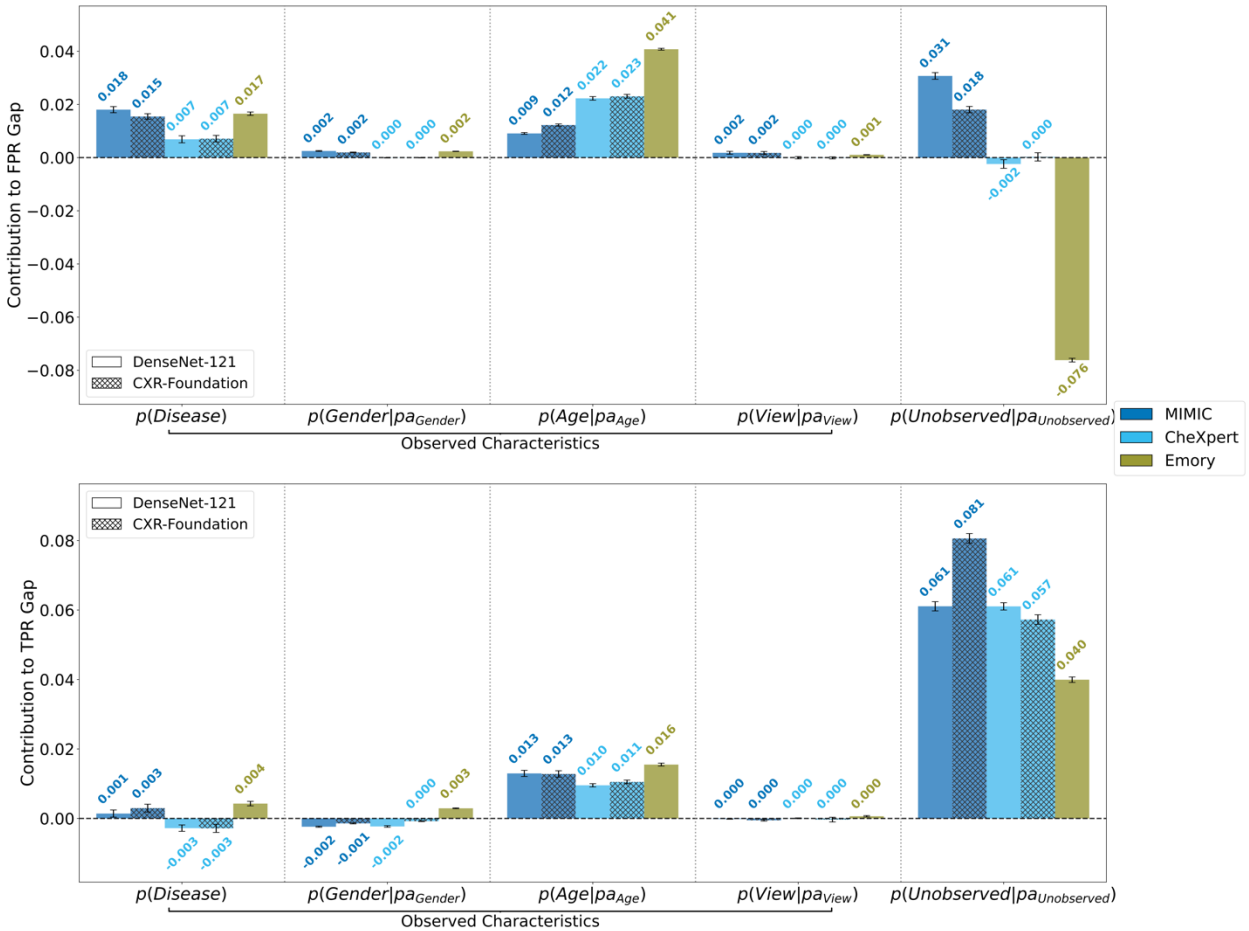
Each factor represents a probabilistic mechanism or dependence structure between variables that can shift across racial subgroups. In other words, the distribution shift in $p(Z, D, C, X)$ from P to Q can be decomposed into four shifts; A shift in $p(Z)$ represents a prevalence shift of pathologies across race (in $Q(Z)$). A shift in $p(D|Z)$ indicates that for a certain disease, the age or gender representation may differ across race. A shift in $p(C|Z, D)$ may arise as a shift in a clinician's acquisition workflow for certain diseases or demographic characteristics, for instance if lateral views are obtained more often for one subgroup of patients. Finally, a shift in $p(X|Z, D, C)$ may be indicative of manifestation shift given a certain age, gender, and pathology. Given that distribution shifts in image-level features are difficult to estimate and interpret, we aim to decompose this error contribution using text-based proxies as well as visualizing distribution shifts in the model's feature representation space using saliency maps (Fig. 1b).

Note that each factor is a distribution conditioned on or adjusting for variables that influence the confounding variable according to the data-generating process. For example, we consider the conditional distribution $p(C|Z, D)$ as the view position may be influenced by patient demographics and pathologies. This enables our attribution to reliably decompose error contributions across multiple confounding factors. A formalism of our approach based on causal models is provided in the Supplementary materials.

Prevalence and Age Shifts Contribute to Diagnostic Disparity. We plot the results of our attribution method, where the x-axis indicates the distribution of each aggregate confounding factor. For each of these distributional factors, we plot the Shapley value on the y-axis. Each Shapley value represents the additive contribution of the factor to the increased FPR of the "No Finding" label in Black patients. For example, an attribution of 0.022 for 'age' indicates that the shift in the conditional distribution of age given gender and disease from White patients to Black patients lead to an increase in FPR of 2.2% for Black patients. Note the relative contribution can be negative, indicating that the distribution shift reduced the FPR gap.

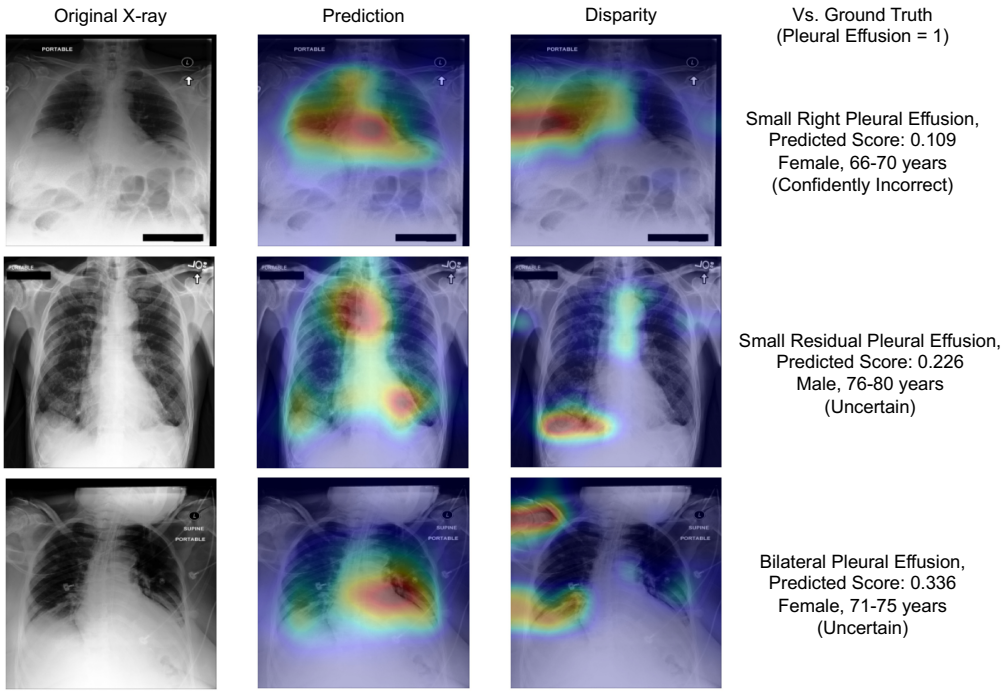
We find that disease prevalence shifts consistently leads to underdiagnosis and unfavorable model performance for Black patients. Across all three datasets, White patients have a higher prevalence of pleural effusion, and Black patients have a higher prevalence for cardiomegaly (Extended Data Fig. 1-3). One plausible explanation is that for each disease, the relative task difficulty may differ inherently, leading to lower underdiagnosis rates in Whites.

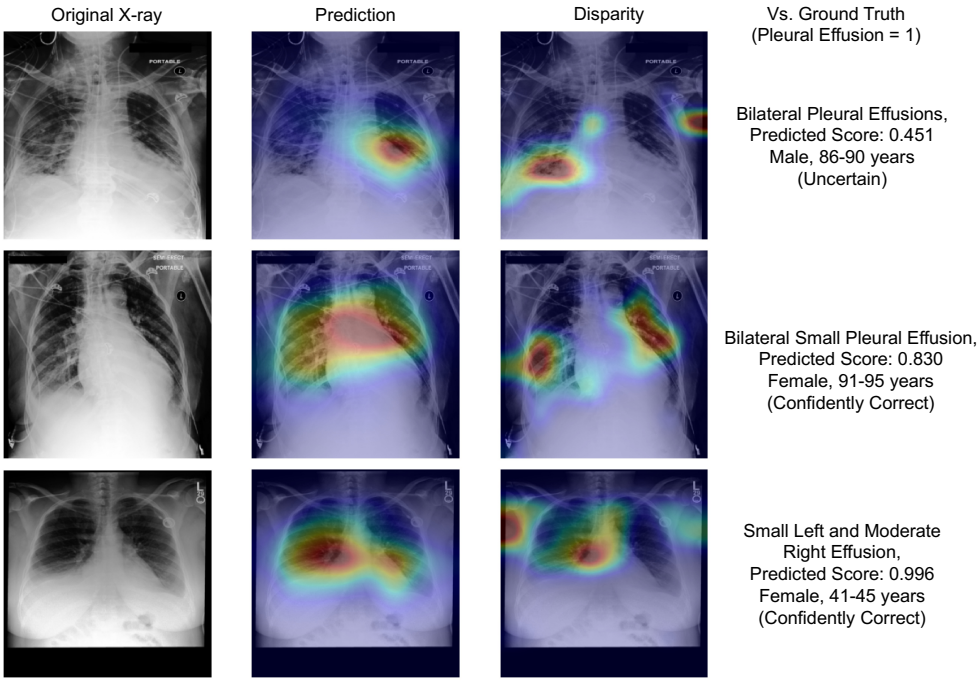
Figure 3: Contribution of each distribution shift mechanism to performance disparity across Black and White patients, error bars represent standard deviation across 100 bootstrap iterations. **a.** A larger positive attribution value (y-axis) indicates a larger contribution to the overall false positive rate (FPR) disparity for the 'No Finding' label. The x-axis shows the five different mechanisms that contribute to overall performance disparity. For example, a shift in $p(Age\ |pa_{Age})$ indicates a shift in the age distribution given parent variables, which in this case are pathology and gender. Marginal disease prevalence shifts, and conditional age shifts are consistent contributors to greater underdiagnosis rates for Black patients. The difference in X-ray attribution across datasets can arise for various reasons, such as sample classification difficulty or label error. **b.** A larger positive attribution value (y-axis) indicates a larger contribution to the total true positive rate (TPR) disparity for the 'Pleural Effusion' label. Distributional differences of age result in lower detection rate among Black patients. However, observed patient characteristics do not fully explain the TPR disparity, indicating the presence of other unobserved or spurious features in the X-ray affecting model performance.



Furthermore, distributional differences in age lead to underdiagnosis and an increase in FPR for Black patients, even after adjusting for disease prevalence. Interestingly, Black patients are consistently younger compared to White patients across all datasets (Extended Data Fig. 4). The model is systematically performing worse on younger patients, potentially due to differences in how the pathology manifests between younger and older patients. For example, the same pathology may be smaller in size and harder to detect in younger patients, if the condition is less severe compared to older patients. We note that the magnitude of disease prevalence and age attributions is similar for CXR-Foundation, indicating that the mechanisms for causing systematic differences in performances are agnostic to model capacity.

Figure 4: Saliency map examples for the MIMIC-CXR dataset, for the task of detecting pleural effusion. We choose test set examples from the underrepresented subgroup (Black patients) across a range of model predictions. We include the predicted score for effusion, showing the relative model uncertainty. We also include the description of pleural effusion found in the corresponding radiology report. The three columns indicate the original chest X-ray image (Original X-ray), the saliency map for the pleural effusion prediction (Prediction), and the saliency map for the most informative features for error disparity across White and Black patients (Disparity).





Lastly, while the X-ray attribution may be due to a combination of many factors such as insufficient model capacity, we hypothesize that one contributing factor may be due to different prevalences of label noise[27,28] across racial subgroups. Given that the labeler is trained on radiology reports in the CheXpert dataset, it is possible that a higher frequency of ambiguous cases is being incorrectly labeled as "No Finding" (a false positive label) in MIMIC-CXR and Emory, leading to larger disparity attribution to unobserved factors. Conversely, low label noise in the CheXpert dataset may explain why there is no

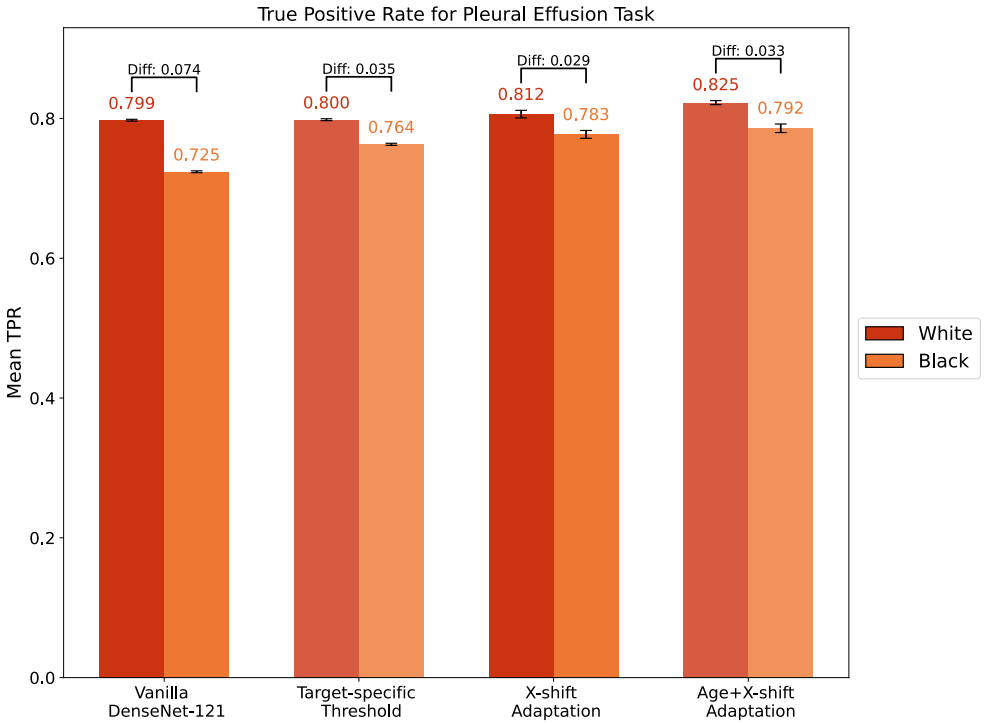
attribution to the unobserved factors. The distributions of predicted probabilities (Fig. S3) also suggest that MIMIC-CXR and Emory contain more false positive samples that are likely to be due to labeler error.

Image-level Feature Shifts Influence Pathology Detection Rates. We plot the attributions for the detection rate disparity in pleural effusion, using the same method and subgroup configurations as the previous section. We find that age differences across race consistently lead to worse model performance among Black patients, indicating that both the DenseNet-121 and CXR-Foundation models are systematically failing to detect effusion in younger patients (Fig. S4). However, across all datasets the performance disparity is still largely unexplained after adjusting for differences in demographic and pathology composition across race, indicating the presence of additional manifestation shifts in the image-level features. For example, pathology manifestation differences including the location, size, and ambiguity of the pleural effusion in the X-ray may influence the ability for the model to detect the disease, but such disease-related characteristics are "unobserved" as they are not recorded as meta-data in the dataset (Fig. 1b).

To this end, we perform a sensitivity analysis using text-based proxies of hypothesized unobserved confounders extracted from radiology reports. Our results (Fig. S1) shows that this age attribution can be partially explained away by effusion characteristics such as size, location, and radiologist uncertainty. This indicates that the model may be influenced by a combination of purely spurious age-related features, as well as "mixed" features that mix the label signal (pleural effusion) and age, such as effusion size and severity. We include a discussion of shortcut learning[29,30,31] and spurious features in the Supplement.

However, the performance disparity is still largely unexplained after correcting for distributional differences in our extracted proxy variables. We hypothesize that there may be types of effusion indications or image features that are underrepresented in the general population but encountered more often in the Black population. To perform attributions with respect to image features, we combine our framework with commonly used saliency methods[32,33] to identify image-based features that explain the component of error disparity arising from manifestation shift. We observe that the predicted probability of effusion is relying on holistic information in the lung regions (Fig. 4). Furthermore, we observe some consistent patterns where the features that are most predictive of disparity are the lower left regions of the lung, potentially indicative of the absence of some indication of pleural effusion. However, spurious structures outside of the lungs such as bone density and structures in the central region may also be informative for the error disparity. Based on the saliency maps, an evaluation team could potentially discern why the model diagnosis rates may differ across these X-rays.

Figure 5: Racial performance disparities for the "Pleural Effusion" task for the DenseNet-121 model in the MIMIC dataset after performing target adaptation, the mean and standard error for TPR is computed for each method. Adapting to shifts in a subset of the image features (X-shift adaptation) increases the TPR for Black patients to 0.783. Additionally adapting to age shifts further increases TPR to 0.792. Interestingly, the detection rates are uniformly higher for the adapted models, showing the benefits of resampling the dataset to improve the representation of specific age groups and image features. We include a baseline (target-specific threshold) which takes the original vanilla DenseNet-121 model and computes an optimal classification threshold for each racial subgroup separately.



Adapting Models to Underrepresented Subpopulations. Inspired by methods in domain adaptation[34], we demonstrate how insights from our evaluation procedure can be used to design granular model fine-tuning strategies. For the pleural effusion detection task, we adapt the model to the distribution of Black patients in a given dataset by targeting two distribution shifts that were contributors to the performance disparity: the manifestation shift in image features and the shift in the age distribution. We resample the original training dataset by upweighting the samples corresponding to the image features and age groups that were represented more frequently in the Black population and retrain the final classification layer on the reweighted dataset. We plot the TPR results for the race-adapted model (Fig. 5), showing that our new models achieve greater overall TPR and lower TPR disparity across race with statistical significance (Table S3). We also note that adapting to specific distribution shifts outperforms models that simply employs a race-adaptive threshold for classification. This suggests that our mitigation strategy may be preferable to other approaches as it simply adapts the model at test-time to the distributional composition of patients in a target patient subgroup, instead of potentially compromising overall predictive performance by regularizing or removing information correlated with sensitive attributes in the model prediction.

DISCUSSION

In this study, we perform a systematic study to understand why underdiagnosis bias arises between racial subgroups for AI-based chest X-ray disease classification models. We show empirically that the mechanisms causing underdiagnosis bias are surprisingly consistent across datasets, and a nuanced understanding of model and dataset biases are required to interpret subgroup-level diagnostic disparities. Our results for CXR-Foundation indicate that while performance disparities may be partially mitigated with state-of-the-art models, the overall patterns of racial underdiagnosis bias and the mechanisms for

causing systematic differences in performances are agnostic to model capacity. Finally, we show that our evaluation method can inform fine-tuning strategies to adapt the model to specific subpopulations.

Our analysis addresses a fundamental limitation of subgroup-level model evaluation, which is increasingly important as large-scale clinical ML models are being evaluated on diverse patient subpopulations. While no prior approach systematically accounts for distributional differences of confounding factors across subpopulations, one commonly employed solution is "slicing" the evaluation dataset to compute model performance on finer subgroups or resampling to balance the representation of confounding variables[4]. However, instead of heuristically choosing which demographic variables to adjust for, we demonstrate that our strategy of jointly modeling 'error contributions' can reveal interactions between confounding attributes and provide granular insights through retrospective evaluation of an ML model. We note that our strategy can flexibly incorporate increased granularity of potential confounding attributes and variable groupings depending on the granularity of meta-data recorded in datasets. The significance of collecting meta-data and patient characteristics for image-based models cannot be overstated, as access to granular confounding attributes is necessary to uncover the mechanisms underlying subgroup-level disparities.

The study also offers a cautionary note for practitioners aiming to evaluate or ensure fairness in AI-based medical imaging diagnosis. Well-intentioned strategies for enforcing error rate parity have been proposed to mitigate performance disparities[35], or to mitigate reliance on spurious attributes or "shortcuts" in the image[36]. However, applying crude adjustment methods without considering implicit distribution shifts have been shown to lead to unintended or harmful consequences. For example, eliminating age-related information in model prediction or model embeddings may be harmful if age is correlated with disease-related characteristics such as the severity, size and location of pleural effusion. Therefore, we stress the importance of disentangling the sources of subpopulation heterogeneity within a target patient population to assess the appropriateness of bias-mitigation strategies. We also propose an alternative mitigation strategy that simply adapts the model to the distributional composition of patients in a target subgroup, instead of removing information correlated with sensitive attributes in the model prediction.

We acknowledge that claiming unfairness in the context of deep learning-based diagnosis of chest X-rays requires further consideration and analysis. We emphasize that our main objective is not to determine whether the deep learning models exhibit unfairness. Various fairness definitions exist in the literature, based on different statistical and causal constraints on model performance metrics[2,37,38]. While our work provides a practical approach for understanding how dataset biases contribute to population-level error rate disparity, understanding fairness violations at increased levels of granularity is generally challenging. For instance, even if the model satisfies equalized error rates across subgroups in a representative and balanced evaluation set, the model may still encode spurious sensitive features in the chest X-ray image to produce individual predictions. However, it is also unclear what constitutes a sensitive feature that should be deemed impermissible and removed, particularly in disease diagnosis. Various works in the fairness literature suggest that the use of characteristics related to subgroup membership may be desired depending on the task and clinical context[39-41], as disease prevalences in demographic subgroups can act as informative priors.

Our approach is flexible and can incorporate recent advancements made in the pipeline for chest X-ray classification. For example, since our attribution method is agnostic to the base disease detection model, we demonstrate its use with a foundation model that has much larger capacity and is pretrained on completely distinct datasets. In addition to the model itself, other improvements such as leveraging large language models (LLMs) for label and confounder proxy extraction from radiology reports can be incorporated to reduce errors. While label noise in training and evaluating the disease detection model is a limitation in our analysis, our insights regarding dataset biases and our methodology for performing granular subgroup analysis remain applicable as better label curation methods are introduced. Additionally, our method can integrate any improvements in saliency mapping and interpretability methods. Finally, while our analysis focuses on disparities in disease classification performance, our framework can be applied to disparities in alternative task definitions. For example, one may be interested in decomposing the disparities in radiology report generation performance by decomposing an evaluation metric related to the likelihood of the correct diagnosis being generated.

METHODS

Attributing Performance Disparity to Observed Distribution Shifts. We use the method from Zhang et.al.[15] to attribute degradation in model performance to shifts from a set of possible distribution shifts. Multiple distributions can shift simultaneously in practice, and some distribution shifts may have no impact on model performance. To attribute the contribution of a given distribution shift to model performance, the method relies on Shapley values[43] to attribute changes in model performance to *individual distributions*. The key distinction from existing work in Shapley based feature importance is that the method from Zhang et. al. is directly able to identify the contribution of distributional changes or mechanisms as opposed to the sensitivity to adding/deleting individual features to explain performance changes, in addition to allowing for this attribution in terms of mechanisms or variables not observed by the model, enabling us to attribute performance differences to systemic causes, as opposed to in-vivo model perturbations. A detailed description of the estimation procedure is provided in the Supplement.

Sensitivity Analysis with Extracted Text-based Proxies. We aim to further decompose the error contribution from the X-ray image features $p(X|Z, D, C)$ using text-based proxies of unobserved confounders (Fig. 1b). These attributions correspond to the error disparity unexplained by distributional differences in the composition of demographics, pathology, and view-position.

For each chest X-ray with pleural effusion, we perform an automated extraction of pleural effusion size (small, moderate, large) and location (left, right, bilateral) characteristics, as well as radiologist uncertainty from the corresponding radiology report using Radgraph[44]. Each characteristic is represented with a binary variable indicating its presence in the radiology report. For cases where the characteristic is unrecorded, we perform conditional mean imputation by training a DenseNet-121 classifier that predicts the missing characteristic of the effusion. The extracted and predicted values are included as an additional aggregate variable to assess the contribution of pathology manifestation differences to error disparity. We analyze the sensitivity of the remaining attributions when effusion characteristics are included as an upstream variable.

Interpreting Shifts in Image-level Features using Saliency Maps.
Finally, we aim to attribute the error disparity to specific image features that are represented more frequently in Black patients compared to their White patient counterparts. Saliency methods have been validated in chest X-ray disease classification and have been shown to localize regions of interest when the model is able to detect the disease successfully[45]. Using the feature embedding learned by the model, we propose an approach to approximate which regions of the image explain the error disparity resulting from manifestation shift in $p(X|Z, D, C)$. We define a variable $\phi(X)_i$ for each dimension i of the chest X-ray feature embedding, where ϕ denotes that the input X-ray has been transformed by the (pretrained) model. We opt for a simple heuristic to detect which variable $\phi(X)_i$ contributed to the error disparity. Given the large dimensionality, we first screen the variables using L1-penalized logistic regression to predict race group from the embedding features. For the selected indices, we introduce the factor $p(\phi(X)_i|Z, D, C)$ to approximate the manifestation shift in the attribution method and select i with the largest error contribution.

To map each of these variables $\phi(x)_i$ to the chest X-ray input image space, we use a Grad-CAM-based saliency method[32,33]. For (x, y)-coordinates in the image space, the saliency map $M_k(x, y)$ is computed for the top-k selected $\phi(x)_i$ variables, and $\sum_k M_k(x, y)$ denotes the sum of the k saliency maps. We select k=3 for our analysis. The saliency method maps the feature embedding to regions in the original image, corresponding to image features that differ across racial groups given the same patient demographics and clinical characteristics.

Adaptation Strategy for Fine-grained Distribution Shifts. To adapt the model to the target patient distribution (Black patients), we retrain the final linear classification layer of the DenseNet-121 model using a resampled version of the training dataset. We compute the weights for resampling the dataset based on an estimation of the density ratios for two distribution shifts across the entire population and the subgroup of Black patients in the training set. For the X-shift, we use the top-k (k=3) chest X-ray feature embeddings extracted and visualized in the saliency method. We use the same method[2] to estimate the density ratio $\frac{p_Q(\phi(x)_i)}{p_P(\phi(x)_i)}$, which upweights the samples containing image features that are represented more frequently in the Black population. The weights for the age distribution shift are obtained in the same manner.

Model Training & Disparity Attribution Experiment Details. The backbone architecture across all experiments for the disease diagnosis model is the 121-layer DenseNet, with the weights initialized using ImageNet pretraining. We applied data augmentation via center crop, random horizontal flip, and 15° random rotation for model training. The optimization is conducted using Adam with default parameters and binary cross-entropy loss. A batch size of 48 is used due to computation limitations. Learning rate is initially set to 0.0005 and drops to half if validation loss does not improve over three epochs. An early stopping condition was implemented if validation loss does not improve over 5 epochs. For binarizing the model prediction, we follow the practice in Seyyed-Kalantari et. al.[5] and use the single best threshold that maximizes the F1 score in the entire test dataset across all demographic groups for each task.

For the CXR-Foundation[22] model, we use the 1376-dimensional embeddings which can be obtained by calling the CXR-Foundation API as outlined in https://github.com/Google-Health/imaging-

research/tree/master/cxr-foundation. We finetune logistic regression models for the downstream tasks. A batch size of 256 is used. The same optimization parameters are used as the DenseNet-121 model.

Since the dataset contains missing values for the demographic variables of interest, we filter out the subset of the data that has any missing demographic attribute for the analysis of performance disparities. We create an aggregated dataset using the filtered patients that belonged in the original training and validation fold for fitting the domain classifiers for importance weighting. We use logistic regression models throughout for estimating importance weights. The output probabilities from each domain classifier are clipped at 0.99, and each final importance weight is also clipped using a threshold of 99.9 percentile. We perform 100 bootstraps with the aggregated training dataset for the domain classifier training to simulate sampling variability. For each bootstrap, we compute the subgroup-level error metrics based on the fixed prediction probabilities and optimal threshold from the model and compute the attributions using a fixed held-out test set. This ensures that the subgroup-level performance disparity based on the error metric is fixed, but variability is introduced in the attributions through the differences in importance weights computed by the domain classifier at every bootstrap iteration.

REFERENCES

1. Chen, I. Y., Pierson, E., Rose, S., Joshi, S., Ferryman, K., & Ghassemi, M. (2021). Ethical machine learning in healthcare. *Annual review of biomedical data science*, *4*(1), 123-144.
2. Plečko, D., & Bareinboim, E. (2024). Causal fairness analysis: a causal toolkit for fair machine learning. *Foundations and Trends® in Machine Learning*, *17*(3), 304-589.
3. Friedler, S. A., Scheidegger, C., & Venkatasubramanian, S. (2021). The (im) possibility of fairness: Different value systems require different mechanisms for fair decision making. *Communications of the ACM*, *64*(4), 136-143.
4. Glocker, B., Jones, C., Bernhardt, M., & Winzeck, S. (2023). Algorithmic encoding of protected characteristics in chest X-ray disease detection models. *Ebiomedicine*, *89*.
5. Seyyed-Kalantari, L., Zhang, H., McDermott, M. B., Chen, I. Y., & Ghassemi, M. (2021). Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. *Nature medicine*, *27*(12), 2176-2182.
6. Drukker, K., Chen, W., Gichoya, J., Gruszauskas, N., Kalpathy-Cramer, J., Koyejo, S., ... & Giger, M. (2023). Toward fairness in artificial intelligence for medical image analysis: identification and mitigation of potential biases in the roadmap from data collection to model deployment. *Journal of Medical Imaging*, *10*(6), 061104-061104.
7. Banerjee, I., Bhattacharjee, K., Burns, J. L., Trivedi, H., Purkayastha, S., Seyyed-Kalantari, L., ... & Gichoya, J. (2023). "Shortcuts" causing bias in radiology artificial intelligence: causes, evaluation and mitigation. *Journal of the American College of Radiology*.
8. Rajpurkar, P., Irvin, J., Ball, R. L., Zhu, K., Yang, B., Mehta, H., ... & Lungren, M. P. (2018). Deep learning for chest radiograph diagnosis: A retrospective comparison of the CheXNeXt algorithm to practicing radiologists. *PLoS medicine*, *15*(11), e1002686.Rimmer, A. (2017).
9. Radiologist shortage leaves patient care at risk, warns royal college. *BMJ: British Medical Journal (Online)*, *359*.
10. Chen, I. Y., Joshi, S., & Ghassemi, M. (2020). Treating health disparities with artificial intelligence. *Nature medicine*, *26*(1), 16-17.
11. Rajkomar, A., Hardt, M., Howell, M. D., Corrado, G., & Chin, M. H. (2018). Ensuring fairness in machine learning to advance health equity. *Annals of internal medicine*, *169*(12), 866-872.
12. Castro, D. C., Walker, I., & Glocker, B. (2020). Causality matters in medical imaging. *Nature Communications*, *11*(1), 3673.
13. Bernhardt, M., Jones, C., & Glocker, B. (2022). Potential sources of dataset bias complicate investigation of underdiagnosis by machine learning algorithms. *Nature Medicine*, *28*(6), 1157-1158.

14. Jones, C., Castro, D. C., Ribeiro, F. D. S., Oktay, O., McCradden, M., & Glocker, B. (2023). No Fair Lunch: A Causal Perspective on Dataset Bias in Machine Learning for Medical Imaging. *arXiv preprint arXiv:2307.16526*.
15. Zhang, H., Singh, H., Ghassemi, M., & Joshi, S. (2023) "Why did the model fail?": Attributing model performance changes to distribution shifts. In *Proceedings of the 40th International Conference on Machine Learning*, Proceedings of Machine Learning Research. PMLR
16. Lundberg, S. (2017). A unified approach to interpreting model predictions. *arXiv preprint arXiv:1705.07874*.
17. Roschewitz, M., Mehta, R., Jones, C., & Glocker, B. (2024). Automatic dataset shift identification to support root cause analysis of AI performance drift. *arXiv preprint arXiv:2411.07940*.
18. Cai, T. T., Namkoong, H., & Yadlowsky, S. (2023). Diagnosing model performance under distribution shift. *arXiv preprint arXiv:2303.02011*.
19. Johnson, A. E., Pollard, T. J., Berkowitz, S. J., Greenbaum, N. R., Lungren, M. P., Deng, C. Y., ... & Horng, S. (2019). MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific data*, *6*(1), 317.
20. Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., ... & Ng, A. Y. (2019). Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI conference on artificial intelligence* (Vol. 33, No. 01, pp. 590-597).
21. Gichoya, J. W., Banerjee, I., Bhimireddy, A. R., Burns, J. L., Celi, L. A., Chen, L. C., ... & Zhang, H. (2022). AI recognition of patient race in medical imaging: a modelling study. *The Lancet Digital Health*, *4*(6), e406-e414.
22. Sellergren, A. B., Chen, C., Nabulsi, Z., Li, Y., Maschinot, A., Sarna, A., ... & Shetty, S. (2022). Simplified transfer learning for chest radiography models using less data. *Radiology*, *305*(2), 454-465.
23. Lotter, W. (2024). Acquisition parameters influence AI recognition of race in chest X-rays and mitigating these factors reduces underdiagnosis bias. *Nature Communications*, *15*(1), 7465.
24. Weng, N., Bigdeli, S., Petersen, E., & Feragen, A. (2023, October). Are Sex-Based Physiological Differences the Cause of Gender Bias for Chest X-Ray Diagnosis?. In *Workshop on Clinical Image-Based Procedures* (pp. 142-152). Cham: Springer Nature Switzerland.
25. Daffner, R. H., & Hartman, M. (2013). *Clinical radiology: the essentials*. Lippincott Williams & Wilkins.
26. Broder J. (2011). Imaging the Chest: The Chest Radiograph. *Diagnostic Imaging for the Emergency Physician*, 185–296.
27. Busby, L. P., Courtier, J. L., & Glastonbury, C. M. (2018). Bias in radiology: the how and why of misses and misinterpretations. *Radiographics*, *38*(1), 236-247.
28. Jain, S., Smit, A., Truong, S. Q., Nguyen, C. D., Huynh, M. T., Jain, M., ... & Rajpurkar, P. (2021, April). VisualCheXbert: addressing the discrepancy between radiology report labels and image labels. In *Proceedings of the Conference on Health, Inference, and Learning* (pp. 105-115).
29. Jabbour, S., Fouhey, D., Kazerooni, E., Sjoding, M. W., & Wiens, J. (2020). Deep learning applied to chest x-rays: Exploiting and preventing shortcuts. In *Machine Learning for Healthcare Conference* (pp. 750-782). PMLR.
30. Brown, A., Tomasev, N., Freyberg, J., Liu, Y., Karthikesalingam, A., & Schrouff, J. (2023). Detecting shortcut learning for fair medical AI using shortcut testing. *Nature Communications*, *14*(1), 4314.
31. Makar, M., Packer, B., Moldovan, D., Blalock, D., Halpern, Y., & D'Amour, A. (2022, May). Causally motivated shortcut removal using auxiliary labels. In *International Conference on Artificial Intelligence and Statistics* (pp. 739-766). PMLR.
32. Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision* (pp. 618-626).
33. Omeiza, D., Speakman, S., Cintas, C., & Weldermariam, K. (2019). Smooth grad-cam++: An enhanced inference level visualization technique for deep convolutional neural network models. *arXiv preprint arXiv:1908.01224*.
34. Tsuboi, Y., Kashima, H., Hido, S., Bickel, S., & Sugiyama, M. (2009). Direct density ratio estimation for large-scale covariate shift adaptation. *Journal of Information Processing*, *17*, 138-155.

35. Zhang, H., Dullerud, N., Roth, K., Oakden-Rayner, L., Pfohl, S., & Ghassemi, M. (2022, April). Improving the fairness of chest X-ray classifiers. In *Conference on health, inference, and learning* (pp. 204-233). PMLR.
36. Yang, Y., Zhang, H., Gichoya, J. W., Katabi, D., & Ghassemi, M. (2024). The limits of fair medical imaging AI in real-world generalization. *Nature Medicine*, 1-11.
37. Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. *Advances in neural information processing systems*, *29*.
38. Caton, S., & Haas, C. (2020). Fairness in machine learning: A survey. *ACM Computing Surveys*.
39. Basu, A. (2023). Use of race in clinical algorithms. *Science Advances*, *9*(21), eadd2704.
40. Coots, M., Saghafian, S., Kent, D., & Goel, S. (2023). Reevaluating the Role of Race and Ethnicity in Diabetes Screening. *arXiv preprint arXiv:2306.10220*.
41. Manski, C. F., Mullahy, J., & Venkataramani, A. S. (2023). Using measures of race to make clinical predictions: Decision making, patient health, and fairness. *Proceedings of the National Academy of Sciences*, *120*(35), e2303370120
42. Weng, W. H., Sellergen, A., Kiraly, A. P., D'Amour, A., Park, J., Pilgrim, R., ... & Liu, Y. (2024). An intentional approach to managing bias in general purpose embedding models. *The Lancet Digital Health*, *6*(2), e126-e130.
43. Shapley, L. S. (1953). A value for n-person games.
44. Jain, S., Agrawal, A., Saporta, A., Truong, S. Q., Duong, D. N., Bui, T., ... & Rajpurkar, P. (2021). Radgraph: Extracting clinical entities and relations from radiology reports. *arXiv preprint arXiv:2106.14463*.
45. Saporta, A., Gui, X., Agrawal, A., Pareek, A., Truong, S. Q., Nguyen, C. D., ... & Rajpurkar, P. (2022). Benchmarking saliency methods for chest X-ray interpretation. *Nature Machine Intelligence*, *4*(10), 867-878.

Extended Data Fig. 1. Prevalence of pathologies across racial groups (Emory)

Pathology	White, N (%)	Black, N (%)
Atelectasis	61436 (21.48%)	37911 (14.58%)
Cardiomegaly	71065 (24.85%)	78501 (30.18%)
Consolidation	17855 (6.24%)	13552 (5.21%)
Edema	32268 (11.28%)	29165 (11.21%)
Enlarged Cardiomediastinum	57097 (19.97%)	96970 (37.28%)
Fracture	9049 (3.16%)	3779 (3.779%)
Lung Lesion	21407 (7.49%)	15288 (5.88%)
Lung Opacity	80324 (28.09%)	64098 (24.64%)
Pleural Effusion	77900 (27.24%)	42089 (16.46%)
Pleural Other	10539 (3.69%)	4965 (1.91%)
Pneumonia	18000 (6.29%)	16404 (6.31%)
Pneumothorax	38732 (13.54%)	47737 (18.35%)
Support Devices	134255 (46.95%)	137822 (52.99%)

Extended Data Fig. 2. Prevalence of pathologies across racial groups (MIMIC)

Pathology	White, N (%)	Black, N (%)
Atelectasis	29049 (32.85%)	4897 (27.56%)
Cardiomegaly	26367 (29.82%)	6604 (37.17%)
Consolidation	6826 (7.72%)	1296 (7.29%)
Edema	18113 (20.48%)	3641 (20.49%)
Enlarged Cardiomediastinum	4356 (4.93%)	761 (4.28%)
Fracture	2231 (2.52%)	175 (0.98%)
Lung Lesion	3129 (3.54%)	589 (3.31%)
Lung Opacity	30805 (34.84%)	6055 (34.08%)
Pleural Effusion	35760 (40.44%)	5271 (26.67%)
Pleural Other	991 (1.12%)	138 (0.78%)
Pneumonia	9207 (10.41%)	1933 (10.88%)
Pneumothorax	6929 (7.84%)	861 (4.85%)
Support Devices	37237 (42.11%)	6402 (36.03%)

Extended Data Fig. 3. Prevalence of pathologies across racial groups (CheXpert)

Pathology	White, N (%)	Black, N (%)
Atelectasis	18886 (16.61%)	1479 (14.00%)
Cardiomegaly	14394 (12.66%)	2338 (22.13%)
Consolidation	8243 (7.25%)	690 (6.53%)
Edema	29198 (25.68%)	3046 (28.84%)
Enlarged Cardiomediastinum	6131 (5.39%)	593 (5.61%)
Fracture	5588 (4.91%)	292 (2.76%)
Lung Lesion	4941 (4.35%)	379 (3.59%)
Lung Opacity	60369 (53.09%)	5380 (50.93%)
Pleural Effusion	49438 (43.48%)	3791 (35.89%)

Pleural Other	2120 (1.86%)	133 (1.26%)
Pneumonia	3354 (2.95%)	359 (3.4%)
Pneumothorax	11441 (10.06%)	687 (6.5%)
Support Devices	60387 (53.11%)	5157 (48.82%)

Extended Data Fig. 4. Distributional differences of age across racial groups, for patients with No Finding = 0 and for patients with Pleural Effusion = 1.

Dataset	Age Group	No Finding = 0		Pleural Effusion = 1	
		White, N (%)	Black, N (%)	White, N (%)	Black, N (%)
MIMIC	0-20	323 (0.37%)	85 (0.48%)	79 (0.22%)	35 (0.66%)
	20-40	5291 (5.98%)	1816 (10.22%)	1622 (4.54%)	444 (8.42%)
	40-60	24087 (27.24%)	6309 (35.51%)	8422 (23.55%)	1611 (30.56%)
	60-80	41415 (46.84%)	7561 (42.55%)	17701 (49.50%)	2445 (46.39%)
	80+	17307 (19.57%)	1997 (11.24%)	7936 (22.19%)	736 (13.86%)
CheXpert	0-20	1058 (0.93%)	232 (2.20%)	341 (0.69%)	90 (2.37%)
	20-40	10228 (9.00%)	1515 (14.34%)	3586 (7.25%)	497 (13.11%)
	40-60	32674 (28.74%)	4525 (42.84%)	13639 (27.59%)	1580 (41.68%)
	60-80	48394 (42.56%)	3364 (31.85%)	22068 (44.64%)	1281 (33.79%)
	80+	21351 (18.78%)	927 (8.78%)	9804 (19.83%)	343 (9.05%)
Emory	0-20	2364 (0.83%)	3039 (1.17%)	262 (0.34%)	233 (0.54%)
	20-40	27716 (9.69%)	52207 (20.07%)	4565 (5.86%)	5398 (12.61%)
	40-60	82368 (28.8%)	104369 (40.13%)	20202 (25.93%)	16151 (37.73%)
	60-80	138174 (48.32%)	86906 (33.41%)	41342 (53.07%)	17959 (41.95%)
	80+	35336 (12.36%)	13569 (5.22%)	11529 (14.80%)	3068 (7.17%)